ABSTRACTING INJUSTICE: AN ANALYSIS OF THE USE OF ARTIFICIAL INTELLIGENCE IN CRIMINAL JUSTICE

Vincent Huynh-Watkins

Computer and Information Science: Departmental Honors Thesis

May 2021

Acknowledgements

First, I would like to thank Professor Bryce Newell, for his tireless willingness to answer my obscure and silly questions about academic writing and academia in general, his superb copyediting skills, his helpful insight into a variety of subjects, and much more. Without his support, I am all but certain this thesis would never have come to fruition.

Also deserving of acknowledgement are my lovely family. My mom and dad, my sister Nora, and her partner Wyatt who have all listened to me go on about my thesis with varying levels of engagement and understanding (usually dependent on how lucid my thoughts are on a given day.)

I would also like to thank my wonderful partner, Katie, who has sat through many weekends of me being engrossed by my laptop, Microsoft Word, and various academic journals with nothing but support and understanding.

Additionally, I would like to thank my good friend Ashley Kim, who inspired me to pursue a thesis in spite of the fact that it was not a required part of my curriculum. Without her encouragement I never would have considered undertaking this task.

Finally, I would like to thank University of Oregon Computer and Information Science faculty Stephen Fickas and Kathleen Freeman, both of whom were instrumental in starting, approving, and encouraging this paper.

Table of Contents

Introduction

As artificial intelligence ("AI") has become commonplace in many aspects of life and society– often seen as a faster, more accurate, and less labor-intensive alternative to human cognition[1]–the use of AI in criminal justice systems has been a naturally occurring phenomenon. There are many potential applications of AI in criminal justice that may seem sensible, with the touted possibility to provide fairer outcomes and increased safety for society. For instance, AI-based facial recognition may help investigators and prosecutors solve previously unsolvable cases.[2] AI may also help law enforcement agencies predict criminal activity prior to its occurrence (known as predictive policing), which may help in resource allocation and targeting areas for increased policing, theoretically leading to reduced rates of crime (Lau 2020). Further, some researchers claim that AI algorithms can provide a more objective and complete analysis of the recidivism risk posed by convicted criminals,[3] therefore providing a better basis for sentencing to prevent repeat offenses and free those who are not repeat threats. These are just some of the potential uses of AI in the justice system, however they are seen as particularly promising applications of technology which have the potential to make society safer and fairer and are beginning to flood the world of criminal justice.

With that said, these technologies can, will, and do pose significant barriers to the pursuit of a truly fair and healthy justice system as well as a functioning democratic society. These barriers raise questions over our society's ethics and challenge beliefs over what we are willing to sacrifice for what some would like to label "safety."

In fact, the drawbacks to the use of AI in criminal justice are already coming to fruition. There are multiple known instances in which facial recognition has incorrectly identified a person as a suspect to a crime and lead to their wrongful arrest (Hill 2020; Williams 2020).

Predictive policing is rife with possibilities to introduce bias and prejudice, and it has been reasoned that this policing methodology could be considered as "treating people as guilty of future crimes for acts they have not committed and may never commit" (Asaro 2019, 1), which stands directly in opposition to criminal justice norms. Additionally, the use of AI algorithms in recidivism risk assessment poses due process and reliability concerns (Hillman 2019) and poses ethical questions about accountability and the parameters being used to determine outcomes.

Clearly these technologies are not silver bullets to solve the issues facing modern criminal justice and could exacerbate existing systemic inequalities. There are a host of ethical and legal questions society must face, such that we avoid further human rights violations as well as to ensure the conditions for a thriving society. This analysis will broadly examine the use of AI in criminal justice and look to conclude on the ways in which its use poses civil rights concerns as well as how we may proceed ethically within the legal systems which outlines our criminal justice. It will build on a host of current works which analyze three areas of AI use in criminal justice (facial recognition, predictive policing, and recidivism risk assessment) as well as in society at large. Work such as Kate Crawford's *Atlas of AI,* Peter Asaro's "AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care," the Honorable Noel. L Hillman's legal analysis, David Leslie's analysis of societal and ethical implications of facial recognition, and the University of Washington Value Sensitive Design lab's large body of materials have been analyzed along with many more. Additionally, I have conducted legal and legislative case analysis to contextualize the current moment in which AI applications are being litigated and legislated and give perspective into how these violations are playing out already—how they are affecting the lives of the marginalized and how continuing down this path only serves to worsen already deep-seated inequalities. Ultimately this is an intersectional and

interdisciplinary analysis of power, computer science, the criminal justice system, and where AI fits into this puzzle; what its role is and how it has shaped and will continue to shape wide-ranging outcomes in a variety of applications.

Methodology

The research approach for this paper was to do a significant literature review of current and important works which overlap and coincide with the subject of AI use in criminal justice. Specifically, literature that interfaces with similar issues to those addressed in this paper and literature that addresses relevant and adjacent topics (i.e., the history of race and criminal justice in the US) have been examined. Resources were accessed using University of Oregon library database resources, Google Scholar, and through the purchase of relevant books. Through the process I searched these databases for literature which addressed issues relevant to my work. Additionally, case studies in the legal and legislative systems have been reviewed in order to ground what can be a process (development and application of artificial intelligence products) that is often abstracted from the real-world consequences which these products have. Using prior works as well as these case studies in the three areas of focus, analysis has been done to conclude on the ethical implications of artificial intelligence use in criminal justice, what this means for our society moving forward, and how we might best proceed.

Background

The Application of AI Within Criminal Justice Contexts

Proponents of AI in criminal justice will list claims such as the ability for AI to overcome human errors and function as "experts," "increase the speed and quality of statutory interpretation…" and "…predict potential victims of violent crime based on associations and behavior..." among

other uses including facial recognition and DNA analysis (Rigano 2019, 3-8). In fact, in his paper for the National Institute of Justice, Christopher Rigano fails to even attempt to interface with the potential pitfalls of the use of AI in these domains, not even mentioning the capabilities that AI systems have to retain and produce bias. As highlighted in David Leslie's (2020) work "Understanding the bias of facial recognition," we must go further than just asking whether bias exists in AI systems, but rather we should question whether such technologies even stand on solid enough moral ground to merit use in certain domains in the first place. Leslie additionally provides evidence that it remains possible to alter the course of what is colloquially known as "Big Tech," making the case that not all hope is lost yet, but that swift action is needed to avoid potentially society altering negative effects (2020). In the following sections, different examples of how AI already is beginning to affect criminal justice will be summarized and discussed to provide reasoning and analysis for how society should proceed; what kind of legislation should be drafted and demanded to restrict the legality of AI use in criminal justice and protect citizens from the pitfalls of such use.

In an ethical discussion of the United States' criminal justice system, the first question that begs asking is what is the purpose of a justice system? Naturally this a broad question that results in a variety of answers. Some criminal justice theories include retribution, deterrence, and rehabilitation– with these different theories manifesting in drastically different outcomes. For instance, in a retributive justice the offender is removed from society and is therefore not a threat but simultaneously is incapable of being a productive member of society (Meyer 1968). Here I propose that the purpose of the United States criminal justice system *should* be to provide equal justice for all, and ultimately create a safer, fairer, and more humane society which rehabilitates and cares for those who are criminal offenders. Currently, however, the justice system in the US

is extremely dysfunctional; the rate of incarceration is vastly greater than that of any other nation (Wagner and Sawyer 2018) and yet crime levels are no better than similarly wealthy and stable nations.[4] This brings about a discussion of the goals of the criminal justice system and how they are being carried out. Is the justice system failing, or is it really fulfilling a covert goal? The stated goals of the United States Criminal Justice System are to "… begin to reduce crime of all sorts and to erase those social conditions associated with crime and delinquency-poverty, unemployment, inferior education, and discrimination" (Conaboy, Smith, and Snyder, n.d.). Fundamentally, though, the United States Justice System is currently operating on a model of extreme punitive justice. The use of AI to further this method could have serious negative effects regarding race and class. Use of any historical data presents the possibility for introduction of bias and, particularly when it comes to the justice system, there are significant concerns over data bias given the history of racism in America. In this context, it is crucial to understand and analyze the historical purpose of policing and criminal justice in the USA and how AI and machine learning (ML) might play into this historical framework.

In "The Racial History of Criminal Justice in America," Heather Ann-Thompson (2019) argues that the size of the current American criminal justice system is historically unprecedented as well as extremely racialized, with African Americans, Latinos, and Indigenous peoples being disproportionately represented in U.S. jails and prisons. None of this should be seen as revelatory, however, the main goal of her paper is to discuss the way in which American criminal justice has *always* been deeply racialized and why this can explain deep injustices that remain today as well as why prison populations rose when they did (such as after the outlaw of slavery and after the civil rights act passed.) Generally, Ann-Thompson highlights the way in which the United States (and its criminal justice) grew from conquest and plunder, using criminalization as

a technique to gain power over the people native to America as well as slaves of African descent. As this conquering and criminalization of Native Americans was occurring, the power of the nation was growing on the backs of African slaves. Prior to the abolishment of slavery, however, the prison system which we now are familiar with was relatively nonexistent. During this time, the nation's ideas of criminality and deviance were already beginning to be deeply racialized, with Native Americans seen as "savages" and African Americans seen as "… 'brutes.'" (Thompson 2019, 222). This lens lead to a collective sentiment that Brown and Black people needed to be "controlled and confined due to their innate and inherent criminal and deviant natures…" (Thompson 2019, 222). Upon the abolition of slavery, penal institutions "began immediately to fill with people of color in numbers well out of proportion with their presence in the population" (Thompson 2019, 222-223). Further, new laws were adopted which targeted newly freed Blacks—in fact these laws were claimed to be about crime control, a frightening parallel to the rhetoric surrounding data-driven practices being introduced into criminal justice today. Moving swiftly through history, we can find that the criminal justice system as we know it in the United States was birthed out of a desire to control the lives of Black and Brown people. From exceptions to the abolishment of slavery for convicted criminals to the war on drugs[5], the systems of policing and incarceration which makeup a significant part of American criminal justice have always sought to create conditions in which social mobility for Black and Brown people is limited. With this as a backdrop, it becomes clear to see that the integration of AI models which rely on historical data and preexisting structures built by the very institutions that have subjugated entire demographics can and will inevitably lead to ethical issues as well as civil rights violations. In the following sections, examples of these issues are highlighted and subsequently discussed.

Predictive Policing

In the criminal justice field of predictive policing, we face great ethical and practical questions. As noted in the work of Peter Asaro, this method of policing may in fact lead to a situation in which people are treated as "guilty of (future) crimes for acts they have not yet committed and may never commit" (Asaro 2019, 1). In "Predictive Policing," a report by Sarah Brayne, Alex Rosenblat, and Danah Boyd (2015) predictive policing is separated into two categories: location-based and person-based. In the former, police may increase their presence in a certain area due to prior incidents. In the latter, individuals are targeted as "most likely to be involved in crimes, either as victims or offenders." (Brayne et al. 2015, 3) This leads to fairly evident ethical and civil rights ramifications. Those who are most at risk according to potentially inaccurate and biased historical data are subjected to greater levels of policing, regardless of whether they genuinely pose a threat. In fact, some who pose no threat and are most *at risk* of being affected by a crime will be subjected to higher surveillance and policing. Higher levels of policing produce an environment where the likelihood of being introduced to the justice system increases; runs ins with the law will inevitably increase and a criminal record will begin, leading to higher likelihood of punishment and yet greater surveillance. Rather than allocating resources towards caring for those at risk, we increase policing and therefore introduce yet more people into the prison system – all at an astounding cost of $80 billion per year (Kearney et al. 2014, 13). This can be viewed as taking what Asaro dubs the "Models of Threat" approach, in which we begin "from the assumption that the world can be classified into clear categories, *i.e.,* threats and non-threats…" (Asaro 2019, 3). As argued by Asaro, there is mounting evidence that this approach has very little positive effect on crime and subjects some to greater levels of privacy invasion as well as greater chance of police encounters[6]. By using these kinds of methodologies,

we are implicitly agreeing that the safety and security of some is more important than the civil liberties and freedom to remain undisturbed by law enforcement of others. Additionally, there is ample space for serious feedback loops to present themselves. As Brayne, Rosenblat, and Boyd discuss (2015), predictive policing can and does lead to new data creation in areas which are flagged for increased policing by predictive models. This increased collection and creation of data then, in turn, supports increased policing in that area and so on. In *Atlas of AI,* Kate Crawford (2021) argues similarly that by using machine learning in justice systems, a feedback loop is constructed in which those who have are introduced to the criminal databases will be surveilled at higher levels. Those surveilled at higher level, Crawford reasons, will have higher likelihoods of information about them being included in crime databases, which in turn will justify yet greater police scrutiny(2021). This is a scary and very real situation which is beginning to be carried out nationwide.

Along with this, data can only reflect what has been historically collected. As previously discussed, in a criminal justice system that has long sought to imprison and control Black and Brown people, it is relatively self-evident then, that predictive policing models would continue to target areas predominantly inhabited by Black and Brown people. Here we encounter a tool which can be seen as abstracting harm. In the case of machine learning, algorithmic, or data driven predictive tools, it is quite often the case that proponents will argue that the results are "just math" or a function of algorithmic statistics which are absolute and unbiased. For instance, a 2016 Wired Article by Oren Etzioni titled "Deep Learning Isn't a Dangerous Magic Genie. It's Just Math" claims exactly what the title would indicate: that deep learning (and machine learning generally) is not dangerous, but rather "simple math on an enormous scale" (Etzioni 2016). The truth of the matter is that ML can be dangerous in holding bias and that bias (an idea that is

difficult to define and quantify) is exceedingly difficult to sniff out in practice making machine learning all the more dangerous. Machine learning algorithms are imbued with countless design decisions–frequently made by anonymous engineers– *and* these algorithms can pick up patterns in data which humans cannot. This can mean that, for instance, in a predictive policing algorithm which has had any race-related data removed from training, the machine learning model could still pick up on race-sensitive factors which may not be evident to humans.[7]

In this case of predictive policing, is entirely possible that we would see broad communities affected, with already 56% of the US incarcerated population being represented by African Americans and Hispanics despite representing only 32% of the US population ("NAACP | Criminal Justice Fact Sheet" n.d.). Add to this already alarming statistic the possibility for a further feedback loop and we see that this is a serious emerging threat to the civil rights of a population which has already seen significant subjugation throughout the history of the United States. As put in "Dirty Data, Bad Predictions" by Richardson, Schultz, and Crawford (2019, 41):

> …it supports a wider culture of suspect police practices and ongoing data manipulation. Add to this the lack of oversight and accountability measures regarding police data collection, analysis, and use, and it becomes clear that any predictive policing system trained on or actively using data from jurisdictions with proven problematic conduct cannot be relied upon to produce valid results…

In their analysis, Richardson, Schultz, and Crawford (2019) examine various case studies (similarly to the analysis of Asaro (2019)). From examining cases of predictive policing in Chicago, New Orleans, and Maricopa county which relied on what they label "dirty" data (data that has been derived from policing practices which created "juked" stats in which police

departments have potentially falsified crime statistics) Ultimately they conclude that using dirty data generated by suspect police departments will create distorted predictive policing models which in turn will elevate risks of "creating lasting consequences that will permeate throughout the criminal justice system and society more widely. (Richardson, Schultz, and Crawford 2019, 48)

In "Predictive Policing and the Platformization of Police Work," Simon Egbert (2019) argues that the trend towards using predictive policing models and algorithms will foster an environment which "…intensifies the need for surveillance techniques and practices… in order to gain actionable intelligence that will allegedly make it possible to fight crime effectively– crime that, in some cases, has not even happened" (Egbert 2019, 87) This combination should be seen as a horrifying prospect to all who cherish privacy and justice. In this system which prioritizes incarcerating Black and Brown people, increased pressure to surveil for the purpose of creating data driven decisions could easily generate uncontrollably biased data which mirrors historical inequalities and targets already downtrodden areas and disenfranchised communities of color. This combination of police departments which lack oversight regarding their data collection practices and pressure for more and more data will conceivably have significantly negative effects in the pursuit of a fair and honest justice system. As analyzed by Crawford (2021), we are entering a period in which the tech industry sees data as there for the taking and exploiting, from mug shots in the NIST Multi Encounter Dataset to the ever growing trove of crime history data which police departments possess, control, and create. Seen in this light, the machine is further grotesque and frightening. As a society we have built a deeply ingrained state of policing which now has awoken to the fact that its data is valuable and is hurtling headfirst at harnessing it, with little consideration for the effects. This is what Cathy O'Neill (2017) labels a

"weapon of math destruction" (or "WMD") in which a model goes onward in its predictions with no correctional inputs resulting in dangers and damaging feedback loops. In fact, if the data being fed to these models is further warped by the unethical practices of police departments, then we might have a super-WMD on our hands.

Recidivism Risk Assessment

On the topic of recidivism risk assessment algorithms, I will first discuss the legal case *Wisconsin v. Loomis* (2016) and the precedent it may set for AI recidivism risk assessment tools, as well as the ongoing case, *Henderson v. Stensberg,* which provides an interesting contrast. Both cases raise concerns over a system used in the Wisconsin judicial system, called COMPAS. COMPAS is used as a recidivism risk assessment tool which uses information gathered from an interview with the offender, as well as information from their criminal history (State of Wisconsin v. Eric L. Loomis, 2016). In the case of Eric Loomis, COMPAS was used in the presentencing investigation report (PSI or PSR, which affects the severity of a sentence) and he was sentenced to six years in prison with five years of extended supervision for charges of attempting to flee a traffic officer as well as operating a motor vehicle without the owner's consent. Loomis then filed a post-conviction motion for relief, making the argument that the use of COMPAS in the sentencing process violated his due process rights under the 14th Amendment. Loomis' argument hinged on the fact that COMPAS's methodology for determining recidivism risk is obfuscated from public view as a result of the COMPAS algorithm being a trade secret. Specifically, Loomis argued that his due process rights to be sentenced based on accurate information and to an individualized sentence were violated and that COMPAS's consideration of gender violated his due process rights. The appeal was denied, and

the decision was upheld by the Supreme Court of Wisconsin. In the Supreme Court decision upholding the denial of appeal, Justice Ann Walsh Bradley found that the use of gender as a parameter for risk assessment "served the nondiscriminatory purpose of promoting accuracy" (State of Wisconsin v. Eric L. Loomis, 2016) and that not enough evidence had been provided by Loomis to prove that the sentencing court had even considered gender in their sentencing. Additionally, the decision by the court held that COMPAS used only publicly available data and data that Loomis provided, meaning that Loomis could reasonably have verified the accuracy of information used in sentencing. Bradley did, however, recognize that the COMPAS algorithm did pose concerns regarding individualized sentencing more generally. In spite of this, she reasoned that, since the COMPAS score was not the only factor considered, the sentencing was adequately individualized. The decision did go on to provide caution for judges when utilizing similar risk assessments.

This decision and its reasoning are concerning. Justice Bradley's assertion that the data being public (and therefore able to be vetted) is equivalent to transparency and due process for the convicted raises concerns over the ability of a convicted defendant to properly examine all necessary information. In the case of an AI model, a great deal of domain knowledge is necessary to understand the implementation details. Further, the model itself has not been made public, raising additional concerns that even in the event that a defendant has access to such a domain expert, their ability to properly examine a model for sources of inaccuracy or bias would be effectively zero. This can be seen clearly as a method which complicates the process of sentencing in a way that not only makes it more difficult to appeal, but in large part abstracts key information about how a decision was made to an autonomously operating algorithm. Granted that a judge ultimately makes the final sentencing decision, but by integrating a machine-

assigned score, an already convoluted and imprecise process becomes even more so. Some might argue that by using an absolute score from a recidivism risk assessment algorithm the process is simplified and clarified. The reality is that this is not true; as Hillman argues, "An algorithm-generated risk assessment score presents itself to the court as a presumptive factual determination. In essence, predictive technology becomes another witness against the defendant without a concomitant opportunity to test the data, assumptions, and even prejudices that underlie the conclusion." (2019, 37) This adds unnecessary opacity to an already complex process rather than creating simplicity. Not only does it alter the sentencing process to include a factor which is difficult for a defendant to challenge, but it gives judges what many claim to be an "objective" measure on which to base their decision. From what we know about the way machine learning and artificial intelligence carries bias, this is nearly impossible to prove. As highlighted by Crawford (2021) and Buolamwini and Raji (2019), AI fairness is difficult to measure, and eliminating bias from AI system is exceedingly challenging. In Buolamwini and Raji's (2019), the authors delve into algorithmic fairness and auditing, following up on a prior study by Gebru and Buolamwini (2018) which had investigated accuracy of commercial face-based gender classification services. The upshot of Buolamwini and Raji (2019) is that companies which had been audited were able to "prioritize issues and yield significant improvements…" (2019, 6) but that non-targeted (by the study) corporations saw significant persistence of subgroup performance disparities. Ultimately, they conclude that while algorithmic fairness "may be approximated through reductions in subgroup error rates…" a transformation in the development and use of facial recognition AI is necessitated in order to avoid potential abuse and weaponization (Buolamwini and Raji 2019, 6).

In the ongoing case of *Henderson v Stensberg,* Titus Henderson alleges that "prison officials discriminate against him and other African American prisoners in various parts of the parole process…" (Henderson v Stensberg 2020), including in the use of COMPAS. Justice James D. Peterson, a Wisconsin District Judge, allowed the filing to proceed under the Equal Protection Clause of the Fourteenth Amendment. Henderson alleged that the creators of COMPAS (Northpointe Inc., now Equivant) are aware of racial bias in the program and won't upgrade it to "make it more accurate" without being paid to do so. He also alleged that defendant Wisconsin Department of Corrections (DOC) employees "supported using COMPAS… knowing that the program is biased against African Americans, and they won't pay Northpointe" to upgrade the software (Henderson v Stensberg 2020, 1). Further, he alleges that defendant DOC employees "allow correctional officers to 'fudge' his parole file… with false negative comments that inmates are not privy to and thus not allowed to challenge" (Henderson v Stensberg 2020, 1). The Northpointe defendants have filed two motions to dismiss Henderson's claims under Federal Rule of Civil Procedure 12(b)(6). They claim that "Henderson has not plausibly explained what they have done to violate his rights" and that the Wisconsin Supreme Court had already concluded that use of COMPAS does not violate a defendant's due process rights (Henderson v Stensberg 2020, 2).

Interestingly, Justice Peterson has denied motions to dismiss, stating among other reasons that, "I cannot conclude, based on *Loomis*, that Henderson fails to state claims against the Northpointe defendants" (Henderson v Stensberg 2020). Although *Henderson* has not been decided and thus does not yet provide any legal precedent, it remains an interesting case, with arguments hinging on the Equal Protection clause of the Fourteenth Amendment, rather than the Due Process clause. It is also notable that Northpointe (Equivant) was unable to successfully

have these allegations dismissed, providing some hope (and potential precedent) that, in a case where such a consequential piece of software is found to be biased or prejudiced, the creators may ultimately be held accountable.

Interestingly, in Angwin (2016), it was found that a risk assessment model produced by Northpointe and used in Broward County, Florida (again, COMPAS) was "somewhat more accurate than a coin flip," and was nearly two times more likely to designate Black defendants as future criminals than white defendants (Angwin 2016). While Northpointe has disagreed with the conclusions offered by ProPublica, the findings of their investigation are still compelling and raise serious concerns regarding the use of such technology. If this is the case, it seems possible that Henderson's Equal Protection complaint could hold water in a legal fight. With mounting evidence that Northpointe's COMPAS is an unjust and racially disparate algorithm, there is reason to believe its use violates the Equal Protection clause of the Fourteenth Amendment.

Even more contrastingly, the case of *Kansas v. Walls* provides a direct counter to the ruling of *Loomis*, albeit in a different state. In this case, John Walls was convicted of criminal threat, pleading no contest. The court used the Level of Service Inventory-Revised (LSI-R) risk assessment tool to evaluate Walls and he was labeled as "high risk," which made him eligible to be supervised by community corrections as opposed to court services. When he requested access to the questions and scores associated with them, his request was denied. Walls appealed, arguing that his due process rights had been violated. Ultimately the appeals court ruled in his favor, reasoning that the denial of access to the complete LSI-R assessment made it impossible for Walls to challenge the accuracy of the information (Re: State of Kansas v. John Keith Walls 2017). This is an encouraging development; however, it does not negate the fact that the results

of Loomis stand in direct contradiction to what has been determined to be a potentially rights-violating use case.

Asaro (2019) lays out two approaches to AI ethics and their relation to criminal justice– specifically on the topic of predictive policing using AI models like those discussed here. Asaro (2019, 3) explains the "Models of Threat" view as beginning from "the assumptions that the world can be classified into clear categories, i.e., threats and non-threats, and that this is the first step in choosing an appropriate action to take." As it stands today and is currently being adopted, the use of recidivism risk assessment models fits perfectly with this idea. We see a process in which AI models are being used in an attempt to determine whether a convicted defendant is likely to be a repeated threat to society. Rather than use our technology to increase overall societal benefit, this approach is punitive and can be seen as ascribing guilt in a predetermined fashion. Asaro (2019) discusses this phenomenon in his paper, saying about guilt in the justice system, "one must actually commit the act for which one is held responsible, and one must have had in mind the intention…" Asaro's (2019) discussion is in relation to predictive policing (as previously discussed), but this topic applies similarly to recidivism risk assessment. Although the affected parties in recidivism risk assessment have been found guilty of a crime, does that permit them to be pre-judged as guilty of future crimes by an AI model? Even in a punitive criminal justice system, should a convicted defendant not be held accountable for their previous actions and rehabilitated to avoid future crime? Under the Models of Threat ideology which is currently beginning to be applied using AI assessment algorithms to seek out potential threats, convicted criminals may have their rights subject to less than transparent processes in the name of identifying potential threats.

In Asaro's (2019) discussion over these two different AI ethics, the Ethics of Care approach is laid out as holistic and taking a "broad, big-picture view of the values and goals of systems design," considering the interactions and interrelations between actions or interventions and "the nature of classifying things and predicting outcomes within specific contexts" (Asaro 2019, 4). Ultimately, the Ethics of Care approach seeks to be cognizant of the complexity inherent to social relations and socio-technical systems. Asaro (2019, 4) states that the Ethics of Care approach "does not expect more and better data to simply solve complex social and institutional problems, but rather to provide opportunities for finding better solutions, better actions, and better policies than what are already considered." This is the opposite of what recidivism risk assessment algorithms attempt to do. Specifically, in using these algorithms, we are indicating that we believe the increase in available data will help to ascribe pre-guilt, solve the problem of recidivism, identify criminals, and predict who deserves higher levels of state surveillance. Instead of seeking to prevent an individual from reoffending by reintegrating them into society and creating a situation in which society as a whole benefit (a rehabilitated and hopefully productive member of society and a lower burden on the tax-funded prison system), this mode attempts to punish recidivism out of existence.

Given this distinction between Models of Threat and Ethics of Care, it seems we are easily able to identify that the current use AI on criminal justice falls clearly under the Models of Threat approach to AI ethics. We must ask then, which of these approaches is better suited to aid in providing a safer, fairer society?

Facial Recognition

There have now been a growing number of cases in which Black men have been misidentified by facial recognition software as suspects for crimes and subsequently arrested (Perkowitz 2021; Leslie 2020) For instance, in January 2020, Robert Williams was arrested in front of his family outside of his Farmington Hills, Michigan, home after a facial recognition algorithm misidentified him as the perpetrator in a case of theft from a watch store. He was held in a detention center for 30 hours and only released when police later realized the algorithm had been incorrectly identified him (Williams 2020). In his opinion piece for the Washington Post in which he details the events of his arrest, Williams goes on to question why these technologies are even being used for such purposes when it is known that facial recognition algorithms misidentify Black and Asian people at up to 100 times worse rates than white people (Williams 2020; Grother, Ngan, and Hanaoka 2019, 2–3). This is a clear example of why these algorithms should not be used for such purposes, however, there are deeper reasons than just the fact that algorithms are potentially unreliable. No model is infallible, but to add to this, law enforcement officials have what Leslie (2020, 26) calls "gateway attitudes." That is to say that they are not actively working to eliminate and counteract sources of bias within their work. Specifically, Leslie refers to the Detroit Police Department and their nonchalance in the face of serious errors such as the arrest of Williams. In fact, *prior* to Williams' arrest, concerns had been highlighted by the ACLU to DPD regarding tendencies for facial recognition technologies to generate disproportionate false positives for minority groups. Reportedly, on the subject of the matter, Assistant Police Chief James White said that the facial recognition would be used as "an investigatory tool that will be used solely to investigate violent crimes, to get violent criminals off the street" (Hunter 2017). The example of Williams is not the only evidence to suggest that

this was not a true statement. In July of 2019, another black man, Michael Oliver, was involved in a similar breach of rights. Facial recognition software had incorrectly identified him as a match with another person who was wanted on suspicion of committing larceny (in spite of non-matching tattoos, skin tone, and facial shape) and he was subsequently arrested and held in detention for two and a half days (N. O'Neill 2020). In his analysis, Leslie (2020, 27) labels these two violations as "part of a systemic pattern of derelict behavior rooted in the apathetic tolerance of discriminatory crime." This is a solid analysis, with tangible evidence to back it up. However, this pattern can be seen as more than just evidence of a harmful "gateway attitude." Although it certainly *can* be seen as such, it can also be seen as the criminal justice system working in the way which it was designed. Specifically in the United States, given that the criminal justice system was born out of a desire for white Americans to secure dominance over the Native people in what is now the United States (Thompson 2019). In fact, as discussed, the birth of the criminal justice system grew from the white perspective of Black and Brown peoples as inherently criminal in nature and needing to be controlled. Later in the history of criminal justice in the USA, when slavery was abolished, criminalization quickly filled the vacuum of control as a method to subjugate Black and Brown people. Notably as Thompson highlights, "the 13th amendment that outlawed slavery also included an exception for anyone convicted of a crime," as well as a clause of the Fourteenth Amendment which robbed convicts of their right to vote (Thompson 2019, 223) Much further down the line in history, the criminal justice system in the USA has remained deeply racialized, with the problem of mass incarceration affecting Black and Brown people at a highly disproportionate level. This is all to say that these "mistakes" by departments such as the DPD, are not simply errors of apathy. While the individual actions of investigators and detectives are potentially apathetic to the bias and discrimination with which

they are operating (internally and within systems such as facial recognition software), the outcomes are precisely in line with the original intent of the system. Black and Brown people (largely Black and Latino men) are being introduced into a system of criminal justice which wants nothing more than to incarcerate and control them. The specific danger with using AI algorithms and models to affect these outcomes is that there is no mechanism of accountability built into this system. While it may be argued that seeking recourse for wrongdoing is already difficult enough, it becomes completely impossible once decisions are abstracted away to a computer. In Crawford (2021), law Professor Andrew Ferguson is quoted as explaining "We are moving to a state where prosecutors and police are going to say, 'the algorithm told me to do it, so I did, I had no idea what I was doing" (Crawford 2021, 197). This prospect is frightening and further reinforces the concept that AI use in criminal justice is just another method of abstracting real-world harms away from responsible individuals– a method of legitimizing undue harm, violence, and control over specific populations. In the cases of Robert Williams and Oliver Michaels, who was held accountable? The answer is no one. Not the software engineers and data scientists who created the model, not the managers who prioritized shipping a "good enough" product which replicated bias, not the curators of the data which trained the model, not the investigators who blindly believed the algorithm in the face of ample evidence to its counter, and ultimately, not the responsible officials who decided these faulty models to be sufficiently accurate to be used for such a delicate application. In all these layers of decision making, each of which could have prevented these violations of two Black men, not a single actor was held responsible. We must then ask, if AI in criminal justice presents these challenges of abstraction of responsibility, how can we justly continue to apply them in situations where "mistakes" have unfathomable consequences?

With this all said, there are certain actions being taken currently which will have a positive effect on outcomes relating to facial recognition. In the United States, for instance, cities such as Portland, San Francisco, and Boston are outlawing the use of surveillance systems which leverage facial recognition (Metz 2020). These movements to ban facial recognition in public places stem from concerns over privacy and bias concerns, which are well-placed given examples such as previously discovered cases. In Portland, two ordinances were passed banning the acquisition of facial recognition surveillance technologies for city bureaus as well as private entities (Metz 2020). Mayor Ted Wheeler is quoted as saying "We must protect the privacy of Portland's residents and visitors, first and foremost. These ordinances are necessary until we see more responsible development of technologies that do not discriminate against Black, Indigenous and other people of color…" (Becker 2020). In San Francisco, an ordinance was passed in 2019 requiring city departments to "disclose any surveillance technologies they currently use or plan to use, and to spell out policies regarding them that the Board of Supervisors must then approve" (Van Sant and Gonzales 2019). These ordinances demonstrate that it is certainly possible to write legislation designed to protect citizens from these kinds of technologies (particularly on the local level). Another encouraging and recent development comes from the state of Massachusetts, which has recently passed a police reform bill requiring the permission of a judge before police are legally allowed to run a facial recognition search. Additionally, it mandates that facial recognition searches be conducted by state police, the FBI, or the state Registry of Motor Vehicles rather than by municipal police officers or detectives (Hill 2021). This is seen as more of a "guard rail" than a prohibition of facial recognition in law enforcement, but critically, it creates a structure which institutionalizes the idea that facial recognition should not be used

lightly. The law also creates a commission to study facial recommendation policies, a necessity for legislators to be able to keep up with rapidly developing technology.

None of these laws or ordinances are perfect—far from it, in fact. Of the three examples, Portland should be seen as a strong instance, however the scope of this facial recognition ban does nothing to address other applications of AI. San Francisco and Massachusetts are both far too liberal in their allowance of these technologies under specified conditions. Neither piece of legislation is enough to ensure that errors won't happen and neither covers a wide enough scope to be particularly useful when it comes to AI in criminal justice more broadly. Regardless, all three examples are a start and with further bolstering, they provide potential frameworks and case-studies for how this kind of legislation can be passed. In the following section I will discuss generally how we might approach further legislation as well as how these examples can be iterated on to create more robust systems which are resistant to the pitfall of AI use in criminal justice.

Discussion

The Challenge: Abstracting Injustice

Naturally, many of the findings discussed earlier are cause for serious concern. It should be clear by this point that the unfettered and corporatized use of this kind of software should be viewed as, at best, standing on tenuous ethical grounds and, at worst, as potentially disastrous and illegal. With all this information, I propose we call the rise in the use of artificial intelligence, machine learning, and data-driven solutions in criminal justice what they are: *abstractions of injustice*. It has become clear that these technologies can no longer be seen as valuable tools for justice but rather as abstractions that muddy already foggy waters. These tools are now being used as a mechanism for individual actors within the criminal justice system to avoid responsibility for

poor or unjust decisions, giving them cover to point at models and numbers in efforts to shift blame onto non-human actors not subject to accountability. This is highly troublesome given the history of injustice in the United States and presents a unique challenge. In the current state of AI and the justice system (as well as governance), we are often told that data-driven decisions are better; more informed, less biased, and best of all automated. However, we know this to be untrue. There is significant and mounting evidence to the contrary. How, then, can we begin to stem the tide on an already massive and growing data-machine in criminal justice? This is the key question, and one I will attempt to provide some answers for hereafter. I intend to add to a growing and large list of calls for change, including those by Crawford (2019; 2021), Asaro (2019), Leslie (2020), Hillman (2019),and many more (see, e.g., O'Neill 2017; Robles Carrillo 2020).

The previously mentioned legal cases, *Loomis* and *Henderson*, can grant an interesting view into how some of the issues associated with recidivism risk assessment are currently playing out in the judicial system and can inform how we ought to move forward. Though the two relate to the same software, these decisions grant a look into separate legal arguments that are unfolding regarding use of recidivism risk assessment algorithms. Both invoke Fourteenth Amendment protections; however, *Loomis* invokes the Due Process Clause and *Henderson* invokes the Equal Protection Clause. As mentioned, given that *Henderson* has not yet been decided in court, it's difficult to predict what the outcome will be, but the dismissal of the Northpointe defendants' multiple motions can provide a look into the reasoning in that case. In the case of *Loomis,* it is worrying to see the Supreme Court of Wisconsin rule that a tool such as COMPAS in sentencing does not violate the Due Process Clause of the Fourteenth Amendment. As argued by the Honorable Noel Hillman, the introduction of predictive AI into sentencing

raises concerns over the presumption that the risk assessment score is fact (Hillman 2019). He argues that these scores may be delivered from "black boxes" with little recourse for defendants to challenge the factuality of their score. He also compares this to the ability for a defense lawyer to question a witness who provided information to the probation officer versus subpoenaing and questioning a software developer to be questioned at a sentencing hearing. In relation to *Loomis,* we come across this almost immediately. The decision dismissed Loomis' appeal partially on the basis that all data used by COMPAS was either publicly available or given by Loomis himself. While this remains true, the proprietary nature of models such as COMPAS makes this significantly more complex than this decision makes it out to be. In general, artificial intelligence and machine learning models are trained on some sort of historical data—although this data might be made publicly available, models will often be trained on extremely large datasets, making vetting of this data time consuming and laborious. Additionally, the nature of some machine learning models can create a situation in which even the developers don't understand how a given model comes to its conclusion (see Appendix 1). Given this information, it becomes increasingly difficult to argue that use of these models does not violate due process rights in some way. It may be technically possible in some situations (not all) for a defendant's legal team to analyze this sort of algorithm and data but, realistically, this is not a feasible method of recourse. Additionally, it could foreseeably lead to deeper inequalities where they already exist. It is not difficult to imagine that defendants with greater means are able to hire teams to do this examination of data and model while those who are of lesser means do not have this method of recourse. Are we as a society prepared to combat this potential civil rights disparity? When it comes to the case of AI being used in criminal justice, it does not seem so.

Further yet, when contrasting *Walls* and *Loomis*, we see differing approaches to a similar issue. In the case of *Walls*, the appeals court ruled with the defendant, reasoning that Walls' due process rights had indeed been violated by the refusal to allow him access to his assessment. In the case of *Loomis*, the opposite argument prevailed at the highest level of Wisconsin State courts. I tend to agree with the *Walls* decision and hope it will be used as persuasive precedent in future cases based on similar facts. If a result is being delivered from what is effectively a black box (an algorithm or model hidden from view, as in these cases) and is considered to represent a factual understanding of a convicted defendant's risk, we run the danger of blindly making consequential decisions that can impact the incarceration of individuals based on potentially faulty and biased information. With no process established to avoid these outcomes, it seems likely that they will occur.

Hillman (2019, 37) additionally argues that the reliability of AI algorithms should be drawn into question, noting that the data used to create a model can result in a situation where there is "garbage in, garbage out." This is particularly notable when discussing the United States criminal justice system (and nearly any other system within the federal government) due to the long history of systemic racism as well as more recently the staggering mass incarceration of Black men (Nellis 2016). When we also consider the historical purpose of the criminal justice system, it then becomes clearly apparent that using historical criminal justice data is problematic. Given these facts, it is easy to see that datasets that may be used (whether publicly available or not) can introduce bias into models unless mitigatory steps are taken in creation, training, and production of such AI models. Even with these kinds of interventions, it can be difficult to know how a machine learning model has learned—i.e., what attributes of the data weigh the heaviest in the model's prediction. This can lead to "clean" data being used as a proxy for other attributes

such as race. As it stands, there is astoundingly little in the way of protections against these kinds of harms. We see this with COMPAS as well as with other instances of AI use in criminal justice, such as facial recognition. In the case of facial recognition, we've seen how this plays out: innocent people are harmed, sometimes in heavily traumatic and damaging ways which can have unknown and long-term effects. In the case of Robert Williams, his two young daughters had to witness their father be arrested in front of their own home, an event that could forever alter their lives. In the case of Michael Oliver, Oliver lost his job and car. When it comes to predictive policing, we see that those who are most vulnerable to crime (either as a potential victim or perpetrator) are subjected to higher levels of surveillance and state control. Left to develop without intervention, it seems likely that the use of AI in criminal justice will continue to unfold along the same lines on which it has begun, exacerbating, and creating disparities whether racial or otherwise.

Using Hillman's (2019) analysis as a starting point, it becomes clear that the use of this kind of technology in the criminal justice system requires great caution. Although the due process appeal of Loomis may have failed, these concerns are relevant nonetheless and need further consideration. The Walls decision gives this argument further strength. Given the three cases presented and the Fourteenth Amendment's guarantee that "all persons in the United States shall enjoy the 'equal protection of the laws'" (Sagal n.d.), the question then remains: how can we enjoy equal protection of the laws if software is being used in legal processes which discriminates against certain populations. To begin with, given the nature of such consequential models, it's reasonable to argue that their contents should be made public and at least be subject to accountability regulations. Although this is in direct tension with the goal of a for-profit entity, in the name of due process it is justified to question why such important algorithms are given the

protection of operating as black boxes, with no room for public critique or examination. While it may be argued, as seen in *Loomis,* that these models are merely a portion of the sentencing process and thus do not violate due process rights, it is fair to question why they would be included at all if their reliability and fairness cannot be guaranteed and verified. Regardless of how much they influence an outcome, they still have an effect, and this should not be overlooked. Granted, the inclusion of risk assessment models, facial recognition, and predictive policing in the justice system may be legal (although somewhat tenuously), but it brings about the question: just because we can, does it mean that we should?

Watch Dogs and Accountability

In the current situation, we see the already complex tasks of criminal justice complicated further by incorporating AI. Rather than simplifying the process, introducing models to these systems erects opaque walls of technical knowledge. AI in predictive policing grants entire police departments license to bolster surveillance of certain areas or individuals, displacing any potential blame from decision-making individuals and teams when things go wrong, and harm is demonstrated. Although not fully removing the role of the judge, the inclusion of a recidivism score can also be seen as abstracting responsibility. Additionally, as we see with the case of Northpointe/Equivant, the creators of such an algorithm are not likely to take responsibility for shortcomings of their work and critically there seems to be no framework for holding them accountable. Facial recognition seemingly simplifies the job of an investigator; however, this may lead to a lack of critical analysis as seen in the cases of Michael Oliver and Robert Williams. Herein lies one part of a proposed solution which I will call *actor accountability*. In order to motivate companies and individuals to act in a just manner, we need more than just the vague notions of "AI Ethics" or "AI for Good" which are becoming popular within the tech

industry. The fact of the matter is that these ideas are not enforceable and are, therefore, useless when profit is on the line.[8] Thus, I propose that it is necessary to create a watchdog entity along with significant legislation which would hold AI vendors and users accountable for the ways in which the technologies are applied. This sort of legislation would be beneficial for the application of AI broadly, but it is AI in the criminal justice system that is the focus here. Ultimately it would be far preferable to avoid misuse of AI in criminal justice (and thus avoid miscarriages of justice relating to AI), which I will discuss later. However, this proposal should be seen as a starting point to disincentivize the gateway attitudes as discussed by Leslie(2020). The proposed watchdog entity would be chiefly responsible for auditing artificial intelligence use in criminal justice, with domain experts being employed to analyze AI development methodologies, training data, and more. It would be important to include key stakeholders in decision making for such an entity and it is here that I recommend the method of stakeholder analysis presented by Young, Magassa, and Friedman (2019) in their paper entitled "Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents" and in which they lay out a framework they call Diverse Voices. The Diverse Voices framework involves amplifying the voices of stakeholders who may not be accounted for in tech policy and integrating their inputs into said policy. By including such a framework, the proposed AI watchdog would be able to draft AI policy that minimizes disparate impacts on marginalized groups by hearing directly from those who are part of these communities.

Along with this proposed AI auditor, I propose that significant new legislation be drafted and passed to increase the accountability of the companies that create AI-based technologies. Such legislation would need to include clauses which make it more difficult for tech companies to avoid responsibility when their technologies cause harm. For instance, under such legislation,

a company such as Northpointe could be found be guilty of harm if their product was found to be harmful and could be sued under grounds of discrimination with evidence such as the ProPublica report. This kind of legislation would evidently necessitate further research into what would be constitutionally legal as well as incorporation of the Diverse Voices framework.

A Broader View: Models of Care

Speaking more broadly, given that we know there are serious ethical and legal implications to the use of AI systems in criminal justice, it deserves consideration to pause their application in the criminal justice context until further research and development is completed. It is not at all clear that their use is doing anything to improve outcomes and, in fact, may be creating injustices. Legally, it remains dubious that these sorts of models are largely obfuscated from any critical, public view, including that of the defendant. While access to the data used by these models is important, it is not a full view into how the algorithms determine their risk scores. This creates a situation rife with opportunity for bias to be introduced without the ability for those being affected to know or have any recourse. Additionally, the usage of AI in determining threats can already be seen as a failure. For instance, in Asaro's (2019) article, we see an example of the Ethics of Care approach outperforming the Models of Threat approach when applied to predictive policing.[9] Given this, how can we continue to use these types of algorithms for criminal investigation, prosecution, and sentencing when we know that they may not be as accurate in their predictions as we would like and that their application in a Models of Threat manner has already been seen to perform worse than in the Ethics of Care counterpart. I then conclude that two things must occur: first, further legal and technical research must be done to provide a set of regulations and stipulations for the use AI in criminal justice. Society and the justice system must establish methods with which we can evaluate and identify biases in these

technologies. This could mean developing additional technologies (openly and transparently) which are able to assess the fairness and bias of an AI model or creating a process of manual inspection and scrutiny—or any number of different critical protocols. Additionally, the use of "trade secret" technology in criminal justice must be ended. This information imbalance is simply impossible to circumvent when attempting oversight without opening algorithm details to be analyzed. I argue that in cases where the fate of a legal or criminal decision is on the line, there is no room for private entities to play such a major role in decision making. This dynamic leads to a long chain of decisions being washed away from any sort of critical view, with product managers, executives, engineering leads, and software engineers making countless decisions behind the veil of a private corporation. These decisions cannot be obscured in such a way, particularly when potentially racialized and class-based decisions are being made by actors who have no verified credibility to be seeking the most just solution. Second, the applications of these technologies must be reconsidered. It is entirely possible to use artificial intelligence and its sub-categories to better society.[1] That said, using AI in a manner that prioritizes threat detection has been shown thus far to not work in criminal justice. Detecting the threat of cancer is helpful, as stopping an aggressive version of the disease before it can inflict great damage can save lives and quality of life. This same logic cannot be applied to humans and particularly cannot be applied to humans while also maintaining civil rights, such as the Fourteenth Amendment. We should not use AI to predict who will or will not be a future criminal and, equally, we cannot determine who will or will not reoffend. We cannot rely on fallible machines and models to identify who *might* be culpable for a crime, particularly in a system that is rife with apathy towards the lives and wellbeing of communities they affect directly. Along those same lines, we cannot rely so heavily on technology to make definitive statements about identity when serious

trauma and consequences are on the line. Let us as a society begin to reframe how we view AI as a tool—one that can have significantly beneficial impacts but equally damaging repercussions. To date, popular discourse surrounding AI is commonly about how well AI systems can perform against human competition, from IBM Watson winning Jeopardy (Best 2013) to Alphago defeating a top ranked Go player (Koch 2016). We must alter society's discourse surrounding AI– how its significant power can be used for benefit and for harm. This is certainly beginning to occur, however the magnitude of the effects of these systems must become widely known. Let us proceed in a manner that leverages the power that artificial intelligence presents with care, rather than as a mechanism for threat detection and punishment.

Conclusions

Limitations

Some limitations to the scope of this research should be noted. Currently there is a rapidly expanding literature on topics in alignment with the research of this paper as well as on topics similar and of interest to this research. Given the time frame and scope of an undergraduate honors thesis, it would not have been possible to read and digest all relevant work. From pieces such as Ruha Benjamin's *Race After Technology* and Sandra G. Mayson's "Bias In, Bias Out" to the Europeans Union's recent proposal for artificial intelligence regulation, there are wide ranging materials which are both relevant and were not able to be covered. Additionally, as a student in computer and information science, there are certain personal limitations on my interpretation of legal and legislative issues.

Further Research

While, as noted, there is a large body of growing work on the subject, further research needs to be pursued. Specifically, the case studies of cities and states which have implemented rules and

regulations against specific technologies must be undertaken to better understand how future, broader policies might be written at the city, state, and even federal levels. Further review and synthesis of current works must also be undertaken to better understand the direction of AI ethics and regulation as a field. This research must also be transformed from theoretical to practical application. Working together with stakeholders and governing bodies, academics must push to implement and study new methods for prevention of AI-related rights violations as well as how these technologies can best be applied. We must also strive to clearly define a set of attributes which might be prohibitive to the use of AI. For example, applications that would rely on data which has been influenced by systemic and historical racism as well as "dirty" data collecting practices may potentially be candidates for prohibition entirely.

Further, significant research must be done on what AI regulation would be constitutionally legal (as mentioned in the discussion) as well as how legislation can accomplish goals of justice and fairness in relation to AI systems in criminal justice. It is necessary that legislation be significantly informed by academic research and after consultation with a variety of stakeholders from those communities affected to domain experts in law, computer science, ethics, and other appropriate disciplines.

Appendix

On Black Boxes and Machine Learning as a Field

Machine learning is an undoubtedly powerful technology. The ability to learn complex patterns which are frequently too complicated for humans to recognize. For instance, neural networks—a type of machine learning algorithm—use potentially many layers of activation "neurons" to mimic biological neurons. Complex neural networks such as convolutional neural networks

(often used for computer vision, another subset of machine learning) have "hidden" layers which self-adjust during model training. This is all simply to say that these models are complex. Not only are they complex to the layman, though. Frequently ML models are created with the goal of being "accurate predictors on a static dataset that may or may not represent how the model would be used in practice" (Rudin and Radin 2019, 3). Additionally, models are often black boxes, created "directly from data by an algorithm, meaning that human, even those who design them, cannot understand how variables are being combined to make predictions" (Rudin and Radin 2019, 2). This creates an inherently disadvantageous system any kind of monitoring for bias or injustice propagated by such systems. If data scientists and engineers don't even understand how their models arrives at a conclusion given a specific dataset, how can anyone understand this? And if no know understands, how can we prove whether a system is biased? The answer is that we can't, really.

However, as Rudin and Radin (2019) argue, it is not necessary for models to be uninterpretable and further it is not necessary to sacrifice model accuracy for interpretability. In their article, the authors assert that the belief that interpretability would diminish accuracy has "allowed companies to market and sell proprietary or complicated black box models for high-stakes decisions when very simple interpretable models exist for the same tasks" (2019, 3). In fact, they go on to discuss how this assumption that interpretable models must sacrifice accuracy is a false choice. For instance, in criminal justice, it has been demonstrated on numerous occasions that complicated black box models for predicting recidivism risk are not any more accurate than simple predictive models (Rudin and Radin, 2019).

Ultimately, within the context of this paper, it is clear that the use of black box models in general is unnecessary and in criminal justice can be potentially disastrous. As a whole field, ML

relies far too heavily on these black box models in a fashion which is both unnecessary and potentially harmful.

References

Angwin, Julia. 2016. “Machine Bias — ProPublica.” May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Asaro, Peter M. 2019. “AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care.” *IEEE Technology and Society Magazine* 38 (2): 40–53. https://doi.org/10.1109/MTS.2019.2915154.

Becker, Tim. 2020. “City Council Approves Ordinances Banning Use of Face Recognition Technologies by City of Portland Bureaus and by Private Entities in Public Spaces.” Portland.Gov. September 9, 2020. https://www.portland.gov/smart-city-pdx/news/2020/9/9/city-council-approves-ordinances-banning-use-face-recognition.

Best, Jo. 2013. “IBM Watson: The inside Story of How the Jeopardy-Winning Supercomputer Was Born, and What It Wants to Do Next.” TechRepublic. September 9, 2013. https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/.

Brayne, Sarah, Alex Rosenblat, and Danah Boyd. 2015. “Predictive Policing,” October, 11.

Buolamwini, Joy, and Timnit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” 15.

Buolamwini, Joy, and Inioluwa Raji. 2019. “Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products.” MIT Media Lab. January 24, 2019. https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products/.

Conaboy, Richard, Henry Smith, and Richard Snyder. n.d. *The Criminal Justice Standards and Goals of the National Advisory Commission Digested from A National Strategy to Reduce Crime*. Pennsylvania Committe for Criminal Justice Standards and Goals.

Crawford, Kate. 2021. *Atlas of AI. Power, Politics, and the Planetary Cost of Artificial Intelligence.* Yale University Press.

Egbert, Simon. 2019. “Predictive Policing and the Platformization of Police Work.” *Surveillance & Society* 17 (1/2): 83–88. https://doi.org/10.24908/ss.v17i1/2.12920.

Esteva, Andre, and Eric Topol. 2019. “Can Skin Cancer Diagnosis Be Transformed by AI?” *The Lancet* 394 (10211): 1795. https://doi.org/10.1016/S0140-6736(19)32726-6.

Etzioni, Oren. 2016. “Deep Learning Isn’t a Dangerous Magic Genie. It’s Just Math.” *Wired*, 2016. https://www.wired.com/2016/06/deep-learning-isnt-dangerous-magic-genie-just-math/.

"Former Jackson County Man Arrested for a 1999 Child Abduction Case." 2017. January 13, 2017. https://www.justice.gov/usao-sdin/pr/former-jackson-county-man-arrested-1999-child-abduction-case.

Grother, Patrick, Mei Ngan, and Kayee Hanaoka. 2019. "Face Recognition Vendor Test Part 3:: Demographic Effects." NIST IR 8280. Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8280.

Henderson v Stensberg. 2020 18-cv-555-jdp. UNITED STATES DISTRICT COURT FOR THE WESTERN DISTRICT OF WISCONSIN.

Hill, Kashmir. 2020. "Wrongfully Accused by an Algorithm." *The New York Times*, June 24, 2020, sec. Technology. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

———. 2021. "How One State Managed to Actually Write Rules on Facial Recognition." *The New York Times*, February 27, 2021, sec. Technology. https://www.nytimes.com/2021/02/27/technology/Massachusetts-facial-recognition-rules.html.

Hillman, Noel. 2019. "The Use of Artificial Intelligence in Gauging the Risk of Recidivism." January 1, 2019. https://www.americanbar.org/groups/judicial/publications/judges_journal/2019/winter/the-use-artificial-intelligence-gauging-risk-recidivism/.

Hunter, George. n.d. "Project Green Light to Add Facial Recognition Software." The Detroit News. Accessed April 13, 2021. https://www.detroitnews.com/story/news/local/detroit-city/2017/10/30/detroit-police-facial-recognition-software/107166498/.

Kearney, Melissa S, Benjamin H Harris, Elisa Jácome, and Lucie Parker. 2014. "Ten Economic Facts about Crime and Incarceration in the United States," May, 28.

Koch, Christof. 2016. "How the Computer Beat the Go Master." Scientific American. March 19, 2016. https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/.

Lau, Tim. 2020. "Predictive Policing Explained | Brennan Center for Justice." April 1, 2020. https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained.

Leslie, David. 2020. "Understanding Bias in Facial Recognition Technologies." *ArXiv:2010.07023 [Cs]*, October. https://doi.org/10.5281/zenodo.4050457.

Lin, Zhiyuan "Jerry," Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. "The Limits of Human Predictions of Recidivism." *Science Advances* 6 (7): eaaz0652. https://doi.org/10.1126/sciadv.aaz0652.

Mandal, Vishal, Abdul Rashid Mussah, Peng Jin, and Yaw Adu-Gyamfi. 2020. "Artificial Intelligence-Enabled Traffic Monitoring System." *Sustainability* 12 (21): 9177. https://doi.org/10.3390/su12219177.

Metz, Rachel. 2020. "Portland Passes Broadest Facial Recognition Ban in the US." CNN. September 9, 2020. https://www.cnn.com/2020/09/09/tech/portland-facial-recognition-ban/index.html.

Meyer, Joel. 1968. "Reflections on Some Theories of Punishment." *The Journal of Criminal Law, Criminology, and Police Science* 59 (4): 595. https://doi.org/10.2307/1141839.

"NAACP | Criminal Justice Fact Sheet." n.d. NAACP. Accessed April 20, 2021. https://www.naacp.org/criminal-justice-fact-sheet/.

Nellis. 2016. "The Color of Justice: Racial and Ethnic Disparity in State Prisons." The Sentencing Project. June 14, 2016. https://www.sentencingproject.org/publications/color-of-justice-racial-and-ethnic-disparity-in-state-prisons/.

O'Neill, Cathy. 2017. *Weapons of Math Destruction*. Penguin Books.

O'Neill, Natalie. 2020. "Faulty Facial Recognition Led to His Arrest—Now He's Suing." September 4, 2020. https://www.vice.com/en/article/bv8k8a/faulty-facial-recognition-led-to-his-arrestnow-hes-suing.

Perkowitz, Sidney. 2021. "The Bias in the Machine: Facial Recognition Technology and Racial Disparities." *MIT Case Studies in Social and Ethical Responsibilities of Computing*, February. https://doi.org/10.21428/2c646de5.62272586.

Re: State of Kansas v. John Keith Walls. 2017. Court of Appeals of the State of Kansas.

Richardson, Rashida, Jason M Schultz, and Kate Crawford. 2019. "DIRTY DATA, BAD PREDICTIONS: HOW CIVIL RIGHTS VIOLATIONS IMPACT POLICE DATA, PREDICTIVE POLICING SYSTEMS, AND JUSTICE" 94 (May): 41.

Rigano, Christopher. 2019. *AI and NIJ*. New York, NY: Basic Books.

Robles Carrillo, Margarita. 2020. "Artificial Intelligence: From Ethics to Law." *Telecommunications Policy*, Artificial intelligence, economy and society, 44 (6): 101937. https://doi.org/10.1016/j.telpol.2020.101937.

Rudin, Cynthia, and Joanna Radin. 2019. "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition." *Harvard Data Science Review* 1 (2). https://doi.org/10.1162/99608f92.5a8a3a3d.

Sagal, Peter. n.d. "Due Process Equal Protection and Disenfranchisement | Equality and the 14th Amendment | Constitution USA | PBS." Due Process Equal Protection and

Disenfranchisement | Equality and the 14th Amendment | Constitution USA | PBS. Accessed April 4, 2021. https://www.pbs.org/tpt/constitution-usa-peter-sagal/equality/due-process-equal-protection-and-disenfranchisement/.

Small, Deborah. 2001. "The War on Drugs Is a War on Racial Justice." *Social Research* 68 (3): 896–903.

State of Wisconsin v. Eric L. Loomis,. 2016, 881 N.W.2d 749. Supreme Court of Wisconsin.

Thompson, Heather Ann. 2019. "THE RACIAL HISTORY OF CRIMINAL JUSTICE IN AMERICA." *Du Bois Review: Social Science Research on Race* 16 (1): 221–41. https://doi.org/10.1017/S1742058X19000183.

United Nations. n.d. "Homicide Rates." Tableau Software. Accessed April 29, 2021. https://public.tableau.com/views/Homiciderates_15826327950430/Homicide-rates?:embed=y&:showVizHome=no&:host_url=https%3A%2F%2Fpublic.tableau.com%2F&:embed_code_version=3&:tabs=no&:toolbar=yes&:animate_transition=yes&:display_static_image=no&:display_spinner=no&:display_overlay=yes&:display_count=yes&:loadOrderID=0.

Van Sant, Shannon, and Richard Gonzales. 2019. "San Francisco Approves Ban On Government's Use Of Facial Recognition Technology." NPR.Org. May 14, 2019. https://www.npr.org/2019/05/14/723193785/san-francisco-considers-ban-on-governments-use-of-facial-recognition-technology.

Wagner, Peter, and Wendy Sawyer. 2018. "States of Incarceration: The Global Context 2018." 2018. https://www.prisonpolicy.org/global/2018.html.

Williams, Robert. 2020. "Opinion | I Was Wrongfully Arrested Because of Facial Recognition. Why Are Police Allowed to Use It?" *Washington Post*, June 24, 2020. https://www.washingtonpost.com/opinions/2020/06/24/i-was-wrongfully-arrested-because-facial-recognition-why-are-police-allowed-use-this-technology/.

Young, Meg, Lassana Magassa, and Batya Friedman. 2019. "Toward Inclusive Tech Policy Design: A Method for Underrepresented Voices to Strengthen Tech Policy Documents." *Ethics and Information Technology* 21 (2): 89–103. https://doi.org/10.1007/s10676-019-09497-z.

End Notes

[1] Some examples of beneficial applications of AI include in the application of deep learning models to automate traffic monitoring (Mandal et al. 2020), and to accelerate cancer research as well as cancer diagnosis. (Esteva and Topol 2019).

[2] For instance, In January of 2017 it was announced by the Indiana Department of Justice that Charles Hollin, an alleged child molester and kidnapper had been found and arrested in Oregon using facial recognition. ("Former Jackson County Man Arrested for a 1999 Child Abduction Case" 2017).

[3] A 2020 study found that when compared side-by-side on certain datasets, algorithms were significantly better at predicting recidivism than humans (Lin et al. 2020). This study followed up on another study which had claimed that recidivism risk assessment algorithms were no better than humans at predicting recidivism and affirmed the prior study's findings, then expanded research to indicate that such algorithms could outperform humans, *under the correct circumstances.* This can be seen as a significant qualification, given that real-world applications can frequently struggle to replicate conditions.

[4] For instance, the US has a rate of roughly five homicides per 100,000 population, nearly double that of Europe as a continent and five times worse than Northern Europe (United Nations n.d.).

[5] In "The War on Drugs Is a War on Racial Justice," Deborah Small argues that the war on drugs has replaced chattel slavery and segregation in perpetuating the historical racial oppression found in America. She asserts that, although "superficially neutral..." drug laws are "enforced in a manner that is massively and pervasively biased" (Small 2001, 897).

[6] Given the events of recent years (2015-2021) it is fair to say that higher levels of police encounters can range from uncomfortable to deadly. In reality, the odds of a deadly police encounter are low, however increased interaction with police can often lead to the beginning of a criminal record, fines, and ultimately the beginning of involvement with law enforcement

[7] For instance, the COMPAS algorithm does not take any racial factors into account, and yet black defendants were still predicted as 77% more likely to be labeled as at higher risk of committing a violent crime in the future (Angwin 2016). This is certainly not conclusive evidence that racial factors were learned by the algorithm, but it provides evidence that just because racial factors are not considered does not mean a model might not pick up on proxy racial factors.

[8] This reasoning is inspired from the previously unmentioned "Artificial intelligence, from ethics to law" (Robles Carrillo 2020).

[9] In two separate studies performed in Chicago when predictive policing was used to grant jobs (congruent with the Ethics of Care approach) to those at risk, gun violence dropped. When

predictive policing was used to increase surveillance and policing on at risk individuals, there was no discernible positive effect (Asaro 2019).