Interpretability Through Interrogation: Fairness and Interpretability in the Context of Criminal Sentencing

Contents

2 Introduction

2.1 Motivation

Machine learning algorithms are useful for everything from predicting rates of car accidents in order to set insurance rates to determining the probability of death from pneumonia.

Criminal sentencing is one area in which algorithms are theoretically useful. Sentencing algorithms might be more accurate than human judgment in predicting recidivism rates or other quantities of interest, just as insurance company algorithms are more accurate than traditional methods in predicting who will get sick or get into a car accident. A second benefit is that they can help eliminate human racial bias in the criminal justice system.

In an attempt to right the wrongs of implicit bias and a racially-charged criminal justice system, many states have begun adopting algorithms that determine sentence length or bail amounts. However, in adopting these machine learning algorithms, which train themselves on historical data, perhaps those states are codifying the racism that they were trying to avoid.

Currently, the Massachusetts Legislature is considering adopting an actuarial risk assessment tool for use in the criminal justice system. Researchers at the Berkman Klein Center at Harvard Law School, among other MIT and Harvard faculty, wrote to urge the legislature to first evaluate the tools and work to mitigate the risk of amplifying bias in the justice system.

As long as the government employs algorithms to alter behavior or human outcomes, many argue that the algorithms used should be transparent and explained to the public. As Kate Crawford, a principal researcher at Microsoft Research, put it in an interview with *The New York Times* [3] , "if you are given a score [by an algorithm] that jeopardizes your ability to get a job, housing or education, you should have the right to see that data, know how it was generated, and be able to correct errors and contest the decision."[4] However, accessibility to relevant code would often not address transparency concerns, as black box algorithms can be difficult or impossible to understand or interpret, even with the code. We must ensure that not only are the algorithms themselves fair and equitable, but also that they are applied in a fair and equitable manner.

Marc Rotenburg, President and Executive Director of the Electronic Privacy Information Center (EPIC), said in an email to the Harvard Political Review that, "there must always be a way to determine how a decision is made," whether through examinations of the model and output, or through third party auditing. If that attribute cannot be established, the algorithm should not be used.[37] In this paper, I will determine if transparency and interpretability of machine learning algorithms are necessary or sufficient for fairness. I will do so in the context of criminal justice risk assessment algorithms used for sentencing.

2.1.1 Context

Algorithms have been introduced to help determine bail and pretrial release as well as sentencing. This was done to combat human biases and improve prediction.

Theorists of criminal punishment have set out four main justifications for criminal punishment in general and penal incarceration in particular: (i) retribution, the idea that that those who commit certain kinds crimes deserve to suffer certain punishments; (ii), deterrence, the idea that punishment prevents future crimes by deterring other would-be offenders; (iii) incapacitation, the idea that incarceration incapacitates the offender so he cannot commit future crimes while incarcerated; and (iv) rehabilitation, the idea that incarceration is a means of rehabilitating the individual so that they do not commit further crimes when released. There is a large research literature debating which of these justifications can in fact ground a system of incarceration, at times the four approaches would lead to divergent results. American criminal law does not expressly endorse any one theory, but instead has elements of each. [31]

From the beginning of the system through which someone is arrested and convicted of a crime, there is human discretion at every step. Many criminal cases begin with a police stop. In New York, for example, black and Latino people make up half the population, and 85% of marijuana arrests, although white people use marijuana at similar rates. In 2016, there were 14 marijuana arrests on the Upper East Side, and 492 in East Harlem, one precinct over. Arrest rates are often not based on citizen behavior, but rather on overpolicing in communities of color.

In fact, black people are on average 3.73 times more likely to be arrested for marijuana possession than white people are, even though both populations use marijuana at similar rates. In 2010, there was a marijuana arrest every 37 seconds. States spent over $3.6 billion enforcing marijuana possession laws. New York and California alone spent over $1 billion in total justice system expenditures for marijuana possession arrests. [6]

The next decision made in the system is the decision to arrest and charge someone. The police may only make an arrest when there is probable cause to believe a crime has been or is being committed, but police otherwise have broad discretion in when to make arrests. Arrest rates for black people in the US are five times higher than for white people. [30]

Although there are many areas of the criminal justice system where predictive algorithms are relevant, including predictive policing and hotspot prediction, the two relevant for this discussion are algorithms used in bail and sentencing. I will address both of these systems below.

Bail

Bail determinations are conventionally made by humans and are therefore subject to

human biases.

Risk assessment tools are now used to determine risk for failing to appear in court. This assessment often uses employment information and history of missing court appointments, as well as criminal history, past drug use, whether the defendant is a local resident, among other information, and uses these factors to make a prediction about whether the defendant will fail to appear, be rearrested, or be rearrested for a violent offense.

The recommendations made are: release, moderate risk of failure to appear, high risk of failure to appear, and not recommended for release. Moderate risk is associated with a 16% risk of ever missing one court date. [12] [17] This decision often serves as justification for setting bail based on the recommendation or denial of pretrial detention.

If someone has bail set that they cannot afford, they are 34% more likely to plead guilty or to be convicted.[34] Pretrial detention increases the chance of conviction by incentivizing the defendant to plead guilty. It makes it harder for defendants to find work later and can increase sentence length.

Timothy Lynch, director of the criminal justice project at the Cato Institute, said, "The truth is that government officials have deliberately engineered the system to assure that the jury trial system established by the Constitution is seldom used." More than 90% of criminal cases are never tried before a jury, as defendants often plead guilty. [2]

Seventy percent of people in local jails are not convicted of any crimes. This population in American jails is larger than most other countries' total incarcerated populations. Ninety-nine percent of the total jail growth in the last 15 years was in pretrial detention. Nationally, in 2009, 34% of defendants were detained pretrial for the inability to post money bail. [33]

Every year, 626,000 people walk out of prison gates, but people go to jail 10.6 million times each year. Jail churn is particularly high because most people in jails have not been convicted. Some have just been arrested and will make bail in the next few hours or days, and others are too poor to make bail and must remain behind bars until their trial. Only 150,000 on any given day have been convicted, generally serving sentences under a year. [42]

Pretrial incarceration imposes high costs on individuals and society, as evident from the case of Kalief Browder.

Kalief Browder was 16 when he was accused of stealing a backpack, a crime he claimed he did not commit. He was born into Child Protective Services and was adopted into a family of seven siblings. Although Kalief did not have a backpack or any of the alleged contents on him, Kalief was arrested in response to a phone call the police received regarding a stolen backpack. Although he was never convicted of a crime, he spent more than one thousand days on Rikers Island awaiting trial because his family could not afford the $3,000 bail. For much of these three years, he spent about two years in solitary confinement and was frequently starved and beaten by prison guards. After his release, Kalief walked an hour each way to a G.E.D. prep class, and passed his G.E.D. on his first try. Two years after his release, he committed suicide, unable to recover from the trauma of prolonged

solitary confinement. [39]

New Jersey Public Safety Assessment

The New Jersey Public Safety Assessment is a bail algorithm that functions as an alternative to the traditional bail analysis that was applied to Kalief Browder and others. In January, New Jersey implemented an algorithm called the Public Safety Assessment (PSA). According to the Laura and John Arnold Foundation, a nonprofit which funds innovative solutions to criminal justice reform, the PSA predicts "the likelihood that an individual will commit a new crime if released before trial, and... the likelihood that [they] will fail to return for a future court hearing. In addition, it flags those defendants who present an elevated risk of committing a violent crime."[5]

The algorithm works by comparing "risks and outcomes in a database of 1.5 million cases from 300 jurisdictions nationwide, producing a score of one to six for the defendant based on the information." It also provides a recommendation for bail hearings. If someone meets the right criteria, they could be released without paying any bail at all.

New Jersey seems to be succeeding with the PSA. The state now sets bail for far fewer people than it once did because the algorithm predicts who is likely to try to run away and who is safe to release. Within the first six months, the number of people held in New Jersey jails awaiting trial dropped by 15%. The number of unconvicted people held in jail dropped by 34.1% between 2015 and 2017.[32]

While some believe that the system allows criminals to roam free, others believe that it is a more equitable system because in the absence of the PSA, bail often allows the wealthy to buy their freedom. The PSA score also acts as a recommendation, and judges don't need to follow it. It is worth noting that any decrease in bail, through algorithmic mechanisms or otherwise, are bound to be successful because of the highly problematic and inefficient nature of the cash bail system.

Recently, there have been attempts to improve the cash bail system by developing algorithmic assessments. The New Jersey system is an improvement on the bail system because it is less punitive and incorporates less bias in its determinations. [9]

Sentencing

The United States is home to 5 percent of the world's population but 25% of the world's prisoners. It has the highest rate of incarceration in the world at 3.2 million people. Prisons consist of mainly males in their 20s and 30s. 38.0% of them are black, and 58.4% of them are white according to the Federal Bureau of Prisons. [18] Some of the inequities in the U.S. prison population stem from the fact that black people are regularly given much harsher sentences than other races for equivalent crimes. In Florida, defendants in criminal prosecution cases are given a judge-calculated score based on the crime committed and previous crimes. Matching scores should logically lead to the same sentence,

but the *Herald Tribune* found that black people get much larger punishments, and that there is little oversight of judges.[40] Meanwhile, the *Washington Post* found that judges in Louisiana gave harsher punishments to defendants following unexpected losses by the Louisiana State University football team, and that these punishments were disproportionately borne by black people.[21] Clearly, the human bias inherent in sentencing is a problem that we must address.

Sentencing Guidelines

The Federal Sentencing Guidelines (FSG) are structured rules creating uniform sentencing for criminals convicted of felonies and serious misdemeanors. The guidelines allow a judge to consider certain elements of a case using a grid or worksheet scoring system, outputting a sentence or sentence range. The goal is that offenders with similar offenses and criminal histories be treated alike. [31]

The guidelines, essentially an algorithm, start with the base offense level and make adjustments based on factors like the defendant's acceptance of responsibility and whether he had a gun. The resulting value is a defendant's offense level, from one to 43. [29]

The guidelines were originally created in 1984 in response to the Sentencing Reform Act (SRA), to increase honesty in sentencing. However, hundreds of federal judges condemned the SRA as unconstitutional. The sentences specified in the Guidelines Manual are now advisory, not mandatory, after the *Booker* ruling in 2005.

The SRA was "perhaps the most dramatic change in sentencing law and practice in our Nation's history." It is also the most disliked. Its system of determinate sentencing calls for parity and predictability. However, there is little interest in the offender's capacity to change: sentences are meant to redress past wrongs, not to influence future conduct. Under the SRA, retribution takes precedence. Prisons are not places for penitence and rehabilitation, but of warehousing and detention. [31]

Sentences associated with crimes under federal jurisdiction are dramatically longer than equivalent crimes charged in state courts. The FSG increased penalties for violent crimes "where the Commission was convinced they were inadequate," without further explanation. For drug offenses, state sentences average thirty-one months while federal sentences average eighty-four. The FSG sentencing for drugs are above the mandatory minimum terms. Sometimes the choice between state and federal prosecution is arbitrary. [31]

According to Dr. J. C. Oleson's paper on the Sentencing Reform Act, "One offender may be charged in state court for cultivating marijuana and receive a $1,000 fine (waived because he's indigent) while his partner, charged in federal court for the identical crime, may receive ten years in prison and eight years of postconviction supervised release."[31] Through the federalization of criminal law, the SRA created disparity between federal and state-level convictions.

Justice Kennedy asserted that federal judges who depart downward from the Guidelines are "courageous, and [are] exercising the independence and the authority of the judiciary

not to follow blindly unjust guidelines."

"The tragedy of mass incarceration," says Judge Stefan Underhill, "has recently focused much attention on the need to reform the federal-sentencing guidelines, which often direct judges to impose excessive sentences." [29]

The FSG is not just unnecessarily severe and rigid. It is also incredibly complex and labyrinthine. It has been said that, "the Guidelines make the federal tax code look like Reader's Digest." It can be difficult to draw meaningful distinctions between overly elaborate sentencing grades. Michael Tonry said that sentencing commissions cannot easily answer the question, "Could we plausibly explain to a judge why a level sixteen crime is more serious than a level fifteen crime?" [31]

There are 102 federal prisons, but the federal system includes the US Marshals Service, the Bureau of Prisons, and Immigration and Customs Enforcement. Nonviolent drug convictions make up a significant portion of the federal system, but the federal system is not nearly as large as the state system, as seen in the figure below.

The jargon of the Guidelines, using terms like "points," "levels," and "scores," creates an appearance of objectivity and analytical precision. But their reasoning is difficult to explain or audit. The Sentencing Commission has a monopoly on federal sentencing data, so nobody can access the data to understand the determinations of the guidelines. More specifically, chief judges of each district are required to submit a written report on each sentence imposed to the Sentencing Commission. Many of these documents are not public record, so there is no meaningful way to examine patterns of federal sentencing outside of the Commission's dataset. [31]

The Guidelines are also susceptible to immense bias, which many attribute to fact that there are significant racial disparities between users of crack cocaine and users of powder cocaine. There is a 100:1 disparity between federal crack and powder-form cocaine sentences.

The federal prison system consists of just a small part of total incarceration, but the federal government can use its financial and ideological power to incentivize better paths forward. Elected sheriffs, district attorneys, and judges, all working within the state system, can also slow the flow of people into the criminal justice system but also need legislative reform to do so.

The state system is where most of the criminal system takes place, as shown in the figure below.

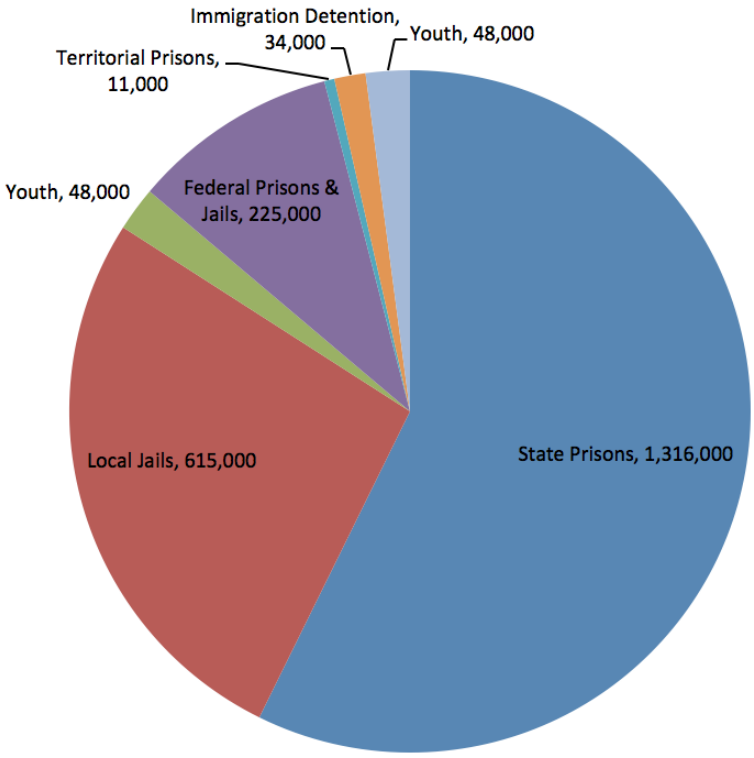


Figure 1: The United States locks up more people, per capita, than any other nation. But grappling with why requires us to first consider the many types of correctional facilities and the reasons that 2.3 million people are confined there. [42]

Local jails hold one out of every three incarcerated people. [42] As seen in Figure One, state and local jails and prisons are a significant portion of the justice system, so state-level risk assessment algorithms would potentially affect millions of lives if implemented.

Many sentencing guidelines mirror the federal one. However, there has been a strong focus on using algorithms to replace these sentencing guidelines, systematizing and simplifying the process with a focus on predictive power and elimination of bias. Below is a table of state usage of different risk assessment tools used for sentencing: [16]

Table 1: State Risk Assessment Tools

State	Use	Type
Alabama	Yes	
Alaska	No, but recommended	
Arizona	PSA	Bail
Arkansas	Parole Risk Assessment Tool	Parole
California	Adaptation of LSI-R	
Colorado	LSI-R	Pre-sentencing report
Connecticut	Salient Factor Score	Parole
Delaware	LSI-R	
Florida	COMPAS	
Georgia	Yes	
Hawaii	LSI-R	
Idaho	Alternatives to incarceration for low risk defendants	Sentencing
Illinois	LSI-R	Probation
Indiana	Indiana Risk Assessment	Sentencing
Iowa	LSI-R recommended	Sentencing
Kansas	Yes	Parole
Kentucky	Yes	Sentencing
Louisiana	LARNA	Parole
Maine	Yes for sex offenders	Sentencing
Maryland	Recommended	Sentencing
Massachusetts	For sex offenders	
Michigan	COMPAS	
Minnesota	For sex offenders, recommended for sentencing	Sentencing
Mississippi	Yes	Sentencing
Missouri	Instrument	Sentencing, Parole, Probation
Montana	Yes	Parole
Nebraska	LS/CMI	Sentencing
Nevada	Yes	Parole
New Hampshire	Yes	Probation and parole
New Jersey	Yes	Parole
New Mexico	COMPAS	
New York	Yes	Parole
North Caroline	Yes	Probation and parole
Ohio	Ohio Risk Assessment	Sentencing, parole, etc
Oklahoma	LSI-R	Alternative sentencing
Oregon	Public safety checklist	

State	Use	Type
Pennsylvania	Risk assessment	Sentencing
Rhode Island	Recommended	Probation
South Carolina	Yes	Probation and parole
South Dakota	Yes	Parole
Tennessee	Yes	Sentencing
Texas	Texas Risk Assessment System	
Utah	LS/RNR	Bail
Vermont	Yes	Every stage
Virginia	Yes	Sentencing, pre-trial services
Washington	LSI-R	
West Virginia	Yes	
Wisconsin	COMPAS	
Wyoming	COMPAS	

In spite of its problems outlined earlier, there is value in structured sentencing. Now that judges must provide a reason for their sentence decision, we have the ability to study differing sentencing for similar offenders with a wealth of data.

As Judge John Coughenour, a Senior United States District Judge of the United States District Court for the Western District of Washington, says, "The thing that really hit me hard with the Sentencing Reform Act was that this art form–I considered sentencing to be an art and not a science–Congress tried to convert it into a science. And it's not a science. It's a human being dealing with other human beings. And it shouldn't be done by computers." [29]

2.1.2 COMPAS

The Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS, contains a risk assessment machine learning algorithm developed by Northpointe Inc. (now Equivant), a private company. The recidivism risk scale predicts a defendant's risk of committing a misdemeanor or felony within 2 years of assessment. This risk is then factored into determinations of bail and in some cases, sentence length.

Since the algorithm is the key to its business, Northpointe does not reveal the details of its code. However, multiple states use the algorithm for risk assessments to determine bail amounts and sentence lengths. Many worry that this risk assessment algorithm, and algorithms like it, have unfair effects on different groups, especially with regards to race.

COMPAS assigns defendants a score from 1 to 10 based on 137 features, including age, sex and criminal history that indicates how likely they are to reoffend. Race, ostensibly, does not factor into the calculus. However, ProPublica, an independent newsroom producing investigative journalism in the public interest, found that the software assessed risk based on information like ZIP codes, educational attainment, and family history of incarceration, all of which can serve as proxies for race. COMPAS has been used to assess more than 1 million offenders since its development in 1998. [13]

There are a few ways to define COMPAS' success. One way is to look at its false positives rate–how many people the algorithm incorrectly labels as being at high risk for recidivism. Another way is to look at false negatives–how many risky people the algorithm misses. The metric that should be applied is context-based. The justice system must decide if it would rather falsely punish more innocent people because of a bad false positive rate or let risky people roam free with a bad false negative rate.

Sixty percent of white defendants who scored a 7 on COMPAS reoffended, and 61 percent of black defendants who scored a 7 reoffended. On the surface, these numbers seem fairly equal in terms of output of true positives. But if you consider the false positives, you see a different story. Among defendants who did not reoffend, 42 percent of black defendants were classified as medium or high risk, compared to only 22 percent of white defendants. In other words, black people were more than twice as likely as whites to be classified as medium or high risk.

These differences highlight the tension between giving longer sentences to ensure less recidivism and giving shorter sentences at the risk of false negatives. This is a problem that we must resolve before we can regulate any algorithm, and it depends on how we define fairness. According to Alexandra Chouldechova [11], whose paper is summarized in the literature review of this thesis and who has studied the COMPAS algorithm, "Fairness itself... is a social and ethical concept, not a statistical one." There is very little legal precedent and regulatory power we can hold over private companies. But if governments are going to use these private algorithms for the public interest, they must first figure out how to ensure they are fair and legitimate.

Other recidivism algorithms

COMPAS is by no means the only recidivism risk assessment tool.

The LSI-R score, or Level of Service Inventory-Revised, is a validated risk/need assessment tool which identifies problem areas in an offender's life and predicts his/her risk of recidivism. LSI-R is commonly used today. It asks fifty-four questions including those about prior criminal history, employment history, educational history as well as those about emotions, attitude, and orientation.

Algorithms include those predicting homicide offender recidivism (Neuilly et al., 2011), predicting serious misconduct among incarcerated prisoners (Berk et al., 2006), forecasting potential murders for criminals on probation or parole (Berk et al., 2009), forecasting

domestic violence and helping to inform court decisions at arraignment (Berk and Sorenson, 2014). Berk et al., 2005 helped the Los Angeles Sheriff's Department to develop a simple and practical screener to forecast domestic violence by using decision trees. [43]

Well-known risk factors for recidivism (Bushway and Piehl, 2007; Crow, 2008) have been used in risk assessment tools since 1928 (Borden (1928), Hinojosa et al. (2005), Berk et al. (2006) and Baradaran (2013)). They include: [43]

- Information about prison release using data from 1994 (e.g. time served and infraction in prison),
- Age at release
- Information from past arrests, sentencing and convictions (e.g. prior arrests 1 and any prior jail time),
- History of substance abuse (e.g. alcohol abuse) and
- Gender
- Prior arrest (for felony, misdemeanor, local ordinance, with firearms involved, with child involved, public order)

Ohio experimented with a risk-assessment tool in the 1960s, and California began using a prediction tool in the early 1970s, as did the federal government. While some states, such as Illinois and California, later stopped using actuarial methods when they abandoned parole, other states, such as Georgia, Iowa, Tennessee, South Carolina, Alabama, and Florida, began using risk-assessment tools in the late 1970s and early 1980s. Soon, many other states followed their lead: Missouri, Michigan, North Dakota, South Dakota, Washington, Arkansas, Colorado, Nevada, Maryland, Connecticut, New Jersey, Ohio, Vermont, Alaska, Idaho, Kentucky, Maine, Montana, Pennsylvania, Texas, Utah, and Delaware.

In 2004, twenty-eight states used risk-assessment tools to guide their parole determinations, approximately 72 percent of states that maintain an active parole system. As a leading parole authority association suggests, "In this day and age, making parole decisions without benefit of a good, research-based risk assessment instrument clearly falls short of accepted best practice." [22]

In 1994, the Virginia legislature directed the state's new sentencing commission to develop a tool to divert low risk nonviolent offenders out of the prison system: to "study the feasibility of using an empirically based risk assessment instrument to select 25% of the lowest risk, incarceration bound, drug and property offenders for placement in alternative (nonprison) sanctions." The commission produced an actuarial instrument, the Risk Assessment Instrument, that was put into effect in pilot sites in 1997. A follow-up study by the National Center for State Courts of 555 diverted offenders in six judicial circuits in Virginia concluded that the program was a success and recommended the statewide expansion of risk assessment at the sentencing stage.

Decreasing pretrial incarceration is a huge step in the direction of justice. Also, 1 in 5 people are locked up for a drug offense, and this is especially prominent in federal prisons. Drug arrests give people in over-policed communities criminal records, which not only makes employment harder, but also makes it more likely to have longer sentences for future offenses, as seen in the Analysis section below. Most incarcerated youth are locked up for nonviolent offenses or "technical violations" of their probation, rather than for a new offense. [32]

Koepke and Robinson, Danger Ahead:
Risk Assessment and the Future of Bail Reform

Pretrial risk assessment tools make "zombie predictions," meaning that predictive machine learning models are trained on data from older bail regimes, and are blind to recent bail reforms. This leads to predictions that overestimate the risk of letting someone out without bail. There is therefore a danger in using these pretrial risk assessment tools because they legitimize unfair practices under the shroud of success, even though the success is solely caused by decreasing bail.

Bail was initially supposed to be used only in cases where there is a risk that the accused will flee the jurisdiction if released. Unofficially, a defendant's predicted dangerousness has always mattered, as judges would set unattainable bail amounts to detain dangerous individuals. In 1966 Congress passed the Bail Reform Act to address jail overcrowding, minimizing the reliance on cash bail. The Act forbid judges from treating a defendant's dangerousness or risk to public safety as a reason for detention, except in capital cases. This all changed under Nixon, when Attorney General John Mitchell argued for the necessity of preventative detention in response to rising crime rates. A 1984 Bail Reform Act made public safety a central concern in the determination of pretrial options. Public safety under this Act included non-violent crimes such as those against property, expanding the meaning of danger. There was a 32% increase in prisoner population during the first year after its passage. [31]

Illinois, New Jersey, Alaska, and New Orleans have eliminated or severely restricted monetary bail. Atlanta eliminated cash bonds for low-level offenses, and New York districts have ordered prosecutors to not request money bail in most cases.

In terms of risk assessment tools, the American Bar Association recommends that judges use actuarial models in making bail determinations. However, Human Rights Watch believes pretrial risk assessment tools should be opposed entirely on the grounds that these tools deny the individuality of each defendant and mask discriminatory practices under the guise of neutrality: "We should reject anything that normalizes racial bias in our criminal justice (or any other) system." [24]

2.1.3 Definitions of Fairness

How do we define whether an algorithm is fair? One metric besides false negative (FNR) and false positive rates (FPR) to measure fairness is the positive predictive value (PPV), which is the probability of recidivism given that the algorithm predicted that person would recidivate. The relationship between the three is below:

$$FPR = \frac{p}{1-p}\frac{1-PPV}{PPV}(1-FNR)$$

PPV is also referred to as precision, and the true positive rate is recall or sensitivity.

Fair Representations Below is a representation of a confusion matrix.

Table 2: Confusion Matrix

	$\hat{y}=1$	$\hat{y}=-1$	
$y=1$	True Positive	False Negative	False Negative Rate: $P(\hat{y}\neq y \mid y=1)$
$y=-1$	False Positive	True Negative	False Positive Rate: $P(\hat{y}\neq y \mid y=-1)$
	False Discovery Rate: $P(\hat{y}\neq y \mid \hat{y}=1)$	False Omission Rate: $P(\hat{y}\neq y \mid \hat{y}=-1)$	Overall Misclassification Rate: $P(\hat{y}\neq y)$

According to ProPublica, the algorithm fails differently for black and white defendants. Although the total accuracy for both races are roughly the same, COMPAS is more likely to flag black defendants as future criminals. 23.5% of white people were labelled higher rsisk but didn't reoffend, while for black people the number is 44.9%.

Northpointe would argue that the proportion of black and white defendants who re-offend is the same so the algorithm is fair. This is the positive predictive value, and Northpointe used this to justify their algorithm's accuracy.

In her paper [11], Alexandra Chouldechova finds that test fairness, or the positive predictive value being equal between different classes, is not dependent on race in COMPAS. COMPAS satisfies test fairness, but there are large discrepancies between the false positive and false negative rates.

Because of the different base rates (the proportion of individuals who recidivate), and because perfect prediction is impossible, it is impossible to satisfy all of the fairness definitions described in Chouldechova's paper. Recidivism rates are so vastly different (not necessarily because one group recidivates more, but perhaps because police officers are more watchful of black people) so not all of the fairness criteria can be satisfied at the same time. This is an impossibility result. Some notions of fairness are fundamentally incompatible with each other. [19]

Statistical parity is the idea that the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population.

Say S is the minority group while T is the majority, then

$$Pr[x \in S|M(x) = 0] = Pr[x \in S]$$

$$Pr[M(x) = 0|x \in S] = Pr[M(x) = 0|x \in T]$$

Group fairness can be abused, and can allow for self-fulfilling prophesies: allowing people to select the smartest students in T and random students in S would allow a company to claim that students in T perform better and should therefore be hired more.

Individual fairness is the idea that similar individuals should be treated similarly. We lack a good metric to compare these individuals, however. [14] Individual fairness means that the existence of a sensitive attribute shouldn't matter. The Federal Sentencing Guidelines succeeded in treating like offenders alike, but they did so by treating unlike offenders alike. Parity in sentencing is an important goal, but should not be the only goal.

In this situation, looking at whether individuals are treated similarly is a nebulous concept. Defining similarity of individuals is not only a tricky task, it cannot be done in a way that would make sense in this context. Because black and white people behave so differently and have different input features, it would be impossible to measure them at face value on the same grounds and get output predictions that made sense.

2.2 Interpretability Literature Review

As machine learning algorithms continue to pervade into areas like the insurance industry, financial markets, and the criminal justice system, it can seem dangerous or problematic that humans are unable to understand these models. Many argue that decision makers should be able to understand the behavior of models so that they can audit them and determine how much to depend on these models, detect potential biases, and further refine them. The **right to explanation** is therefore a compelling legal lens with which to view the necessity of interpretability.

The European Union's General Data Protection Regulation (GDPR) restricts automated decision-making, creating a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made about them. For example, if someone is given a sentence based on a recommendation from COMPAS, the defendant could potentially look at the decision-making process the algorithm underwent to understand the sentence. [38]

However, by nature, machine learning models are often uninterpretable.

2.2.1 Introduction to Machine Interpretability

In the creation of our protocol, we attempt to create an interpretable explanation, or a simple model in an interpretable feature space, using decision trees. However, the decision-

making process itself remains uninterpretable within decision trees. This is apparent when comparing defendants aged 22 versus 25, with 0 priors: the 22 year olds are given a high risk score, while the 25 year olds are not in COMPAS. This is explored in the Analysis section below.

Instead, we introduce a new system to provide a persuasive explanation strategy. This does not attempt to achieve explanatory power, but incorporates user judgement and intuition into the protocol for the purpose of simulatability. Simulatability is highly subjective, and has two subtypes: one based on the total size of the model, and another based on the computation required to perform inference.

A caveat is that simulatability as a notion of interpretability requires that inputs themselves be individually interpretable. The scope of the type of model we will be examining in this thesis will be those that have this feature of interpretable inputs.

Perhaps instead of focusing on human interpretability we can design algorithms that help check their work, or a machine that works with humans to interpret itself.

In the construction in Chapter 3, the auditor wants to ensure that the model does not rely on something that we may consider undesirable or illegal, such as a protected class. The prover assists the auditor in their interpretation by attempting to prove their algorithm interpretable. The auditor chooses a decision that seems spurious, and the algorithm delves into it to see if the decision makes sense.

While extensive literature on interpretability exists, many of them do not focus on the issue of defining interpretability, or whether interpretability is necessary or sufficient for fairness. Rather, the literature focuses on increasing or improving interpretability through human intelligence tasks and extraction of decision lists, decision trees, and decision sets. Some papers equate interpretability with understandability, trust, or intelligibility.

What is interpretability and why is it important? Interpretability is often associated with elements such as trust, transferability, informativeness and transparency. These elements can be broken down into different ideas and interpreted differently. Below I will attempt to use different perspectives on interpretability to inform the literature review.

I will use a framework for interpretability based on Zachary Lipton's paper, "The Mythos of Model Interpretability." [27]. Through this literature review we will identify gaps in the literature surrounding interpretability and attempt to fill those gaps through our protocol. We will also attempt to define different versions of interpretability and look at the merits of different definitions through this review.

2.2.2 Transferability

Machine learning algorithms are judged by their ability to generalize outside of the data they are given to train on. Overfitting is when a model fits too closely to the data it trains on, lacking the ability to generalize to new data. Humans are very good at transferring knowledge and skills to unfamiliar situations, and machines are not. The validity and strength of a machine learning algorithm is important for interpretability because if its

predictive strength is weak or easily game-able it is not a useful predictive algorithm, and its explanatory power is useless.

Interpretability includes the premise of being able to trust a model, and transferability is a large part of this.

Caruana et al. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Perhaps the clearest example of the need for interpretability can be seen through this paper. Here, Caruana looks at an algorithm with rule extraction that misinterprets the relationship between some features and the outcome, which is probability of death. Without human intervention, this interpretation could have been potentially fatal. In the analytical portion of this thesis, I will attempt to apply different methods for the same purpose–understanding how a decision is made, and trying to determine if there is something problematic about this process. Caruana finds spurious relationships that require domain knowledge or at least human reasoning to fix.

In Caruana's paper, he looks at a pneumonia risk prediction study in which an intelligible model allows misleading patterns to be removed. A machine learning algorithm attempted to predict the probability of death for patients with pneumonia. The motivation was to identify and focus on high-risk patients while treating low-risk patients as outpatients.

The study ended up using a more human-interpretable algorithm at the price of increased accuracy, using a logistic regression instead of the more accurate but less interpretable neural net.

Generalized additive models (GAMs) are an additive modeling technique in which relationships between individual predictors and the dependent variable follow smooth patterns that can be linear or nonlinear. These relationships are estimated simultaneously, and the GAM is then predicted by adding the estimations up. GAMs can capture nonlinear patterns that a classic linear model would miss.

GAMs have the form

$$g(E[y]) = \beta_0 + \sum f_j(x_j)$$

g is the link function and for each term f_j, $E[f_j] = 0$.

GA^{2}Ms add pairwise interactions to improve accuracy. These pairwise interactions are intelligible (according to Caruana's definition of intelligibility) because they can be visualized as a heat map, as shown below.

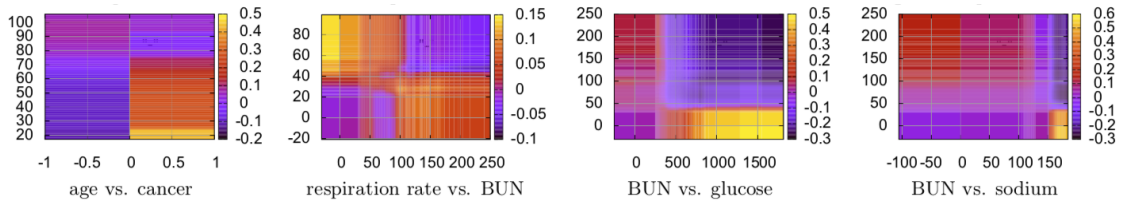


Figure 2: 4 of 56 components of the model trained on the pneumonia data. These heat maps were created by the GA^{2}M using pairwise interactions noted in the captions of each graph. The x axis represents the probability of death while the y axes are the two variables in question.

GA^{2}Ms have the form

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)$$

The GA^{2}M builds the most accurate GAM and then detects and ranks all possible pairs of interactions. The GA^{2}M then keeps the top k pairs, as determined by cross-validation. Both the GAM and GA^{2}M models are trained on data using 100 rounds of bagging to reduce overfitting and provide pseudo-confidence intervals.

The machine learning algorithm output the following information: the algorithm determined that those with asthma are at lower risk of death from pneumonia, when in fact asthmatics have much higher risk. This is because asthmatics are typically given much more aggressive care, so the treatment they receive significantly lowers their risk of dying from pneumonia compared to the general population.

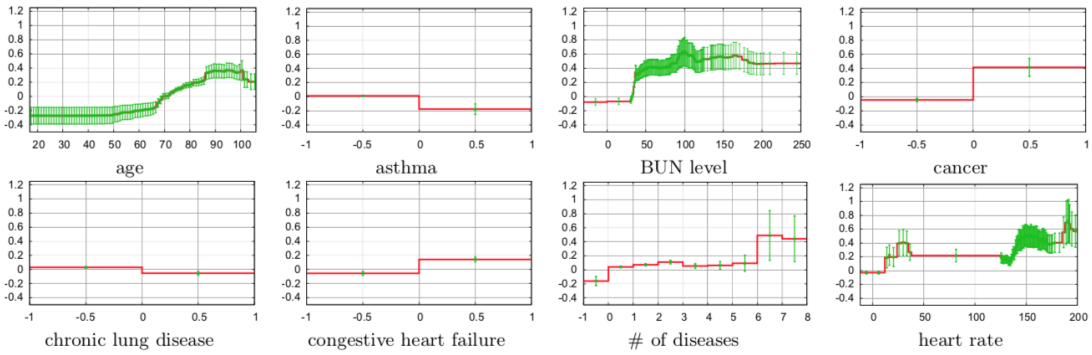


Figure 3: 8 of 56 components of the model trained on the pneumonia data. The green errorbars are pseudo-errorbars from bagging. Asthma, cancer, lung disease, and congestive heart failure are booleans. The vertical axis is probability of death. Asthma has an inverse relationship with probability of death, although having asthma increases probability of death from pneumonia.

In this situation, if the model were not interpreted or investigated, the use of a model might have altered the environment, invalidating their future predictions. Specifically, if doctors had used this information, they would have focused less on asthmatic patients thinking they needed less care. This would have increased death and invalidated the model.

Upon further investigation, attributes other than asthma have similar incorrect relationships with probability of death, including chronic lung disease and a history of chest pain. These patients may receive care earlier, or may receive more aggressive care, which the model does not reflect.

Caruana suggests manually finding incorrect relationships found in the model and manually changing them. Using domain knowledge and human expertise, he recommends redrawing the graph such that the risk score for having asthma is positive for probability of death.

Caruana defines his notion of intelligibility by one that is interpretable by humans. For example, the algorithm learned the rule "HasAsthma(x) $\rightarrow$ LowerRisk(x)" which is something that a human can understand. He also implies that an interpretable model is one that can be visualized easily through graphs as seen above, similar to Ribeiro's claims. His suggestions for using interpretability are reliant on human judgment and domain expertise. For example, without knowing that asthma increases the probability of death from pneumonia, someone without domain knowledge would not realize that the model incorrectly made assumptions based on the data it was given. It is therefore difficult to find these incorrect assumptions without understanding the science behind pneumonia. [10]

Another problem with this paper is that it misinterprets regression coefficients. There is a belief that parameters must be positive or negative to make sense in a model, but this is more a function of the other parameters in the model, and how they interact. Trying to make sense of Supervised Learning results is a fool's errand, because SL algorithms are not meant to make sense. They find patterns in data and predict based on the assumption that the distribution of data will not stray too far away from the data it trained on. Its patterns don't need to make sense, in fact, the computer does not know or care about the patterns themselves, so they can be as obscure or unintelligible as they need to be to get a good model, depending on the hyperparameters and parameters the user includes.

The ability to manually flip an input, like in the asthma case, or generally manipulating features to make the output different, defeats the purpose of supervised learning. The positive correlation with asthma might be positive or negative depending on what other features exist in the training set, or even with parameter tuning.

However, Caruana's study shows that deploying a model might alter their environment, invalidating future predictions. If the model was misunderstood and asthma patients were given less treatment, adding this data to the training set would invalidate the predictive model.

2.2.3 Trust

Trust could be thought of as confidence that a model will perform well. However, as explored above, models can perform well for certain cases and poorly for others. In the case of the criminal justice system, we care about how often our predictions are correct, and which examples are correct.

Trust could be a host of things, including faith in a model's performance, robustness, or a low-level understanding of the model. In ProPublica's original assessment of the COMPAS algorithm, Julia Angwin looked at not only how often the model is right but also for which examples it is right, which are two important factors of trust in an algorithm.

Ribeiro et al. "Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Ribeiro argues that trust is fundamental if one plans to take action based on a prediction, saying, "if the users do not trust a model or a prediction, they will not use it." He makes the distinction between models and predictions because a model is a general algorithm while predictions are on an individual basis. However, what happens when users blindly trust a model or prediction and use it although they perhaps should not? Or, if a user trusts the algorithm but those who it is acting upon does not, such as in the case of the criminal justice system?

Ribeiro attempts to explain a prediction by presenting textual or visual artifacts to provide qualitative understanding to the model's prediction.

He defines an interpretable algorithm as one that provides qualitative understanding between the input and output of the algorithm. This further implies that explanations should be easy to understand, which is dependent on the target audience and on the complexity of the inputs and of the model itself.

Ribeiro's algorithm, Local Interpretable Model-Agnostic Explanations (LIME), attempts to identify a interpretable model over the interpretable representation that is locally faithful to the classifier. $\omega(g)$ is a measure of complexity, whether that is the depth of the tree or the number of non-zero weights, depending on the type of algorithm. $\mathcal{L}(f, g, \pi_x)$ is a measure of how unfaithful g is at approximating f in the locality defined by π_x. We must therefore minimize $\mathcal{L}$ while having ω be low enough to be human-interpretable. [36]

However, LIME and understanding feature dependence is not useful for the same reason that looking at Caruana's individual feature correlation is not. These weights can change based on the fragility of the model, the features used, and the parameters tuned. They are not necessarily indicative of how a decision was made, but rather how the specific decision in that particular situation was weighted.

Sarah Tan et al. Detecting Bias in Black-Box Models
Using Transparent Model Distillation

In this paper, researchers explore the idea of model distillation from a large model to a simpler model without significant loss in prediction accuracy. Researchers compared transparent student models trained to mimic COMPAS to transparent models trained on the true recidivism labels, allowing them to identify biases in COMPAS that do not appear to be justified by the data, raising specific questions about COMPAS that warrant further investigation.This allows us to detect bias.

y^S is a risk score and y^O is the actual outcome the risk score was intended to predict. The labels used in this study were both the risk score outputted by COMPAS and the actual recidivism of each defendant.

The researchers created two sets of models: one that mimicked the risk assessment tools where the labels are the risk scores, and a more transparent model trained on actual outcomes of recidivism. This second model contained race, age, and gender as features, so that the researchers could see correlation between these features and other proxies that would normally be included in a model. The researchers claim that removing protected features, which are highly correlated with features typically included in models, has no benefit other than the ability to claim that these features were not included. In other words, the model will learn the bias of these features anyway.

Let $y^S = r^S(x)$ be the true risk scoring model.

$x_i = (x_{i1}, ..., x_{ip})$ is a vector of p features for person i.

The student model of the COMPAS risk score teacher is r_S is

$$y^S = f^S(x)$$

The model of the actual outcome, where g is the logistic link function is

$$g(y^O) = f^O(x)$$

The transparent model uses bagged, short trees learned using gradient boosting. Its trees are shallow and only operate on one feature, added to trees that include pairwise interactions, for transparency. In other words, they are GA^2Ms like those in Caruana's paper on pneumonia. The researchers chose this over decision trees because decision trees tend to have instability that could allow for spurious inferences of bias.

This allows the model to detect regions of the feature space where the risk scoring model significantly differs from the actual outcome, or where y^O and y^S vary systematically.

The researchers assessed fidelity, or how close the student model was to the teacher model in terms of predictions. They also assessed the accuracy of the outcome. Using two transparent models allowed the researchers to find biases without first knowing which biases to look for. This is a major difference from the pneumonia paper, which requires heavy amounts of domain knowledge.

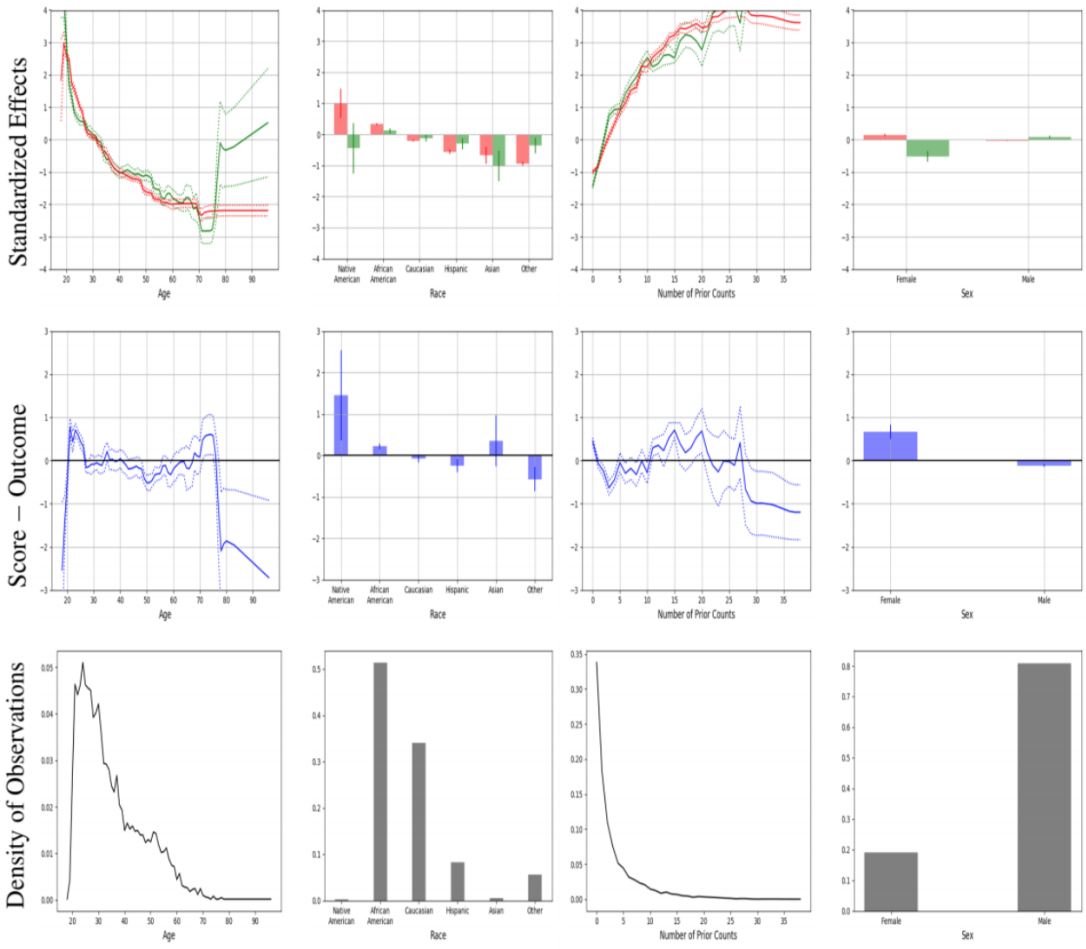


Figure 4: Shape plots for four of six features for recidivism prediction. Top row: Red lines: effect of feature on COMPAS risk score. Green lines: standardized effect of feature on actual recidivism outcome. Categorical terms ordered in decreasing predicted risk of the score. All plots mean-centered on the vertical axes to allow terms to be easily added or subtracted from the model. Middle row: Blue lines: difference between score and outcome models (score - outcome). Bottom row: Density histogram of number of observations at each feature value.

Similarly to Caruana's paper on pneumonia, Tan uses visualizations to interpret the GA^2Ms. The researchers found that COMPAS is biased for certain age and race groups. For example, there are fewer samples for those older than 70 years old, so the variance is large. COMPAS also predicts low risk for very young offenders although when trained on the true recidivism labels, risk appears highest for these young offenders. COMPAS lastly predicts women are more likely to recidivate, and men are less likely, than the true labels

predict.

While locating biases is helpful in understanding COMPAS, we cannot necessarily see why the model has biases such as those described above. [41]

2.2.4 Transparency

Algorithmic transparency can take many forms. You could have transparency at the level of the entire level, which is simulatability. Transparency could also be at the level of individual components, called decomposability. Lastly, at the level of the training algorithm, transparency is called algorithmic transparency.

Ribeiro and Caruana both examined the idea of human interpretability, but did not specify what exactly it means for a human to interpret something.

Simulatability and Human Interpretability

Lipton defines simulatability as something with an explanation simple enough to be contemplated by a human all at once. This of course is dependent on user expectations and expertise. Simulatability can also be defined as transparency at the level of the entire model, while decomposability is transparency at the level of individual parameters.

Julia Dressel and Hany Farid. The Accuracy, Fairness, and Limits of Predicting Recidivism

Both this paper and the Herman paper look at human interpretability. Simulatability is the ability for a human to contemplate an entire model at once, which suggests that an interpretable model is a simple one. A human, according to simulatability, should be able to take in input data and reason through steps required to produce a prediction. Dressel in this paper does this with a large group of people, to see if human predictions can compare to those of the algorithm itself.

In this paper, the researchers compare COMPAS accuracy and fairness to predictions made by people with little or no criminal justice experience.

Using Amazon Turk, the researchers provided participants with a short description of a defendant that included the defendant's sex, age, and criminal history. This information did not include their race. These nonexperts had an overall accuracy rate of 67%, and comparable accuracies per race, for false positives, and for false negatives of each race. Participants' predictions were in agreement with COMPAS's 69.2% of the time.

One problem with the paper is that its criticism of COMPAS as unnecessarily complex is valid, but for the wrong reasons. A logistic regression on two features yields similar prediction accuracy as COMPAS. However, using more features, or something other than a logistic regression, can often be more interpretable. In the legal system, it is important to consider the procedure when looking at the outcome: even if accuracy remains the same, the procedure can have varying levels of clarity. The representation of a model matters.

Another more logistical problem is the general claim that the participants have equivalent accuracies. The accuracy of the 20 *median* participant accuracies is compared to the COMPAS accuracy. Even then, the accuracy is 2.4 points lower, which has a p-value of 0.045, just within standard conceptions of rejection of the null hypothesis, which is a flawed statistical test anyway. If it were the case that we were to look at majority rules criterion such as the 20 median scores, we cannot compare this value to COMPAS or to the judicial system, in which majority rules is not the way decision are made.

The accuracy rates for black and white people, as well as the false positive and false negative rates of both races are comparable between human prediction and COMPAS. Perhaps this justifies COMPAS usage in a way that Dressel did not intend in writing this paper–COMPAS effectively simulates human decision-making, and does it slightly better, with much less manpower. Isn't that all we can expect from a machine learning algorithm trained on human behavior? Or should we not use this because it legitimizes racist discrepancies in predictions through technology? It seems as though since humans cannot do any better than COMPAS, we can at least use the algorithm to double check our work.

Bernease Herman. The Promise and Peril of Human Evaluation for Model Interpretability

Herman argues that using human interpretability as a metric can introduce implicit human cognitive bias into the system. We can even fail to satisfy our ethical goals of fair and sufficiently accurate algorithms by relying on human understanding.

Herman defines an "interpretable explanation" as "a simple model, visualization, or text description that lies in an interpretable feature space and approximates a more complex model." Model complexity is "a measure of the amount of information contained in a model."

Descriptive explanations, according to Herman, best satisfy the ethical goal of transparency as they provide humans with information about inner workings of the system. This type of explanation is one that "generates explanations with maximum model fidelity for a particular explanation vehicle and underlying machine learning model." In contrast, a persuasive explanation strategy incorporates user preferences, knowledge, or characteristics. This type of explanation balances accuracy with being convincing to the user. This is the type of explanation we seek in this thesis.

To combat implicit human cognitive bias, Herman introduces two research directions. The first is separation between descriptive and persuasive tasks. Explanation complexity can be considered a persuasive strategy, as in which the explanation is projected into an interpretable feature space, and the explanation is then altered to become more persuasive. Then, we incorporate user preferences and expertise into the explanation. This allows us to alter the explanation in the final step depending on different users and applications, and allows for flexibility based on human interpretability.

The second direction is explicit inclusion of cognitive features and expertise, allowing human cognition to influence the measures of interpretability. However, it is unclear how one would be able to first measure the knowledge of an individual or a group to inform decision-making based on expertise.

Human cognition can include user conviction, or trust in a person's own judgement of a classification model above that of an interpretable explanation. [23]

Manuel Blum, Sampath Kannan. Designing Programs That Check Their Work

In this paper, Blum defines a program checker, that checks its work. He distinguishes between checking and verification, which has to do with formal proofs of correctness. This is more expensive. Checking is easier to do, as it verifies that a given program returns a correct answer on a given input, rather than on all inputs.

This could be solved by checking if the checker C is correct, which is sometimes easier than proving the original program correct. Although the paper does not deal directly with this idea, it presents it as an option. Otherwise, we can also try to make the checker independent of the program it checks, in which case C has the little oh property with respect to P if and only if the expected running time of C is little oh of the running time of P. This ensures that the checker is programmed differently from the program it checks. Little oh is a strict upper bound on time and space complexity.

Let π denote a decision or search problem. For x, an input to π, let $\pi(x)$ denote the output of π. Let P be a program for π that halts on all instances of π. P has a bug if for some instances x of π, $P(x) \neq \pi(x)$.

An efficient program checker C_π for a problem π is defined as follows:

$C_\pi^P(I;k)$ is any probabilistic oracle Turing machine that satisfies the following conditions for any program P that halts on all instances of π, for any instance I of π, and for any positive integer k, which is the security parameter:

1. If P has no bugs, $P(x) = \pi(x)$ for all instances x of π, then with probability greater or equal to $1 - \frac{1}{2}^k$, $C_\pi^P(I;k) = CORRECT$.

2. If $P(I) \neq \pi(I)$, then with probability greater or equal to $1 - \frac{1}{2}^k$, $C_\pi^P(I;k) = BUGGY$.

This system of program checking allows for the possibility of an incorrect answer because of its probabilistic interactive protocol. This allows a program to have problems with computation, just as if the computations were done by hand. [7]

3 Construction

We have three goals for this construction:

1. This construction does not require interpretability to be a "one pass" operation in which an algorithm is devised and immediately interpretable. Instead, the act of interpreting is interactive between the auditor and the algorithm that builds the classifier. This algorithm may be invoked many times during the interaction.

2. The interrogation of the algorithm will be on increasingly simpler datasets. This idea has its roots in complexity theory.

3. The interrogation involves having a prover guess the values of two different classes, as described below.

The construction also allows the legal system to define interpretability how it chooses, and allows for user discretion and subjectivity in defining fairness and interpretability.

In this protocol, the prover is trying to prove their algorithm interpretable, and attempting to illuminate underlying decisions in the decision tree. The algorithm defined below attempts to understand specific classification decisions at the discretion of the auditor.

Before defining the construction, we will explain what it does. The construction begins with a dataset and an algorithm Gen that generates classifiers. In our application section, these classifiers are decision trees.

Gen generates a classifier that needs interpreting. The auditor uses her discretion to locate a spurious rule in the classifier that needs further interpretation. She then gives this rule to the prover, and using Gen generates a new classifier trained on that rule. If the prover can distinguish between the classes in the new classifier, it is interpretable, but also reveals the inner workings of the original classifier. This sheds light onto the original classifier in ways we will discuss in the conclusion.

This method of interacting with an interpreter on an uninterpretable model works when starting with a decision tree. The decision tree returns a classifier using simple rules. If there are two groups that one would guess should be treated the same but the classifier gives them different outcomes, it might warrant interpreting. The protocol restricts the input space to the union of those two cases and finds another decision tree based on this new space. The secondary decision tree attempts to understand what made the treatment of these two groups different.

Finding these groups that need interpreting is subjective. But it narrows the problem, simplifying it so that an interpreter can look at a simpler problem.

At the end of the interactive process, we hope to understand how our classifier worked on a particular split, and to audit this procedure for interpretability. We do this because in the legal system, one would need to be able to understand how the algorithm reached a specific decision, rather than how the model as a whole works.

Examples

Imagine a situation in which a machine learning algorithm predicting violent behavior finds that the most predictive feature is whether the person's jacket is red or not. While there is no immediate explanation for this feature being predictive, since it is highly predictive based on the algorithm, it is used. Is this fair? The jacket could be correlated with race. Or, red jacket colors could go out of style and then would be incorrect. But it could also be predictive of gang membership whose gang color is red, in which case the model picked up on something useful and fair.

Using red jackets without understanding why, but with the knowledge that red jackets are predictive, is a world in which the outcome matters more than the procedure. However, in a legal system that requires auditing and interpretation for the sake of fairness and justice, this will not be sufficient. Using this model in a legal system would additionally require understanding why the model penalizes red jackets.

Another example is the one we will discuss in the application section. Let us say we have an algorithm, the COMPAS algorithm, which makes predictions trained on recidivism. We have two classes: the defendant did recidivate, and the defendant did not recidivate.

The auditor looks at this decision tree and finds a split between people younger than 22.5 and people older than 22.5. This seems arbitrary or at least uninterpretable to the auditor. This split occurred for people who had fewer than 2 priors, and who were younger than 25.5. Restricting our dataset to these people, the auditor makes a new decision tree in which the tree is trained on two classes: people younger than 22.5, and people older than 22.5 but younger than 25.5.

The auditor then notifies the prover, who based on this secondary tree, tries to predict when a person is younger than 22.5, and when they are older on the decision tree (see Game Example section). This provides us insight about the types of people in the dataset and why the original decision tree made the decision it made to split on that uninterpretable value.

Through this protocol we are able to see which features made their way into the original decision tree without being explicitly in the decision-making system of the tree. For example, after training on age, we see race in the tree, implying that age and race are correlated in the data set, and therefore questioning the inclusion of age in the input space.

Overview of the construction

We have created an interactive protocol to understand why the model makes the decisions it makes. We single out uninterpretable decision splits and focus on training on these splits, creating "simplified" decision trees. We choose to use decision trees because they are commonly accepted as the most interpretable models. [8] These decision trees are not necessarily more interpretable as independent decision trees, but they attempt to understand a simpler question, or a single decision split, rather than the original prediction of predicting the original true label. We are therefore satisfying our second goal of narrowing our question to something simpler so that we can interpret a simpler problem.

This proof is interactive, between an auditor and a prover. The auditor is attempting to audit the model in question by asking the prover questions about its interpretability. The auditor sends to the prover an unlabeled decision tree, and the prover tries to predict which nodes correspond to which labels. In the analysis section below, the prover would try to predict, given two labels (recidivated versus did not recidivate), which class belongs to the class of people who recidivated, and which belongs to those who did not. If the prover can correctly do so for a certain number of rounds, the auditor is satisfied that the model is interpretable.

When narrowing in on a specific split, we restrict the input space to all data points that fall into the subtree. Since the new decision tree now trains on the label of the decision split, this leads to an entirely different tree without the expectation that it looks at all like the original tree.

The Gen function generates a decision tree based on data points that all either have labels y_0 or y_1. In the case of our application, the Gen function would generate a decision tree with inputs that are defendants who either recidivated (y_0) or did not recidivate (y_1).

Proving an algorithm interpretable:

Compute M = Gen(y_0, y_1)
Do k times:

Identify a decision split that needs interpreting, splitting the restricted input space based on a new label, z. z is the class of the decision split, which the new decision tree will train on.

Create a new model M' trained on that decision split, $M' = Gen(z_0, z_1)$, where z_0 and z_1 are the two values z can take on.

If the prover is able to correctly match the classes with the nodes on the decision tree, M is INTERPRETABLE. Else return UNINTERPRETABLE.

Theorem: The program interpreter runs efficiently and works correctly as specified.

Let M be any decision tree that halts on all inputs and always outputs 0 or 1. Let y_0 and y_1 be any nodes. Let k be a positive integer.
If M is interpretable, then I will definitely output INTERPRETABLE.
If M is uninterpretable, then

$$\Pr[I(M(y_0, y_1)) = \text{INTERPRETABLE}] \leq (\frac{1}{2})^k$$

Assumptions:

If M is fully interpretable, then the algorithm constructs an interpretable subtree M' and the prover correctly outputs INTERPRETABLE.

If M seems interpretable but its subtree M' is not, then the prover has a $\frac{1}{2}$ chance of proving it interpretable anyway for each round of the game.

If M is an uninterpretable algorithm, then the probability that the algorithm outputs INTERPRETABLE is at most 1/2 because the prover can only do so randomly. Since the class distribution is equal (there are two equal possibilities for classes), the prover correctly assigns the classes to the colors by chance.

The Further Research section discusses the guarantees that would be needed for this construction.

4 Analysis

The Game Example section of the Analysis provides an application for the construction described above.

4.1 Understanding the Data

Our analysis is based on a database of 2013-2014 pretrial defendants from Broward County, Florida, provided by ProPublica.

We are given the following features:

- Name
- Compas screening date
- Sex
- Age
- Race
- Juvenile felony count
- Juvenile misdemeanor count
- Juvenile other count

Labels provided in the dataset include:

- decile score: the COMPAS score
- is violent recid
- v decile score: violent recidivism score by COMPAS

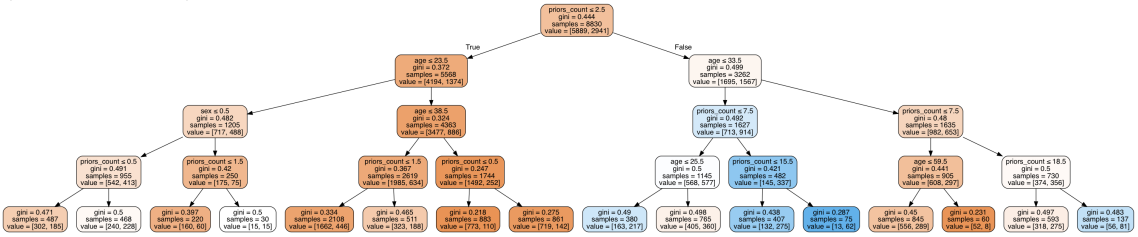
We first reverse engineered the COMPAS algorithm using the labels and decile scores provided in ProPublica's analysis and GitHub. Our purpose was to create models trained on the true recidivism labels (whether the defendant recidivated or not) and compare it to the decile score labels provided by the COMPAS algorithm. We include race and sex,

protected attributes, to see how the model learns. In learning a decision tree, we do not see race, but we do see proxy variables and will examine this below. We also then removed race from the dataset to see how this would change accuracy and interpretability.

We first compared a decision tree using the recidivism data (where the label is whether or not each defendant recidivated) to the decision tree using the decile score (where the label is the score that the COMPAS algorithm provided.

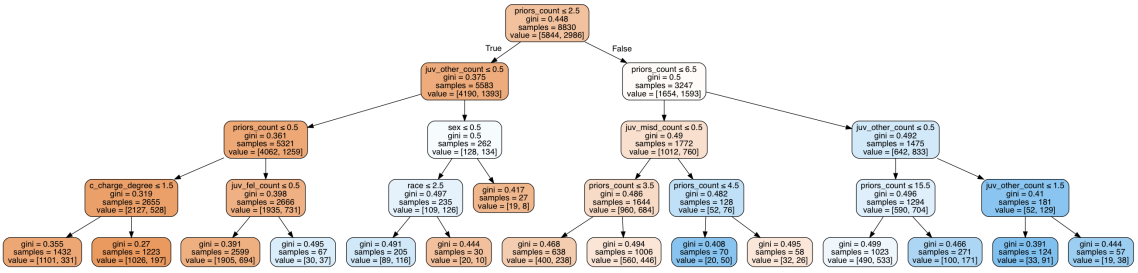
It is worth noting that using the label "is recid" is problematic for a few reasons. For one, it is apparent that black people are policed at far higher rates than are white people. Therefore, the recidivism rates for black people, ceteris paribus, are bound to be higher. For another, we do not know of the people who were jailed and would have recidivated had they not been jailed. We also do not know of the people who were not jailed, reoffended, but were not caught. Therefore, these numbers are not perfect, but are the best we can do.

The following is the decision tree trained on whether or not the defendants recidivated (the true label).

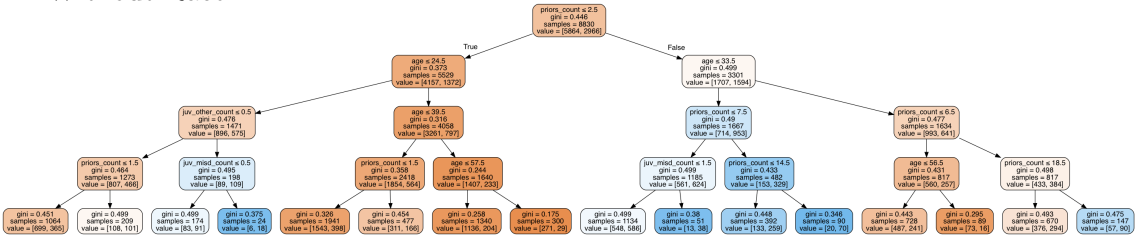


This was controlled to have a maximum depth of 4. The following trees are also trained on whether the defendants recidivated (the true label), but with age, race, or sex removed from the dataset.

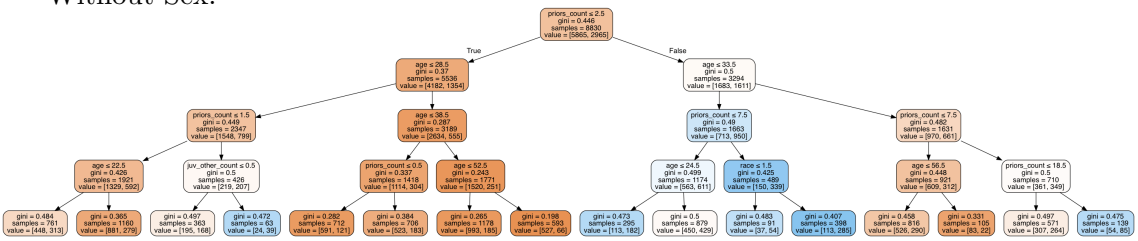
Without Age:



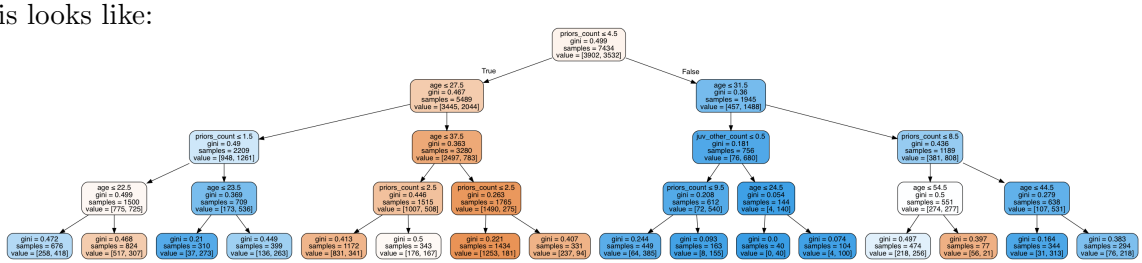
Without Race:



Without Sex:



We will be using the decision tree trained on whether or not the defendants recidivated, with all features, but only looking at white and black people for clarity. We will also remove the charge degree "other," keeping just those who committed felonies or misdemeanors. This looks like:



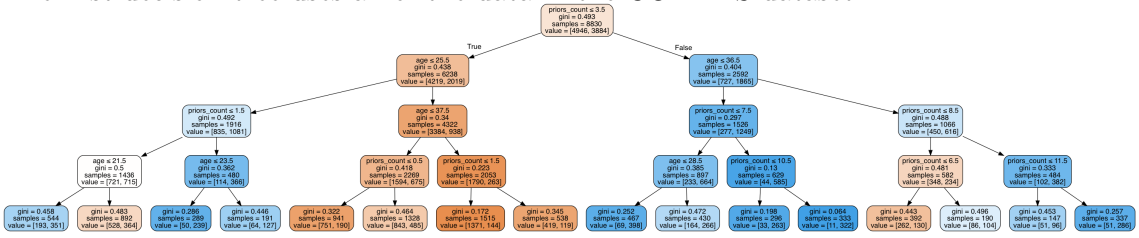
Accuracy: 0.677038626609442

Table 3: Recidivism as label, only black and white for race

	precision	**recall**	**f1-score**	**support**
0	0.71	0.87	0.78	1222
1	0.56	0.31	0.40	642
avg / total	0.65	0.68	0.65	1864

The next set of decision trees are trained on the COMPAS decile scores of each defendant. To make this a Boolean value, we assigned every decile score less than 5 to be 0, and every decile score greater than or equal to 6 to be 1, so that the data could be comparable to the previous decision trees.

The first decision tree uses all of the data in the COMPAS dataset:

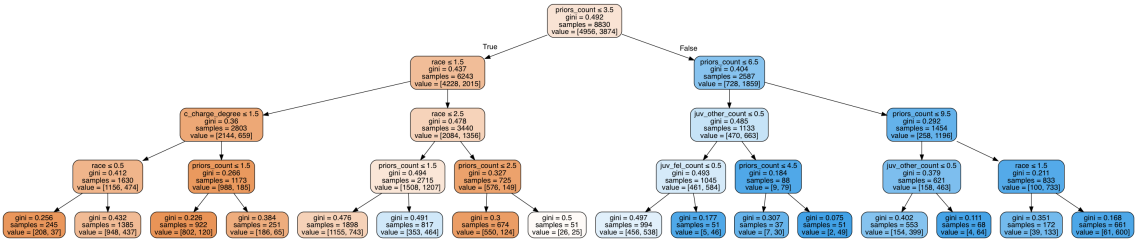


The accuracy is: 0.7586050724637681

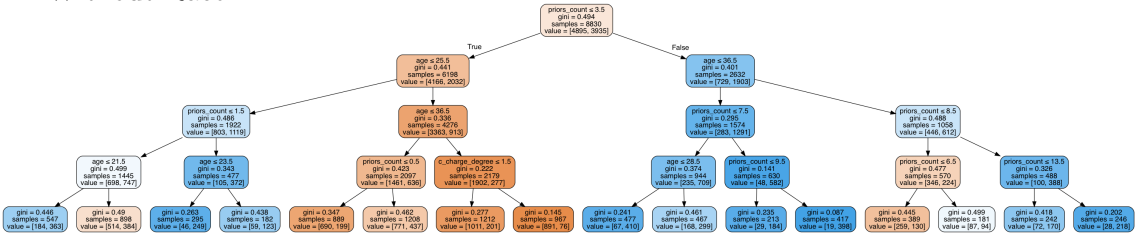
Table 4: COMPAS as label, all races

	precision	recall	f1-score	support
0	0.77	0.81	0.79	1218
1	0.75	0.70	0.72	990
avg / total	0.76	0.76	0.76	2208

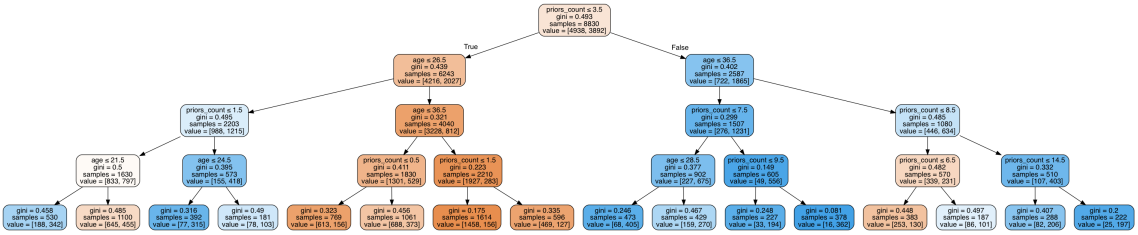
Without Age:



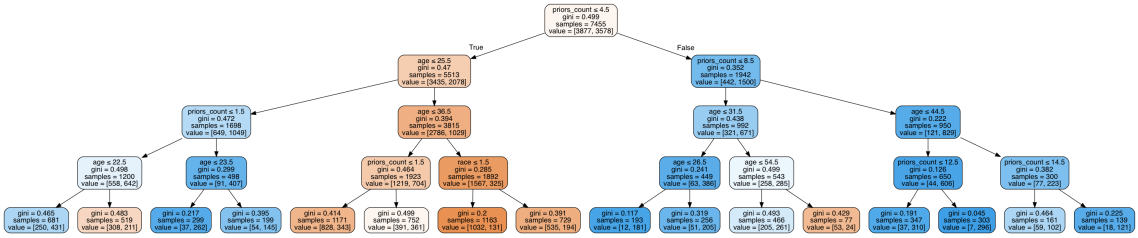
Without Race:



Without Sex:



We will be using the COMPAS decision tree using all of the data, except only using white and black for clarity. This looks like:



Accuracy: 0.7398068669527897

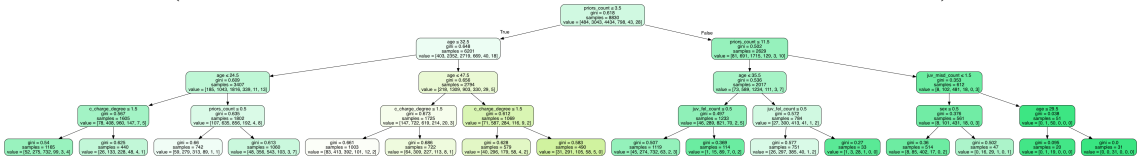
Table 5: Decile score as label, only black and white for race

	precision	recall	f1-score	support
0	0.73	0.81	0.77	992
1	0.75	0.66	0.70	872
avg / total	0.74	0.74	0.74	1864

Let us now try to understand our data. Would a decision tree tell us something about race?

Let us try to train on race, that is, our label will be race and we will see how well a decision tree can distinguish between the races.

All races (Other, Caucasian, Black, Hispanic, Asian, Native American):



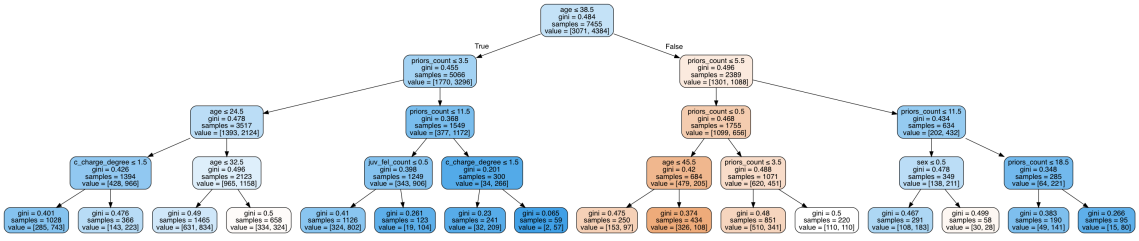
Accuracy: 0.5380434782608695

Confusion matrix

Table 6: Predicting Race (all races)

	precision	recall	f1-score	support
0	0.00	0.00	0.00	124
1	0.47	0.34	0.39	778
2	0.56	0.84	0.67	1101
3	0.00	0.00	0.00	186
4	0.00	0.00	0.00	13
5	0.00	0.00	0.00	6
Avg/total	0.44	0.54	0.47	2208

This model is not great at predicting races other than white and black. When just training on those:



Accuracy: 0.6314377682403434

Table 7: Predicting race (black and white)

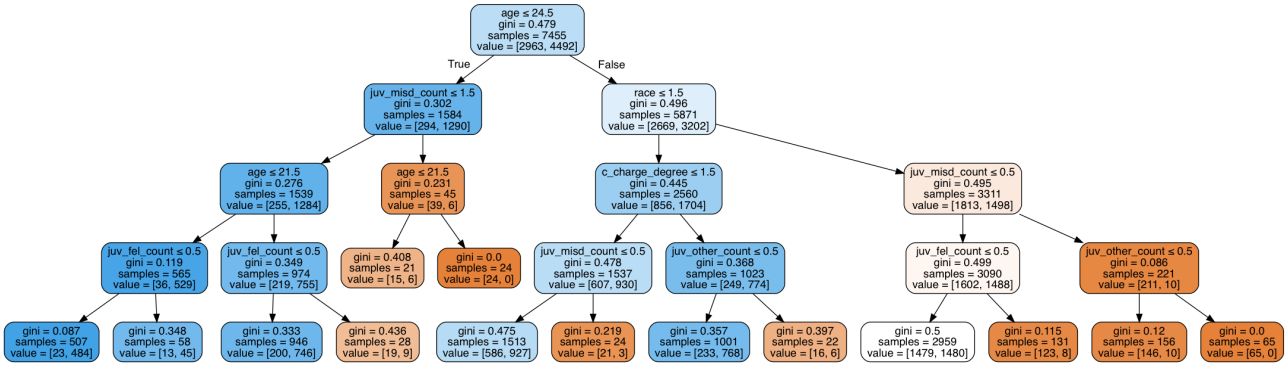
	precision	recall	f1-score	support
0	0.59	0.34	0.43	766
1	0.65	0.83	0.73	1098
Avg/total	0.62	0.63	0.61	1864

This has better accuracy, but still not great. Note the large difference in recall (true positive) between black and white defendants, as well as the F1 score.

Young people and those with high priors counts are often black (blue boxes) while those with lower prior counts and/or those who are older are white.

This distinction suggests that priors count and age can act as proxies for race, perhaps. Let us now train our model with priors count as our y label (still only looking at black and white people):

Accuracy: 0.6496781115879828



Where black is given the label of 2, and white is given 1.

Table 8: Priors count as label (black and white only)

	precision	recall	f1-score	support
0	0.87	0.16	0.27	754
1	0.63	0.98	0.77	1110
avg / total	0.73	0.65	0.57	1864

4.2 Analytical Application

In this section we see how our construction would be applied with real data. First, we train a decision tree on the true label. Then, the auditor locates problematic portions of the decision tree, focusing on one specific decision split. Then, a new decision tree is created

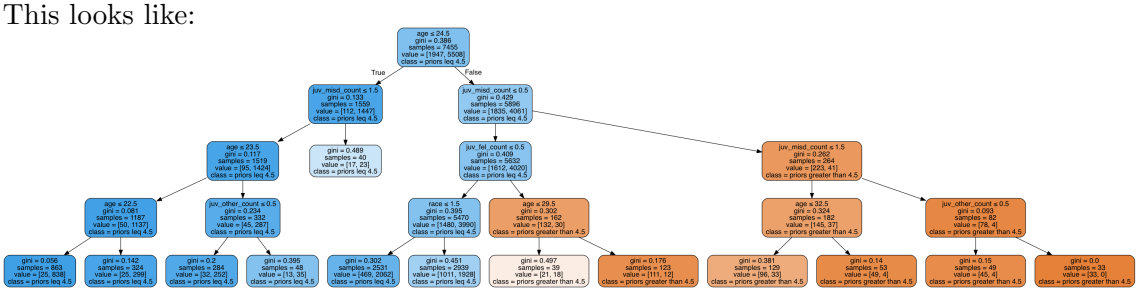
training on that decision split, instead of on the true label. It is the task of the prover to be able to interpret that new decision tree.

To avoid overfitting, we set certain parameters on our decision trees. We have a random state set to 100, and a minimum number of samples per leaf of 20.

The first split in the original COMPAS decision tree is the priors count. If the priors count is less than or equal to 4.5, they are immediately less likely to recidivate.

To analyze this in our protocol we instead train on priors count. Our label is a binary, where 0 is when $priorscount > 4.5$ and 1 is when $priorscount \leq 4.5$.

This looks like:



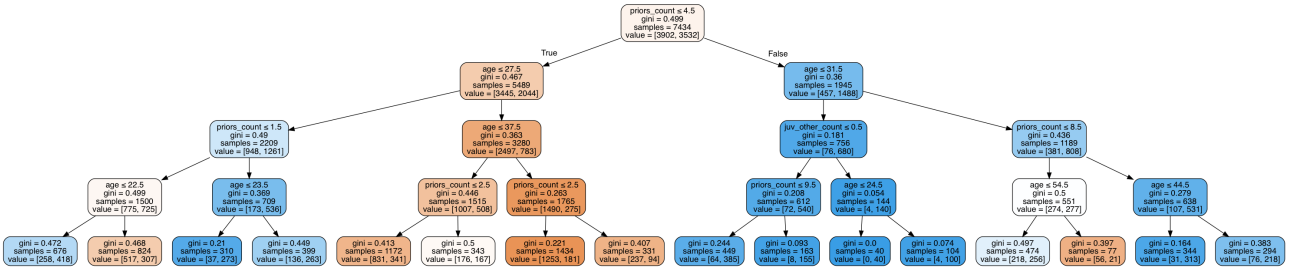
The accuracy is 0.7795064377682404.

Table 9: Priors count as label (black and white only)

	precision	**recall**	**f1-score**	**support**
0	0.86	0.18	0.30	484
1	0.77	0.99	0.87	1380
avg / total	0.80	0.78	0.72	1864

Robert Cannon, a 22-year-old black male, never committed a crime in his life. He was arrested in 2013 for petty theft, but was given a decile score of 6 by COMPAS: he is at medium risk for recidivating.

Let us go back to our original decision tree, trained on the COMPAS decile score.

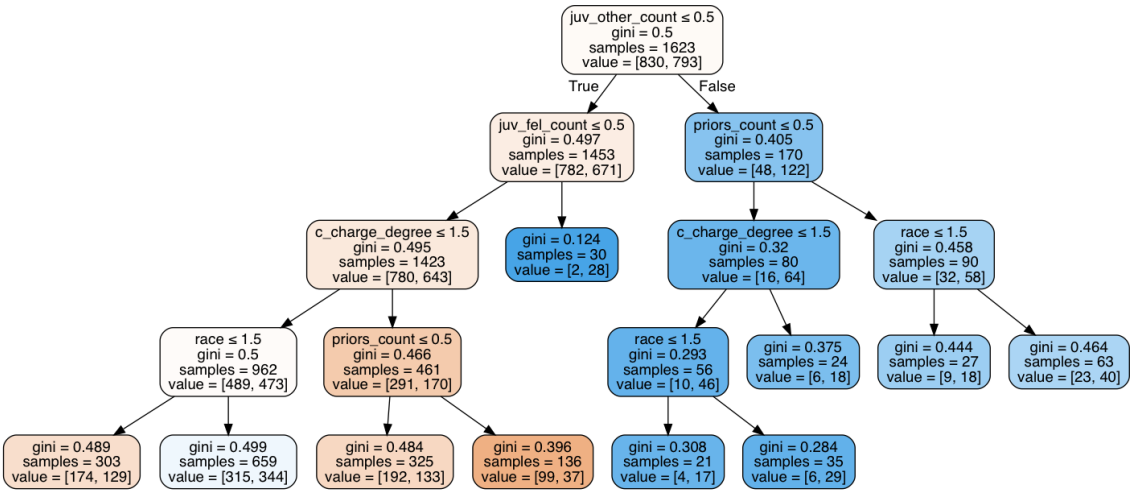


We see this on the decision tree: someone who has a priors count of 0 could be given a high risk score if they are 22.5 years old or younger.

Therefore, the last two leaves on the left seem to need interpreting. What is it about 22 year olds that make them riskier? Is this a fair decision tree split?

We now assign people from 22.5 to 26.5 the label 0, and those younger than 22.5 the label 1. The decision tree still trains on all other predictors except for age and the decile score of the original classifier. It is only trained on the subset of data points that fall into the subtree we are considering.

This provides us with the following decision tree:



The accuracy of this tree is 0.591133.

The matrix is:

Table 10: Leftmost leaves

	precision	**recall**	**f1-score**	**support**
0	0.62	0.62	0.62	217
1	0.56	0.56	0.56	189
avg / total	0.59	0.59	0.59	406

Lastly, looking at decision trees with different depths, and more complex decision trees with more features, can gives us an idea about how interpretability would look with larger, more complex models.

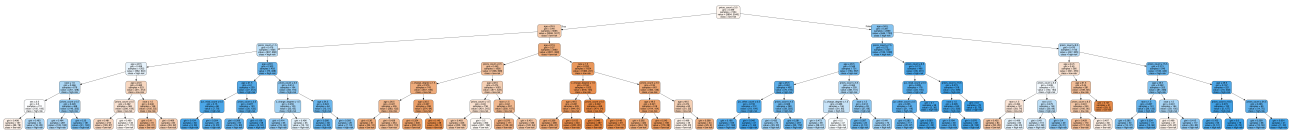


Table 11: Training on recidivism, Depth = 6

	precision	recall	f1-score	support
0	0.71	0.84	0.77	960
1	0.79	0.64	0.71	899
avg / total	0.75	0.74	0.74	1859

Since the auditor must find uninterpretable groups, larger decision trees may be difficult to interpret in this protocol. This tree has a depth of 6, with the same number of features, eight.

Although this does not affect accuracy very much (0.7444862829478214). But it is easy to see that with a greater number of samples, larger numbers of interpretability rounds can be played, and the algorithm can be more fully audited.

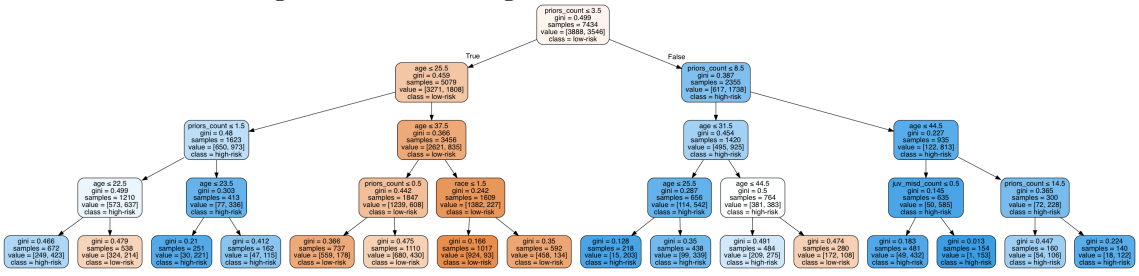
4.3 Game Example

Let us first clarify our variables:

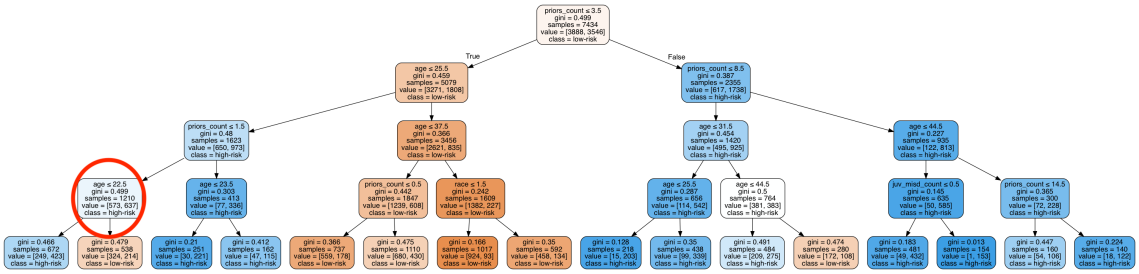
- We only keep black and white defendants. White defendants are given a label of 1, and black defendants are given 2.
- Charge degree: felonies are given 1, and misdemeanors are given 2.
- Recidivism: the true label

When we show the colors of the decision tree to the prover, there are only two colors that they can choose between, because there were only two labels. The shades on the decision tree correspond to the gini score, which is negligible for the game.

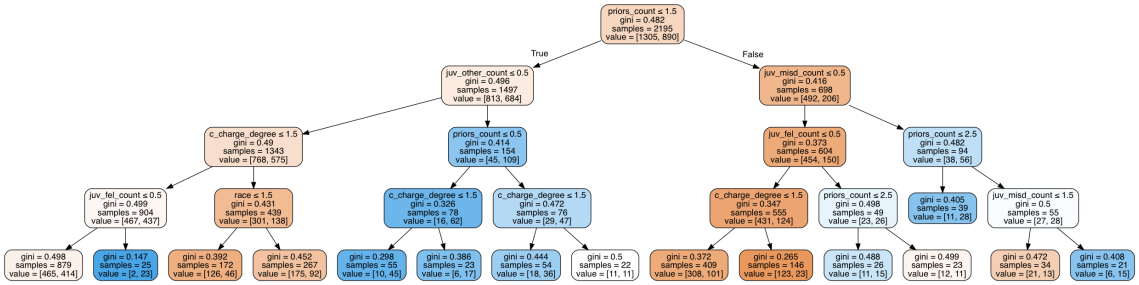
We start with the original COMPAS algorithm:



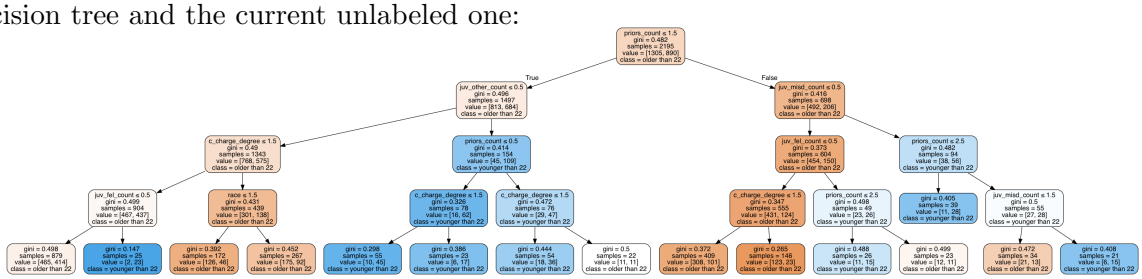
The auditor then sends the prover a decision split that needs interpreting:



And creates a model trained on that split:



The prover attempts to guess which class corresponds to which color, given the original decision tree and the current unlabeled one:



The prover has a $\frac{1}{2}$ chance of guessing it right each time, for each of k games. The auditor can now see race as a determining factor in the decision tree. Although this is not necessarily unfair (in this case it seems to be unfair), it helps illuminate this, and allows us to determine if we want to use this classifier.

5 Conclusion

5.1 Interactive protocol

We propose a method to interactively interpret a black box model by identifying unclear decision splits and training on these splits to understand the underlying factors. Through this method, we are able to simulate a two player game in which an auditor and prover work together to interpret a classifier. This serves as a framework for auditing black box models because it exposes underlying proxies and uninterpretable decisions that, if found to be uninterpretable after n iterations, or unfair because of its use of protected attributes,

can be rejected as a model that we should use. The interactive protocol highlights potential biases in the model by using human judgment and machine interpretation in tandem.

Our game does not ask about the truth, it rather asks if the model learned something strange or uninterpretable.

The auditor does not need to initially determine ground truth about whether an algorithm is interpretable through this protocol.

It attempts to determine if a human is able to distinguish between a correct and incorrect decision tree. This provides a framework for auditing an algorithm: is this something a human can understand? Are the decision splits sensible enough for a human to reason around?

This interactive protocol therefore does not work for models that are not human-interpretable. But it has the ability to serve us well in a legal framework that requires accountability. An auditor must be able to determine if the procedure, not just the outcome, is fair and interpretable by judges. This procedure is where our interactive protocol could be useful.

Mapping our original decision tree to a latent space in which different nodes are compared to each other and latent models train on subclasses, we learn about the internal structure of the original model. We are able to understand why a model made the decision it did by looking at the individual decision and training on it.

Therefore, the protocol allows for added interpretability from the two player game, but also adds potential explanatory power from the interpretability algorithm.

This is a probabilistic protocol, which means that it fails within a certain probability. This probability is bounded by randomness: for each of the k plays, there is a $\frac{1}{2}$ chance the prover guessed. This leaves us with a probability of being correct of $1 - (\frac{1}{2})^k$.

5.2 Policy Recommendations

The question I set out to answer was whether interpretability of machine learning algorithms was necessary and/or sufficient for fairness, especially when the algorithm is used to directly impact human lives like in the criminal justice system.

My answer is that it is necessary but not sufficient, and can be extremely helpful if applied correctly.

Interpretability is just one part of the field of fairness and accountability in machine learning. But if an algorithm is not interpretable, it should not be used in the legal system or in systems that directly impact human decision-making processes in systems such as insurance or the justice system.

Therefore, this tool is only useful for auditing purposes. It is not guaranteed to prove that an algorithm is interpretable, or guaranteed to provide an explanation of the decision-making process of a black box model. It is, however, the bare minimum that an auditor should ask for in terms of interpretability.

From 1985 to 1987, six people died after using the Therac-25, which attempted to destroy tumors with minimal impact on surrounding healthy tissue. According Nancy Leveson, who investigated the accidents:

> "Most previous accounts of the Therac-25 accidents blamed them on a software error and stopped there. This is not very useful and, in fact, can be misleading and dangerous: If we are to prevent such accidents in the future, we must dig deeper. Most accidents involving complex technology are caused by a combination of organizational, managerial, technical, and sometimes, sociological or political factors. Preventing accidents requires paying attention to *all* the root causes, not just the precipitating event in a particular circumstance." [26]

Penn Professor Richard Berk says, "I'm not trying to explain criminal behavior, I'm trying to forecast it."[35] But we need to understand fairness in the broader context through which our software will operate if we are going to be forecasting criminal behavior. How can we declare that a recommendation is fair if we don't know what the conditions of the recommendation mean for the individual?

The idea of compounding injustice means that by the time a defendant gets scored by COMPAS, they have either benefited from or been disadvantaged by a lifetime of systemic discrimination and injustice. Perhaps this should also be taken into account, depending on our values of what we want the criminal justice system to do for society.

By using actuarial tools like COMPAS, we need to worry about whether we are legitimizing and codifying unfair practices. Are we undermining reformers who want sweeping policy changes while focusing on the fairness of risk assessment tools? How much fairness are we bringing to the system?

Furthermore, it is an unquestioned assumption that those with high risk scores deserve to spend more time in prison, and a further assumption that this is better for society. Kalief Browder, and many others, were clearly not harmful or risks to society, and prison often has a negative impact on their lives. Is the prison system designed to reform criminals, punish them for crimes, or neither? In the case of Kalief Browder, and many others stuck in prison awaiting their trials because they could not post bail, young black men are punished for crimes they did not commit, and even if they did commit them, does anybody deserve to spend time in the conditions that we see in prison? Our thoughts on this issue should be influenced by the fact that 20% of pretrial detainees are not charged with anything and are awaiting trial.

This paper has shown how hidden biases can exist in algorithms. This implies that algorithms require a human component because of inherent algorithmic bias and because of the interrogations needed for these systems. The people who write the algorithms should be held accountable to greater transparency and laws that control the unbridled power that technocrats see today.

To ensure due process, individuals must be aware of the algorithms helping inform

the length of their sentence and must be reasonably satisfied that the algorithms are not enforcing racial bias, through a system of auditing and interactive interpretation.

5.3 Areas of future study

Further research includes preventing spurious interpretations. There are no guarantees on accuracy preservation through composition. Performing this interactive protocol may limit the input space so severely that the accuracy makes the classifier unintelligible. What happens when we create a secondary decision tree that is at odds with the previous decision tree?

We must find a way to force guarantees about interpretability and/or accuracy. We lose accuracy when we restrict our data set to a subtree of points. We must also determine guarantees for cases in which this protocol works.

In the field of machine learning and human interaction, we can also apply this interactive protocol to models other than decision trees, expecting that this would work similarly. Instead of using decision splits, there is extensive literature on rule extraction methods that could be used for less human-interpretable models. We can see how interpretability works when applied to human behavior, by administering our protocol on real people. This could be done by playing the interactive game on Amazon Turk, or with judges.

To delve into statistical robustness, it may also be useful to re-evaluate the COMPAS algorithm without case-wise deletion on charge degree being "other," and other such deletion that ProPublica and I performed. Re-evaluation would take place with imputation or multiple imputation of the data. We can also look at the protocol for data that has more than two potential classes.

5.4 Implications

The original hypothesis of this paper was that proving the interpreted decision tree interpretable is often easier than proving the original decision tree is interpretable. This is not necessarily true: the complexity of an underlying decision split is independent. However, the protocol still allows us to gain insight on how decisions were made in a model. Although interpretability is not all one needs, it is a necessary part of the fairness of certain systems like the justice system.

Interpretability does not imply fairness. But it can be used to spot unfairness in a system. This is not sufficient or perfect but it should be one of many tools in an auditing toolkit.

References

[1] Alexander, Michelle. The new Jim Crow: Mass incarceration in the age of colorblindness. The New Press, 2012.

[2] Alexander, Michelle. "Go to Trial: Crash the Justice System." The New York Times, The New York Times, 10 Mar. 2012, www.nytimes.com/2012/03/11/opinion/sunday/go-to-trial-crash-the-justice-system.html.

[3] Angwin, Julia. "Make Algorithms Accountable." The New York Times, The New York Times, 1 Aug. 2016, www.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html.

[4] Angwin, Julia. "Making Algorithms Accountable." ProPublica, www.propublica.org/article/making-algorithms-accountable.

[5] http://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf

[6] Bunting, W. C., Lynda Garcia, and Ezekiel Edwards. "The War on Marijuana in Black and White." (2013).

[7] Blum, Manuel, and Sampath Kannan. "Designing programs that check their work." Journal of the ACM (JACM) 42.1 (1995): 269-291.

[8] Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical science 16.3 (2001): 199-231.

[9] Calaway, Wendy, and Jennifer Kinsley. "Rethinking Bail Reform." (2017).

[10] Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

[11] Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." arXiv preprint arXiv:1703.00056 (2017).

[12] Cohen, Thomas H., and Scott W. Van Benschoten. "Does the risk of recidivism for supervised offenders improve over time: Examining changes in the dynamic risk characteristics for offenders under federal supervision." Fed. Probation 78 (2014): 41.

[13] Dressel, Julia, and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism." Science Advances 4.1 (2018): eaao5580.

[14] Dwork, Cynthia, et al. "Fairness through awareness." Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ACM, 2012.

[15] Eckholm, Erik. "Public Defenders, Bolstered by a Work Analysis and Rulings, Push Back Against a Tide of Cases." The New York Times, The New York Times, 18 Feb. 2014, www.nytimes.com/2014/02/19/us/public-defenders-turn-to-lawmakers-to-try-to-ease-caseloads.html.

[16] EPIC - Algorithms in the Criminal Justice System.? Electronic Privacy Information Center, epic.org/algorithmic-transparency/crim-justice/.

[17] FAT Conference 2018: Understanding the Context and Consequences of Pre-trial Detention. February 23, 2018. Elizabeth Bender (Decarceration Project at The Legal Aid Society of NYC), Kristian Lum (Human Rights Data Analysis Group), and Terrence Wilkerson (entrepreneur)

[18] Federal Bureau of Prisons. BOP Statistics: Inmate Race

[19] Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." arXiv preprint arXiv:1609.07236 (2016).

[20] Fortnow, Lance. "Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. SIAM journal on computing, vol. 18 (1989), pp. 186:208. Oded Goldreich, Silvio Micali, and Avi Wigderson. Proofs that release minimum knowledge. Mathematical foundations of computer science 1986, Proceedings of the 12th symposium, Bratislava, Czechoslovakia, August 25?29, 1986, edited by J. Gruska, B. Rovan, and J. Wiedermann, Lecture notes in computer science, vol. 233, Springer-Verlag, Berlin ...." The Journal of Symbolic Logic 56.3 (1991): 1092-1094.

[21] Guo, Jeff. "Black Defendants Suffer When a Judge's Favorite Football Team Loses.? The Washington Post, WP Company, 7 Sept. 2016.

[22] Harcourt, Bernard E. Against prediction: Profiling, policing, and punishing in an actuarial age. University of Chicago Press, 2008. Page 9-10.

[23] Herman, Bernease. "The Promise and Peril of Human Evaluation for Model Interpretability." arXiv preprint arXiv:1711.07414 (2017).

[24] https://www.hrw.org/news/2017/07/17/human-rights-watch-advises-against-using-profile-based-risk-assessment-bail-reform

[25] Katz, Jonathan, and Yehuda Lindell. Introduction to modern cryptography. CRC press, 2014.

[26] Leveson, Nancy G., and Clark S. Turner. "An investigation of the Therac-25 accidents." Computer 26.7 (1993): 18-41.

[27] Lipton, Zachary C. "The mythos of model interpretability." arXiv preprint arXiv:1606.03490 (2016).

[28] Lopez, German. "Why You Can't Blame Mass Incarceration on the War on Drugs." Vox, Vox, 30 May 2017, www.vox.com/policy-and-politics/2017/5/30/15591700/mass-incarceration-john-pfaff-locked-in.

[29] Meter, Matthew Van. "One Judge Makes the Case for Judgment." The Atlantic, Atlantic Media Company, 25 Feb. 2016, www.theatlantic.com/politics/archive/2016/02/one-judge-makes-the-case-for-judgment/463380/.

[30] "Criminal Justice Fact Sheet." NAACP, www.naacp.org/criminal-justice-fact-sheet/.

[31] Oleson, J. C. "Blowing Out All the Candles: A Few Thoughts on the Twenty-Fifth Birthday of the Sentencing Reform Act of 1984." U. Rich. L. Rev. 45 (2010): 693.

[32] Pretrial Justice Institute. "The State of Pretrial Justice in America." November 2017.

[33] https://www.prisonpolicy.org/reports/incomejails.html

[34] "The Path from Arrest to Pretrial Detention.? The Path from Arrest to Pretrial Detention — Prison Policy Initiative

[35] Popp, Trey. "Black Box Justice.? The Pennsylvania Gazette, 28 Aug. 2017, thepenngazette.com/black-box-justice/.

[36] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[37] Rotenburg, Marc. Message to the author. 29 March 2017. E-mail.

[38] Selbst, Andrew D., and Julia Powles. "Meaningful information and the right to explanation." International Data Privacy Law 7.4 (2017): 233-242.

[39] Schwirtz, Michael, and Michael Winerip. "Kalief Browder, Held at Rikers Island for 3 Years Without Trial, Commits Suicide." The New York Times, The New York Times, 8 June 2015, www.nytimes.com/2015/06/09/nyregion/kalief-browder-held-at-rikers-island-for-3-years-without-trial-commits-suicide.html.

[40] Salman, Josh, et al. ?Same Background. Same Crime. Different Race. Different Sentence.? Bias on the Bench — Sarasota Herald-Tribune Media Group — Sentencing, projects.heraldtribune.com/bias/sentencing/.

[41] Tan, Sarah, et al. "Detecting Bias in Black-Box Models Using Transparent Model Distillation." arXiv preprint arXiv:1710.06169 (2017).

[42] Wagner, Peter, et al. "Mass Incarceration: The Whole Pie 2018. Mass Incarceration: The Whole Pie 2018 — Prison Policy Initiative, www.prisonpolicy.org/reports/pie2018.html.

[43] Zeng, Jiaming, Berk Ustun, and Cynthia Rudin. "Interpretable classification models for recidivism prediction." Journal of the Royal Statistical Society: Series A (Statistics in Society) 180.3 (2017): 689-722.