Discrimination in the U.S. Criminal Justice System from Recidivism Score Algorithms

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Grace Kisly

Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

MC Forelle, Department of Engineering and Society

Introduction:

The United States has the highest incarceration rate in the world, with incarceration rates significantly higher for Blacks and Latinos than for whites (Population Reference Bureau). Every US state incarcerates more people per capita than any independent democracy on earth, with the US incarceration rate as a whole being 5 to 20 times higher than every founding NATO country (Herring & Widra, 2021). If every state was an independent nation, 24 states would have the highest incarceration rate in the world, higher than even the United States (Herring & Widra, 2021). When examining the discrepancies by race, Blacks make up 13.9% of the US population, yet they account for 38.5% of the federal prison population and 34% of the state prison population (Federal Bureau of Prisons, 2023; Herring, Sawyer, et al, 2022; US Census Bureau, 2023). Despite 18.9% of the US population identifying as Hispanic, they comprise 30.2% of federal prison inmates and 21% of state prison inmates (Federal Bureau of Prisons, 2023; Herring, Sawyer, et al, 2022; US Census Bureau, 2023). The 1997 US Department of Justice published that, among men, blacks (28.5%) are about twice as likely as Hispanics (16.0%) and 6 times more likely than whites (4.4%) to be admitted to prison during their life (Beck & Bonczar). For women, 3.6% of blacks, 1.5% of Hispanics, and 0.5% of whites will enter prison at least once (Beck & Bonczar, 1997).

While the implementation of artificial intelligence strives to reduce crime by identifying suspects more efficiently and to remove prejudice by providing objective insight on criminal data, its application has had adverse effects. By using historical data, minority groups that have higher rates of incarceration are more likely to be identified as a criminal, perpetuating the disproportionate rates (Lee & Lai, 2022). The criminal risk assessment tool, Correctional Offender Management Profiling for Alternative Sactions (COMPAS), was created by the for-

profit company Northpointe to score defendants on a scale of 1 to 10 to indicate whether they will commit offenses upon release, as an aid to determine sentencing. The results can persuade officials to sentence individuals and assign stricter or more lenient probation and parole requirements (Chohlas-Wood, 2020). Algorithms like COMPAS are highly popular in courtrooms across the Unites States, and in 2021 they were used in 46 states (Mesa, 2021). The outcome of the assessment can propagate false information and uphold biased rulings, demonstrating how the derived meaning from the output of an algorithm can drastically affect those whose livelihoods depend on the score they receive. Utilizing artificial intelligence to score individuals on their risk of misconduct can be extremely detrimental, as erroneous assessments of defendants can cause undeserving suspects to face harsher sentences.

To successfully automate decision-making for recidivism likelihood, it is not enough to simply improve the algorithms that run COMPAS; rather, what is required is better development, oversight, and regulation of these AI tools. Prior research has established the problems with artificial intelligence in the criminal justice system, and how COMPAS is inaccurate and unfair. I will utilize this overview as a framework to analyze why discrimination manifests in the algorithm in order to understand what must be addressed to resolve its issues. To support this, evidence in academic journals and media from investigative, technological, and scientific articles from the last 5 years will be synthesized to perform a discourse analysis. The findings of the analysis will highlight that the US government and criminal justice system lack policy and national AI regulation to support the use of COMPAS, and developers of COMPAS do not currently possess the resources to create an unbiased algorithm. In order to employ recidivism score algorithms successfully, national AI governance, AI regulation, adequate training of

criminal justice officials on how to incorporate AI, and diversified data and software developer teams are necessary for more equitable applications.

Literature Review:

Although the introduction of AI in the criminal justice system is intended to resolve issues of biased decision-making, its application instead propagates inequity. AI is being incorporated in various aspects of the legal system, from facial recognition tools that identify suspects to algorithms that shorten criminal sentences. While these technologies may seem like a solution to improving the criminal justice system in isolation, once incorporated into practice unexpected issues arise. Its use may overcomplicate what it was attempting to reform, having potentially dire social repercussions. A study from the American Civil Liberties Union (ACLU) comparing members of Congress against a database of 25,000 mugshots with Amazon's facial recognition software found that 39% of members of Congress were incorrectly matched with people who had been arrested, with a disproportionate number of Congress members of color misidentified (Altun & Humble, 2020). Utilizing facial recognition tools that have high rates of misidentification to help identify suspects could potentially lead to many cases of the wrong suspects being accused, illustrating how the outcome does not match the intent. Further, the National Department of Justice created the PATTERN algorithm as part of Congress' First Step Act that is meant to shorten criminal sentences and improve prison conditions (Federal Bureau of Prisons, 2018). PATTERN plays the role of identifying candidates eligible for early release, aiming to reduce the prison population for low-risk nonviolent offenders. However, the Department of Justice published a review of PATTERN in 2021 that the algorithm overpredicts recidivism among minority inmates at rates of 2% to 8% higher than white inmates, meaning that

people of color are less likely to be selected for early release (Lee & Lai, 2022). While the goal of artificial intelligence tools is to alleviate inconsistency and inaccuracy in judicial decisions, in practice this is not the case.

When examining COMPAS' use in the criminal justice system specifically, the algorithm's outputs that determine a convict's likelihood of reoffending are not accurate enough to be utilized in courts. One statistical analysis from the investigative journalism site ProPublica calculated that COMPAS is only accurate for two out of every three cases (Angwin et al, 2016). The data encompassed more than 7,000 individuals arrested in Broward County, Florida between 2013 and 2014 who had been scored on the COMPAS general recidivism risk scale. In addition, researchers Julia Dressel and Hany Farid at Dartmouth College issued an article on their two studies for the Science Advances scientific journal published by the American Association for the Advancement of Science, advocating against the use of COMPAS in courts throughout the US, as it is ineffective at being more accurate and unbiased than humans (2018). For the first study, people from a popular online crowdsourcing marketplace, who could be reasonably assumed to have little to no expertise in criminal justice, were given a short description of the defendant including their sex, age, and previous criminal history, but not their race (Dressel & Farid, 2018). Participants predicted whether this person would recidivate within 2 years of their most recent crime, using 1,000 defendant descriptions divided into 20 subsets of 50 each, where each participant was assigned one subset (Dressel & Farid, 2018). The mean and median accuracy for the 20 predictions was 62.1% and 64%, whereas when the COMPAS algorithm was fed these 20 sets of 50 descriptions, the median accuracy was 65.2% (Dressel & Farid, 2018). Thus, the small crowd of non-experts, only having 7 of the features compared to COMPAS' 137, were as accurate as COMPAS at predicting recidivism. To further prove their point, they also found that a simple linear predictor with only age and previous number of convictions had

approximately the same accuracy as COMPAS with a 66.8% overall accuracy (Dressel & Farid, 2018).

Not only is the algorithm inaccurate, but also it discriminates against marginalized groups and can perpetuate further injustice. The aforementioned ProPublica statistical analysis found that Black defendants were often predicted to be at a higher risk than they actually were while white defendants were often predicted to be less risky than they were (Angwin et al, 2016). The study found that over a two-year period the black defendants who did not reoffend were almost twice as likely to be misclassified as higher risk compared to their white counterparts, while white defendants who did reoffend within that period were mislabeled as low risk nearly twice as often as black reoffenders, as seen in Table 1 (Angwin et al, 2016). The same disproportionate ratios were true when looking at violent recidivists, with black defendants 77% more likely to be assigned higher risk scores than white defendants and white violent recidivists were 63% more likely to have been classified as low risk of violent recidivism (Angwin et al, 2016).

Table 1

	White	Black
Labeled High Risk, but did not reoffend	23.5%	44.9%
Labeled Low Risk, but did reoffend	47.7%	28.0%

Note: Prediction failure rates by race. Adapted from Machine Bias Risk Assessments in Criminal Sentencing, In *ProPublica,* Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Another piece published by Melissa Hamilton in the *Behavioral Sciences & the Law* academic journal found that the COMPAS risk assessment tool is sexist, utilizing the same Broward County dataset, as the results systematically over-classify women in higher risk groupings

(2019). COMPAS's developers do realize this bias and offer gender-specific versions; however, in the cases when agencies fail to incorporate gendered scoring, such as in Broward County, the COMPAS algorithm disproportionately impacts women.

The existing research proves that the use of artificial intelligence in the criminal justice system in general can be extremely detrimental, and it disproves the legitimacy of COMPAS specifically in terms of accuracy and fairness. This racial and gender bias does not only exist in the case studies, but its tangible applications also permeate the US court and prison systems with its harmful consequences. The present literature agrees that COMPAS' use perpetuates discrimination, but scholars fall short in adequately examining all of the causes at the root. One must analyze the sources of the problems to understand how to effectively implement COMPAS, or to reveal whether COMPAS can effectively be implemented at all.

The implementation of this technology reshapes the network of actors involved in the criminal justice system, impacting criminals and redefining the roles of judiciaries alike. COMPAS's impact is molded by the various actors involved in legal systems, software companies, and government institutions, functioning in a complex web. Evaluating how the introduction of a new actor into a system connects to other human and non-human actors is a valuable lens to determine whether a technology is just in its use. Michael Callon, Madeleine Akrich, Bruno Latour, and John Law's Actor-Network Theory (ANT) will be employed to understand how the use of COMPAS risk assessment critically impacts a variety of actors in the criminal justice network. ANT is summarized in Darryl Cressman's overview as "reducible neither to an actor nor to a network… An actor-network is simultaneously an actor whose activity is networking heterogeneous elements and a network that is able to redefine and transform what it is made of" (Callon 1987, p. 93, as cited in Cressman, 2009, p. 3). This view of the recidivism scoring algorithm serving as an actor in a network will not only highlight how

small biases can manifest vastly throughout the criminal justice process, but also propose how systemic change is necessary to approach more ethical artificial intelligence use. Instead of solely scrutinizing the data quality and statistical models to reduce discrimination, one must examine the entire context of the organizational structures that form the ways where discriminatory practices may or may not be produced (Schwarting & Ulbricht, 2022). This way, technologists can mitigate biases and protect historically marginalized groups to move toward equity in the digital age (Lee & Lai, 2022).

Methods:

The question I strive to answer is- How can the bias in the COMPAS risk assessment algorithm be mitigated to reduce the perpetuation of discrimination in the criminal justice system? I want to understand what systemic change must occur in order to facilitate the ethical use of recidivism scores in the criminal justice system, and if it is even plausible. To best approach researching this, I will examine existing literature through published papers as well as research articles in technology journalism. Each piece of evidence will focus on a different group's relationship to COMPAS, representing a branch in the large web of actors influencing and being influenced by the recidivism score algorithm. By analyzing articles on each of these groups, one can understand where the issues lie and the areas that each actor needs to address. These articles will center around an exposé on COMPAS published by ProPublica in 2016, and the resulting articles from the past 5 years that examine COMPAS further. Delving into these recently published articles in tech journalism will also reveal the current voices advocating for change, or possible solution pathways. In my review of this evidence, I will examine a discourse analysis through the lens of ANT, with a focus on COMPAS functioning as an actor in a network and its impacts on all human and non-human connections. I will first see how the group

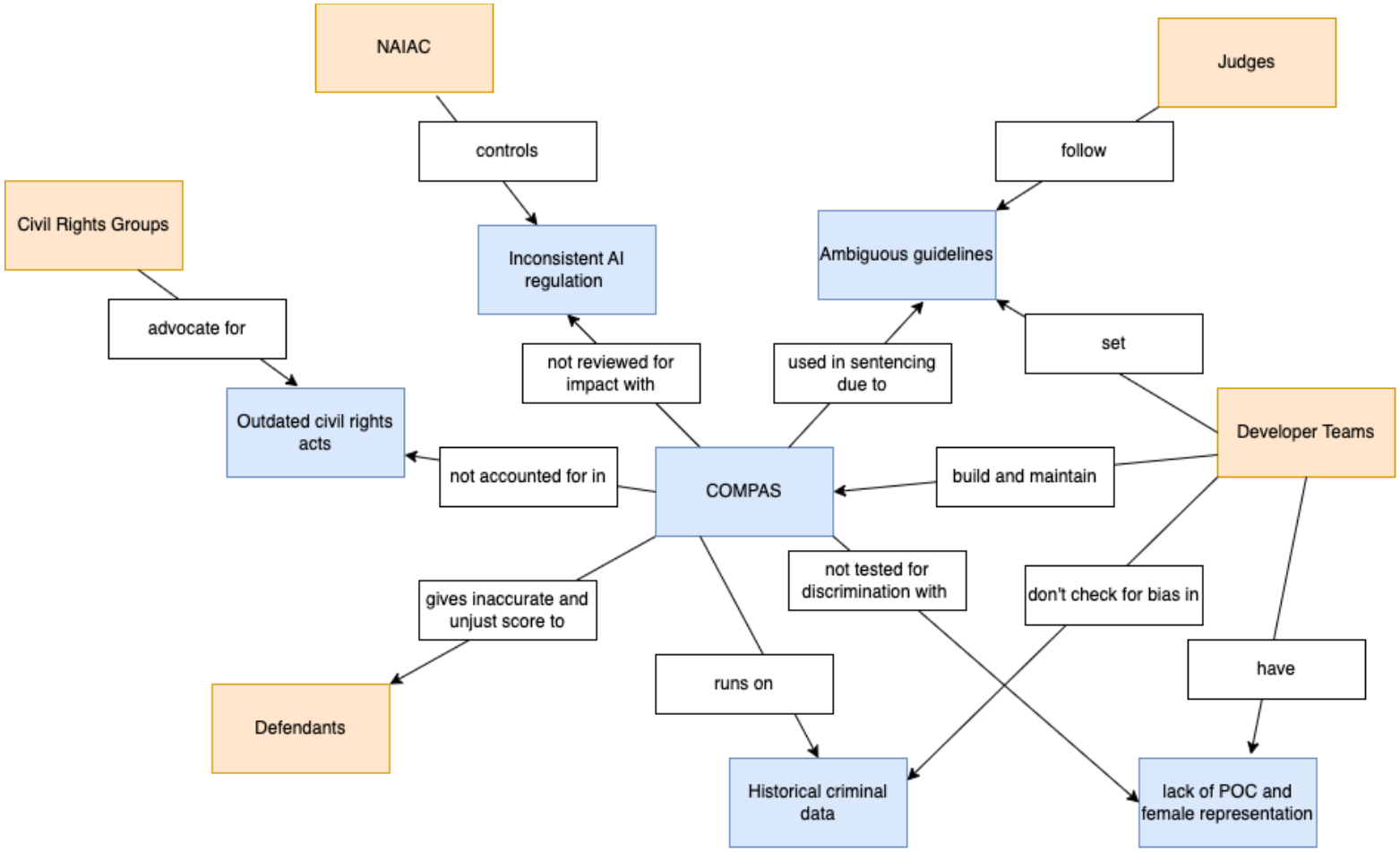
who is most directly influenced by COMPAS, the defendants, are affected, which will highlight the urgency necessary for addressing this issue and justify the problem's importance. Then, I will turn to the actors that influence COMPAS, to determine where the issues emerge. This will reveal the agents of change, whether it be technological solutions, policy changes, diversified data, or shifts in the role of risk assessments.

Analysis:

Given the recidivism score algorithm is neither accurate nor fair, the use of COMPAS infringes on the rights of the defendants. The artificial intelligence used has violated the clauses from the international law framework and the UN, including the International Convention on the Elimination of All Forms of Racial Discrimination, International Covenant on Civil and Political Rights, and the Universal Declaration of Human Rights (Altun & Humble, 2020). This violation occurs as "ethnic minorities and women already discriminated against due to existing biases have these biases exacerbated by the increased use of AI technology" (Altun & Humble, 2020). Female defendants and defendants of color are directly harmed as the use of COMPAS results in discrimination. Those who receive inaccurate high-risk scores can face unjust punishment, deeply hurting the defendants and the defendants' families. For example, Paul Zilly, who had been convicted of stealing a push lawnmower and some tools, heard his score during his sentencing hearing on Feb. 15, 2013 in court in Barron County, Wisconsin (Angwin et al, 2016). Despite the prosecutor recommending a year in county jail and follow-up supervision and his lawyer agreeing to the plea deal, Judge James Babler received results that Zilly had been rated as high risk for future violent crime and medium risk for general recidivism. Babler, stating in court the scores were "about as bad as it could be," overturned the plea deal and imposed two years in state prison and three years of supervision (Angwin et al, 2016). Babler had directly cited

COMPAS as the reason for the higher sentence, saying in an appeals hearing that “had I not had the COMPAS, I believe it would likely be that I would have given one year, six months” (Angwin et al, 2016). The results of the algorithm gave Zilly a higher sentence, demonstrating how his life was altered by COMPAS. In the ProPublica interview with Zilly, he had said that the score did not account for all the changes he was making in his life, from converting to Christianity to his battle of quitting drugs to be more available for his son, stating “not that I’m innocent, but I just believe people do change” (Angwin et al, 2016). Being reduced to a number dehumanizes the defendants, leaving them to feel futile as their fate is determined by an algorithm. Consequently, the reasons that COMPAS produces inaccurate and unjust results must be brought to light to understand how this can be resolved.

Figure 1



Note: Actor-Network Theory analysis of the discrimination resulting from the COMPAS recidivism score algorithm. COMPAS is used to determine sentencing, it gives harsher scores to people of color and women, and it continues the disproportionate rates of incarceration for minorities. This figure visualizes the different actors involved, with orange boxes representing human actors and blue boxes representing non-human actors.

The biased historical data from the US prison system utilized by the COMPAS algorithm sustains the enduring discrimination against minorities in criminal sentencing. COMPAS is formulated through historical criminal data, meaning that the data serves as a non-human actor, as seen in Figure 1. When one traces back to the root of the problem of AI biases, they result from the data that brings forth the discrimination of the past: “If we are using historically biased data sets or if we are using data sets that are challenged because of bias and discrimination and systematic racism, it means the accuracy of our decision-making systems are being undermined”

(Lee et al, 2023). The US has a history of disproportionately imprisoning people of color, with Black Americans are incarcerated in state prisons at nearly 5 times the rate of white Americans (Nellis, 2021). Therefore, utilizing prisoner history datasets to find correlations between characteristics leading to higher probability of recidivism will reflect the characteristics of people of color. Karen Hao, a journalist for the MIT Technology Review, criticizes recidivism score algorithms, explaining that machine learning algorithms find patterns in data (2019). Those patterns are statistical correlations, not the same as causations, but Hao states that these algorithms make them synonymous, turning "correlative insights into causal scoring mechanisms" (2019). Thus, the use of COMPAS can amplify these unintentional biases and generate more biased data, creating a cycle. This cyclical creation of biased data illustrates that it is not just the way the algorithm was built that brings forth bias, but also the broken system that it exists in. COMPAS will not be able to function equitably using criminal data until the racist practices in prisons, minority targeting by law enforcement, and prejudice in sentencing are eradicated. Without a resolve to these issues, there needs to be improvements to data scrutiny. Existing bias in the prison system resulting in biased data is not the only issue resulting in COMPAS' perpetuation of discrimination, though, as there are other systemic changes necessary to eradicate bias.

Lack of diversity and perspective in the teams developing risk assessment algorithms unintentionally leads to neglect in addressing AI bias. The design teams create COMPAS, and thus they serve as a valuable actor that can introduce bias into the system, see Figure 1. Utilizing artificial intelligence to assess the risk criminals pose to society illustrates how developers prescribe meaning from the output of an algorithm. These software development companies assigning meaning to AI outputs have a lack of diversity in design teams, with small numbers women to people of color. Women account for only 27.6% of the tech workforce, and Black

professionals account for only 7.4% of the tech workforce (AnitaB.org, 2022; US Equal Employment Opportunity Commission). In 2021, the Brookings Institution Center for Technology Innovation (CTI) convened a group of stakeholders to better understand and discuss the US's evolving positions on artificial intelligence. Dr. Nicol Turner Lee, the director of the CTI, and Samantha Lai, a research analyst for CTI, published after the convention that "the lack of diversity in tech spaces means that machine learning algorithms and other autonomous systems are being developed without the lived experiences necessary to avert poor data treatment, or create much better products or services" (Lee & Lai, 2022). Developers do not possess the perspectives to create and test COMPAS with the cognizance of equity for women and Blacks, unconsciously letting bias into the algorithm. One of Northpointe's founders, Tim Brennan, published a validation study for COMPAS in 2009, finding the accuracy rate of 68%; however, it did not examine racial disparities beyond that, including whether some groups were more likely to be wrongly labeled higher risk (Angwin et al, 2016). By not being racially conscious and testing for how the results affect different groups, Northpointe failed to catch its discriminatory applications prior to its introduction into the criminal justice system. Not taking into account the perspective of the people who are impacted the most by its outputs, the defendants themselves, can lead to these unbeknownst biases.

Moreover, ambiguous instruction on how to implement COMPAS results into court rulings propagates disparities in judges' decisions. The guidelines for COMPAS serve as an influencing actor as they establish how the algorithm's results will be applied, see Figure 1. In 2020, Eugenie Jackson and Christina Mendoza published an article for the MIT Press to clarify how COMPAS functions as response to the criticisms following the ProPublica article. Jackson and Mendoza explain that it is a widely accepted position that courtroom decisions should not be based solely on a recidivism risk score, arguing that the guidelines for how risk assessments

should be used issued by the National Center for State Courts (NCSC) clearly state this (2020). According to the two authors, the first NCSC guideline in summary states that a person's risk scores should not have any effect on the sentence imposed (Jackson & Mendoza, 2020). Although they argue that there are effective guidelines instructing judges on how to use COMPAS, there are many studies that disprove this. Further, Jackson and Mendoza are both employed by Northpointe to do statistical analyses and validation studies, and in the acknowledgements thanked Tim Brennan for his suggestions for the article, discrediting their stance (2020). Despite their belief that the standards for COMPAS' use are effective, there are many discrepancies in its application that do not match the intended use. The algorithm is incorporated into state judicial systems drastically differently, from certain states like Wisconsin using it for each step in the prison system from sentencing to parole, to other states not using it at all (Mesa, 2021). This incongruity between states leads to no clear standard on the score's value, rendering the guidelines ineffective. Further, research conducted by the Cologne Journal for Sociology and Social Psychology examining how the COMPAS score is embedded in courts formally and informally found that the score represents an ambiguous and redundant source of information for judges (Schwarting & Ulbricht, 2022). Therefore, the NCSC's guidelines do not successfully convey that COMPAS is only an aid to sentencing, leading to ambiguous values and disparate applications.

Given the ineffectiveness of the current standards, judges misuse COMPAS. As judges use the results to determine sentencing, they bring into fruition the consequences of the score that the algorithm produces, and improper use can perpetuate COMPAS's biases, depicted in Figure 1. While COMPAS's developers trusted that judges are proficient at using COMPAS and can effectively receive scores without influencing decisions, in practice it remains unclear for judges how the score should be utilized and the weight the results should hold, leading to further

inefficiencies without clear direction. And despite intentions to have the score determine which defendants are eligible for probation or treatment programs, not to give longer sentences to higher risk score defendants, in reality this does not occur. There have been specific cases where judges explicitly cite COMPAS, such as the case with Eric Loomis who was charged with driving a stolen vehicle and fleeing police (Angwin et al, 2016). In 2013, Judge Scott Horne in Wisconsin declared that the defendant Loomis had been "identified, through the COMPAS assessment, as an individual who is at high risk to the community," sentencing him to 8.5 years in prison (Angwin et al, 2016). This undermines Jackson and Mendoza's claim that a person's risk scores should not have any effect on the sentence imposed, and it illustrates how judges improperly using COMPAS carries out harsh sentences (2020). Further, judges are determining these sentences based on an algorithm that they have no access to and do not understand. When Loomis brought his case to the Wisconsin Supreme Court saying that COMPAS' use violated due process as the risk assessment tool was never disclosed, the court defended Horne's use of the score despite the algorithm being never being made available to the court or Loomis (Greeenstein, 2022). Therefore, judges are being influenced by software that they can neither access nor know how it works, as Northpointe had refused following this court case to make the software available, citing that it was proprietary and a core business secret (Greenstein, 2022). Given that judges are using the results of confidential software, they cannot effectively make independent decisions. Therefore, the use of COMPAS undermines the professional decision autonomy of the judges, and it can lead to discrepancies without guidelines or training on its application. Not only are court level standards necessary, but also national AI policy.

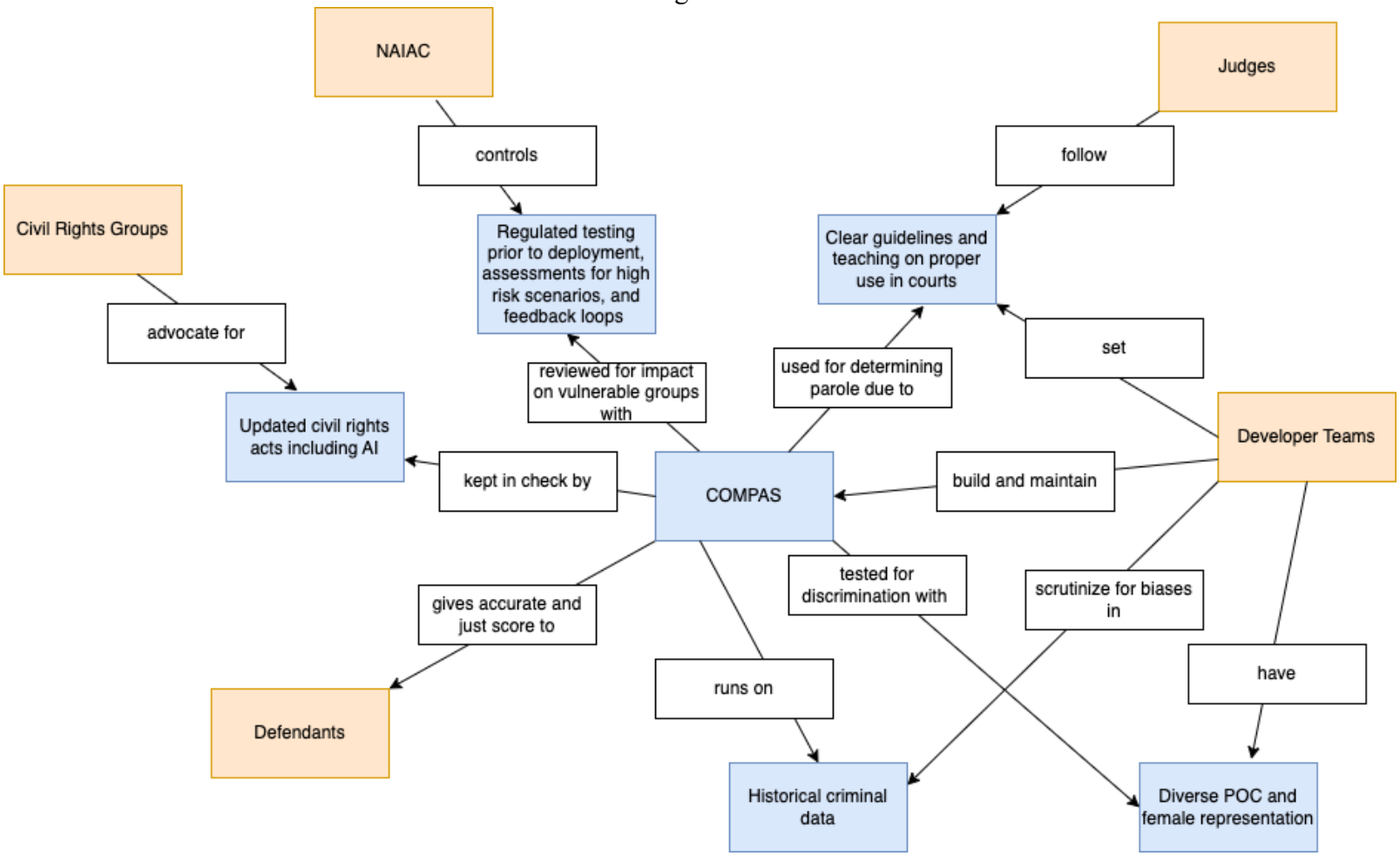
Without a national AI governance strategy, the bias found in COMPAS cannot be controlled and mitigated. The US government establishes policy for AI governance which in turn determines the criteria recidivism score algorithms must meet, see Figure 1. Although some

technological enthusiasts may argue that artificial intelligence should not be restrained by stringent US regulation as software freedom in the US has led to impressive AI growth in the past years, the aforementioned problems with COMPAS indicate a need for reform. The new National AI Advisory Committee (NAIAC) formed in 2020 has been established to keep AI companies in check (2023). However, there still exists a lack of interdisciplinary experts and social activists having their voices heard in discussions of AI policy. According to Lee and Lai (2022) following the meeting with AI stakeholders, civil rights activists have fought for equity for years, but the use of AI can reverse such progress and not support those hurt by discrimination. The current civil rights regime according to them is ill-equipped and outdated, with various acts that are meant to protect from discrimination not accounting for automation in that area (Lee & Lai, 2022). In addition, Lee and Lai (2022) argue that current infrastructure lacks disclosures for those most impacted by these technologies. Even for sensitive algorithms that cannot be disclosed, such as COMPAS, there is no consistent reviewing process to evaluate the long-term impacts on more vulnerable groups (Lee & Lai, 2022). Although the formation of the NAIAC was a step in the right direction, the US government lacks adequate involvement in order to prevent recidivism score algorithms from propagating inequity. An absence of national AI governance results in a lack of protection of marginalized groups from the potential bias that is perpetuated, as seen with COMPAS.

Conclusion:

The complex web of actors connected to the COMPAS algorithm reveal where each group lacks, contributing to perpetuation of discrimination with COMPAS' use. However, by understanding each actor's faults, we can create solutions to address them and prevent the harm caused to defendants, depicted in Figure 2.

Figure 2



Note: Actor-Network Theory analysis of the potential solutions to mitigate discrimination resulting from the COMPAS recidivism score algorithm. COMPAS would only be used to determine parole eligibility, it would continuously be in check for appropriate and unbiased use, and it would help reduce the prison population of low-risk non-violent offenders to combat disproportionate incarceration rates. This figure visualizes the different actors involved, with orange boxes representing human actors and blue boxes representing non-human actors.

While race is not an explicit variable in the COMPAS algorithm, variables that are correlated with race are, such as poverty and unemployment, and being race blind leads to the cyclical generation of biased data. By acknowledging that historical criminal data is biased, and by understanding where this prejudice derives from in the prison system, developers can work toward scrutinizing for data biases and improving which datasets and variables are utilized.

Moreover, the diversification of developer teams can aid in gaining new perspectives that would be more conscious of discrimination and ambiguity in the algorithm. Incorporating input of those who are not only using the results but also who are affected by the results can also contribute new insight for developer teams. Implementing anti-racist approaches to all steps of design, such as regular testing for racial and gender disparities, can help improve recidivism score algorithms from the company and developer side.

On the court side, ensuring that the score is only used for determining probation conditions and not lengthening sentences through adequate training and more clear guidelines would help prevent judges from being wrongly influenced by the score. The professional decision autonomy of judges should be maintained, and it should be standard that judges do not rely on the score. Judge Mark Boessenecker trains judges around the state of Florida on evidence-based sentencing, and he cautions them to not use the score as an indicator if a person is dangerous or should go to prison (Angwin et al, 2016). Having the judges who train others ensure that the score is decoupled from the defendant would prevent the abuse of COMPAS to sentence prisoners.

In terms of national AI governance, there needs to be guardrails in place for systems that can replicate its inequities, such as the criminal justice system. The NAIAC could be a vessel to incorporate frameworks that would classify use cases by varying degrees of risk to determine appropriate levels of regulation (Lee & Lai, 2022). To manage biased AI, there needs to be identification procedures and assessments in place for addressing high risk scenarios like discrimination in COMPAS. Not only should the NAIAC improve regulation, but also involve participation of civil rights activists and defendants affected alike. To protect the civil rights of historically marginalized groups, an evaluation of the existing civil rights regime is a step toward more responsible AI and greater equity, as handling AI discrimination can be incorporated into

equity acts. Furthermore, at present direct feedback to those creating recidivism score algorithms is limited. By mandating for testing in a regulated environment prior to deployment and requiring feedback loops to get input on algorithms that pose substantive risk to citizens could be used in efforts to remove bias (Lee & Lai, 2022).

The aforementioned solutions could be applied by policymakers, developers for recidivism score algorithms, judiciary authorities, and advocacy groups against rapid and unchecked AI growth. Future researchers could utilize these findings as a jumping point to explore in more depth how potential solution pathways can mitigate the discrimination perpetuated by COMPAS. Their future research could model how implementing national policy, development team improvements, data scrutiny, and court guidelines could produce more accurate and just results for recidivism score algorithms, and research can help determine the effectiveness of the solutions as opposed to elimination of its use altogether. The analysis of the ethical issues of the COMPAS algorithm generates an appreciation for the complexity of introducing AI into different systems in society, and how its application is directly and indirectly shaped by a diverse range of actors. Although the systemic change necessary is a great feat, with joint efforts from legislators, court systems, tech companies, civil rights groups, and social activists, AI bias could be mitigated, creating a fairer legal system and achieving pivotal prison reform.

References:

Altun, D. & Humble, K. P. (2020). Artificial intelligence and the threat to human rights. *Journal of Internet Law,* 24(3), 1-18. Retrieved from https://gala.gre.ac.uk/id/eprint/30040/

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 23). *Machine bias risk assessments in criminal sentencing.* ProPublica. Retrieved from https://www.propublica.org/article /ma chine-bias-risk-assessments-in-criminal-sentencing

AnitaB.org. *2022 top companies for women technologists.* (2022, December 13). Retrieved from https://anitab.org/research-and-impact/top-companies/2022-results/

Beck, A., & Bonczar, T. (1997, March). Lifetime likelihood of going to state or federal prison. *Bureau of Justice Statistics.* Retrieved from https://bjs.ojp.gov/content/pub/pdf/ Llgsfp.pdf

Cressman, D. (2009, April). *A brief overview of Actor-Network Theory: punctualization, heterogeneous engineering & translation*. CT Lab/Centre for Policy Research on Science & Technology *(CPROST)*, 1-17. Retrieved from https://summit.sfu.ca/item/13593

Chohlas-Wood, A. (2020, June 19). *Understanding risk assessment instruments in criminal justice.* The Brookings Institution. Retrieved from https://www.brookings.edu/research/ understanding-risk-assessment-instruments-in-criminal-justice/

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC57 77393/

Federal Bureau of Prisons. BOP. (n.d.). Retrieved from https://www.bop.gov/inmates/fsa/ overview.jsp

Federal Bureau of Prisons. Inmate ethnicity. (2023, April 22). Retrieved from https://www.bop.gov/about/statistics/statistics_inmate_ethnicity.jsp

Federal Bureau of Prisons. Inmate race. (2023, April 22). Retrieved from https://www.bop.gov/about/statistics/statistics_inmate_race.jsp

Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artificial Intelligence and Law*, 30(3), 291-323. Retrieved from https://link.springer.com/article/10.1007/s10506-021-09294-4

Hamilton, M. (2019) The sexist algorithm. *Behav Sci Law,* 37(1), 145– 157. Retrieved from https://pubmed.ncbi.nlm.nih.gov/30931534/

Hao, K. (2019, January 21). *AI is sending people to jail- and getting it wrong*. MIT Technology Review. Retrieved from https://www.technologyreview.com/2019/01/21/137783/algorithms -criminal-justice-ai/

Herring, T., & Widra, E. (2021, September). *States of incarceration: The global context of 2021.* Prison Policy Initiative. Retrieved from https://www.prisonpolicy.org/global/2021.html

Herring, T., Sawyer, W., Wang, L., & Widra, E. (2022, April). *Beyond the count: A deep dive into state prison populations.* Prison Policy Initiative. Retrieved from https://www.prisonpolicy.org/reports/beyondthecount.html#:~:text=Race%20and%20ethnicity,-Nationwide%2C%20state%20prison&text=In%20state%20prisons%2C%2034%25%20of,Hawaiian%20or%20other%20Pacific%20Islander.

Jackson, E. & Mendoza, C. (2022, Mar 31). *Setting the record straight: What the COMPAS core risk and need assessment is and is not.* MIT Press. Retrieved from https://hdsr.mitpress.mit.edu/pub/hzwo7ax4/release/7

Lee, N. T., Cummings, R., & Rice, L. (2023, February 8). *TechTank podcast episode 39: Civil rights and artificial intelligence - can the two concepts coexist?* The Brookings Institution. Retrieved March 6, 2023, from https://www.brookings.edu/blog/techtank/2022/03/07/techtank-podcast-episode-39-civil-rights-and-artificial-intelligence-can-the-two-concepts-coexist/

Lee, T. L., & Lai, S. (2022, May 17). *The U.S. can improve its AI governance strategy by addressing online biases.* The Brookings Institution. Retrieved from https://www.brookings.ed u/blog/techtank/2022/05/17/the-u-s-can-improve-its-ai-governance-strategy-by-addressing-online-biases/

Mesa, N. (2021, May 13). *Can the criminal justice system's artificial intelligence ever be truly fair?* Massive Science. Retrieved from https://massivesci.com/articles/machine-learning-compas-racism-policing-fairness/

Nellis, A. (2021, October 13). *The color of justice: Racial and ethnic disparities in state prisons.* The Sentencing Project. Retrieved from https://www.sentencingproject.org/reports/the-color-of-justice-racial-and-ethnic-disparity-in-state-prisons-the-sentencing-project/

Population Reference Bureau. (n.d.). *U.S. has world's highest incarceration rate.* Retrieved from https://www.prb.org/resources/u-s-has-worlds-highest-incarceration-rate/

Roa Avella, M. del P., Sanabria-Moyano, JE, & Dinas-Hurtado, K. (2022). Use of the COMPAS algorithm in criminal proceedings and risks to human rights. *Brazilian Journal of Criminal Procedural Law*, 8 (1). Retrieved from https://www.scielo.br/j/rbdpp/a/6W9b8CHYbXcsc6 qczDxCSfr/abstract/?lang=en

Schwarting, R., Ulbricht, L. (2022) *Why organization matters in "algorithmic discrimination".* Köln Z Soziol 74(Suppl 1), 307–330. Retrieved from https://ideas.repec.org/a/zbw/espost/261200.html

The National AI Advisory Committee (NAIAC). National Artificial Intelligence Initiative. (2023, March 1). Retrieved from https://www.ai.gov/naiac/

U.S. Census Bureau. (2023). *Quick facts: United States.* Retrieved from https://www.census.gov/quickfacts/fact/table/US/RHI225221#qf-headnote-a

U.S. Equal Employment Opportunity Commission. *Diversity in high tech*. US EEOC. (n.d.). Retrieved from https://www.eeoc.gov/special-report/diversity-high-tech