



IMSIS

International Master
Security, Intelligence
& Strategic Studies



**Erasmus
Mundus**

Title: Machine Learning Applications in the United States

Criminal Justice System

Subtitle: A Critical Content Analysis of the COMPAS

Recidivism Risk Assessment

August 2021

GUID: 24870703B

DCUID: 19108346

CUID: 51456294

**Presented in partial fulfilment of the requirements for the Degree of
International Master in Security, Intelligence and Strategic Studies**

Word Count: 21,348

Supervisor: Dr. Petr Špelda

Date of Submission: 02/08/2021



CHARLES UNIVERSITY

Abstract

Artificial intelligence and machine learning (AI/ML) models are increasingly utilised in every aspect of life and society due to their superhuman abilities to digest large amounts of data and find obscure patterns and correlations. One contentious area of this technological application is in the criminal justice system, where AI/ML is used as a recommendation or decision-making support tool. These applications are particularly popular in the United States of America (USA), the nation with the highest rate of incarceration and correctional budget, to aid in managing overcrowded and overspending facilities. Angwin et al.'s (2016) ground-breaking study found the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) model to be biased against Black defendants and sparked an influential academic debate around algorithmic bias and fairness. This study aims to fill the gap in the scholarship by focusing on the content of COMPAS's recidivism risk assessment questionnaire through a qualitative content analysis within the conceptual framework of Critical Race Theory (CRT). The findings presented in this research are twofold: (1) almost half of the COMPAS questions were opinion-based, thus reducing quantitative neutrality, and (2) there were significant proxy factors for race that could have led to biased results in the model. Implications of these findings are discussed.

Keywords

Algorithmic fairness, AI/ML models in policing, AI/ML models in the criminal justice system, Policing in the USA, Recidivism risk assessments, COMPAS assessment, Algorithmic bias, Disparate impact, Critical Race Theory and AI/ML, Critical content analysis

Table of Contents

1. Introduction	5
2. Background.....	7
2.1.1 Machine Learning Algorithmic Models	8
2.1.1 Definitions and AI/ML Background	8
2.1.2 The Quantitative Search for ‘Fairness’	12
2.1.3 Sources of Bias.....	16
2.1.4 The AI/ML ‘Black Box’ Mystery.....	21
2.2 Policing and ML/AI: The Case of COMPAS.....	22
2.3 Policing and Black America.....	28
2.3.1 Cementing Racial Hierarchies.....	29
2.3.2 The Thirteenth Amendment’s Powerful Loophole	31
2.3.3 Lasting Impacts of Legalised Segregation and the ‘Justice Apartheid’.....	34
3. Conceptual Framework.....	38
3.1 Critical Race Theory.....	39
4. Research Design and Methodology.....	44
4.1 Methodology:	45
4.2 Data Collection.....	47
4.3 Data Analysis	48
5. Findings	51
5.1 Current Charges, Criminal History, Non-Compliance, and Family Criminality	52
5.2 Gang Membership.....	57
5.3 Peers.....	60
5.4 Substance Abuse	62
5.5 Residence/Stability.....	63
5.6 Social Environment	65
5.7 Education	67
5.8 Vocation (Work)	69
5.9 Leisure/Recreation, Social Isolation, Criminal Personality, Anger, and Criminal Attitudes.....	72
6. Conclusion.....	74

7. Bibliography	77
Appendix A	115
Appendix B	129

1. Introduction

The United States of America (USA) has a corrections problem. With a rate of 639 incarcerations per 100,000 individuals, the USA holds the highest global incarceration rate per capita in recent history (Bureau of Justice Statistics, 2021). At any given point in time, there are at least two million people in American prison or jail and over three million people on parole (The Sentencing Project, 2020a). The USA also has a big recidivism issue; within five years of release, 76% of ex-convicts will be once again behind bars (Smalls et al., 2020). Not only are there a disproportionate amount of Americans incarcerated, but an even higher asymmetrical number of Black Americans are incarcerated or imprisoned every year. Despite making up 12% of the adult American national demographic, Black Americans constitute 33% of the incarcerated population, being on a national average five times more likely to be imprisoned than White Americans – although several states like Wisconsin or New Jersey increase this likelihood up to eleven or twelve times (Gramlich, 2020; The Sentencing Project, 2020b). Increased imprisonment has been linked in the literature to policing attitudes that are deleterious to People of Colour (POC), where they experience excessive suspicion, targeting, and aggressive or violent treatment (Bilotta, et al., 2019). These hostile attitudes stem from as far back as the establishment of the first colonies on the North American terrain, where a long and deeply institutionalised history of enslaved labour, legal segregation, and mass incarceration have instilled fundamentally racialised roots in all areas of American society and governance.

In 2019, it was calculated that US states had collectively spent \$56.6 billion in corrections expenditures (The Sentencing Project, 2020a). To avoid overcrowding and more effectively allocate policing resources, law enforcement authorities have been implementing

machine learning tools to provide a range of assessments, from mapping out areas with more expected crime to judging the likelihood of prisoner recidivism (Shapiro, 2017). These algorithmic assessments are influential tools that guide decision-makers through every level of the criminal justice system, from location of officer deployment to bond-setting and sentencing (Angwin et al., 2016; Courtland, 2018). However, the study of its social effects remains lacking and in contention whilst several key authors argue that potential biases found within data could lead to harmful effects for marginalised groups. Indeed, one key investigation into potential algorithmic biases found that COMPAS, a recidivism risk model applied for setting bail in Broward County, Florida, was classifying Black arrestees as higher risk than White arrestees (Angwin et al., 2016). The COMPAS assessment was also statistically incorrect, as the investigation demonstrated that Black arrestees classified as high risk were half as likely of recidivate, whilst White arrestees listed as low risk were conversely twice as likely to recidivate (Angwin et al., 2016). This prominent study sparked several notable debates in the literature, from how to measure 'bias' and 'fairness' in algorithmic models to the ethical implications of algorithmic model application in socially significant areas and the use of historical data in these models.

This research aims to offer a deeper analytical context of the COMPAS model to contribute to this ongoing discussion in the literature. This study will examine the COMPAS questionnaire, which provides a list of questions posed to arrestees that are then inputted into an algorithmic model to gauge their probability of recidivism. Although there has been significant literature debating Angwin et al.'s (2016) definitions of 'fairness' in algorithmic modelling and the paper's accompanying mathematical calculations, there have been, to the knowledge of the researcher, no efforts in the literature to examine this key questionnaire

used in the COMPAS algorithmic model. The researcher therefore aims to fill this gap by carrying out a qualitative critical content analysis of the COMPAS questionnaire within the conceptual framework of critical race theory, in order to bridge the scholarships on algorithmic sources of bias and biased policing and justice practices in the USA. Furthermore, this research finds itself in a highly relevant environment given the increasing popularity of the implementation of these algorithmic tools, and, furthermore, falls within wider conversation surrounding the policing and profiling of black communities in the wake of the Black Lives Matter movement.

Chapter 2 of this research will provide a relevant contextual background to the question at hand. It will first discuss issues in the AI/ML academic literature pertinent to this research and then delve deeper into outlining the subject of study: the COMPAS recidivism risk assessment. Lastly, this chapter will make a case, built on the well-documented literature, for the relevance of American history in its foundational institutionalisation of racism in the criminal justice system and what impacts this could incur to algorithmic models operating within US correctional departments. Chapter 3 will introduce the conceptual background that will be implemented as part of this study's analysis: Critical Race Theory. Chapter 4 will outline the methodology applied to the analysis and how the data was codified. Chapter 5 will present the findings of the qualitative critical content analysis performed on the COMPAS questionnaire. Lastly, Chapter 6 will conclude on the implications of these findings and future research.

2. Background

This research finds itself at the nexus of algorithmic bias, algorithmic models used in the criminal justice system, and systematic racial inequalities in the USA within the criminal justice system. Section 2.1 of

the literature review will examine the current academic standing on AI/ML algorithmic models, providing definitional clarity, background on the ‘fairness debate,’ potential sources and explanations of bias, and outline issues of inexplicability. Section 2.2 will then delve specifically into the COMPAS model and why it is appropriate to examine for this study. Lastly, Section 2.3 provides a contextual history of Black American relations with the US criminal justice system in order to define a proper background understanding for Chapters 3 and 5.

2.1.1 Machine Learning Algorithmic Models

This first section will explore the literature relevant to the research question at hand within the field of artificial intelligence and machine learning.

2.1.1 Definitions and AI/ML Background

Although often interchangeably utilised¹, machine learning falls under the category of artificial intelligence. With many different definitions across the literature, the author has opted for Bundage et al.’s definitions due to their succinctness within the scope of this research (2018). The term ‘artificial intelligence’ concerns “the use of digital technology to create systems that are capable of performing tasks commonly thought to require intelligence,” and ‘machine learning’ refers to “digital systems that improve their performance on a given task over time through experience” (Brundage et al., 2018, p.9). These systems will discover patterns in the data based on the problem that the computer scientist has defined and will automate this pattern-discovering process to reveal an “accumulated set of discovered relationships [that] is commonly called a ‘model,’” which will then subsequently be applied to new data (Barocas and Selbst, 2016, p.677). For example, an AI/ML system could

¹ As this research will indeed do, by referring to these terms mutually as AI/ML, as in “artificial intelligence and machine learning.”

be created to enhance online consumer experience by learning over time what purchases are associated with each other to recommend relevant products to customers. This model would find that certain purchases tend to go with one another or may be particularly popular during certain times of the day. These algorithmic models today can do what no human-controlled statistical formula could have done years ago; they can take immense quantities of data, also known as 'big data,' and 'mine' it for useful correlations, that over time will become more and more accurate as more data is fed into the model. These discovered correlations become strengthened with more clicks on algorithmically-recommended purchases or friends you may know on social media applications, for example. As outlined in Brundage et al.'s (2018) definition, these models learn their rules and behaviours for processing data through 'experience,' what is known in the AI/ML field as 'training data.' This is used when building the model to ensure that it is picking up on the correct observations within the training dataset (Barocas and Selbst, 2016). Machines may learn via supervision, where the data is labelled (such as 'human face' versus 'not human face') or may be unsupervised and thrown into the dataset to learn groupings and associations on their own (Levendowski, 2018). Algorithmic models require significant amounts of training data to pinpoint correlations and patterns within large datasets. They will implicitly infer rules based on these discovered patterns, and these rules will become formalised in the model and thus systematically applied throughout afterwards when new data is inputted (Barocas and Selbst, 2016). The definitions above are meant for the reader to obtain a basic understanding of AI/ML functioning as it relates to this research. As such, the various methods for teaching machine learning systems

and discussions around them fall out of the scope of this research and will not be addressed².

By being able to find patterns in vast amount of data, AI/ML applications have quickly been developed in seemingly every sector, from optimising Wi-Fi performance (Krishnan et al., 2018) to detecting phishing scams (Bahnsen et al., 2018), and from clinical-decision making during the COVID-19 pandemic (Debnath, et al., 2020) to reducing human error in the detection of breast cancer (Wang et al., 2016). Not only can these models detect patterns and anomalies in enormous amounts of data oftentimes better than humans (Topol, 2019), but proponents of these applications argue that they may also reduce human error and bias in decision-making (Chiao, 2019). In the same vein, supporters of machine learning applications in policing bring up the highly notable point that decisions made in the criminal justice system, potentially life-changing decisions at that, are made by humans with their own implicit biases and decision-making heuristics (Chiao, 2019). Indeed, previous studies have shown the influence of heuristics, unconscious biases, and extraneous factors in judges and judicial decisions (Englich, 2009; Goodman-Delahunty and Sporer, 2010; Danziger, Levav, and Avnaim-Pesso, 2011). Therefore, this camp in the literature argues that, if anything, machine learning models may be fruitful in resolving issues of arbitrary sentencing and biased judgment that may occur in human decision-making (Chiao, 2019). For instance, Kleinberg et al. (2017) built a model using over one million bond court cases and found high potential for welfare gains in the criminal justice system, where a simulation reduced a hypothetical jail population by 42% with no increase in crime, evidencing that machine learning's

² For an introduction to AI/ML, see: Nilsson, 1998. For a more in-depth and contemporary understanding of AI/ML methods, see: Shalev-Schwartz and Ben-David, 2014. For a review of formal AI/ML methods, see: Urban and Miné, 2021.

accuracy based on large sources of data has the potential to mitigate against human errors. However, it is important to note that even though Kleinberg and colleagues support the application of AI/ML in socially significant decision-making, they argue, that this emerging technology must be kept on a short leash and that its proposed value cannot be realised without proper policies, education, and careful constrictions in place (Kleinberg et al., 2017; Kleinberg et al., 2018b; Abebe et al., 2019). Although promising in its desiderata, there have also been plenty of evidenced instances of AI/ML models demonstrating discriminatory decision-making or converging into homogeneous recommendations³ that reduce social welfare (Kleinberg and Raghavan, 2021). As will be thoroughly discussed in Section 2.1.3, machine learning is not without bias.

The other camp of this debate contends that AI/ML applications may indeed be helpful to experts in certain situations, such as providing medical second opinions (Raghu et al. 2018), but that the use of AI/ML in such high-stakes social issues have significant ethical, moral, and political consequences (Berman and Hirschman, 2018; Sareen, Saltelli, and Rommetveit, 2020). Indeed, this camp argues that by making social decisions through numbers, we are morally undermining the value of cornerstone concepts in society, such as criminal justice. Barman (2016) and Bigo, Isin, and Ruppert (2019) reflect on the political process of quantifying society and societal values. Not only does this process involve the politicisation of definitions and calculations of the quantified issue, but also of ensuring that all the appropriate stakeholders are involved. Additionally, by quantifying social issues, we run the risk of obfuscating issues that are not as clearly quantifiable but still present

³ Also known as ‘algorithmic monoculture’ or ‘algorithmic curation,’ this concept reflects algorithmic models’ tendency towards greatest optimisation, therefore leaving out diversity (Kleinberg and Raghavan, 2021).

(Sareen, Saltelli, and Rommetveit, 2020). O’Neil (2016), for example, developed a rubric measuring the destructiveness of an algorithm, with the three main tests of opaqueness, scalability, and damage as the deciding factors for its applicability in society and noted that even though algorithmic models tend to be highly opaque, they are still being increasingly applied in decision and policy-making processes. Markham, Tiidenberg, and Herman (2018) warn of a “crisis in accountability” from the ubiquitous use of data in making “societal interventions” (Crawford and Schultz, 2019). Indeed, AI/ML’s burgeoning applications at all levels of society leave many unanswered questions about who is accountable for potential faults.

Overall, the above survey of literature showcased the ongoing debate regarding AI/ML applications in society. The following section will discuss the ongoing debate in the literature regarding the algorithmic definition and calculation of ‘fairness’ and will introduce the implications of such debate.

2.1.2 The Quantitative Search for ‘Fairness’

A final algorithmic definition, or lack thereof, that is of particular relevance to this research is that of ‘fairness.’ As will be outlined in Section 2.1.3, AI/ML models are prone to inherit human biases, thus producing, in certain circumstances, ‘unfair’ results towards groups that may have been over or under-represented in the data. This issue of ‘fairness’ and its definition, not only within AI/ML, but also in philosophy and ethics, has been long debated and is of high contention particularly within the new and quickly evolving field of algorithmic fairness. Indeed, at the first Fairness, Accountability and Transparency Conference (FAT), Narayanan (2018) presented 21 different definitions of fairness used in computer science, which are nowhere near close to the comprehensive list of definitions found in the literature. In the context of racial

discrimination in the United States, legal definitions of equal opportunity and discrimination can be found in the 1964 Civil Rights Act, for example, but they fail to indicate to the computer scientist curating the dataset or writing the code how to define the problem and determine the values and their respective weights to an AI/ML system. An early example of this precise question can be found in St. George's Hospital, London, where staff decided in 1979 that they were receiving too many medical school applications and wanted to utilise a computer program to filter out applicants at the first level of the recruitment process to reduce workload (Lowry and Macpherson, 1988). So, the Hospital administrative staff developed a computer program for this very question that was 90-95% correlated with the historical selection that had previously been made since the staff wanted the program to continue making the same decisions as before but in an automated fashion. By 1982, the program filtered all initial applications, and it was not until 1986 that the staff noticed, what had once been a relatively diverse hospital, was now interviewing White males and filtering out women and male applicants with non-European sounding names of equal credentials, despite race not being a datapoint in this program (Lowry and Macpherson, 1988). Although this program was not designed to discriminate against certain applicants, it ended up unintentionally creating discriminatory practices, also known as having 'disparate impact,' a notion born in US anti-discrimination laws, which Feldman et al. (2015) introduced and quantified into the computer science literature. Several explanations for this disparate impact will be discussed in the following Section 2.1.3, however, this example, and many after it, bring up fundamental questions about our societal biases and institutional fairness.

One recommendation after this fault was discovered was to add race as a datapoint to ensure diversity in the interview pool, but there have been many other efforts at discrimination mitigation and correction

in AI/ML since then (Lowry and Macpherson, 1988). If we treat everyone equally in the model, then we have gender and racial disparities, as demonstrated at St. George's Hospital. On the other hand, if we ensure in the model that groups that were left out are included *ex post*, we are essentially applying some form of affirmative action⁴ to the model, which has its own realm of debate regarding the ethics of 'positive discrimination' (Barocas and Selbst, 2016). One prominent definition in the literature that is relevant to this research is that of 'statistical parity,' which argues for group fairness and is achieved when the 'protected' group, or the discriminated demographic, has the same probability of being classified in the model as the unprotected group (Fish, Kun, and Lelkes, 2016). However, Dwork, et al. (2011) demonstrate that statistical parity is not compatible with individual fairness and therefore do not deem it a 'fair' calculation of 'fairness.' Another key definition argued for in the literature is 'predictive parity,' which considers the predicted outcomes and states that for a model to have predictive parity, the percentage of correct predictions should be equal across groups (Dieterich, Mendoza, and Brennan, 2016). Flores et al. (2016) critique predictive parity and employ the notion of 'calibration,' which asserts that, unlike predictive parity, there should be equal classification of risk irrespective to group membership for a model to be fair. On the other hand, looking at incorrect predictions, there is the notion proposed in the scholarship of 'error rate balance,' stating that both groups must receive equal false positive and false negative rates⁵ for the model to be 'fair' to both groups (Chouldechova, 2017). Other notable definitions of 'fairness'

⁴ A highly contentious topic in the US, affirmative action aims to positively favour minorities where they may be under-represented. For a comprehensive discussion on the nature of this debate and its application in different areas of American society, see: Leiter and Leiter, 2011; Moses, 2016; Carter and Lippard, 2020.

⁵ A false positive rate is the percentage of 'negative' individuals incorrectly classified as 'positive' (so, those with low-risk of re-arrest classified as high-risk) and conversely a false negative rate is the percentage of 'positive' individuals classified as 'negative' (those with high-risk of re-arrest being classified as low-risk) (Verma and Rubin, 2018).

found in the literature are: Fish, Kun, and Lelkes' (2016) 'resilience to random bias,' that deems a model fair if it can recover from a random feature being introduced, Kilbertus et al.'s (2017) use causal reasoning to remove unresolved discrimination in models, and Datta et al.'s (2017) notion of 'proxy non-discrimination' which judges models by the use of proxies of protected groups that are strongly correlated to those groups (further showcased in Chapter 5). These top definitions of 'fairness' have been at the centre of the debate regarding the COMPAS algorithmic tool for assessing re-arrest, which is the focus of this research and will be thoroughly discussed in Section 2.2. Studies have mathematically proven that these prominent algorithmic definitions of 'fairness' in the literature cannot be simultaneously satisfied, thus establishing unavoidable trade-offs regarding which notion of 'fairness' to choose, social implications for protected groups, and model accuracy, in what has been dubbed as the 'impossibility theorem' of AI/ML 'fairness' (see: Kleinberg, Mullainathan, and Raghavan, 2016; Chouldechova, 2017; Corbett-Dabies et al., 2017; Saravamakumar, 2021). As demonstrated by the review of this debate found in the literature, the jury is still out for an accepted definition and measurement of 'fairness' in AI/ML. Additionally, the scope of this research does not allow to delve further into of 'fairness' definitions and calculations⁶.

Computer scientists remain plagued with questions regarding the definition and measurement of fairness. If we do want certain groups to have equal representation, how do we measure this? And how do we ensure that other groups are also fairly treated? As the above-summarised literature debate regarding this subject demonstrates, "fairness is a value-driven concept and not a technical feature of ML

⁶ For further discussion on the definition and understanding of the concept of 'fairness' both in social and computer sciences, see: Rabin, 1993; Young, 1995; Roemer, 2000; Rawls, 2001; Bansal and Sviridenko, 2006; Kleinberg et al., 2018a; Verma and Rubin, 2018; Barocas, Hardt, and Narayanan, 2021.

models,” so each definition will depend on the intended use of the model and its context, trade-offs, disparate impact, and stakeholders (Miron et al., 2020, p.114; Narayanan, 2018). Additionally, with regards to these types of models that affect decision-making in key parts of society, there exist moral qualms with leaving these important and life-changing decisions up to the calculations of a model which may or may not be biased, such as the criminal justice system (Chiao, 2019). It is therefore important for computer scientists to consistently remember that “the goal is to build algorithmic systems that further human values, which cannot be reduced to a formula,” and, importantly, to ensure that they do not contribute to any societal harm (Narayanan, 2018). The following section will delve into the specifics of AI/ML bias and the academic discussion on its causes.

2.1.3 Sources of Bias

Despite the positive points in favour of machine learning applications outlined above, AI/ML suffer from bias. Every so often we see a story in the news about technology demonstrating discriminatory practices or reflecting social bias, such as Google Translate using gender stereotypes when translating from languages with gender-neutral pronouns or Amazon’s same-day delivery not covering zip codes of predominantly Black neighbourhoods (Ingold and Soper, 2016). Brought together, these disparities are even more stark at the intersectional level, as studies conducted by Klare et al. (2012) and Boulamwini and Gebru (2018), for example, find that women of darker complexions are the most misclassified group in facial analysis algorithms, whilst males of lighter complexions are most accurately classified. A large percentage of the literature on machine learning applications has thus dedicated itself to investigating why these biases occur and developing methods to mitigate them, both along the machine learning pipeline and in the quality of the training data given to the system (Miron et al., 2020). Overall,

algorithmic systems are complex and vary heavily on the context of their design, calculations, and level of sophistication. Thus, not one source of bias can be attributed to a model or be generalised for all models; to investigate a model's bias, the study must be context specific (Gangadharan and Niklas, 2019). Therefore, the following sections will only discuss the literature relevant to the research question at hand. This section will cover the discussion of bias within training data in the literature, and Section 2.1.4 will discuss the literature on bias along the model learning pipeline and why it is so difficult to discern if there is bias in this area or not.

One of the main identified sources of machine learning bias is within the training data that is provided for the model's learning. Put simply, our models are producing biased results because we are providing them with biased data, or, as a well-known computer science adage eloquently summarises this issue: "garbage in, garbage out" (Barocas and Selbst, 2016). This section will discuss the types of biases found within the policing system, as this best fits the scope of the research.

As discussed in Section 2.1.1, algorithmic models learn their rules from vast amounts of data that are fed to them, and the stronger these models perceive a correlation or pattern to be, the more hardwired it will become as a rule in the model. However, this can potentially develop what is known as a 'feedback loop,' as demonstrated in the literature that a model will be significantly influenced by the feedback received from the previous iteration and thus repeat it until the model becomes a self-predicting loop (O'Neil, 2016; Ensign et al., 2018). This is because the outcome determines the feedback that is received in the model. Thus, if a white man is hired, the model learns from this positive feedback to select white men, or if crime is found in a particular area, the model

learns to send police there to find more crime. In the US, crime rates vary by region, and this is often historically linked to socioeconomics, neighbourhood segregation, and historical policing trends, which will be further discussed in Section 2.3. Lum and Isaac (2016) applied a popular predictive policing system, PredPol, to the city of Oakland, California, and found that it consistently sent police to Black neighbourhoods and diverged from the true crime rate in the area. This phenomenon has long been observed in the literature: wherever you send officers to find crime, crime will be found, and thus making that area statistically prone to crime when basing predictions on historical data (Marvell and Moody, 1996; Eterno, Verma, and Silverman, 2016). Another important point with regards to training data for AI/ML policing applications is that reported and discovered incidents are not representative of true crime rates, which are often skewed, distorted, or missing (Bayley, 1983; Frank, Brantingham, and Farrel, 2012). There have long been issues of reporting biases within police departments addressed in the literature, such as: crime underreporting by victims (Allen, 2007; Pezzella, Fetzer, and Keller, 2019; Comino, Mastrobuoni, and Nicolo, 2020), lack of uniformity in crime reporting between police departments (Levitt, 1998; Rosenfeld, 2007), lack of timeliness in reporting (Kleiman and Lukoff, 1981; Rodríguez-Ortega et al., 2020), or the manipulation of data by departments due to political pressure to reduce crime rates (Seidman and Couzens, 1974; Eterno, Verma, and Silverman, 2016). However, these remain pervasive issues in policing.

Another important source of bias the base rates of certain measurements that can be traced back to 'structural bias,' or the everyday attitudes, norms, and policies towards a social group reflected at all levels of society (McIntosh, 1988; Short and Wilton, 2016). There has been research in the literature on its widespread impact, from mental health and sexual health issues (Hall, 2016; Wilton, 2016) to media

negative reinforcement of stereotypes in the media (Williams, Short, and Ghiraj). Racist structural bias is particularly visible in the United States, given its history (See: Section 2.3). One empirical example of structural violence illustrated in policing is Voigt et al. (2017), who analysed police body camera footage using natural language processing algorithmic models and found that there was a consistent difference in respectful language and attitudes towards White and Black citizens. Another example is expounded by Goel, Rao, and Shroff (2016), who analysed three million New York City police department stops over a period of five years, under the department's controversial 'stop-and-frisk' policy, which allows officers to stop and conduct a body search on any individual without a warrant if they have 'reasonable suspicion' of weapons or contraband. The authors found, not only was this policy not effective, as only 1% of those stopped carried weapons, but that Black and Hispanic citizens were disproportionately stopped by police officers (Goel, Rao, and Shroff, 2016, p.356). In another study, Richardson, Schultz and Crawford (2019) outline the long history of policing's manipulation of statistics, erasure of data, and corruption of data, in what Kim et al. (2003) coin as 'dirty data.' The authors then evidence cases of different jurisdictional areas in the US that incorporated this dirty data into their predictive and policing systems. This was all while this "unlawful," "unconstitutional," and "racially biased" data was under federal investigation, thus tainting the predictive models with statistically incorrect and discriminatory data (Richardson, Schultz and Crawford, 2019, pp. 192-3).

Even without being so outwardly nefarious, the literature has long considered 'implicit bias,' a type of unconscious cognitive bias, as a strong psychological element to the way in which we make decisions every day, from doctors to jurors (Chapman, Kaatz, and Carnes, 2013; Cardi, Hans, and Parks, 2020). Indeed, in the wake of a disproportionate

amount of police brutality against POC in US media, efforts such as body cameras and implicit bias training have been implemented to combat these issues, however, they have proven to not be an effective solution (Onyeador, Hudson, and Lewis, 2021). Miron et al. (2020) demonstrate this is not only a problem in the US, where their study of sources of bias in recidivism risk assessment methods for juveniles in Catalonia found a disparity of 23% in the base rates of recidivism risk scores for ethnic-Maghrebi and ethnic-Spanish males without ethnicity being a datapoint in this algorithmic model (p.132). When these types of biases are so rampant in data that they are easily picked out by proxy factors, they risk the threat of further entrenching negative biases and attitudes towards a particular group in the algorithmic model (Chouldechova, 2017; Miron et al., 2020). In sum, authors on the AI/ML cautious side argue that the mix of perceptive algorithmic models applied to high-stakes decision-making based on biased data is a dangerous application and may lead to harmful societal effects and disparate impact.

With regards to developments in the literature to mitigate these biases found in training data, there have been numerous efforts from training data pre-processing to in-processing during training, and in post-processing when modifying outcomes (Miron et al., 2020). Although much progress is being made in finding methods to mitigate against certain biases in machine learning models (See: Zemel et al., 2013; Johndrow and Lum, 2019; Zafar et al., 2017; Agarwal et al., 2018; Kleinberg and Raghavan, 2018), another camp in the literature contends that no true fairness will be achieved even with mitigation techniques applied, and that even if it were the case, models are so context-specific that these mitigation techniques would need to be adjusted on a case-by-case basis (See: Corbett-Davies et al., 2017; Ensign et al., 2018; Kallus & Zhou, 2018; Liu et al., 2018; Lipton, Chouldechova and McAuley, 2019; Miron et al., 2020). Civil society efforts to mitigate bias,

such as the *statactivisme* movement in France aim to ‘fight’ with and against numbers to shed light on statistical errors and the vacuity of traditional methods and metrics that have deep effects on society (Bruno et al., 2014). Although issues of biased data and inheritance of biases in algorithmic models are worrying, the media has driven attention over the past few years to this topic, sparking a promising conversation around the topic with experts in the field (Metz, 2019; Smith, 2019). Overall, this section discussed potential sources of bias found in training data by the literature, such as feedback loops, reporting bias, ‘dirty data,’ structural bias, and the efforts to mitigate them. The following section will briefly address potential bias found along the AI/ML learning pipeline and the debate within the literature regarding this topic.

2.1.4 The AI/ML ‘Black Box’ Mystery

Although machine learning models base their correlations and conclusions on the training data we provide them with (which we are open to investigate), one area where we remain in the dark is in understanding how the model reached a certain assessment or conclusion (Chiao, 2019). We may be able to infer as to how data influenced the model, but we cannot know the model’s ‘thinking’ process, otherwise known as ‘explainability’ (Gilpin et al., 2018). In this sense, the model is a ‘black box,’ where, even though we can see the inputs and outputs of the algorithm, we cannot see the reasoning that connected the two and whether there was bias during the process (Reisman et al. 2018). Much like how neuroscientists are still attempting to understand human decision-making, so are computer scientists regarding artificial neural networks (Castelvecchi, 2016). However, authors such as Castelvecchi (2016) and Chiao (2019) extend this similarity even further and contend that, even though we do not fully understand human brain function, we still trust human judgment in all societal aspects, therefore, why should we not allow highly sophisticated machine learning models

to be met with the same acceptance? Although a very fair point, machine learning models are still a nascent technology and face several challenges to match human intelligence, such as lack of contextual awareness, adversarial examples, and feedback loops, for example (Brundage et al, 2018; Nguyen, Yosinki and Clune, 2015; O’Neil, 2016). Clearly, a large number of questions regarding sources of AI/ML bias within the black box and lack of explainability remain to be addressed by the literature.

A final point salient to this research is the fact that private companies are increasingly providing these algorithmic tools for public services, such as police departments, making it harder to determine the correlation between models’ inputs and outputs (Levendowski, 2018). This is because these algorithmic models are considered trade secrets, and therefore intellectual property, which cannot be disclosed even in trial, as exemplified by *People v. Chubbs* (2015). This lack of algorithmic transparency further entrenches problems within the “black box” issue to investigators and researchers who try to answer these difficult questions by removing access to source code and training data (Wexler, 2017). Legal protection of the models further obfuscate their mechanics, not only for those attempting to appeal a decision that affects them, but also to those scrutinising the technology (Levendowski, 2018; Wexler, 2017). The following section will discuss a particular application of AI/ML that will be the focus of this study: the COMPAS recidivism risk assessment.

2.2 Policing and ML/AI: The Case of COMPAS

Justice departments have long used statistical and empirical methods as credible foundations for criminal behaviour prediction, assessment, and decision-making (Porter, 1996; Berk and Bleich, 2013; Miron et al., 2020). Indeed, this academic field can be traced back to the 1960s and its applications are now widely used across the US, with over sixty

different types of algorithmic assessment tools employed in decision-making support (Schapiro, 2017; Barry-Jester, Casselman, and Goldstein, 2015). Since the introduction of machine learning models to enhance these criminology and policing methods, they have come under renewed scrutiny, given the bias and statistical limitations mentioned in Section 2.1.2. This section will dive into the subject of investigation within this research: the COMPAS recidivism risk assessment tool and why this tool has been critical to the study of AI/ML fairness.

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a software that uses an algorithmic model to quantify the probability of recidivism (future re-arrest or violation of probation or parole) in an arrested individual. This recidivism risk assessment tool was owned and developed by the private company Northpointe at the relevant time of study (now called Equivant) and has been employed in several police departments in New York, Wisconsin, Florida, and California. The tool, according to Northpointe, bases itself on “a range of theoretically relevant criminogenic factors and key factors emerging from meta-analytic studies of recidivism” (Brennan, Dieterich, and Ehret, 2009). Although many factors are considered in an individualised and context-specific manner for a sentencing decision, recidivism has increased in importance as a factor of consideration, as an individual having a high risk of re-arrest would imply a threat to public safety and a strain on public resources. Indeed, the 2007 Conference of Chief Justices and State Court Administrators declared in a resolution that, “the best research evidence has shown that use of validated ‘offender risk and need assessment tools’ is critical in reducing recidivism,” which “promotes public safety while making effective use of taxpayer dollars.” The US’s overcrowded and underfunded criminal justice system has led to a strong and, surprisingly, bipartisan support for data-driven criminal justice reform (Barry-Jester, Casselman, and

Goldstein, 2015). For example, in Broward County, Florida, the correctional facility capacity rate was at 91.5% at the time of Angwin et al.'s (2016) publication (Florida Department of Corrections, 2014). These tools appeal to both political sides as a cost-saving solution to reducing crime, and potentially, mitigating against cognitive biases from decision-makers. The authors quickly bring up the moral elephant in the room: should judges be making decisions based on future crimes that have not occurred yet and statistical generalisations (Barry-Jester, Casselman, and Goldstein, 2015)? Does this not contradict the US's legal bases of 'innocent until proven guilty' and 'rights to due process'? This debate, like the one on algorithmic fairness, falls out of the scope of this research, but is key for policy and lawmakers to keep in mind with the increasing implementation of these algorithmic tools in the criminal justice system.

COMPAS is meant to "provide decision support to correctional agencies" at many levels of the criminal justice system: from setting bail to final sentencing and from recommending rehabilitation programmes to probation/parole terms (Brennan, Dieterich, and Ehret, 2009). Like most risk assessment tools, it utilises a questionnaire based on certain factors that are believed to be linked with recidivism, such as having a prior criminal record, substance abuse struggles, or being in a gang. The questionnaire is run through the algorithmic model, which then produces a statistical probability of the defendant's likelihood of recidivism based on a set of statistical scales developed by the private company which are not available to the public. COMPAS is a 137-question interview with the defendant filled out by a correctional screener, and looks to calculate the risk of re-arrest, violent crime, or failure to appear for court for that individual (See: Appendix A; Northpointe, 2012). Those classified as 'low risk' are given less bail, sentencing time, or less strict probation/parole terms, and vice versa for those classified as 'high risk.'

A pivotal legal decision regarding the use of COMPAS in the criminal justice system is *State v. Loomis* (2017), where Eric Loomis pleaded guilty to fleeing police officers in a vehicle in 2013. His presentencing investigation report included a COMPAS estimate of Loomis' recidivism risk, which was mentioned in his sentencing determination (Harvard Law Review, 2017, p. 1531). Loomis was sentenced to six years in prison and appealed to the state supreme court, arguing that the fact that COMPAS's methodology is legally a trade secret and therefore not open as opposing evidence in this case violated his rights to due process⁷ (Miron et al., 2020). This decision was appealed to the Wisconsin Supreme Court, where Justice Bradley rejected the appeal, stating that the COMPAS tool and the presentencing investigation report are not the only factors considered by the judge in the sentencing decision, thus Loomis has a fair and impartial individualised trial (Harvard Law Review, 2017, p. 1532). However, both Justices Bradley and Abramson provided a word of "caution" and "concern" to sentencing courts utilising risk assessments, for judges to better understand the nature of algorithmic risk assessment tools, and for the courts to file for more evidence-based information on the strengths and weaknesses of the risk assessment tools (Harvard Law Review, 2017, p. 1533). Although formal caution was warned unanimously by the Wisconsin Supreme Court Justices, the ruling still upheld COMPAS's defence as a tool to be consulted in court.

In 2016, Angwin et al. published a landmark investigative report in ProPublica on the use of COMPAS in Broward County, Florida, claiming that it is biased against Black Americans. The authors obtained

⁷ In the US, the right to due process is the legal right to a fair trial with certain established procedures guaranteed to the defendant. It is protected by the Fifth and Fourteenth Amendments to the Constitution. The right most salient to Loomis' case would be the right to know opposing evidence against the defendant. For a deeper dive into procedural rights of due process, see: Friendly, 1975.

seven-thousand assigned risks scores of individuals arrested in Broward County in 2013-2014 through freedom of information requests. Following the COMPAS recidivism benchmark of two years, Angwin et al. (2016) monitored whether the ex-offenders had been arrested again within the two-year period. This benchmark time frame is likely based on a majority consensus in the literature that recidivism is prone to occur within three years of release (Alper, Durose and Markman, 2018). This report then found that the error rate balance between White and Black defendants in this risk assessment system was highly unfavourably skewed against Black defendants, meaning that twice the number of Black Americans were false positives, while, conversely, half the number of White Americans were false negatives (Angwin et al., 2016). In layman's terms, a Black American in the Broward County COMPAS system was twice as likely of being deemed 'high risk' and not being a recidivist within two years, while a White American who was a recidivist during that time was twice as likely of being classified as 'low risk.' Since this report was published, there has been massive attention in the literature to the arguments and implications of this report. As discussed in Section 2.1.2, there has been a relentless debate regarding the methodology used to justify bias and 'fairness,' with several camps divided on the subject⁸. As previously discussed, definitions of bias and fairness vary among the scholarship, so, even though COMPAS showed variation in error rate balance amongst groups, the model's base probability for recidivism amongst Black and White defendants was well-calibrated, thus making it 'fair' for proponents of that definition (Kleinberg, Mullainathan and Raghavan, 2016). The ProPublica report also paved the way for a

⁸ For those authors in favour of Angwin et al.'s methods and conclusions, see: Larson et al., 2016; Chouldechova, 2017; Dressel and Farid, 2018; Rudin, Wang, and Coker, 2020. For those authors who critique Angwin et al.'s methods and conclusions, see: Dieterich, Mendoza, and Brennan, 2016; Flores et al., 2016; Gong, 2016; Jackson and Mendoza, 2020.

discussion into the ethics of AI/ML applications playing significant roles in society, such as the criminal justice system and has been cited and studied in countless papers and conferences (See, for example: Grgić-Hlača, 2016; Corbett-Davies et al., 2017; Žliobaite, 2017; Miron et al., 2020; Barocas, Hardt, and Narayanan, 2021). Indeed, it was joked at the Association for Computing Machinery's first Conference on Fairness, Accountability, and Transparency that this case study has become synonymous with the debate on AI/ML bias and 'fairness' and no conference on the subject could begin without its mention (Narayanan, 2018).

Although much attention has been brought to COMPAS after the ProPublica investigation, most of the debate in the literature has been on its statistical and methodological validity, with some mention on its broad societal and judicial implications (See, for example: Barocas et al., 2017; Green and Hu, 2018; Levendowski, 2018; Benthall and Haynes, 2019; Hertweck, Heitz, and Loi, 2021). However, to the researcher's knowledge, there has been no in-depth investigation into the contents of the COMPAS questionnaire itself (See: Appendix A). The researcher finds this gap telling of the mathematical focus of the literature within this subject. Although methodological debate is important, it is equally relevant to the question of algorithmic fairness to precisely examine the data being inputted into the model, as seen in Section 2.1.3's discussion of bias in data. The researcher thus aims to fill this gap by examining the COMPAS questionnaire used in Broward County, Florida, to supplement the work done in the literature and to provide a point of view different from methodological and definitional debate. The research design and methodology of this investigation can be found in Chapter 4, and the findings and discussion in Chapter 5. The following section will provide further pertinent background to this study and the racial bias found in the ProPublica study by offering an overview of policing and Black America.

2.3 Policing and Black America

“There is not a country in world history in which racism has been more important, for so long a time, as the United States” (Zinn, 2015, p.23). This section aims to contextualise the history of Black Americans within the US criminal justice system, in order to provide a proper foundational understanding for Chapter 3’s conceptual framework and Chapter 5’s discussion. The author understands that this is a lofty ambition and does not claim by any means to provide a comprehensive history of the Black American experience and identity in the USA and its criminal justice system⁹. However, given that the history of policing in the USA is inextricably linked with race, it is fundamentally necessary to provide at least a broad contextualisation of this phenomenon.

As above-mentioned, the USA has a globally unparalleled proportion of incarcerated individuals. This not only takes a toll on the communities affected, but on public funds and resources, thus sparking an interest in optimising big-data and automation for criminal justice reform and the employment of algorithmic tools such as COMPAS. The history of the USA’s criminal justice system is deeply racialised and these attitudes go back as far as the first arrival of ships filled with enslaved people. Indeed, “although today’s rate of incarceration is both historically unprecedented and internationally unparalleled, its racially discriminatory character is not” (Thompson, 2019, p. 221). The following section will thus outline key political and legal developments in this such history to provide context for potential biases in data found in a historically biased nation. Due to scope limitations of this research, the following sections will only provide an overview of the racial history of the

⁹ For a comprehensive history and literature review see: Meltzer, 1964-1967; Aptheker, 1951-1994; Gray-Ray et al., 1995; Zinn, 2015; Hinton and Cook; 2021. For a more extensive discussion on race identity and politics in the USA see, for example: Schaefer, 2004; Neblo, 2009; Georges-Abeyie, 2010; Collins, 2010; Fiske and Hancock, 2016.

US criminal justice system *vis á vis* Black Americans and will not include the criminalisation and incarceration of mainland Natives, Hawaiians, Mexicans, or Puerto Ricans by White Americans (Thompson, 2019).

2.3.1 Cementing Racial Hierarchies

The first permanent colonial settlement in what is now called the USA was in Jamestown, Virginia, in 1607 and the first ship trading Black slaves to this settlement would soon arrive in 1619 (Zinn, 2015). After a period of brutal starvation during the winter of 1609-1610, the survivors were left desperate for agricultural labour, and “naturally” considered bringing in enslaved labourers from the African continent, as this had already been practiced in Europe, the Caribbean, and South America for over half a century (Zinn, 2015, p.26). By the early 1700s, the slave trade had seeped into everyday colonial American life, employing Whites from all social classes directly and indirectly, as white-collar administrators for slave enterprises, lumber workers and seamen for the manufacturing and running of slave ships, slave catchers and plantation overseers, and a booming aristocracy from the forced manual labour on cotton and tobacco plantations, for example (Feagin, 2013, p.27). Slave labour created such an economic surplus that it was able to support Colonial America’s fight for independence against the British Empire (1775-1783). The newly founded state, established in its Constitution - one of the most important and influential legal documents in the US – that the definition of an enslaved person was equal to three-fifths of a person, that the slave trade could not be abolished before the year 1808, that the return of fugitives was protected to enslavers, and that Congress had the power to suppress revolts and insurrections (National Archives, 1787). These constitutional provisions would remain in place until 1865. Thus, the transatlantic slave trade laid the foundation of the USA’s economic and political development, as well as entrenched deep racial hierarchies at many levels of socioeconomic life.

The newly independent United States of America was a nation that thrived off the fruits of slave labour. As part of the justification and reconciliation with the exploitation of humans for profit, entire academic fields of pseudosciences were developed, such as anthropometry/Bertillonage, physiognomy, physical anthropology, craniometry, and phrenology, to argue that Black people were intellectually inferior to Whites and prone to violence and criminality (Browne 2015; Taslitz, 2010). Backed by these pseudoscientific fields, Black bodies were literally and metaphorically 'branded' as "bestial," "primitive," "alien," "dangerous," "animalesque," "overly sexual" "rebellious," and predisposed to criminality and savagery (Carter, 2010, p.267) to justify corporeal, mental, and sexual abuse from the part of the enslaver (Feagin 2013, Browne 2015). To further legitimise this racial hierarchy in society, laws and bureaucratic processes were put in place to ensure a distinction between different groups, with the ruling White Protestants on top. Feagin (2013) argues that one of the reasons why American society is so deeply racialised is because its foundational public and legal systems were designed during the era of the slave trade with racial subjugation overtly at the forefront of US politics and economy before democracy, thus making these systems fundamentally racist.

Racial hierarchies in legal and policing systems predated formal institutions in antebellum America. Prior to the abolition of slavery, the criminal justice system as we know it today was non-existent, with no prisons and no formal criminal courts (Thompson, 2019). If an enslaved individual ran away or did not comply with the enslavers' demands, punishments would be enacted at the will of the enslaver. As enslaved people were legally considered property, they had no rights to individual freedoms or protections whatsoever, even though they represented 20% of the US population by 1787 (Feagin, 2013, p. 30). Slave patrols, town constables, and state militias were the first examples of formalised

policing in the US and were present in certain Southern areas as early as 1704 with the objective to capture suspected fugitives and people of colour (Dulaney, 1996; Hadden, 2003; Dunbar-Ortiz, 2014). Slave Codes were laws across the states that related to enslaved people, some included prohibiting enslaved people to read and write, sell and buy commerce, marry people from other plantations, or leave the plantation (Hadden, 2003; Thompson, 2019). The more profitable slave labour became, the stricter these codes evolved and the harder these policing efforts intensified to keep this racial division cemented in society (Berlin and Morgan, 1991). The Slave Codes thus established a legal framework for the treatment of Black enslaved and freed people and solidified the racial lines along which American society would interact.

2.3.2 The Thirteenth Amendment's Powerful Loophole

On the eve of the American Civil War (1861), slavery had been abolished in the English, French, and Spanish empires for more than half a century, yet the US was at existential grips over the issue (Zinn, 2015). Indeed, eleven Southern states had seceded from the union to form the Confederate States of America, which would fight to continue to enjoy the benefits of slave labour, against newly elected President Abraham Lincoln's abolitionist campaign that represented the North's political and economic transition towards emancipation (Feagin, 2013). The war was won by the North in 1865 and, ten generations after the first Africans were forced onto ships headed towards Jamestown, the institution of slavery was officially abolished by the Thirteenth Amendment (1865) to the Constitution of the United States, which states:

“1. Neither slavery nor involuntary servitude, except as a punishment for crime whereof the party shall have been duly convicted, shall exist within the United States, or any place subject to their jurisdiction.

2. Congress shall have power to enforce this article by appropriate legislation.”

(National Archives, 1865)

Even though the Thirteenth Amendment provided former enslaved people with freedom, nullifying the Slave Codes under which they had been beholden to, Section 1 of the Amendment provided enslavers with a loophole for involuntary servitude if that person had been “duly convicted” of a crime (Taslitz, 2010, p.248). This was further enforced by the Fourteenth Amendment to the Constitution, ratified in 1868, which allowed Black individuals to be considered citizens and hold the right to vote, unless they had been convicted of a crime or of rebellion (National Archives, 1868). This is in fact still in place to this day; in most states, convicted felons are barred from having the right to vote in many states. Due to such laws, 2.27% of the whole US population - around seven million people - is disenfranchised. Within the Black American population this stretches to 6.26% - nearly three million people (The Sentencing Project, 2020b). Indeed, historic texts note that this was strategically put in place in Southern states, where former enslavers now found themselves outnumbered by free Black citizens and saw the threat of being outvoted by these newly enfranchised communities (Alexander, 2019).

After losing the Civil War, Southern states woke up to amassed war debt and no profits coming in from forced free labour. As virtually all of the southern White economy depended on Black slave labour, this, in turn, led to the implementation of policing practices as tactics of suppression of freedoms granted to formerly enslaved people to apply this constitutional loophole, and thus the newly emancipated generation experienced a surge in arrest and incarceration (Lichtenstein, 1996). The Slave Codes were transformed into the ‘Black Codes’ seemingly

overnight, and the “resurrection of the chain gang and the rise of the convict lease system [were] *de facto* ways to re-create aspects of slavery” after the Civil War (Taslitz, 2010, p.248). As W.E.B. Du Bois, the most influential Black American scholar and activist during the first half of the 20th Century stated, “the slave went free; stood a brief moment in the sun; then moved back again toward slavery” (1935, p. 30).

The Black Codes were a series of norms that were codified into laws by the Southern States immediately after the war and thus the end of formal slavery, though essentially a reincarnation of the Slave Codes. Like their legal predecessors, certain Black Codes included the prohibition for Black individuals to own land, bear arms, practice freedom of speech, self-defence, or gather for worship, for example (Muhammad, 2010). Through the historical and institutional categorisation of Black individuals as “dangerous” and “aggressive,” there was also the justification for “constant suspicion and monitoring,” by newfound policing bodies, so Black individuals would be arrested for loitering, for example (Carter, 2010, p.267). This ‘vagrancy law,’ which also prohibited unemployment or homelessness for Black citizens was known as a ‘Pig Law,’ a trivial offense that when committed by a Black individual was treated as a serious felony - one of the defining features of the new Black Codes (Muhammad, 2010). Lastly, the convict leasing system, which would last from 1844 until 1941, was the direct manifestation of the return to antebellum slavery through legal and constitutional loopholes: it allowed prisons to lease out prisoners for penal labour under their ‘duly convicted’ crime, and they would be fully subject to the control of the company that leased them. It was “one of the harshest and most exploitative labour systems known in American history,” and was essentially a form of legal slave labour for Southern enterprise and plantation owners (Macini, 1996, p.2). Notably, by 1898, 73% of Alabama’s state revenue came from this convict lease system

(Perkinson, 2010, p.105). All these efforts to suppress Black freedoms and rights through laws and severe policing were manifestations of the effect of the slave labour economy and racial hierarchy ingrained in the US's institutions and society. Indeed, Curtin (2000) illustrates the use of this loophole in the state of Alabama, where the Black prison population in the year 1850 was 1%, and twenty years after the abolition of slavery it had exploded to 85%. Even in the North, where states have claimed to be more abolitionist-friendly, historians demonstrate an overnight shift in policing strategies from monitoring White immigrants to Black freed-people after 1865 (Gross, 2006; Muhammad, 2010). This institutionalisation and legalisation of discrimination against Black individuals provided by the Black Codes paved the way for the Jim Crow Era.

2.3.3 Lasting Impacts of Legalised Segregation and the 'Justice Apartheid'

The Jim Crow Era was that of institutionalised racial apartheid mostly in Southern US states between White and Black Americans, already being implemented in the 1870s in some areas, but officially upheld by law in 1896 with the Plessy v. Ferguson Supreme Court Case of 'separate but equal,' and ending with the 1954 Civil Rights Act and the 1965 Voting Rights Act. Plessy v. Ferguson was a landmark case in that it was deemed legally and constitutionally acceptable to treat White and Black individuals differently based on their race. As such, this period continued to be demarcated by racial difference, and decades after the abolishment of slavery, "the notion of white supremacy proved far more durable than the institution that gave birth to it" (Alexandra, 2019, p.29).

Although this legal mandate supposed that access to services would be 'separate but equal,' this was far from the truth, as facilities, services, and opportunities available to Black communities would be

heavily underfunded or non-existent. This era not only turned a blind eye to racially motivated violence and discrimination, but also cemented and institutionalised disadvantages faced by Black Americans. One key example is the policy of 'redlining,' which was the systematic denial of services to a geographic area - usually a Black area - and has been an active policy in many cities as far back as the 1860s. Services such as health care (Nardone et al., 2020), supermarkets (Eisenhauer, 2001), retail (D'Rozario and Williams, 2005), insurance (Squires, 2016), affordable housing (Hillier, 2003), or banking services such as loans and mortgages (Hernandez, 2009; Cohen-Cole, 2011; Aaronson, Hartley, and Mazumder, 2020) have historically been denied in majority Black neighbourhoods across the USA from the moment they were segregated. To this day, cities across the US remain segregated by race, particularly in the South, where the narrative of racial division was historically stronger (Acharya, Blackwell, and Sen, 2018; Benton, 2018). Benton (2018, p.1122) notably highlights the strategic ways in which the city of St. Louis, Missouri, a city that at the date of publication was 65% segregated by racial demographic, zoned and re-zoned residential and industrial areas to keep Black and White residents apart, even after the Civil Rights Act. The author also discusses the effect that city segregation policies have had towards well-documented negative racial attitudes and animosity in the city (Benton, 2018). As demonstrated by the literature, the concentration of Black communities in poorly serviced, isolated, and low-income urban areas due to racist urban segregation tactics has impactfully disadvantaged economic, social, and health outcomes for Black residents, such as higher unemployment (Zehou and Bocard, 2000).

The population of imprisoned Black Americans continued to explode throughout the 20th Century: "From 1926 to 1986 the recorded black percentage among admissions to State and Federal prisons more

than doubled from 21% in 1926 to 44% in 1986. Importantly, this growth is not explained by general population trends. The number of blacks relative to the general population was about the same in both years: 10% in 1926 and 12% in 1986” (Langan, 1991, p. 6). This leads to a final key historical component, which is the ‘War on Drugs’ and the era of mass incarceration that led to the ballooning of the American prison population by 500% over the past 40 years (The Sentencing Project, 2020a). These correctional policies began in the 1970s, and still have profound effects on marginalised communities today (Huebner, DeJong, and Cobbina, 2010). Indeed, such has been the impact of the War on Drugs on the Black American community, that Small (2001) characterised it as a “system of apartheid justice” (p.897) and Alexander (2019) dubbed it ‘The New Jim Crow.’ When US President Richard Nixon gave his famous ‘War on Drugs’ announcement in 1971, he initiated a global and brutal crackdown on drug production, distribution, and consumption (Small, 2001). Not only did arrests increase, but so did laws that make sentencing longer and stricter, with a five-fold increase of life sentences since 1984; today, one in seven inmates is serving a life sentence (The Sentencing Project, 2020a). In Broward County – where the COMPAS application is examined - the length of new sentences increased from an average of 5.6 years in 2018 to 6.42 years in 2019 (Office of Economic & Demographic Research, 2020). In addition to this are ‘zero tolerance’ laws that ban offenders with drug-related felonies from accessing public services, such as housing and college financial aid, making it all the more difficult to integrate back into non-criminal life (Moore and Elkavich, 2008). Furthermore, arrests as a result of this policy have been extremely detrimental to Black communities; by 1996, 62.6% of drug offenders were Black even though they only composed 13% of drug users in 2000 (Moore and Elkavich, 2008; Small, 2001). Western and Wildeman (2009) have highlighted the observed generational effects of

mass incarceration on Black communities in the scholarship. In 1999, for example, 30% of all Black men were incarcerated, and the scholarship statistically noted as they left families behind, with low incomes and high unemployment rates in Black urban areas, less schooling resources and less available services due to redlining, this increased the effects of children being trapped in the same cycle of criminality.

With regards to Florida today, it is the US state with the 10th highest imprisonment rate (444 incarcerations per 100,000 residents at any point in time) and the 4th highest felony disenfranchisement rate in the country (7,690 felons without voting rights for life per 100,000 residents) (The Sentencing Project, 2020b). In terms of racial disparities in incarceration, Florida has a lower disparity than the national average, with 3.6 Black imprisonments for every White imprisonment (The Sentencing Project, 2020b). However, Black felony disenfranchisement covers 15.42% of the Black Floridian population, while the overall disenfranchisement rate in Florida is 7.69% (The Sentencing Project, 2020b). Looking at Broward, County, Florida, where this example of COMPAS was applied, it is the second most populous county in Florida home to 9.1% of the state's population (Office of Economic & Demographic Research, 2020). It is the ninth most crime prone county in the state of Florida, with 3,000 crimes per 100,000 people and is the county with the sixth highest commitments to prison (Office of Economic & Demographic Research, 2020). The top crimes in Broward County are burglary and drug-related crimes (Office of Economic & Demographic Research, 2020).

Finally, as a bridge to the above literature review on AI/ML, in 2014, then-US Attorney General Eric Holder stated as a warning to legal practitioners on the use of algorithmic models at the annual meeting for the National Association of Criminal Defense Lawyers: "By basing

sentencing decisions on static factors and immutable characteristics – like the defendant’s education level, socioeconomic background, or neighbourhood – [the algorithmic models] may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society” (Department of Justice, 2014). Indeed, Miron et al. (2020) found in their study that static data features (education, neighbourhood) had a higher correlation than dynamic features (substance abuse, hostile behaviour) to protected features (sex, nationality, religion, race), thus demonstrating higher disparities within protected features and showing potential proxies for bias given this higher correlation.

Overall, this section has drawn attention to the vast literature surrounding the generational social, political, and economic damage, as well as disenfranchisement, in Black communities as a result of these institutionally racist policies throughout history (Washington, 2018). Plainly put, it is “impossible to overstate the significance of race in defining the basic structure of American society” and the American criminal justice system (Alexander, 2019, p.29).

3. Conceptual Framework

The following section will provide a background to the conceptual framework that will be applied to this research. It will first present an introduction to critical race theory and its key theoretical tenets. It will then discuss why this theoretical framework is relevant and appropriate for this study. Lastly, this section will briefly outline the limited literature found in the budding academic field that combines critical race theory and algorithmic fairness.

3.1 Critical Race Theory

In order to consider potential structural and proxy biases in the data, as discussed in Section 2.1.3, this research will situate itself in the relevant theoretical paradigm of critical race theory¹⁰. Critical race theory (CRT) was first developed by US scholars of colour in the 1970s as a movement addressing racial identity, racism, and power relations within, as the theoretical framework argues, institutional structures (See: Crenshaw et al., 1995; Ladson-Billings, 2013, Delgado and Stefancic, 2017). It was first developed as a legal response and critique of civil rights law, then education, and now has been increasingly adopted by higher education academics as an outright anti-racist challenge to existing narratives and constructions (Cabrera, 2018). The fundamental tenets of critical race theory are outlined below (See: Delgado and Stefancic, 2017; Cabrera, 2018; Ogbonnaya-Ogburu et al., 2020):

- (1) *Race is a socially constructed concept.* As discussed in Section 2.3.1, debunked pseudosciences such as eugenics and phrenology demonstrate that racial categories derived from physical traits do not represent absolute genetic, biological, or behavioural truths (Roth, 2017). CRT is a 'constructivist' paradigm, meaning that it contends that societies create, or 'construct' different notions to live by, which, over time, societies will hold to be true and intrinsic (such as social classes and caste systems, or gender roles and stereotypes, for example) (Hacking, 1999). Therefore, it argues that not only has history shown that racial categories are fluid and socially constructed, but also that no group possesses inherent

¹⁰ This section only aims to provide an introductory and context-relevant background of the theoretical framework applied to this research. For a deeper dive into critical race theory, its history, and the different academic discussions within the theoretical paradigm, see: Crenshaw et al., 1995; Delgado and Stefancic, 1998; Delgado and Stefancic, 2017; Burrell-Craft, 2020.

characteristics based on its assigned racial category (Bowker and Star, 2000; Roth, 2016).

(2) *Racism is commonplace for People of Colour (POC)*. Racism is not a stand-alone aberrant incident; it is a pervasive and systematic life-long experience for marginalised groups. From outright insults and attacks to more subtle race-driven jokes, comments, assumptions, stereotypes and negative connotations, marginalised groups must navigate through a daily bombardment of aggressions and 'microaggressions' (Yosso et al., 2009; Sue, 2010).

(3) *Identity is intersectional*. Intersectionality is a concept in CRT and feminist theory that, in sum, permits the contextualisation of individual lived experiences and removes 'essentialism,' or the idea that because you identify as a certain race, gender or socio-economic class, you will present inherent traits, following Tenet 1 (Telles and Lim, 1998; Gillborn, 2015). By acknowledging that everyone has different facets of themselves that they identify with, we can pick up on mutually reinforcing forms of oppression (such as being both Black and female or being both homosexual and disabled). For example, combining these first three CRT tenets, someone who identifies as a man, Hispanic American, homosexual and middle-class will have a different lived experiences and instances of oppression and discrimination than someone who identifies as a woman, White American, heterosexual, and lower-class.

(4) *Voices of Colour are Unique and Must be Heard*. Following the above-established tenets, minority and marginalised groups have a markedly different lived experience from White Americans and

critical race theorists establish that there is a uniqueness and truth to their stories and perspectives in discussions of racism and race (Johnson, 1991). Therefore, listening, amplifying, and showcasing POC voices are all fundamental anti-racist and anti-essentialist requirements.

(5) *Advances for POC are subject to White 'interest convergence.'* A final key tenet¹¹ posits that racism benefits some groups in society and that any anti-racist change will not be made unless it is in the interest of those beneficiary groups. Indeed, the nascence of CRT stemmed from critiques of milestones in the US Civil Rights Movement, which theorists argued were not advanced until it became of financial and political interest to White Americans in power. In the case of *Brown v. Board of Education* (1954), critical race theorists point to the poor economic development in the racially segregated South hindering American Cold War objectives and the need to quell growing domestic unrest from maltreated Black Second World War veterans as reasons behind this cornerstone Supreme Court decision, not as a moment of political moral breakthrough (Bell, 1980).

Contrary to its name, critical race theory is not a 'theory' *per se*, but rather a conceptual framework within which one can examine how race, racial bias, and racism are expressed and presented at all levels of society (Delgado and Stefancic, 2017). CRT has also become quite controversial in mainstream American society in recent years and is an issue that delineates strong partisan division (Cabrera, 2018, p.210).

¹¹ There are other proposed tenets debated in the literature, such as liberalism's detrimental colour-blindness, racism as a permanent feature of society, and Whiteness as a function of property rights. For a further discussion regarding these, see: Bell, 1992; Harris, 1993; Stoll and Klein, 2018; Harris, 2020; Clark, 2021.

Relevant to this research's geographic scope, critical race theory has been banned from being taught in Florida public schools starting June 2021, with Republican Governor Ron DeSantis stating that CRT is "state-sanctioned racism and has no place in Florida schools" (Lugo, 2021). One key factor to this reactivity is that, as the above-listed key tenets of CRT summarise, this framework maintains that White Americans participate in racism, as it is an everyday occurrence entrenched in hierarchical levels of society. When a nation built on the foundations of racialised slave labour and segregation decides hundreds of years later that its citizens are equal¹², it cannot simply erase the legacy of these complex systems of identity and power in society that revolve around race from one day to the next. As discussed in Section 2.1.3, researchers have demonstrated algorithmic models inheriting societal bias in language, gender, and race. Therefore, a society with such complicated race and police relations as the USA may likely reflect these difficult historical issues in policing data, statistics, methods, and applied algorithmic models. Ergo, the focus of this research will be to critically investigate the COMPAS recidivism risk assessment questionnaire within a CRT conceptual framework and within the contextualisation of the US's history of race and policing in order to add to the debate in the literature of racial bias in algorithmic models.

This theoretical paradigm is also not without controversy, contention, and debate in the literature (See: for example, Kennedy, 1989; Treviño, Harris and Wallace, 2008; Driver, 2011; Cabrera, 2018). One main criticism of CRT relevant to this research is that of 'black exceptionalism,' or the black-white binary that the paradigm paints of society, racism, and identity (Delgado and Stefancic, 2017). Over time,

¹² This highly generalised point only serves for the Black American experience. This does not include other race-related and xenophobic atrocities; see, for example: Disha, Cavendish and King, 2011; Cameron and Phan, 2018; Novak et al., 2018; Nagata, Kim, and Wu, 2019.

this conceptual framework has evolved to include other minorities or splintered off into the development of other minority-centred critical theories, including Latino, Asian, disabled, and LGBT+ experiences, for example (See: Harper et al., 1997; Chang and Gotanda, 2007; Annamma, Ferri and Connor, 2018). This research will note as a limitation that it is focusing solely on Black and White Americans affected by COMPAS, despite one-third of Broward County's population identifying as Latino or Hispanic (United States Census Bureau, 2019). This limitation is due to the ProPublica investigation's own limited scope, which focused only on comparing algorithmic bias in COMPAS risk assessments between White and Black Americans in Broward County.

Critical race theory is challenging claims of racial neutrality and objectivity in various academic fields, from law to education, and from health sciences to computer science (see, for example: Lynn and Dixon, 2013; Bracey, 2015; Bridges, Keel and Obasogie, 2017). However, the combination of CRT and AI/ML is lacking in the literature, with notable exceptions (see: Benthall and Haynes, 2018; Hanna et al., 2020; Ogbonnaya-Ogburu et al., 2020). These few key publications are paving the way for the consideration of the framework and its anti-racist desiderata in the conceptualisation of computer science problems, the assessment of training data, and the explanations for bias or unwanted results. Benthall and Haynes (2018) highlight the lack of attention in the literature that has been given to developing and defining 'protected' classifications of race in computer science and the political implications of these classifiers. Hanna et al. (2020) and Ogbonnaya-Ogburu et al. (2020) specifically focus on applying critical race theory to AI/ML and discuss why a potential critical race methodology for algorithmic models is necessary and what it would look like. Although there has been a surge of publications in the past few years regarding race and AI/ML, they have mostly highlighted racial disparities in artificial intelligence and machine

learning outputs and have not been race-conscious at the larger institutional level (Miron et al., 2020). Ogbonnaya-Ogburu et al. (2020) found that less than one percent of the Association for Computing Machinery's prestigious Conference on Human Factors in Computing System's (CHI) publications addressed racial discrimination. As Hanna et al. (2020) summarise the issue, when computer scientists are used to "treating race as an attribute, rather than a structural, institutional, and relational phenomenon," this can lead to the literature "to minimize the structural aspects of algorithmic unfairness" (Hannah et al., 2020). Thus, this research will contribute to the literature in this inchoate field by investigating the expressions and narratives of race in the COMPAS questionnaire and potential risks to algorithmic bias.

As a final point regarding the application of this conceptual framework to academic research, CRT mostly utilises qualitative approaches to investigate narratives of oppression. There have been, however, quantitative developments in the literature, such as QuantCrit, to investigate large sources of data that can pinpoint structural injustices caused to marginalised groups (Stage, 2007). This study will employ qualitative methods in the assessment of the COMPAS recidivism risk assessment in Broward County, Florida. The following chapter will discuss the selected methodology, the sample that will be analysed, and the tools of analysis under the chosen methodology.

4. Research Design and Methodology

This study proposes to apply critical race theory in the case of COMPAS in Broward County, Florida to investigate potential racial narratives and inheritances carried over into the algorithmic bias that Angwin et al. (2016) found to assess Black individuals as a higher risk of recidivism than Whites. This chapter will address the empirical framework utilised

to explore this question. It will first discuss qualitative critical content analysis, the methodology of choice for this research. Then, it will outline the data collection for this research and finally, will outline the data analysis performed under the selected methodology. Following Chapter 3's note on conceptual limitations, the following chapter will also reflect on this research's methodological limitations throughout.

4.1 Methodology

In order to investigate potential racial bias within algorithmic model data, the researcher has opted for a methodology that is exploratory in nature and flexible enough to allow for the potential discovery of nuance, interpretation, and meaning in the COMPAS questionnaire that may help explain observed bias (Angwin et al., 2016). Thus, the researcher has opted to perform content analysis, defined as the "systematic description of data through coding," for a "latent and more context-dependent meaning" of the data that is inputted into the COMPAS algorithmic model (Schreier, 2014, p.173). This methodology will thus enable the researcher to analyse the COMPAS questionnaire and the nuance within its questions in depth. Content analysis broadly consists of "unitising" the data within a relevant coding frame and then subsequently analysing this segmented data within an established conceptual framework – this will be further outlined in Section 4.1 (White and Marsh, 2006, p.28).

Furthermore, this research aims to provide a closer reading of the document at hand within a recontextualised lens of critical race theory, thus employing critical content analysis. Epistemologically, qualitative content analysis is a well-suited choice for a constructivist critical race theory interpretation of content (Graneheim, Lindgren and Lundman, 2017). As Beach, Rogers, and Short elucidate, "what makes the [critical content analysis] study critical is not the methodology but the framework

used to think within, through, and beyond the text” (2009, p. 2). The content analysis in this research is critical because it employs the epistemological foundations of critical race theory to approach the subject of potential sources of training data bias in algorithmic models within the COMPAS questionnaire. Thus, this research’s focus on linking, in part, sources of algorithmic bias with American socio-economic issues rooted at the institutional level, organically led to the adoption of critical content analysis as the methodology of choice.

One key methodological challenge of qualitative content analysis that Graneheim, Lindgren, and Lundman (2017) discuss is the increased abstract and interpretative approach to this methodology since its inception, which proves challenging to the researcher’s analytic credibility. Furthermore, given that critical race theory is a constructivist conceptual framework, the author would like to clarify, in agreement with White and Marsh (2006, p.37), that this type of qualitative work does not seek to “describe reality objectively,” but to paint a context-specific picture of the phenomenon at hand as a valuable addition to the academic discussion within the chosen theoretical paradigm (Elder-Vass, 2012). Therefore, this research will employ a deductive or ‘concept-driven’ approach to the methodology, applying the existing critical race theory paradigm to the phenomenon under investigation (Schreier, 2012). Additionally, in order to further heighten ‘credibility’ in this highly constructivist domain and provide ‘truth value’ to this study, the researcher will triangulate findings with other sources and perspectives from the appropriate literature in Chapter 5 (Guba and Lincoln, 1981, p.246). The following sections will discuss the research’s data collection and data analysis process for this qualitative critical content analysis.

4.2 Data Collection

This qualitative critical content analysis will investigate the COMPAS recidivism risk assessment questionnaire used between 2013-2014 in Broward County, Florida (See: Appendix A). This questionnaire is comprised of 137 questions and is asked to every arrestee by a screener. Self-reporting interviews are a popular source of information in correctional methods to supplement information that may be lacking in official reports regarding the defendant's background (Maxfield, Weiler, and Widom, 2000). The questions, argued by Northpointe to be based on empirical criminality literature, range from criminal history to family history and leisure (See Chapter 5). This questionnaire is then fed into the COMPAS algorithmic model, which will then provide the Broward County authorities with a risk assessment the likelihood of an individual being re-arrested. This questionnaire was made available by ProPublica and is the only publicly available document of its kind at the date of this research's writing. The tools and code used in private companies for recidivism risk assessment are deemed to be intellectual property and thus not available to the public. Therefore, this research will focus on this sole document, as it is the only available of its kind at the time of writing.

Regarding potential sample size limitations, White and Marsh (2006) find that, unlike quantitative content analysis, the sample size of a qualitative content analysis will be limited by nature due to the qualitative approach's unique investigation of the data, with more nuance and search for multiple interpretations within the text. The sample in a qualitative content analysis must be purposeful in the research's intention of "characterizing a phenomenon," and thus is not required to meet a sample size threshold if the research question is adequately addressed with the data at hand (White and Marsh, 2006, p.35). Therefore, the author deems that the limited size of this sample, the sole

available recidivism risk assessment questionnaire, does not pose detriment to the research as it fits precisely within the scope of the research question as the main document inputted in the COMPAS model. Furthermore, if more algorithmic models' questionnaires were to be added, this would fundamentally alter the scope of this study. The following section will outline how this document was analysed under the selected qualitative critical content analysis methodology.

4.3 Data Analysis

The design for data analysis used in this study seeks to provide the researcher with an in-depth view of the phenomenon at hand that "carefully incorporates the context, including the population, the situation(s), and the theoretical construct" (White and Marsh, 2006, p.38). Based on Schreier's (2014) model for qualitative content analysis, all the data must be exhaustively segmented into a 'coding frame' built of two or three unidimensional 'main' categories, which seek information for the researcher regarding the material at hand. Within those main categories lie mutually exclusive subcategories that will be more specific to the context of the material in the main categories (Schreier, 2014). The goal of qualitative content analysis is to have every piece of data coded so that the researcher can then assess the data in a purposefully constructed and thematic manner.

The researcher will thus categorise the questions found in the COMPAS questionnaire into one of two main classifications: (1) fact-based questions and (2) opinion-based questions. These main coding frames have been chosen to investigate the number of questions with hermeneutic implications that may affect, on one side, the respondent and/or the screener and, on the other hand, the COMPAS model's interpretation of the data. Given that algorithmic models' main selling

point in the US criminal justice system is to be a neutral, fact-based, and empirical solution to remove human heuristics and biases in decision-making (See: Section 2.1.2), the researcher found that the level of 'objectivity' of the questionnaire would be key to assessing the validity of this point. By coding distinctions between questions that are opinion-based versus fact-based, the researcher found this coding frame opportune to investigate the extent of which the COMPAS questionnaire may be affected by subjective questions, notions, and constructions. The survey literature has long demonstrated that questionnaire wording and format affects responses, among other factors such as questionnaire context, complexity, and length, for example (Kalton and Schuman, 1982; Schaeffer and Dykema, 2011). For a quantitative recidivism risk assessment, it is therefore arguable that the questions for data input must be as precise as possible as this risk assessment will have highly impactful consequences on the respondents of the questionnaire. Therefore, the researcher has defined 'fact-based questions' as those which could be clearly quantified or defined in both the question and answer options. On the other hand, 'opinion-based questions' are deemed to be those with imprecise wording and answers, whose interpretation both from the screener and respondent's view could vary.

The research will then further segment the objective and subjective questions within a set of mutually exclusive subsections that have already been delineated by the COMPAS questionnaire to ensure consistency (See: Appendix A): (a) current charges, (b) criminal history, (c) non-compliance, (d) family criminality, (e) peers, (f) substance abuse, (g) residence/stability, (h) social environment, (i) education, (j) vocation, (k) leisure/recreation, (l) social isolation, (m) criminal personality, (n) anger, and (o) criminal attitudes. These sections are divided by risk

factors of recidivism and criminality¹³ found in the literature, according to Northpointe (2012).

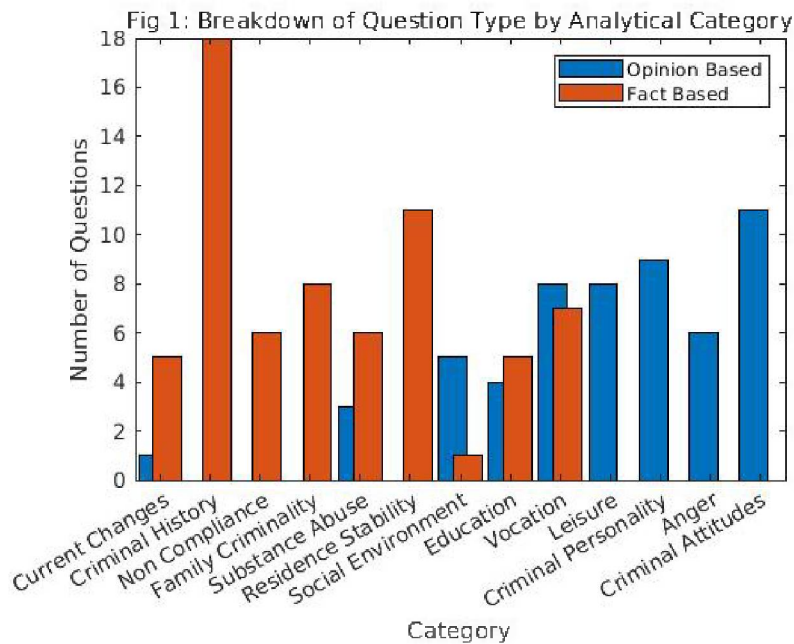
According to Rustemeyer (1992), the coding frame for a qualitative content analysis should be built either on formal or thematic criteria. Given that the COMPAS questionnaire already provides thematic segmentation that will be utilised in the subsections of this analysis, the researcher has opted for a formal unit of coding for the main sections, which are defined as “the inherent structure of the material,” and are units of separation for data analysis that allow the researcher to easily segment the content given its definitional clarity, thus whether they are fact-based or opinion-based (Schreier, 2014, p.178). This coding system allows the researcher to separate between strict formal coding for the main segments and thematic coding for the subsegments without applying critical analysis from the beginning, thus aiming to reduce potential issues of abstractionism, as highlighted by Graneheim, Lindgren, and Lundman (2017). Once the data has been fully coded, the researcher will then apply “analytical constructs” (Krippendorff, 2004, p.173) or “rules of inference” (White and Marsh, 2006, p.27) to investigate meaning and interpretation of these coded questions within the selected theoretical framework: critical race theory. By analysing the COMPAS questionnaire data in this way, the researcher aims to provide a meaningful interpretation of the questions posed to arrestees in a twofold manner. First, by allowing this content analysis methodology to provide a clear view of the subjective or objective nature of the questionnaire, which may pose questions regarding the neutrality of this algorithmic model. Secondly, by interpreting this coded analysis through

¹³ Northpointe states that it incorporates “key scales from several of the most informative theoretical explanations of crime and delinquency including General Theory of Crime, Criminal Opportunity/Lifestyle Theories, Social Learning Theory, Subculture Theory, Social Control Theory, Criminal Opportunities/Routine Activities Theory, and Strain Theory” into the COMPAS model (2012, p. 2)

critical race theory, the author hopes to add a further level of analysis relative to implied narratives, societal structures, and proxy factors that may further affect the algorithmic model. The following section will summarise the findings of this qualitative critical content analysis on the COMPAS questionnaire.

5. Findings

Of the 137 questions in the COMPAS recidivism risk assessment questionnaire, 68, or 49.64% of the questions, have been categorised as opinion-based questions under the qualitative content analysis methodology developed in Section 4.3 (See: Figure 1). Most of the questions classified as opinion-based were at the end of the questionnaire, where the final sections regarding recreation/leisure, social isolation, criminal personality, anger, criminal attitudes were fully comprised of subjective statements for the respondents to agree or disagree with (See: Appendix A). The following Findings and Discussion Section will examine the questionnaire according to its thematic subsections.



5.1 Current Charges, Criminal History, Non-Compliance, and Family Criminality

The first four sections of the COMPAS questionnaire have to do with current charges, criminal history, non-compliance, and family criminality. These sections are mostly fact-based questions, as the screener and respondent can only give specific answers regarding the number of previous convictions or parole violations, for example. With regards to the literature behind the questions in the current charges and criminal history subsections, there is a well-documented and clear correlation between factors like prior arrests and jail/prison time, and recidivism (See: McCord, 1980; Farrington, 1998; Nagin and Paternoster, 2000; Pyrooz et al., 2021). This section will touch upon such literature, before considering theories used by Northpointe (2012) to motivate the questions included in the non-compliance and family criminality subsections.

To begin, influential factors within the subsections of current charges and criminal history include the number of arrests, number of times in prison or jail, duration of past incarceration, if there were parole conditions to past releases, and if those conditions were fulfilled, all of which are well-established elements in the quantitative-backed literature concerning recidivism (Skeem and Lowenkamp, 2016; Pyrooz et al., 2021). Further, there is a branch of criminality theory, 'state dependence,' that helps establish why the questions in these subsections may provide information useful to determining such re-offense risk, the main application of COMPAS. State dependence theories argue that key life events can accelerate or decrease the likelihood of criminal behaviour (Amirault and Lussier, 2011) The theory argues that the criminal justice system in the US affects the incarcerated individual so profoundly and is so detrimental to their life post-incarceration, that it ultimately increases their likelihood of recidivism due to the socio-economic consequences of having been imprisoned (Bushway, Brame and Paternoster, 1999). The literature behind this set of theories provides several layers of explanation: on the one hand there are factors that may increase criminality such as, fraternising and affiliating with other prisoners, or acclimation to comforts provided by prison, such as housing and food (Nagin and Paternoster, 2000). Then there are issues of adjustment to life after incarceration, such as: devaluation of social status and denial of access to previous opportunities like employment, housing, mobility, or voting rights, for example (Sampson and Laub, 2005). Therefore, the first two subsections of the COMPAS questionnaire are analysed as clearly fact-based and grounded by literature acceptably demonstrating that if an arrested individual has been arrested before or has spent time in jail/prison, their likelihood of recidivism is significantly higher than that of newly arrested individual. State dependence provides further theoretical

support for the types of questions asked in the first two sections of COMPAS in their relevance to determining recidivism risk.

The non-compliance subsection of the questionnaire inquires upon issues such as violations of parole/probation terms, and failing to appear for court, for example. The family criminality section is concerned with questions such as which guardians were legally responsible for the inmate during early childhood, and whether these guardians (or other family members) themselves exhibited criminal behaviour. There is a main branch of criminality theory attempting to explain innate criminality which provides insight into the nature of such questions included in these two subsections: 'population heterogeneity' (Nagin and Paternoster, 2000).

Population heterogeneity considers that individuals may have more stable propensities towards criminal behaviour due to an early-onset personality or behavioural trait which will last throughout their lifetimes, such that those with psychopathy, for example, will consistently be more prone to impulsive and risk-taking behaviours (Nagin and Paternoster, 2000; Yildirim and Derksen, 2015). Sources of this propensity include poor socialisation at youth, a negative upbringing, or even by biological or cognitive factors caused at birth (Wilson and Herrnstein, 1985; Caspi, Moffitt, and Silva, 1994). One relevant theory in this camp is the General Theory of Crime, listed as one of several key criminality theories that Northpointe incorporated into COMPAS's "theory-based assessment approach" in the making of the questionnaire (2012). Gottfredson and Hirschi's (1990) seminal work, *A General Theory of Crime* postulated that the main factor for criminality in an individual is due to poor, neglectful, or abusive child-rearing that failed to instil self-control in children below the age of eight. The main thesis of this theory is that self-control is taught and cemented at a young age,

and that those without it do not think about consequences and long-term impacts of their actions (Gottfredson and Hirschi, 1990). Although very popular when it was first published, General Theory of Crime has received some critique in the literature, the largest of which is that the authors never provided a definition for 'low self-control,' the foundational concept in their argument (Gottfredson and Hirschi, 1990; Arneklev, Elis, and Medlicott, 2006). Although it has been argued that this theory is rooted in tautological assumptions and disregards some deeper aspects of abusive households, most scholars agree with the broader idea that there is some link between low self-control and higher propensity for criminality (Schulz, 2005; Miller and Burack, 2008; Malouf et al., 2014; Nagin and Paternoster, 2000). It is possible, then, that the COMPAS questionnaire aims to provide some insight into an inmate's propensity to reoffend when inquiring about compulsive behaviour in the non-compliance section, or upbringing in the family criminality section. On a moral note, the implications of this argument in the literature includes the notion that no matter the individual's circumstances nor rehabilitation, if they were not taught self-control at an early age, they will be automatically predisposed to crime. It is unknown whether the COMPAS model was set with this conclusion in mind; still, the fact that this theory was cited as a source of motivation for the questions included in this questionnaire brings up worrying ramifications for the concept of 'innocent until proven guilty.' Furthermore, despite General Theory of Crime being listed by Northpointe, there are no questions in the COMPAS survey containing the term 'self-control' or inquiring about abuse or neglect in the home, or whether the individual thinks about consequences to their actions. Family neglect and abuse is a correlator for crime, and thus is surprising to the researcher that no questions are directed towards this recidivism-relevant issue (Maxfield, Weiler, and Widom, 2000). Even more so, clinically and academically validated

methods, such as the Brief Self-Control Measure, could have been implemented in the questionnaire for a clearer investigation into a defendant's self-control levels (Malouf et al., 2014).

This section has provided a literature-backed analysis of the first four subsections in the COMPAS questionnaire, highlighting established quantitative relationships and criminality theories where relevant. Though such relationships and theories may very well contribute to shaping a Broward County inmate's chances of re-offend, the extent of such potential contributions continues to be debated in the literature (Piquero, Farrington, and Blumstein, 2003; Amirault and Lussier, 2011). Even more so, when analysed through a critical race theory lens, although fact-based, the literature argues of potential racial proxy biases within these subsections (Huebner, DeJong, and Cobbina, 2010). The most notable subsection as risk of proxy bias is Family Criminality, where, for example Questions 33 "Was your father (or father figure who principally raised you) ever arrested, that you know of?" and 38 "Was one of your parents (or parent figure who raised you) ever sent to jail or prison?" inquire about the defendant's guardian's history with the criminal justice system. As discussed in Section 2.3.3, the disproportionate mass incarceration of Black Americans led to one in three Black men having been in imprisoned at some point in their lifetime by the 1990s (Western and Wildeman, 2009). Broward County is a highly urbanised county with 18.5% of all incarcerations being drug-related, and as mentioned previously, drug-related arrests have a much higher rate of Black incarceration relative to White incarceration (Office of Economic & Demographic Research, 2020). Thus, the chances of a young Black offender having an ex-convicted guardian or family member are disproportionately high relative to any other racial group undertaking the survey. As a consequence, an algorithmic weight that was based on theory (having on offender guardian poses a higher risk to recidivism)

can turn into a proxy for bias (a higher percentage of Black individuals have offender guardians, therefore they will be weighed as higher risk). The following section will discuss a final question in these three sections related to gang membership, its implications, and two further gang membership questions found in the questionnaire.

5.2 Gang Membership

Despite the majority of fact-based questions in the above-mentioned four sections, there was one opinion-based question in the first Section, Current Charges, and it was a stand-out question to the researcher. Question 4 asks, “Based on the screener’s observations, is this person a suspected or admitted gang member?” (Appendix A). The following section will discuss this question in depth as well as the other two questions, found in the questionnaire’s fourth section on family criminality, directly investigating gang membership. Although these questions fall out of the thematic categorisation provided by the COMPAS questionnaire, the subsequent critical analysis of this question considers them to be so analogous and key to the discussion that they will be presented together.

Question 4 is a clear example of an opinion-based question, given that it requires the screener to make a ‘statement of belief’ based on, as far as the researcher knows, only visual and behavioural observations of the defendant (Kuhn, Cheney, and Weinstock, 2000). This question thus generates potential for heuristics or biases to arise as part of the screener’s cognitive process. In a criminal justice setting, Weinstock and Cronin (2003) found that a decision-maker’s ‘epistemological level,’ or a “person’s conception of what counts as knowledge, and how certain one has to be to say that one knows” was more important in jurors than other factors debated in the literature, such as age, gender, or education level (p.161-2).

This small epistemological discussion is needed because this question, if answered by someone with lower epistemological level, opens the door to biases, assumptions, and stereotypes that the screener may hold to be knowledge. This perceived knowledge in the form of bias may be 'explicit,' therefore an individual being consciously aware of an association, attitude or belief tied to logical thinking, or 'implicit,' which is unconscious, internalised, and will likely influence decisions under stress or pressure and linked to heuristic thinking (Dividio, Kawakami, and Gaertner, 2002). Implicit bias training has come under the spotlight in recent years given its lack of efficacy and statistics that show that Black and Latino men continue to be disproportionately suspected and stopped, and with more force inflicted upon them than White men (Barvosa, 2014; Bilotta et al., 2019). Scholars have noted this behaviour across different state criminal justice systems with regards to gang membership suspicion (Toch, 2007, p.277; Piquero, 2008; Tapia, 2011), where correctional officers will easily misidentify Black and Latino men and male juveniles of colour as gang members, in some cases as much as 90% of the time (Kassel, 1998). Although there are varying jurisdictional policies for identifying gang members, such as: identification of gang symbolism, being arrested with another known gang member, reliable informants, or positive identification from the US gang intelligence database, there is no information available to the researcher's knowledge on which policy or method was applied with this question, and therefore no way of knowing to what extent this question was answered based on facts alone (Huff and Barrows, 2015; Scott, 2020). As such, for an algorithmic model sold on claims that it aims to reduce human decision-making biases and heuristics, this question completely relies on them (Kumar, 2020).

Although it may seem counterintuitive to receive a veracious response to this type of question from the defendant, in what is known

as 'self-nomination,' or 'self-identification,' it is, in fact, a valid and common method in the literature (Decker, et al., 2014; Huff and Barrows, 2015; Pyrooz, Decker, and Owens, 2020). It has also been found to be an equally reliable method when quantitatively compared to official statistics (Scott, 2020; Pyrooz et al., 2021). Although gang membership both in and out of prison has been shown in the literature to be a statistically significant risk factor for recidivism (See: Tapia, 2011; Dooley, Seals, and Skarbek, 2014; Pyrooz et al., 2021), this type of investigation into the respondent's history with gangs could have followed a more fact-based self-identification direction, as seen in Questions 43, "Have you ever been a gang member?," and 44, "Are you now a gang member?" When this research is brought under the lens of critical race theory, Question 4 becomes highly problematic, as it relies on the screener's determination of potential gang membership based on a first impression, which could be open to racial stereotyping, bias, or skewed perception of gang activity. Toch (2007) reflects that this type of accusation based on appearance or allegations of potential gang affiliation is akin to a "witch trial" (p.275). When inputted into the COMPAS model, those who have been suspected of gang membership will have an extra weight in favour of a recidivist assessment with no objective backing to that deliberation. On a final legal note, Toch's assessment of the literature surrounding gang membership classification highlights that not only does this assessment have severe consequences, but also that the most popular methods for this classification are covert, such as informally interviewing other inmates, officer assumption based on behaviour, and the use of deception for self-implication, which completely limits the individual's right to due process (2007).

Overall, this section has discussed the implications of Question 4, classified as opinion-based by the selected methodology and further

discussed as problematic and prone to bias by its second critical race theory analysis. It provided evidence in the literature for this assessment and contrasted it with the other two questions regarding gang membership, numbers 44 and 45. According to the existing literature, these two questions classified as fact-based are equally indicative of gang membership and do not carry the potential for bias that Question 4 has. Therefore, the researcher finds that if Question 4 were omitted from the COMPAS questionnaire, it would have reduced a significant source of potential bias in the survey whilst maintaining statistically sufficient information on gang membership, an important risk factor for recidivism.

5.3 Peers

This subsection focuses on the arrestees' friends and acquaintances and their involvement in gang activity, delinquency, and illegal drug use. Indeed, similarly to the questions regarding gang membership and family criminality, most studies demonstrate that one's social environment is impactful and influential in the offender's propensity to recidivate (Shapiro et al., 2010; Wall, Howells, and Delfabbro, 2011; Smalls et al., 2020). This is backed by theories listed in Northpointe's (2012) theoretically-based frameworks: Social Learning Theory (Bandura, 1977), Social Control Theory (Gibbs, 1989), and Subculture Theory (Cohen, 1955). As a whole, these theories posit that cognitive, behavioural, and self-control traits and value-systems are shaped and learned by our social environment, even if they are not directly enforced therefore making individuals more or less likely to commit crimes depending on the behaviours that were observed growing up from their social circles.

Two out of six questions in this section of the COMPAS survey were classified as fact-based. They are Questions 44 and 45, relating to gang membership, and already discussed in the above subsection, therefore,

they will not be further considered here. The remaining four of six questions in this subsection were classified as opinion-based questions given their lack of specificity, declivity to recall bias, and imprecise answer options. For example, Question 39, “How many of your friends/acquaintances have ever been arrested?” requires the respondent to assess details about their peers’ history with criminality in what is suggested by the question as the peers’ entire life, given that there is no concrete time frame in the word “ever,” thus making it open to interpretation to the respondent and thus lacking precision (Dillman, Smyth, and Christian, 2014). Furthermore, the answer options available to the respondent in the opinion-based questions of this subsection are quantitatively ambiguous: “None,” “Few,” “Half,” and “Most,” and force the respondent to estimate a response according to this given criteria and does not give the respondent room to provide an exact numerical answer if they were to have one (Schaeffer and Dykema, 2011). Lastly, from a CRT point of view, these questions are also prone to proxy bias, where Topel et al.’s (2018) Florida study correlated low-income neighbourhoods with race, due to historically racialised neighbourhood inequality, and high incarceration rates, in what the authors term as ‘prison cycling,’ supported by other studies in the literature conducted in other states (Massoglia, Firebaugh, and Warner, 2013; Western et al., 2021). Furthermore, Black citizens are overall several times more likely to be stopped and arrested by police. A report by the Brennan Center for Justice (2009) showed that 80% of highway patrols in Jacksonville, Florida, stopped people of colour, even though they only made up 5% of highway traffic. Therefore, not only may these questions yield skewed results in terms of recidivism risk due to their ambiguous nature that yield answers prone to heuristics such as rounding and generalising, but further, this proxy factor may inadvertently be biased against POC as

they will statistically be more likely to be arrested or socialise with people who have been arrested.

5.4 Substance Abuse

This following subsection will examine defendant substance abuse, its link to recidivism, and the literature on the subject. Of the nine questions in this section, six were classified as fact-based and three were classified as opinion-based. Questions such as number 53, “Did you use heroin, cocaine, crack or methamphetamines as a juvenile?” reflect the literature that links early onset drug use with juvenile and adult substance abuse and criminality and is an acceptable indicator for an individual’s history with drugs (Huebner, DeJong, and Cobbina, 2010; Stein, Deberard, and Homan, 2013; Belenko, 2019). Under Criminal Lifestyle Theory, one of the listed Northpointe (2012) conceptual frameworks, life and personality-altering issues such as substance abuse, gambling, or other addictions, make an individual more susceptible to risky situations and behaviours (Walters, 2017). Other fact-based questions within this subsection such as Question 48 “Are you currently in formal treatment for alcohol or drugs such as counselling, outpatient, inpatient, residential?” exemplifies a positive factor found in the literature to reduce recidivism: the completion of formal treatment for substance abuse (Stein, Deberard, and Homan, 2013).

However, within the opinion-based camp there are examples of leading questions such as Question 45, “Do you think your current/past legal problems are partly because of alcohol or drugs?” and Questions 51 and 52: “Do you think you would benefit from getting treatment for alcohol[/drugs]?” Given that there is no clarity as to the intent of these questions, if they are to gauge remorse, responsibility, or extent of substance abuse history, it is open to interpretation and therefore lacks quantitative precision. Furthermore, Sullivan and Artino (2017) note that,

“whenever the question topic concerns values...respondents may be more likely to choose more socially acceptable answers (social desirability response bias),” so, perhaps respondents may be more swayed to answer what is socially acceptable, posing a potential methodological limitation to such questions. Lastly, when analysed through CRT lens, this paper found that within the literature although Black communities are asymmetrically target for drug related-crimes, rates of drug use within White and Black communities tend to be equal overall when including marijuana (Moore and Elkovich, 2008). Therefore, although the usage may be similar, the fact that there is such overwhelming historical data linking Black offenders with drug arrests could mean that the model may use this factor as a proxy (Office of Economic & Demographic Research, 2020). Additionally, the highly subjective and leading nature of the opinion-based questions within this subsection call to question the neutrality of this questionnaire.

5.5 Residence/Stability

The next subsection asked respondents about their living situations and relations to family members. All ten questions in this subsection were classified as fact-based questions as both the questions and answers were clear, comprehensible, and provided the respondent with a specific time frame or context to retrieve from, all qualities of a methodologically satisfactory survey question (Schaeffer and Dykema, 2011; Dillman, Smyth, and Christian, 2014). Some examples of questions in this subsection include: 55. “How often have you moved in the last twelve months?,” 60. “How long have you been living in that community or neighborhood?,” and 63. “Do you live alone?” Answer options are either discrete numbers, yes or no, or a specific time frames with appropriate scale lengths that are not overlapping, and thus are excellent examples (Sullivan and Artino, 2017).

With regards to the content of this subsection, the scholarship suggests that place of residence post-release and frequency of contact with family and friends matters as factor towards recidivism. Housing instability is a prevalent problem for ex-convicts and is a key issue addressed by state dependence theories, for example. To elaborate, given poor chances of employment due to felony disclosures in applications, poor financial resources, prohibitions to public services or housing, and a lowered social standing, newly released individuals often face eviction or homelessness if they cannot find temporary stay (Harding, Morenoff, and Herbert, 2013). Further, after release, ex-convicts tend to go back to their original place of residence, with 60% of ex-offenders returning to a 5-mile radius from where they last resided (Harding, Morenoff, and Herbert, 2013). Kirk et al. (2017) found that ex-convicts in Michigan who moved away from their neighbourhoods after being released were statistically less likely to recidivate, one important factor of which is that they are leaving their pre-prison social environment, which may have been prone to criminality (Breetzke and Polaschek, 2018). Certain progressive policies, such as Maryland's MOVE initiative, which provides free housing for ex-convicts subject to parole regulations, are some of the few experimental policies that significantly reduce recidivism; however, these schemes remain the exception, not the norm (Kirk et al., 2018). While this section relies on fact-based questions, they are still prone to sources of inaccuracies for the overall re-offense risk assessment. In particular, family relations as a source of stability are extremely context-dependent: the family may support the ex-offender and be a source of positive influence, or relations may have frayed under the stain of incarceration and cause a further source of social rejection of the individual (Harding, Morenoff, and Herbert, 2013). On the whole, these questions are fact-based and established within a theoretical framework.

5.6 Social Environment

This subsection will examine social environment questions, which ask about the overall effects of crime on the defendant's neighbourhood of residence. Routine Activities Theory, another Northpointe-approved (2012) theoretical framework, argues that if crime is an everyday or normalised occurrence, then this may increase the likelihood of an individual with poor self-control to commit crime (Cohen and Felson, 1979). This also falls in line with the popular and controversial 'broken windows policing theory,' which argues that community bonds are broken by disorder and lack of neighbourhood safety, thus propelling the propensity to commit crime as individuals distance themselves from social influences (Davis, 2017). Extended to recidivism, these questions suggest that a return to a neighbourhood with a high propensity for crime may incite the ex-offender to recidivate. For example, Chauhan, Reppucci, and Turkheimer (2009) found that exposure to violence in delinquent juveniles was linked to recidivism. Of the six questions in this subsection, only one has been classified as fact-based: Question 70 "Are there gangs in your neighborhood?" The assumed reasoning for this question falls in line with what was discussed in Section 5.2, and the questionnaire may want to gauge if an active or ex-gang member may be tempted to join if there is gang presence. One criticism that this researcher has for this question is that, despite it being a clearly constructed question, the two answers available are "yes" or "no," which does not allow for the respondent to not know.

Examples of the remaining questions, classified as opinion-based are: 65. "Is there much crime in your neighborhood?," 68. "Do some of the people in your neighborhood feel they need to carry a weapon for protection?," and 69. "Is it easy to get drugs in your neighborhood?" These questions were classified as opinion-based as they contain vague quantities, such as "much" and "some," leading wording, such as "easy,"

and undefined large concepts, such as “crime” (Dillman, Smyth, and Christian, 2014). Therefore, the effect of this question construction leaves them open to interpretation to the respondent and, thus reducing quantitative precision. As Biemer (2017) asserts, to reduce skew, questions formulated for big data and machine learning use ought to be as accessible, concise, and definite as possible to reduce comprehension, definition, and numerical uncertainties, which will provide different answers. However, critical readings of these theories, particularly broken windows theory, assert that these theories have led to ‘zero-tolerance’ and ‘stop-and-frisk’ approaches that disproportionately target Black communities to police (Goel, Rao, and Shroff, 2016; Davis, 2017). A New York City court, for example, found that crime rates did not vary by racial composition of a neighbourhood, but in fact, the stop rates by police officers did significantly increase in Black neighbourhoods (Davis, 2017). However, when left up to the defendant’s interpretation, they may suffer from measurement errors or response bias (Dillman, Smyth, and Christian, 2014). Regarding response bias, the respondent may experience ‘recency effects,’ meaning that if there was crime recently in their neighbourhood or if an acquaintance purchased a gun in recent memory, the respondent may be inclined to answer that yes, there is usually crime, even if statistically it is not a high crime neighbourhood (Davelaar et al., 2005).

When analysed through a critical race theory, these types of questions may be subject to bias, as a Black American neighbourhood is statistically more likely to have patrols and arrests, thus making residents of these neighbourhood experience heavier police presence than others (Davis, 2017). Furthermore, although violence exposure is linked to recidivism, Chauhan, Reppucci, and Turkheimer (2009) find an important racial distinction in their study: Black delinquents were more likely to recidivate from being exposed to neighbourhood violence, whilst

White delinquents were more likely to recidivate from being exposed to parental physical abuse. Similarly to the above subsection that discusses self-control theories and their links with parental abuse and criminal behaviour, the researcher once again remarks the absence of questions touching upon this issue in the COMPAS survey. The following subsection will discuss defendant educational attainment.

5.7 Education

COMPAS measures 'educational attainment,' of the final year of education achieved by the respondent, which has been shown to be a significant factor for recidivism risk in the literature (Berg and Huebner, 2011; Watt, Howells, and Delfabbro, 2011; Pyrooz et al., 2021). Out of three factors tested, Walters (2014) found that educational attainment was the largest predictor of recidivism from the main theoretical traditions in his study. Of the nine questions in this subsection, five were classified as fact-based and four were classified as opinion based. Important educational attainment questions such as number 72 "What was your final grade completed in school?" are backed with the above-mentioned well-documented evidence linking educational achievements with propensity for recidivism. Other fact-based questions, such as numbers 73 "What were your usual grades in high school?" and 75 "Did you fail or repeat a grade level?" seem to link average academic performance with recidivism, however, to the knowledge of the researcher, there is no such established link in the literature for the time being, therefore finding these questions puzzling.

Regarding potential issues of proxy bias for this topic, there is a well-established racial divide that links low educational attainment, sex, and recidivism, with Black males holding low educational attainment being the most likely to be in prison (Everett et al., 2011). Further, Bolander

and Shuttleworth's (1998) demographic study spanning forty years of Black and White movement and education found that Black American communities remained directly correlated to lower educational attainment. Another study found that school boards and local governments are much more responsive to low scoring in predominantly White schools than Black schools, therefore increasing resources to support White school educational performance and positively contributing to their educational attainment (Hartney and Flavin, 2014). This link between lower educational attainment, race, and recidivism could mean that lower education levels may act as a proxy for race and thus lead to Black defendants being weighed as higher risks within this factor. The literature indeed notes that the effects of segregation and discriminative policies remain influential in school systems to this day, with American schools experiencing *de facto* segregation in demographics, resources received, and academic scoring (Card and Rothstein, 2007; Reardon and Owens, 2014). Although this gap in national educational attainment is narrowing over time, both for women and Black defendants, the effects of structural segregation and educational discrimination remain visible in demographics today, thus potentially creating a significant source of proxy bias within the COMPAS model (Everett et al., 2011).

Opinion-based questions had to do with juvenile aggression and behaviour in the school setting, such as Question 76 "How often did you have conflicts with teachers at school?" and Question 79 "How often did you get in fights while at school?" The available answers for Questions 76, 77, and 79 are: "Never," "Sometimes," and "Often." These questions contained undefined concepts, such as "conflicts" and "fights" and the available answers are difficult to measure, vague, and open to interpretation (Dillman, Smyth, and Christian, 2014). Furthermore, Question 78, which asks respondents to agree or disagree with the

statement “I always behaved myself in school” contains further issues such as the absolute term “always” and the undefined concept of “behaved,” that are considered to be poor constructions of survey questions (Sullivan and Artino, 2017). Although juvenile aggression and self-control are factors that contribute towards criminality, as established in theories of crime, these open opinion-based questions lack quantitative and definitional clarity.

Lastly, one factor that is lacking in this subsection is participation and/or completion of correctional educational programmes. With a high percentage of US correctional facilities offering educational programmes (84%), the absence of questions inquiring about the defendant’s participations in such programmes is notable to the researcher (Williamson, 1992; Cecil et al., 2000; Walk et al., 2012). These programmes not only aid in supporting inmates to achieve academic diplomas, such as the GED, but often provide practical vocational programmes to increase inmate employability (Cecil et al., 2000). In fact, Walk et al. (2021) found that education programmes teaching practical employable skills were statistically significantly higher at deterring recidivism than regular basic education programmes, whose recidivism-prevention are up to debate (Cecil et al., 2000; Cho and Tyler, 2010). Indeed, employment and employable skills are key deterrents for recidivism, which will be discussed in the subsection below.

5.8 Vocation (Work)

Vocational skills, previous work experience, and financial stability are important factors considered to reduce the risk of recidivism (Hannon, 2002; Berg and Huebner, 2011; Skeem and Lowenkamp, 2016; Denver, Siwach, and Bushway, 2017). The theory of criminality listed by Northpointe (2012) that best suits this section is Strain Theory (Agnew, 2012), which argues that negative structural or personal factors may

compel an individual to commit crime, such as homelessness, joblessness, bankruptcy, violence, or tragedy. COMPAS measures immediate employment after release exemplified by Question 83 “How much have you worked or been enrolled in school in the last 12 months?” However, Berg and Huebner (2011) also note that previous work experience before incarceration is equally significant, as those with little experience are much less likely to be employed in the future (Visher, Debus-Sherrill, and Yahner, 2011). In this regard, COMPAS only asks the respondent if they have ever been fired and if so, how many times, not how many jobs they have held and for how long, for example (Questions 84 and 85).

Of the fifteen questions in this subsection, eight have been classified as opinion-based. Some questions were found to have leading wording such as ‘survival’ in 94 “How often do you worry about financial survival?,” ‘trouble’ in 92 “How often do you have trouble paying bills?,” ‘barely enough’ in 92 “How often do you have barely enough money to get by?,” and ‘hard’ in 89 “How hard is it for you to find a job ABOVE (sic.) minimum wage compared to others?” The wording utilised in these questions is leading in that it is all negative, thus instead of asking the defendant how many times they were able to pay their bills on time in [time frame range], the use of the word ‘trouble’ insinuates that the defendant struggles to pay their bills and therefore lacks neutrality in its construction (Sullivan and Artino, 2017). Furthermore, the lack of neutral language and specific time periods in these opinion-based questions risk potential recency effects, where they may have had unsuccessful financial or employment experiences recently that may taint their recollection and thus their response of “how often” (Holbrook et al., 2007). Other imprecise wording examples in this subsection’s opinion-based questions include ‘frequently,’ ‘conflicts,’ and ‘successful,’ that leave the definition up to the defendant. Furthermore, the lack of flexible

answers available such as 'often,' 'sometimes,' and 'never' do not allow the respondent to be specific with financial or employment struggles.

These recidivism risk factors of employment and financial stability, however, are considered a strong proxy for minority races in the literature (Skeem and Lowenkamp, 2016). Overall, White American families have higher net wealth than Black American families, and furthermore, Baker (2017) found that White families are more likely to support their children financially than Black families, where the inverse effect of children financially supporting parents is identified, thus reducing chances for post-release financial support. Furthermore, a study performed in Florida empirically demonstrated that Black American recidivism was significantly influenced by unemployment rates unlike White recidivism, with Black ex-prisoners returning to high unemployment areas being more susceptible to recidivism (Wang, Mears, and Bales, 2010). One line of reasoning examined by the authors was that White ex-inmates have higher social capital upon their return homes and therefore were less affected by the unemployment rate of their area of residence, unlike Black Americans, who had more "accumulated disadvantages," and therefore were more susceptible to detrimental unemployment effects (Wang, Mears, and Bales, 2010, p.1198; Baker, 2017). At the time of the ProPublica investigation, the average unemployment rate in Broward County was 6.5%; however, when observed through racial categories, we note that within this time unemployment for Whites was 3.9% and on the other hand, 11.1% for Black Americans (Broward County, 2017). Therefore, the disproportionately high Black unemployment rate compared to White within Broward County alone supports the scholarship on the problematic nature of this risk factor as a race proxy. Lastly, despite American anti-discrimination laws, employment tends to favourably skew towards males and Whites, with studies demonstrating that women and minorities may be as much as 50% less likely to be

called to interview (Everett et al., 2011). Therefore, in addition to the already difficult task for gaining employment post-release, this statistically demonstrated discrimination adds another hurdle to Black ex-prisoners, thus strengthening the correlation in the algorithmic model between race and recidivism due to recent lack of employment. Finally, from an intersectional perspective, it is important to note that, even though Black women are the majority breadwinners in their households, a 2017 study calculated those Black women earned 21% less than White women on average (Banks, 2019). This study also reflected the disproportionate number of Black women working minimum and sub-minimum wage jobs compared to their White counterparts, which could certainly be seen as a proxy factor in Question 89. The following subsection will discuss the final five themes in the COMPAS questionnaire, which were grouped together due to their stark similarities.

5.9 Leisure/Recreation, Social Isolation, Criminal Personality, Anger, and Criminal Attitudes

The following section has grouped the final five subdivisions since they have all been classified as opinion-based and share a similar question format. The Leisure/Recreation subdivision aims to assess the defendant's recent lifestyle and feelings of boredom, which have been shown in juveniles to sometimes be a reason for offending (Putninš, 2010). Social Isolation aims to address the difficult process of social reintegration after prison, where social isolation has been demonstrated in the scholarship to "considerably increase the risk of recidivism" (Sung, 2011, p.219). Regarding Criminal Personality, Anger, and Criminal Attitudes, the conceptual frameworks and literature that base the construction of these questions can be found in Criminal Lifestyle

Theory, Criminal Opportunity Theory, and the above-discussed Routines Activities Theory, and self-control and innate criminality theories. In brief, Criminal Lifestyle Theory contends that an individual's lifestyle over time develops toward a propensity for criminal activity with increased incentives for crime and opportunities to commit crime, and, ultimately, active decision-making towards criminal activity (Walters, 2017). With regards to recidivism, studies have shown that prison lifestyle change programmes positively reduce recidivism in comparison to inmates who did not partake in these programmes (Walters, 2005). Criminal Opportunity Theory is another theoretical example of the argument that environment helps shape individual criminality and helps put context to the decision-making aspect of crime, where an individual concluded that the context and motivations were opportune to commit an offense (Hannon, 2002; Sacco et al., 2004). Thus, these theories and studies suggest that if released individuals that potentially have low self-control or propensity towards violence or crime return to their pre-incarceration lifestyle, attitudes, and mentalities, they will be more likely to recidivate (Vrućinić, 2019).

This final grouping of COMPAS questionnaire sections mainly contains a list of statements that the respondent can agree or disagree with. Of the eight Leisure/Recreation questions, five ask about how 'often' the defendant feels bored. By asking so many similarly redundant questions in a row (95 "How often did you feel bored?," 100 "Do you often become bored with your usual activities?" and 101 "Do you feel that the things you do are boring or dull?") the researcher questions if this may lead the respondent to agree after being primed with thinking about boredom (Schaeffer and Dykema, 2011). Furthermore, questions such as number 95, which asks the defendant how often they felt bored within the past six months and provided available answers 'never,' 'several times per month,' 'several times per week,' and 'daily,' may induce 'recall

bias,' where the respondent may not remember accurately and thus will not provide a correct answer (Holbrook et al., 2007).

All of the following subdivisions provide statements and the following answers for the respondent to identify with: 'strongly disagree,' 'disagree,' 'not sure,' 'agree,' and 'strongly agree.' An example question for Social Isolation is: 111 "I have never felt sad about things in my life," an example within Criminal Personality is: 113 "I always practice what I preach," for Anger, we can showcase: 122 "I get into trouble because I do things without thinking," lastly, some illustrative questions from Criminal Attitudes are: 127 "A hungry person has a right to steal," and 132 "I have felt very angry at someone or at something." These questions and many others included contain absolute wording such as 'always' and 'never,' and overall, use negatively suggestive or leading wording such as 'trouble,' 'feel left out,' 'unfeeling,' 'violent,' and 'short temper.' Furthermore, questions such as 111 and 132 (see: above) seem to be unrealistic markers for criminal attitudes and social isolation, as it may be difficult to find a person who has never felt sad or angry. There is an asymmetry in negatively worded and positively worded statements, which may skew the respondent to answer negatively. Further, with the questionnaire being 137 questions long, these final statements may have the respondent suffering from survey fatigue, thus further interfering with respondent accuracy and cognitive effort exerted into reflecting and answering the question.

6. Conclusion

This investigation of the US racial inequality-algorithmic bias nexus has underscored the importance of algorithmic programme piloting, implicit bias in data, and inheritance of discriminatory practices and attitudes that may be reflected in the algorithmic model's assessment. This research has provided extensive literature reviews, both from the sides of

algorithmic fairness and bias and the history and discriminatory practices of policing and race in the USA to provide a robust contextualisation of the question at hand. Through a qualitative critical content analysis of the COMPAS recidivism risk assessment questionnaire, this research has identified two main potential sources of bias or skewed data.

First, despite the COMPAS questionnaire's overall theoretical backing, almost half of it is comprised of opinion-based questions, which call to question the survey's quantitative and empirical neutrality to counter human heuristics. Furthermore, many of the opinion-based questions were constructed poorly according to the literature, with ambiguous and absolute terms, suggestive wording, and lack of appropriate response options, thus welcoming heuristics into the survey. Furthermore, the respondents answering this survey may have suffered from survey fatigue, recall bias, response bias, and social desirability response bias due to the inadequate questionnaire construction.

Second, this study found potential for proxy bias within the questionnaire from its additional analysis employing critical race theory. Indeed, factors such as financial stability, family criminality, employment, education, and propensity of arrest have been linked to race; Sections 2.3 and 5 provided ample evidence of this. These proxies thus could reflect racial bias in the algorithmic model and perhaps even have the dangerous possibility of their proxy correlations becoming further cemented in the algorithm. Because we do not know how much each of these factors weighs in the COMPAS model, it is unknown how the COMPAS recidivism risk assessment is actually generated and therefore how much the proxy factors may be considered in this recidivism calculation. However, this research has identified a number of problematic, or methodologically limited, questions in the survey that may be in danger of carrying over implicit biases to these calculations, such as Question

4, which asks for the screener's assessment of the defendant's possible involvement in gang activity.

Notably, COMPAS missed some recidivism deterrent factors and resources from the literature that could have aided in the questionnaire, such as family neglect and abuse, the Brief Self-Control Measure, and inmate vocational programme attainment. Furthermore, the survey was excessively long, taking between 45 minutes and an hour, according to Northpointe (2012), which authors argue could lead to survey fatigue encouraging respondents to "answer carelessly just to finish" and recency effects due to poor recollection from fatigue (Holbrook et al., 2007; Sullivan and Artino, 2017). Other issues found in this analysis of COMPAS are that it was not piloted prior to deployment, (Gehlbach and Brinkworth; 2011; Dillman, Smyth, and Christian, 2014; Rickards, Magee, and Artino, 2020) a fundamental step in evaluating assessment reliability, particularly when big data is involved in socially significant applications (Biemer, 2017).

Broader implications include lack of transparency from these AI/ML services, and legal and ethical issues. It is not known whether the screener or correctional officer has informed the defendant about what the purpose of this questionnaire is and what the COMPAS risk score will imply. Smith (2014) argues that lack of participant awareness of the process is at danger of violating the defendant's rights to due process, which is a point of concern for the researcher.

This research's contribution towards the algorithmic bias debate hopes to encourage future research into data bias inheritance and proxy bias. By expanding the algorithmic bias debate to include qualitative critical content analysis, the author has found that a majority of COMPAS's questions were not only opinion-based, but also potential sources of proxy bias. These findings are worrying, as they imply that

algorithmic models in the criminal justice system could potentially continue to perpetuate cycles of incarceration for people of colour. Thus, the researcher also hopes with the burgeoning field of critical algorithmic fairness that the lack of transparency and testing observed with these socially-impactful applications is addressed by policy-makers, as they have the potential of causing disparate impact in marginalised communities. Lastly, before we allow algorithmic models to take over our criminal justice system and judicial decisions, we must first address the centuries of inequality and discriminatory practices cemented within our institutions and our data that will undoubtedly be inherited by machine learning.

7. Bibliography

Aaronson, D., Hartley, D.A., Mazumder, B. (2020) 'The Effects of the 1930s HOLC 'Redlining Maps,' *Federal Reserve Bank of Chicago*, Working Paper No. WP-2017-12.

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., Robinson, D.G. (2020) 'Roles for Computing in Social Change,' In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 252-260.

Acharya, A., Blackwell, M. and Sen, M. (2018) *Deep roots: how slavery still shapes Southern politics*, Princeton, NJ: Princeton University Press.

Agarwal A., Beygelzimer A., Dudík M., Langford J., Wallach H. (2018) "A reductions approach to fair classification", Preprint, available at: <https://arxiv.org/pdf/1803.02453.pdf>.

Agnew, R. (2012) "Reflection on "A Revised Strain Theory of Delinquency", " *Social forces*, vol. 91, no. 1, pp. 33-38.

- Akers, R.L. (2000) *Criminological theories: Introduction and evaluation* (3rd ed.) Los Angeles, CA: Roxbury.
- Alexander, M. (2019) *The New Jim Crow: Mass Incarceration in the Age of Colourblindness*, London: Penguin Books.
- Allen, W.D. (2007) "The Reporting and Underreporting of Rape," *Southern Economic Journal*, vol. 73, no. 3, pp. 623-641.
- Alper, M., Durose, M.R. and Markman, J. (2018) '2018 Update on Prisoner Recidivism: A 9-Year Follow-up Period (2005-2014),' *Bureau of Justice Statistics*, available at: <https://www.bjs.gov/content/pub/pdf/18upr9yfup0514.pdf>.
- Amirault, J. and Lussier, P. (2011) "Population heterogeneity, state dependence and sexual offender recidivism: The aging process and the lost predictive impact of prior criminal charges over time," *Journal of criminal justice*, vol. 39, no. 4, pp. 344-354.
- Andersen, S.H., Andersen, L.H. and Skov, P.E. (2015) "Effect of Marriage and Spousal Criminality on Recidivism," *Journal of marriage and family*, vol. 77, no. 2, pp. 496-509.
- Angwin, J., Larson, J., Mattu, S., Kirschner, L. (2016) "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica*, May 23, 2016.
- Annamma, S.A., Ferri, B.A., Connor, D.J. (2018) "Disability Critical Race Theory: Exploring the Intersectional Lineage, Emergence, and Potential Futures of DisCrit in Education," *Review of Research in Education*, vol. 42, no. 1, pp. 46-71.
- Aptheker, H. (ed.) (1951-1994) *A Documentary History of the Negro People in the United States*, Secaucus, NJ: Citadel Press, 7 vols.

Arneklev, B.J., Elis, L., and Medicott, S. (2006) "Testing the General Theory of Crime: Comparing the Effects of 'Imprudent Behavior' and an Attitudinal Indicator of 'Low Self-Control,'" *Western Criminology Review*, vol. 7, no. 4, pp. 41-55.

Austin, J., and Irwin, J. (2001) *It's about time: America's imprisonment binge* Belmont, CA: Wadsworth.

Bahnsen, A.C., Torroledo, I., Camacho, L.D., Villegas, S. (2018) 'DeepPhish: Simulating malicious AI', *APWG Symposium on Electronic Crime Research*, pp. 1–9.

Baker, E.S. (2017) "Explaining the black–white gap in returns to education: it's not a black-and-white issue," *Monthly labor review*, pp. 1-2.

Bandura, A. (1977) *Social Learning Theory*, Oxford, England: Prentice-Hall.

Banks, N. (2019) "Black women's labor market history reveals deep-seated race and gender discrimination," *Economic Policy Institute*.

Bansal, N. and Sviridenko, M. (2006) "The Santa Claus Problem", In: *Proceedings on the 38th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 31-40.

Barman, E. (2016) *Caring Capitalism: The Meaning and Measure of Social Value*, New York: Cambridge University Press.

Barocas, S., Bradley, E., Hanover, B., Provost, F. (2017) "Big Data, Data Science, and Civil Rights," Preprint, available at: <https://arxiv.org/abs/1706.03102>.

Barocas, S. and Selbst, A.D. (2016) "Big Data's Disparate Impact", *California Law Review*, vol. 104, no. 3, pp. 671-732.

Barocas, S., Hardt, M., Narayanan, A. (2021) *Fairness and Machine Learning: Limitations and Opportunities*, available at: <https://fairmlbook.org/>.

Barry-Jester, A.M., Casselman, B., Goldstein, D. (2015) 'The New Science of Sentencing: Should Prison Sentences be Based on Crimes that haven't been Committed yet?' *The Marshall Project*, available at: <https://perma.cc/Z5SV-TRLG#gcg29KxNS>.

Barvosa, E. (2014) "Unconscious bias in the suppressive policing of Black and Latino men and boys: neuroscience, Borderlands theory, and the policymaking quest for just policing," *Politics, groups & identities*, vol. 2, no. 2, pp. 260-283.

Bayley, D.H. (1982'3) 'Knowledge of the police,' In Punch, M. (ed.) *Control in the Police Organization*, Cambridge, MA: MIT Press, pp. 18-35.

Beach, S.R., Rogers, R., and Short, K.G. (2009). "Exploring the "critical" in Critical Content Analysis of Children's Literature," In: *58th Yearbook of the National Reading Conference*.

Belenko, S. (2019) "The role of drug courts in promoting desistance and recovery: a merging of therapy and accountability," *Addiction research & theory*, vol. 27, no. 1, pp. 3-15.

Bell, D.A. (1980) "Brown v. Board of Education and the Interest-Convergence Dilemma," *Harvard Law Review*, vol. 93, no. 3, pp. 518-533.

Bell, D.A. (1992) *Faces at the Bottom of the Well: The Permanence of Racism*, New York: Basic Books.

Benda, B.B., Toombs, N.J. and Peacock, M. (2003) "An Empirical Examination of Competing Theories in Predicting Recidivism of Adult

Offenders Five Years After Graduation from Boot Camp,” *Journal of offender rehabilitation*, vol. 37, no. 2, pp. 43-75.

Benthall, S. and Haynes, B.D. (2018) “Racial categories in machine learning,” Preprint, available at: <https://arxiv.org/abs/1811.11668>.

Benton, M. (2018) “‘Just the Way Things Are Around Here’: Racial Segregation, Critical Junctures, and Path Dependence in Saint Louis,” *Journal of urban history*, vol. 44, no. 6, pp. 1113-1130.

Berg, M.T. and Huebne, B.M. (2011) “Reentry and the Ties That Bind: An Examination of Social Ties, Employment, and Recidivism,” *Justice Quarterly*, vol. 28, no. 2, pp. 382-410.

Berk, R.A. and Bleich, J. (2013) “Statistical Procedures for Forecasting Criminal Behavior,” *Criminology & Public Policy*, vol. 12, no. 3, pp. 513-544.

Berlin, I. and Morgan, P.D. (1991) *The Slaves' Economy: Independent Production by Slaves in the Americas*, London: Routledge.

Berman, E.P. and Hirschman, D. (2018) ‘The Sociology of Quantification: Where are We Now?’, *Contemporary Sociology*, vol. 47, no. 3, pp. 257-266.

Biemer, P.P. (2017) *Total survey error in practice*, Hoboken, NJ: Wiley.

Bigo, D., Isin, E. and Ruppert, E. (eds.) (2019) *Data Politics: Worlds, Subjects, and Rights*, London: Routledge.

Bilotta, I., Corrington, A., Mendoza, S.A., Watson, I., King, E. (2019) “How Subtle Bias Infects the Law,” *Annual review of law and social science*, vol. 15, no. 1, pp. 227-245.

Bolander, R.C. and Shuttleworth, M. (1998) “Differentials of Race and Educational Attainment in Residential Centralization Patterns in Three

Ohio Metropolitan Areas, 1950-1990," *Sociological focus*, vol. 31, no. 3, pp. 283-293.

Bolukbasi, T. Chang, K.W., Zou, J., Saligrama, V., Kalai, A. (2016) "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." Preprint, available at: <https://arxiv.org/abs/1607.06520>.

Boulamwini, J. and Gebru, T. (2018) "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*), *Proceedings of Machine Learning Research* no. 81, pp.77-91.

Bowker, G.C. and Star, S.L. (2000) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Bracey, G.E. (2015) "Toward a Critical Race Theory of State," *Critical Sociology*, vol. 41, no. 3, pp. 553-572.

Breetzke, G. and Polaschek, D. (2018) "Moving Home: Examining the Independent Effects of Individual- and Neighborhood-Level Residential Mobility on Recidivism in High-Risk Parolees," *International journal of offender therapy and comparative criminology*, vol. 62, no. 10, pp. 2982-3005.

Brennan Center for Justice (2009) 'Racial Bias in Florida's Electoral System,' *Brennan Center for Justice*, available at: <https://brennancenter.org>.

Brennan, T., Dieterich, W., Ehret, B. (2009) "Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System," *Criminal Justice and Behavior*, vol. 36, no.1, pp. 21-40.

Bridges, K.M, Keel, T., and Obasogie, O.K. (2017) "Introduction: Critical Race Theory and the Health Sciences," *American Journal of Law & Medicine*, vol. 43, no. 23, pp. 179-182.

Broward County (2017) 'Broward County Employment Trends,' *Broward by the Numbers*, no. 2017-01.

Brown, S. (2015) *Dark Matters: On the Surveillance of Blackness*. Durham: Duke University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., Ó hÉigearthaigh, S., Beard, S., Belfied, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., and Amodei, D. (2018) "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation", Preprint, available at: <https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf>.

Bruno, I., Didier, E., Prévieux, J., Tasset, C. (2014) *Statactivisme: Comment lutter avec des nombres*, Paris: La Découverte.

Bureau of Justice Statistics (2021) 'Key Statistics,' *Bureau of Justice Statistics*, available at: <https://bjs.ojp.gov/data/key-statistics>.

Burrell-Craft, K. (2020) "Are (We) Going Deep Enough? A Narrative Literature Review Addressing Critical Race Theory, Racial Space Theory, and Black Identity Development," *Taboo: The Journal of Culture and Education*, vol. 19, no. 4, pp. 9-26.

Bushway, S., Brame, R., and Paternoster, R. (1999) "Assessing stability and change in criminal offending: A comparison of random effects, semiparametric, and fixed effects modeling strategies," *Journal of Quantitative Criminology*, vol. 15, pp. 23–61.

Cabrera, N.L. (2018) "Where is the Racial Theory in Critical Race Theory?: A Constructive Criticism of the Critics," *The Review of Higher Education*, vol. 42, no. 1, pp. 209-233.

Cameron, S.C. and Phan, L.T. (2018) "Ten Stages of American Indian Genocide," *Interamerican Journal of Psychology*, vol. 52, no. 1, pp. 25-44.

Castelvecchi, D. (2016) "Can we open the black box of AI?," *Nature*, vol. 538, pp. 20-23.

Card, D. and Rothstein, J. (2007) "Racial segregation and the black-white test score gap," *Journal of public economics*, vol. 91, no. 11, pp. 2158-2184.

Cardi, J., Hans, V.P., and Parks, G. (2020) "Do Black Injuries Matter? Implicit Bias and Jury Decision Making in Tort Cases," *Southern California Law Review*, vol. 93, no. 3.

Carter, J.S. and Lippard, C.D. (2020) *The death of affirmative action?: racialized framing and the fight against racial preference in college admissions*, Bristol University Press, Bristol.

Caspi, A., Lynam, D.R., Moffitt, T.E., and Silva, P.A. (1994) "Are some people crime prone? Replications on the personality-crime relationship across countries, genders, races, and methods," *Criminology*, vol. 32, pp. 163-195.

Cecil, D.K., Drapkin, D.A., MacKenzie, L., Hickman, L.J. (2000) "The Effectiveness of Adult Basic Education and Life-Skills Programs In Reducing Recidivism: A Review and Assessment of the Research," *Journal of correctional education*, vol. 51, no. 2, pp. 207-226.

Chang, R.S. and Gotanda, N. (2007) "The Race Question In LatCrit Theory and Asian American Jurisprudence," *Nevada Law Journal*, vol. 7, pp. 1012-1029.

Chapman, E.N., Kaatz, A., and Carnes, M. (2013) "Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Health Care Disparities," *Journal of General Internal Medicine*, vol. 28, no. 11.

Chauhan, P., Reppucci, N.D. and Turkheimer, E.N. (2009) "Racial differences in the associations of neighborhood disadvantage, exposure to violence, and criminal recidivism among female juvenile offenders," *Behavioral sciences & the law*, vol. 27, no. 4, pp. 531-552.

Chiao, V. (2019) "Fairness, Accountability and Transparency: Notes on Algorithmic Decision-making in Criminal Justice", *International Journal of Law in Context*, vol. 15, no. 2, pp. 126-139.

Cho, R.M. and Tyler, J.H. (2013) "Does Prison-Based Adult Basic Education Improve Postrelease Outcomes for Male Prisoners in Florida?," *Crime and delinquency*, vol. 59, no. 7, pp. 975-1005.

Chouldechova, A. (2017) "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", *Big Data*, vol. 5, no. 2, pp.153-163.

Clark, L.D. (2021) "A Critique of Professor Derrick A. Bell's Thesis of Permanence of Racism and His Strategy of Confrontation," *Denver University Law Review*, vol. 73, no. 1, pp.23-50.

Cohen, A.K. (1955) *Delinquent Boys: The Culture of the Gang*, New York, NY: Free Press.

Cohen-Cole, E. (2011) "Credit Card Redlining," *The Review of Economics and Statistics*, vol. 93, no. 2, pp. 700-713.

Cohen, L.E. and Felson, M. (1979) "Social Change and Crime Rate Trends: A Routine Activities Approach," *American sociological review*, vol. 44, no. 4, pp. 588.

Collins, P.H. (2010) "Like one of the family: race, ethnicity, and the paradox of US national identity," *Ethnic and Racial Studies*, vol. 24, no. 1, pp. 3-28.

Comino, S., Mastrobuoni, G. and Nicolò, A. (2020) "Silence of the Innocents: Undocumented Immigrants' Underreporting of Crime and their Victimization," *Journal of Policy Analysis and Management*, vol. 39, no. 4, pp. 1214-1245.

Conference of Chief Justices & Conference of State Court Administrators (2007) 'Resolution 12: In Support of Sentencing Practices that Promote Public Safety and Reduce Recidivism,' *National Center for State Court*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. (2017) "Algorithmic Decision Making and the Cost of Fairness", In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 797-806.

Courtland, R. (2018) "Bias detectives: the researchers striving to make algorithms fair," *Nature*, vol. 558, pp. 357-360.

Crawford, K. and Schultz, J. (2019) "AI systems as state actors," *Columbia Law Review*, vol. 119, no. 7.

Crenshaw, K., Gotanda, N., Peller, G., Thomas, K. (1995) *Critical Race Theory: The Key Writings That Formed the Movement*, New York: The New Press.

Curtin, M.E. (2000) *Black Prisoners and Their World, Alabama, 1865-1900*, Charlottesville, VA: University Press of Virginia.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011) "Extraneous factors in judicial decisions," *Proceedings of the National Academy of Sciences - PNAS*, vol. 108, no. 17, pp. 6889-6892.

Davelaar, E.J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H.J., Usher, M. (2005) "The Demise of Short-Term Memory Revisited: Empirical and Computational Investigations of Recency Effects," *Psychological review*, vol. 112, no. 1, pp. 3-42.

Davis, H.E. (2017) "Broken and Disordered: Selected Critical Readings on Broken Windows Policing," *Legal reference services quarterly*, vol. 36, no. 3-4, pp. 166-189.

Debnath, S., Barnaby, D.P., Coppa, K, Makhnevich, A., Kim, E.J., Chatterjee, S., Toth, V., Levy, T.J., Paradis, M.D., Cohen, S.L., Hirsch, J.S., Zanos, T.P., Northwell COVID-19 Research Consortium. (2020) "Machine learning to assist clinical decision-making during the COVID-19 pandemic", *Bioelectronic Medicine*, vol. 6, no. 14.

Decker, S.H., Pyrooz, D.C., Sweeten, G., Moule, R.K. (2014) "Validating Self-Nomination in Gang Research: Assessing Differences in Gang Embeddedness Across Non-, Current, and Former Gang Members," *Journal of Quantitative Criminology*, vol. 30, no. 4, pp. 577-598.

Delgado, R. and Stefancic, J. (1998) "Critical Race Theory: Past, Present, and Future," *Current Legal Problems*, vol. 51, no. 1, pp. 467-491.

Delgado, R. and Stefancic, J. (2017) *Critical Race Theory: An Introduction* (3 ed.) New York: New York University Press.

Denver, M., Siwach, G. and Bushway, S.D. (2017) "A New Look At the Employment and Recidivism Relationship Through the Lens of a Criminal Background Check," *Criminology*, vol. 55, no. 1, pp. 174-204.

Department of Justice (2014) 'Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference,' *Office of Public Affaris*, available at: <https://justice.gov/opa>.

Dhondt, G. (2012) "The bluntness of incarceration: crime and punishment in Tallahassee neighborhoods, 1995 to 2002," *Crime, law, and social change*, vol. 57, no. 5, pp. 521-538.

Dieterich, W., Mendoza, C. and Brennan, T. (2016) "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," *Northpointe Inc. Research Department*.

Dillman, D.A., Smyth, J.D. and Christian, L.M. (2014) *Internet, phone, mail, and mixed-mode surveys: the tailored design method*, (4th ed.) Hoboken, NJ: Wiley.

Disha, I.D., Canvendish, J.C., and King, R.D. (2011) "Historical Events and Spaces of Hate: Hate Crimes against Arabs and Muslims in Post-9/11 America," *Social Problems*, vol. 58, no. 1, pp. 21-46.

Dovidio, J.F., Kawakami, K., and Gaertner, S.L. (2002) "Implicit and explicit prejudice and interracial interaction," *Journal of Personality and Social Psychology*, vol. 82, no. 1, pp. 62–68.

Dooley, B.D., Seals, A., and Skarbek, D. (2014) "The effect of prison gang membership on recidivism," *Journal of Criminal Justice*, vol. 42, no. 3, pp. 267-275.

Dressel, J. & Farid, H. 2018, "The accuracy, fairness, and limits of predicting recidivism", *Science Advances*, vol. 4, no. 1, pp. eaao5580.

Driver, J. (2011) "Rethinking the Interest-Convergence Thesis," *Northwestern University Law Review*, vol. 105, no. 1, pp. 149-198.

D’Rozario, D. and Williams, J.D. (2005) “Retail Redlining: Definition, Typology and Measurement,” *Journal of Macromarketing*, vol. 25, no. 2, pp. 175-186.

Du Bois, W.E.B. (1935) *Black Reconstruction: An Essay Toward a History of the Part Which Black Folk Played in the Attempt to Reconstruct Democracy in America, 1860-1880*, New York: Harcourt, Brace and Company.

Dulaney, M.W. (1996) *Black Police in America*, Bloomington, IN: Indiana University Press.

Dunbar-Ortiz, R. (2014) *An Indigenous Peoples’ History of the United States*, Boston: Beacon Press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2011) “Fairness Through Awareness,” Preprint, available at: <https://arxiv.org/abs/1104.3913>.

Eisenhauer, E. (2001) “In poor health: Supermarket redlining and urban nutrition,” *GeoJournal*, vol. 53, pp. 125-133.

Elder-Vass, D. (2012) “Towards a realist social constructionism,” *Sociología, problemas e prácticas*, vol. 70, pg. 9-24.

Englich, B. (2009) ‘Heuristic strategies and persistent biases in sentencing decisions,’ in Oswald, M.E., Bieneck, S. and Hupfeld-Heinemann, J. (eds), *Social Psychology of Punishment of Crime*. Hoboken: Wiley-Blackwell, pp. 295–314.

Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C., Venkatasubramanian, S. (2018) “Runaway Feedback Loops in Predictive Policing,” In: *Proceedings of Machine Learning Research 81*, pp. 1-12.

Eterno, J.A., Verman, A., and Silverman, E.B. (2016) "Police Manipulations of Crime Reporting: Insiders' Revelations," *Justice Quarterly*, vol. 33, no. 5, pp. 811-835.

Everett, B.G., Rogers, R.G., Hummer, R.A., Krueger, P.M. (2011) "Trends in educational attainment by race/ethnicity, nativity, and sex in the United States, 1989-2005," *Ethnic and racial studies*, vol. 34, no. 9, pp. 1543-1566.

Farrington, D.P. (1998) 'Predictors, causes, and correlates of male youth violence,' In Tonry, M. and Moore, M.H. (eds.) *Youth and Violence: Crime and Justice, An Annual Review of Research*, Chicago, IL: University of Chicago Press.

Feagin, J.R. (2013) *The White Racial Frame: Centuries of Racial Framing and Counter-Framing*, London: Taylor & Francis Group.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., Vankatasubramanian, S. (2015) "Certifying and Removing Disparate Impact," In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovering and Data Mining*, pp. 259-268.

Fish, B., Kun, J. and Lelkes, A.D. (2016) "A Confidence-Based Approach for Balancing Fairness and Accuracy," Preprint, available at: <https://arxiv.org/abs/1601.05764>.

Fiske, J. and Hancock, B.H. (2016) *Media Matters: Race & Gender in US Politics*, London: Routledge.

Flores, A.W., Bechtel, K. and Lowenkamp, C.T. (2016) "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks,'" *Federal Probation*, vol. 80, no. 2, pp-38-46.

Florida Department of Corrections (2014) '2014 Annual Jail Capacity Survey,' *Florida Department of Corrections*, available at: <https://dc.state.fl.us/pub/jails/2014/2014AnnualJailCapacitySurvey.pdf>.

Frank, R., Brantingham, P.L., and Farrel, G. (2012) "Estimating the True Rate of Repeat Victimization from Police Recording Crime Data," *Canadian Journal of Criminology and Criminal Justice*, vol. 54, no. 4, pp. 481-494.

Friendly, H.J. (1975) "Some Kind of Hearing," *University of Pennsylvania Law Review*, vol. 123, pp. 1267-1317.

Gangadharan, S.P. and Niklas, J. (2019) "Decentering technology in discourse on discrimination," *Information, Communication & Society*, vol. 22, no. 7, 882-899.

Garg, N., Schiebinger, L., Jurafsky, D., Zou, J. (2017) "World Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." Preprint, available at: <https://arxiv.org/abs/1711.08412v1>.

Gehlbach, H. and Brinkworth, M.E. (2011) "Measure Twice, Cut Down Error: A Process for Enhancing the Validity of Survey Scales," *Review of general psychology*, vol. 15, no. 4, pp. 380-387.

Georges-Abeyie, D.E. (2010) 'Race, Crime and Criminal Justice in the United States of America,' In Kalunta-Cumpton, A. (ed.) *Race, Crime and Criminal Justice*, London: Palgrave Macmillan, pp. 286-305.

Gibbs, J.P. (1989) *Control: Sociology's Central Notion*, Urbana, IL: University of Illinois Press.

Gillborn, D. (2015) "Intersectionality, Critical Race Theory, and the Primacy of Racism: Race, Class, Gender, and Disability in Education," *Qualitative Inquiry*, vol. 21, no. 3, pp. 277-287.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., and Kagal, L. (2018) "Explaining Explanations: An Overview of Interpretability of Machine Learning", *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Turin, Italy, pp. 80-89.

Goel, S., Rao, J.M., and Shroff, R. (2016) "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-frisk Policy," *The Annals of Applied Statistics*, vol. 10, no. 4, pp. 365-394.

Gong, A. (2016) 'Ethics for powerful algorithms (1 of 4),' *Medium*, July 12, available at: <https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84>.

Goodman-Delahunty, J. and Sporer, S.L. (2010) "Unconscious influences in sentencing decisions: a research review of psychological sources of disparity," *Australian Journal of Forensic Sciences*, vol. 42, pp. 19–36.

Gottfredson, M.R. and Hirschi, T. (1990) *A General Theory of Crime*, Palo Alto, CA: Stanford University Press.

Gottlieb, A. and Sugie, N.F. (2019) "Marriage, Cohabitation, and Crime: Differentiating Associations by Partnership Stage," *Justice Quarterly*, vol. 36, no. 3, pp. 503-531.

Gramlich, J. (2020) 'Black imprisonment rate in the U.S. has fallen by a third since 2006', *Pew Research Center*, available at: <https://www.pewresearch.org/fact-tank/2020/05/06/share-of-black-white-hispanic-americans-in-prison-2018-vs-2006/>.

Graneheim, U.H., Lindgren, B.M., and Lundman, B. (2017) "Methodological challenges in qualitative content analysis: A discussion paper," *Nurse Education Today*, vol. 56, pp. 29-34.

Gray-Ray, P., Ray, M.C., Rutland, S., Turner, S. (1995) "African Americans and the Criminal Justice System," *Humboldt Journal of Social Relations*, vol. 21, no. 2, pp. 105-117.

Green, B. and Hu, L. (2018) 'The myth in the methodology: towards a recontextualization of fairness in machine learning,' In: *Proceedings of Machine Learning: the Debates Workshop* at the 35th International Conference on Machine Learning (ICML).

Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., Weller, A. (2016) "The Care for Process Fairness in Learning: Feature Selection for Fair Decision Making," In *Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems*.

Grinevičius, J. and Akavickaitė, A. (2021) 'People Tested How Google Translates From Gender Neutral Languages And Shared The "Sexist" Results,' *Bored Panda*, available at: https://www.boredpanda.com/google-translate-sexist/?utm_source=bing&utm_medium=organic&utm_campaign=organic.

Gross, K. (2006) *Colored Amazons: Crime, Violence, and Black Women in the City of Brotherly Love, 1880-1910*, Chapel Hill, NC: Duke University Press.

Guba, E.G., and Lincoln, Y.S. (1981) "Epistemological and methodological bases of natural inquiry," *Educational Communication & Technology Journal*, vol. 30, no. 4, pp. 233–252.

Hacking, I. (1999) *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Hadden, S.E. (2003) *Slave Patrols: Law and Violence in Virginia and the Carolinas*, Cambridge, MA: Harvard University Press.

Hall, S.P. (2016) 'Between Rage and a Hard Place: A Cautionary Tale of Colin Ferguson, Racial Politics, and Caribbean American Mental Health,' In, Short, E.L. and Wilton, L. (eds.) *Talking About Structural Inequalities in Everyday Life: New Politics of Race in Groups, Organizations, and Social Systems*, Charlotte, NC: Information Age Publishing, pp. 3-24.

Hanna, A., Denton, E., Smart, A., Smith-Loud, J. (2020) "Towards a Critical Race Methodology in Algorithmic Fairness." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 501-512.

Hannon, L. (2002) "Criminal Opportunity Theory and the Relationship between Poverty and Property Crime," *Sociological spectrum*, vol. 22, no. 3, pp. 363-381.

Harding, D.J., Morenoff, J.D. and Herbert, C.W. (2013) "Home Is Hard to Find: Neighborhoods, Institutions, and the Residential Trajectories of Returning Prisoners," *The Annals of the American Academy of Political and Social Science*, vol. 647, no. 1, pp. 214-236.

Hardt, M., Price, E., and Srebro, N. (2016) "Equality of Opportunity in Supervised Learning," Preprint, available at: <https://arxiv.org/abs/1610.02413v1>.

Harper, P.B., McClintock, A., Esteban Muñoz, J., Rosen, T. (1997) "Queer Transexions of Race, Nation, and Gender: An Introduction," *Social Text*, no. 52/53, pp. 1-4.

Harris, C.I. (1993) "Whiteness as Property," *Harvard Law Review*, vol. 106, no. 8, pp. 1707-1719.

Harris, C.I. (2020) "Reflections on 'Whiteness as Property,'" *Harvard Law Review Forum*, vol. 134, no. 1.

Hartney, M.T. and Flavin, P. (2014) "The Political Foundations of the Black–White Education Achievement Gap," *American politics research*, vol. 42, no. 1, pp. 3-33.

Hernandez, J. (2009) "Redlining Revisited: Mortgage Lending Patters in Sacramento 1930-2004," *International Journal of Urban and Regional Research*, vol. 32, no. 2, pp. 291-313.

Hertweck, C., Heitz, C., and Loi, M. (2021) "On the Moral Justification of Statistical Parity," In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 747-757.

Hillier, A.E. (2003) "Spatial Analysis of Historical Redlining: A Methodological Exploration," *Journal of Housing research*, vol. 14, no. 1, pp. 137-167.

Hofer, B.K. and Pintrich, P.R. (1997) "The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning," *Review of Educational Research*, vol. 67, pp. 88-140.

Holbrook, A.L., Krosnick, J.A., Moore, D., Tourangeau, R. (2007) "Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes" *Public opinion quarterly*, vol. 71, no. 3, pp. 325-348.

Huebner, B.M., DeJong, C. and Cobbina, J. (2010) "Women Coming Home: Long-Term Patterns of Recidivism," *Justice quarterly*, vol. 27, no. 2, pp. 225-254.

Huff, C.R. and Barrows, J. (2015) "Documenting gang activity: intelligence databases," In Decker, S.H. and Pyrooz, D.C. (eds.) *The Handbook of Gangs*, Chichester, West Sussex: John Wiley & Sons, pp. 59-77.

Ingold, D. and Soper, S. 'Amazon Doesn't Consider the Race of Its Customers. Should it?' *Bloomberg* April 21, 2016. Available at: <https://www.bloomberg.com/graphics/2016-amazon-same-day/>.

Jackson, E. and Mendoza, C. (2020) "Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not," *Harvard Data Science Review*, vol. 2, no. 1.

Johndrow, J.E. and Lum, K. (2019) "An algorithm for removing sensitive information: application to race-independent recidivism prediction," *The Annals for Applied Statistics*, vol. 13, no. 1, pp. 189-220.

Johnson, A.M. (1991) "The New Voice of Color," *The Yale Law Journal*, vol. 100, pp. 2007-2063.

Jung J., Concannon C., Shrof R., Goel S., Goldstein D.G. (2017) "Simple rules for complex decisions." Preprint, available at: <https://arxiv.org/abs/1702.04690>.

Kallus N., Zhou A. (2018) "Residual unfairness in fair machine learning from prejudiced data," arXiv preprint, available at: <https://arxiv.org/abs/1806.02887>.

Kalton, G. and Schuman, H. (1982) "The Effect of the Question on Survey Responses: A Review," *Journal of the Royal Statistical Society*, vol. 145, no. 1, pp. 42-73.

Kassel, P. (1998) "The gang crackdown in Massachusetts prisons: Arbitrary and harsh treatment can only make matters worse," *New England Journal on Criminal and Civil Confinement*, vol. 24, pp. 37-63.

Kennedy, R. (1989) "Racial Critiques of Legal Academia," *Harvard Law Review*, vol. 102, no. 8, pp. 1745-1819.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B. (2017) "Avoiding Discrimination through Causal

Reasoning,” In: *31st Conference on Neural Information Processing Systems (NIPS)*.

Kim, W., Choi, B.J., Hong, E.K., Kim, S.K., Lee, D. (2003) “A Taxonomy of Dirty Data,” *Data Mining and Knowledge Discovery*, vol. 7, pp. 81-99.

Kirk, D.S., Barnes, G.C., Hyatt, J.M., Kearley, B.W. (2018) “The impact of residential change and housing stability on recidivism: pilot results from the Maryland Opportunities through Vouchers Experiment (MOVE),” *Journal of experimental criminology*, vol. 14, no. 2, pp. 213-226.

Klare B.F., Burge M.J., Klontz J.C., Vorder Bruegge R.W., Jain A.K. (2012) “Face Recognition Performance: Role of Demographic Information,” in *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789-1801.

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z. (2015) “Prediction Policy Problems,” *American Economic Review: Papers and Proceedings*, vol. 105, no. 5, pp. 491-495.

Kleinberg, J., Mullainathan, S., Raghavan, M. (2016) “Inherent Trade-Offs in the Fair Determination of Risk Scores,” Preprint, available at: <https://arxiv.org/abs/1609.05807>.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2017) ‘Human Decisions and Machine Predictions,’ *National Bureau of Economic Research*, Working Paper 23180.

Kleinberg, J. and Raghavan, M. (2018) “Selection Problems in the Presence of Implicit Bias,” Preprint, available at: <https://arxiv.org/abs/1801.03533>.

Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A. (2018a) "Algorithmic Fairness," *AEA Papers and Proceedings*, vol. 108, pp. 22-27.

Kleinberg, J., Ludwig, J., Mullainathan, S., Sunstein, C.R. (2018b) "Discrimination in the Age of Algorithms," *Journal of Legal Analysis*, vol. 10, no. 2018, pp. 113-174.

Kleinberg, J. and Raghavan, M. (2021) 'Algorithmic Monoculture and Social Welfare' Preprint, available at: <https://arxiv.org/abs/2101.05853>.

Kleinman, P.H. and Lukoff, I.F. (1981) "Official Crime Data: Lag in Recording Time as a Threat to Validity," *Criminology* vol. 19, no. 3.

Krippendorff, K. (2004) *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: SAGE.

Krishnan, N.N., Torkildson, E., Mandayam, N., Raychaudhuri, D., Rantala, E.H., Doppler, K. (2018) "Optimizing Throughput Performance in Distributed MIMO Wi-Fi Networks using Deep Reinforcement Learning", Preprint, available at: <https://arxiv.org/abs/1812.06885>.

Kumar, C. (2020) "The Automated Tipster: How Implicit Bias Turns Suspicion Algorithms into BBQ Beckys," *Federal communications law journal*, vol. 72, no. 1, pp. 97.

Ladson-Billings, G. (2013) 'Critical Race Theory – What it is not!,' In Lynn, M. and Dixson, A.D. (eds.) *Handbook of Critical Race Theory in Education*, New York: Routledge, pp. 34-47.

Langan, P. (1991) 'Race of Prisoners Admitted to State and Federal Institutions, 1926-86,' *United States Department of Justice, Bureau of Justice Statistics*.

Larson, J., Mattu, S., Kirchner, L., Angwin, J. (2016) "How we analyzed the COMPAS recidivism algorithm" *ProPublica*, May 23, 2016.

Leiter, W.M. and Leiter, S. (2011) *Affirmative action in antidiscrimination law and policy: an overview and synthesis*, 2nd edn, State University of New York Press, Albany.

Levendowski, A. (2018) "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem", *Washington Law Review*, vol. 93, no. 2, pp. 579-630.

Levitt, S.D. (1998) "The Relationship Between Crime Reporting and Police: Implications for the Use of Uniform Crime Reports," *Journal of Quantitative Criminology*, vol. 14, no. 1, pp. 61-81.

Lichtenstein, A. (1996) *Twice the Work of Free Labor: The Political Economy of Convict Labor in the New South*, New York: Verso.

Lipton, Z.C., Chouldechova, A., McAuley, J. (2018) "Does mitigating ML's impact disparity require treatment disparity?" In: *32nd Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada.

Liu L.T., Dean S., Rolf E., Simchowicz M., Hardt M. (2018) "Delayed impact of fair machine learning", Preprint, available at: <https://arxiv.org/abs/1803.04383v1>.

Long, L.J. (2018) *Perpetual Suspects: A Critical Race Theory of Black and Mixed-Race Experiences of Policing*, Palgrave Macmillan, Cham.

Lowry, S. & Macpherson, G. (1988) "A blot on the profession", *British Medical Journal (Clinical research ed.)*, vol. 296, no. 6623, pp. 657-658.

Lugo, A. (2021) 'Why Florida's Ban On Critical Race Theory Won't Affect Alachua County Public Schools,' *WUFT News*, (June 30, 2021), available at: <https://www.wuft.org/news/2021/06/30/why-floridas-ban-on-critical-race-theory-wont-affect-alachua-county-public-school/>.

- Lum, K. and Isaac, W. (2016) "To predict and serve?" *Significance*, vol. 13, no. 5, pp. 14-19.
- Lynn, M. and Dixson, A.D. (2013) *Handbook of Critical Race Theory in Education*, New York: Routledge.
- Malouf, E.T., Schaefer, K.E., Witt, E.A., Moore, K.E., Stuewig, J., Tangney, J.P. (2014) "The Brief Self-Control Scale Predicts Jail Inmates' Recidivism, Substance Dependence, and Post-Release Adjustment," *Personality & social psychology bulletin*, vol. 40, no. 3, pp. 334-347.
- Mancini, M.J. (1996) *One Dies, Get Another: Convict Leasing in the American South, 1866-1928*, Columbia, SC: University of South Carolina.
- Markham, A.N., Tiidenberg, K., and Herman, A. (2018) "Ethics as Methods: Doing Ethics in the Era of Big Data Research – Introduction," *Social Media & Society*, vol. 4, no. 3.
- Marvel, T.B. and Moody, C.E. (1996) "Specification problems, police levels, and crime rates," *Criminology*, vol. 34, pp. 609-646.
- Massoglia, M., Firebaugh, G. and Warner, C. (2013) "Racial Variation in the Effect of Incarceration on Neighborhood Attainment," *American sociological review*, vol. 78, no. 1, pp. 142-165.
- Maxfield, M.G., Weiler, B.L. and Widom, C.S. (2000) "Comparing Self-Reports and Official Records of Arrests," *Journal of quantitative criminology*, vol. 16, no. 1, pp. 87-110.
- McClintock, N. (2011) 'From Industrial Garden to Food Desert,' In Alkon, A.H. and Agyeman, J. (eds.) *Cultivating Food Justice: Race, Class, and Sustainability*, Cambridge, MA: MIT Press, pp. 89-120.

McCord, J. (1980) 'Patterns of deviance,' In Wells, S. B., Crandall, R., Roff, M., Strauss, J. S., Pollin, W. (eds.), *Human Functioning in Longitudinal Perspective*, Baltimore, MD: Williams and Wilkins.

McIntosh, P. (1988) 'White Privilege and Male Privilege: A Personal Account of Coming to See Correspondences Through Work in Women's Studies,' In, McIntosh, P. (ed.) *On Privilege, Fraudulence, and Teaching as Learning: Selected Essays (1981-2019)*, New York: Routledge.

Meltzer, M. (1964-1967) *In Their Own Words: A History of the American Negro*, New York: Crowell, 3 vols.

Metz, C. (2019) 'We Teach A.I. Systems Everything, Including Our Biases,' *New York Times*, Nov 11, available at: <https://www.nytimes.com/2019/11/11/technology/artificial-intelligence-bias.html?action=click&module=RelatedLinks&pgtype=Article>.

Miller, S.L. and Burack, C. (2008) "A Critique of Gottfredson and Hirschi's General Theory of Crime," *Women & Criminal Justice*, vol. 4, no. 2, pp. 115-134.

Miron, M., Tolan, S., Gómez, E. & Castillo, C. (2021) "Evaluating causes of algorithmic bias in juvenile criminal recidivism", *Artificial Intelligence and Law*, vol. 29, no. 2, pp. 111-147.

Moore, L.D. and Elkavich, A. (2008) "Who's using and who's doing time: incarceration, the war on drugs, and public health," *American journal of public health*, vol. 98, no. 9, pp. S176-S180.

Morgan, E.S. (1975) *American Slavery, American Freedom: The Ordeal of Virginia*, New York: Norton.

Moses, M.S. (2016) *Living with moral disagreement: the enduring controversy about affirmative action*, The University of Chicago Press, Chicago.

Muhammad, K.G. (2010) *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America*, Cambridge, MA: Harvard University Press.

Nagata, D.K., Kim, J.H.J., and Wu, K. (2019) "The Japanese American Wartime Incarceration: Examining the Scope of Racial Trauma." *American Psychologist*, vol. 74, no. 1, pp. 36-48.

Nagin, D. and Paternoster, R. (2000) "Population Heterogeneity and State Dependence: State of the Evidence and Directions for Future Research," *Journal of Quantitative Criminology*, vol. 16, no. 2, pp. 117-44.

Narayanan, A. (2018) "21 definitions of fairness and their politics," In: Conference on Fairness, Accountability and Transparency (FAT*), 23-24 February, New York.

Nardone, A.L., Casey, J.A., Rudolph, K.A., Karasek, D., Mujahid, M., Morello-Frosch, R. (2020) "Associations between historical redlining and birth outcomes from 2006 through 2015 in California," *PLoS ONE*, vol. 15, no. 8.

National Archives (1787) 'The Constitution of the United States: A Transcription,' *National Archives*, available at: <https://archives.org/founding-docs/constitution-transcript>.

National Archives (1865) 'The Constitution: Amendments 11-27,' *National Archives*, available at: <https://archives.org/founding-docs/amendments-11-27/#toc-amendment-xiii>.

National Archives (1868) 'The Constitution: Amendments 11-27,' *National Archives*, available at: <https://archives.org/founding-docs/amendments-11-27/#toc-amendment-xiv>.

Neblo, M.A. (2009) "Meaning and Measurement: Reorienting the Race Politics Debate," *Political Research Quarterly*, vol. 62, no. 3, pp. 474-484.

Ngo, F.T., Paternoster, R., Curran, J. and MacKenzie, D.L. (2011) "Role-Taking and Recidivism: A Test of Differential Social Control Theory," *Justice quarterly*, vol. 28, no. 5, pp. 667-697.

Nguyen, A., Yosinski, J., & Clune, J. (2015) "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", Preprint, available at: <https://arxiv.org/abs/1412.1897>.

Nilsson, N.J. (1998) *Introduction to Machine Learning*. Stanford: Stanford University.

Northpointe (2012) "COMPAS Risk & Need Assessment System: Selection Questions Posed by Inquiring Agencies," available at: https://northpointeinc.com/files/downloads/FAQ_Document.pdf.

Novak, N.L., Lira, N., O'Connor, K.E., Harlow, S.D., Kardia, S.L.R., Stern, A.M. (2018) "Disproportionate Sterilization of Latinos Under California's Eugenic Sterilization Program, 1920-1945," *American Journal of Public Health*, vol. 108, no. 5, pp. 611-613.

Office of Economic & Demographic Research (2020) 'Criminal Justice Profile – Broward County,' *Office of Economic & Demographic Research*, available at: <https://edr.state.fl.us/content/area-profiles/criminal-justice-county/broward.pdf>.

Office of Economic & Demographic Research (2021) 'Broward County,' *Office of Economic & Demographic Research*, available at: <https://edr.state.fl.us/content/area-profiles/broward.pdf>.

Ogbonnaya-Ogburu, I.F., Smith, A.D.R., To, A., Toyama, K. (2020) "Critical Race Theory for HCI," In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 1-16.

O'Neil, C. (2016) *Weapons of Math Destruction*. New York: Crown.

Onyeador, I.N., Hudson, S.T.J., Lewis, N.A. (2021) "Moving Beyond Implicit Bias Training: Policy Insights for Increasing Organizational Diversity," *Policy Insights from the Behavioral and Brain Sciences*, vol. 8, no. 1, pp. 19-26.

Perkinson, R. (2010) *Texas Tough: The Rise of America's Prison Empire*, New York: Henry Holt and Company.

Pezzella, F.S., Fetzer, M.D., and Keller, T. (2019) "The Dark Figure of Hate Crime Underreporting," *American Behavioral Scientist*, pp. 1-24.

Piquero, A.R. (2008) "Disproportionate Minority Contact," *The Future of Children Journal*, vol. 18, pp. 59-79.

Piquero, A.R., Farrington, D.P., and Blumstein, A. (2003) "The criminal career paradigm," *Crime and Justice*, vol. 30, pp. 359–506.

Porter, T. (1996) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton: Princeton University Press.

Putnins, A.L. (2010) "An exploratory study of young offenders' self-reported reasons for offending," *The journal of forensic psychiatry & psychology*, vol. 21, no. 6, pp. 950-965.

Pyrooz, D.C., Clark, K.J., Tostlebe, J.J., Decker, S.H., Orrick, E. (2021) "Gang Affiliation and Prisoner Reentry: Discrete-Time Variation in

Recidivism by Current, Former, and Non-Gang Status,” *The Journal of Research in Crime and Delinquency*, vol. 58, no. 2, pp. 192-234.

Pyrooz, D.C., Decker, S.H., and Owens, E. (2020) “Do Prison Administrative and Survey Data Sources Tell the Same Story? A Multi-Trait, Multi-Method Examination with Application to Gangs,” *Crime & Delinquency*, vol. 66, no. 5, pp. 627-662.

Rabin, M. (1993) “Incorporating fairness into game theory and economics,” *The American Economic Review*, vol. 83, pp. 1281-1302.

Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, R., Mullainathan, S., Kleinberg, J. (2018) “Direct Uncertainty Prediction for Medical Second Opinions”, Preprint, available at: <https://arxiv.org/abs/1807.01771>.

Rawls, J. (2001) *Justice a Fairness, A Restatement*, Cambridge: Belknap Press.

Reardon, S.F. and Owens, A. (2014) “60 Years After “Brown”: Trends and Consequences of School Segregation,” *Annual review of sociology*, vol. 40, pp. 199-218.

Reisman, D., Schultz, J., Crawford, K. & Whittaker, M. (2018) ‘Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability’, available at: <https://ainowinstitute.org/aiareport2018.pdf> (Accessed: 15 January 2021)

Richardson, R., Schultz, J.M., Crawford, K. (2019) “Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice,” *New York University Law Review Online*, pp. 192-233.

Rickards, G., Magee, C. and Artino, J., Anthony R (2012) "You Can't Fix by Analysis What You've Spoiled by Design: Developing Survey Instruments and Collecting Validity Evidence," *Journal of graduate medical education*, vol. 4, no. 4, pp. 407-410.

Rodríguez-Ortega, J.A., Florez-Ruiz, J.F., Alvarado Suárez, Y.M., Alba Álvarez, G.H. (2020) "La dificultad analítica del rezago temporal en la denuncia y su relevancia en el análisis de los índices de criminalidad en Colombia (2005-2018)," *Revista Criminalidad*, vol. 62, no. 3.

Roemer, J.E. (2000) *Equality of Opportunity*, Cambridge: Harvard University Press.

Rosenfeld, R. (2007) "Transfer the Uniform Crime Reporting Program from the FBI to the Bureau of Justice Statistics," *Criminology & Public Policy*, vol. 6, no. 4, pp. 825-833.

Roth, W.D. (2016) "The multiple dimensions of race," *Ethnic and Racial Studies*, vol. 39, no. 8, pp. 1310-1338.

Roth, W.D. (2017) "Methodological pitfalls of measuring race: international comparisons and repurposing of statistical categories," *Ethnic and Racial Studies*, vol. 40, no. 13, pp. 2347-2353.

Rudin, C., Wang, C., and Coker, B. (2020) "The Age of Secrecy and Unfairness in Recidivism Prediction," *Harvard Data Science Review*, vol. 2, no. 1.

Rustemeyer, R. (1992) *Praktisch-methodische Schritte der Inhaltsanalyse*. Münster: Aschendorff.

Sacco, V.F., Wilcox, P., Land, K.C., Hunt, S.A. (2004) "Criminal Circumstance: A Dynamic Multicontextual Criminal Opportunity Theory," *Canadian journal of sociology*, vol. 29, no. 1, pp. 160.

Sampson, R.J., and Laub, J.H. (2005) "A life-course view of the development of crime," *Annals of the American Academy of Political and Social Science*, vol. 602, pp. 12–45.

Saravamakumar, K.K., 2021, "The Impossibility Theorem of Machine Fairness: A Causal Perspective," Preprint, available at: <https://arxiv.org/pdf/2007.06024.pdf>.

Sareen, S., Saltelli, A. and Rommetveit, K. (2020) "Ethics of quantification: illumination, obfuscation and performative legitimation," *Palgrave Communications*, vol. 6, no. 20.

Schaefer, R.T. (2004) *Racial and Ethnic Groups*, Upper Saddle River, NJ: Pearson-Prentice Hall.

Schaeffer, N.C. and Dykema, J. (2011) "Questions for Surveys: Current Trends and Future Directions," *Public Opinion Quarterly*, vol. 75, no. 5, pp. 909-961.

Schalev-Schwartz, S. and Ben-David, S. (2014) *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambridge University Press.

Schlesinger, A., Edwards, W.K., Grinter, R.E. (2017) "Intersectional HCI: Engaging identity through gender, race, and class," In *Conference on Human Factors in Computing Systems – Proceedings*, vol. 2017-May. ACM Press, New York, New York, USA.

Schreier, M. (2012) *Qualitative Content Analysis in Practice*. London: SAGE.

Schreier, M. (2014) 'Qualitative Content Analysis' in Flick, W. (ed) *The SAGE Handbook of Qualitative Data Analysis*. London: SAGE, pp.170-183.

Schulz, S. (2005) *Beyond Self-Control: Analysis and Critique of Gottfredson and Hirschi's General Theory of Crime (1990)*, Berlin: Duncker & Humboldt.

Scott, D. (2020) "Regional differences in gang member identification methods among law enforcement jurisdictions in the United States" *Policing: An International Journal*, vol. 43, no. 5, pp. 723-740.

Seidman, D. and Couzens, M. (1974) "Getting the crime rate down: Political pressure and crime reporting," *Law & Society Review*, vol. 8, pp. 457-494.

Sells, D., Curtis, A., Abdur-Raheem, J., Klimczak, M., Barber, C., Meaden, C., Hasson, J., Fallon, P., Emigh-Guy, M. (2020) "Peer-Mentored Community Reentry Reduces Recidivism," *Criminal justice and behavior*, vol. 47, no. 4, pp. 437-456.

Shapiro, A. (2017) 'Reform predictive policing,' *Nature*, vol. 541, pp. 458-460.

Shapiro, C.J., Smith, B.H., Malone, P.S. and Collaro, A.L. (2010) "Natural Experiment in Deviant Peer Exposure and Youth Recidivism," *Journal of clinical child and adolescent psychology*, vol. 39, no. 2, pp. 242-251.

Short, E.L. and Wilton, L. (eds.) (2016) *Talking About Structural Inequalities in Everyday Life: New Politics of Race in Groups, Organizations, and Social Systems*, Charlotte, NC: Information Age Publishing.

Skardhamar, T., Savolainen, J., Aase, K.N., Lyngstad, T.H. (2015) "Does Marriage Reduce Crime?," *Crime and Justice*, vol. 44, no. 1, pp. 385-446.

Skeem, J.L and Lowenkamp, C.T. (2016) "Risk, Race and Recidivism: Predictive Bias and Disparate Impact," *Criminology*, vol. 54, no. 4, pp. 680-712.

Small, D. (2001) "The War on Drugs Is a War on Racial Justice," *Social research*, vol. 68, no. 3, pp. 896-903.

Smith, C.S. (2019) 'Dealing With Bias in Artificial Intelligence,' *The New York Times*, Nov 19, available at: <https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html>.

Smith, K. (2014) "Fifty-six percent success is still a failing grade: reducing recidivism and ensuring due process rights in drug courts," *University of La Verne law review*, vol. 35, no. 2, pp. 315.

Spjeldnes, S., Yamatani, H. and McGowan Davis, M. (2015) "Child Support Conviction and Recidivism: A Statistical Interaction Pattern by Race," *Journal of evidence-informed social work*, vol. 12, no. 6, pp. 628.

Squires, G.D. (2016) "Racial Profiling, Insurance Style: Insurance Redlining and the Uneven Development of Metropolitan Areas," *Journal of Urban Affairs*, vol. 25, pp. 391-410.

Stage, F.K. (2007) "Answering critical questions using quantitative data," *New Directions for Institutional Research*, vol. 2007, no. 133, pp. 5-16.

Stein, D.M., Deberard, S., and Homan, K. (2013) "Predicting success and failure in juvenile drug treatment court: A meta-analytic review," *Journal of substance abuse treatment*, vol. 44, no. 2, pp. 159-168.

Stoll, L.C. and Klein, M. (2019) "'Not in my Backyard': How Abstract Liberalism and Colorblind Diversity Undermines Racial Justice,' In Embrick, D.G., Collins, S.M. and Dodson, M.S. (eds.) *Challenging the*

Status Quo: Diversity, Democracy, and Equality in the 21st Century, Leiden: Brill, pp. 217-240.

Sue, D.W. (2010) *Microaggressions in everyday life: Race, gender, and sexual orientation*, Hoboken: John Wiley & Sons.

Sullivan, G.M. and Artino, J., Anthony R (2017) "How to Create a Bad Survey Instrument," i, vol. 9, no. 4, pp. 411-415.

Sung, H. (2011) "From Diversion to Reentry: Recidivism Risks Among Graduates of an Alternative to Incarceration Program," *Criminal justice policy review*, vol. 22, no. 2, pp. 219-234.

Tapia, M. (2011) "Gang Membership and Race as Risk Factors for Juvenile Arrest," *The Journal of Research in Crime and Delinquency*, vol. 48, no. 3, pp. 364-395.

Taslitz, A.E. (2010) 'The Slave Power Undead: Criminal Justice Successes and Failures of the Thirteenth Amendment,' in Tsesis, A. (ed.) *The Promises of Liberty: The History and Contemporary Relevance of the Thirteenth Amendment*. New York: Columbia University Press, pp. 245-265.

Telles, E.E. and Lim, N. (1998) "Does it matter who answers the race question? Racial classification and income inequality in Brazil," *Demography*, vol. 35, no. 4, pp. 465-474.

The Sentencing Project (2020a) 'Criminal Justice Facts,' *The Sentencing Project*, available at: <https://sentencingproject.org/criminal-justice-facts/>.

The Sentencing Project (2020b) 'State-by-state Data,' *The Sentencing Project*, available at: <https://sentencingproject.org/the-facts/#map>.

Thompson, H.A. (2019) "The Racial History of Criminal Justice in America," *Du Bois Review*, vol. 16, no. 1, pp. 221-241.

Toch, H. (2007) "Sequestering Gang Members, Burning Witches, and Subverting Due Process," *Criminal Justice and Behavior*, vol. 34, no. 2, pp. 274-88.

Topol, E. (2019) "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56.

Treviño, A.J., Harris, M.A., and Wallace, D. (2008) "What's so critical about critical race theory?," *Contemporary Justice Review*, vol. 11, no. 1, pp. 7-10.

United States Census Bureau (2019) "Quick Facts: Broward County, Florida," available at: <https://www.census.gov/quickfacts/browardcountyflorida>.

Urban, C. and Miné, A. (2021) "A Review of Formal Methods applied to Machine Learning," Preprint, available at: <https://arxiv.org/abs/2104.02466>.

Verma, S. and Rubin, J. (2018) "Fairness Definitions Explained," In: *ACM/IEEE International Workshop on Software Fairness*.

Visher, C.A., Debus-Sherrill, S.A. and Yahner, J. (2011) "Employment after Prison: A Longitudinal Study of Former Prisoners," *Justice Quarterly*, vol. 28, no. 5, pp. 698-718.

Voigt, R., Camp, N.P., Prabhakaran, V., Hamilton, W.L., Hetey, R.C., Griffiths, C.M., Jurgens, D., Jurafsky, D., Eberhardt, J.L. (2017) "Language from police body camera footage shows racial disparities in officer respect," In *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 25, pp. 6521-6526.

Walk, D., Haviv, N., Hasisi, B., Weisburd, D. (2021) "The role of employment as a mediator in correctional education's impact on

recidivism: A quasi-experimental study of multiple programs," *Journal of criminal justice*, vol. 74.

Walters, G.D. (2005) "Recidivism in Released Lifestyle Change Program Participants," *Criminal justice and behavior*, vol. 32, no. 1, pp. 50-68.

Walters, G.D. (2014) "Relationships among Race, Education, Criminal Thinking, and Recidivism: Moderator and Mediator Effects," *Assessment*, vol. 21, no. 1, pp. 82-91.

Walters, G.D. (2017) *Modelling the Criminal Lifestyle: Theorizing at the Edge of Chaos*, Cham, Switzerland: Springer International Publishing.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H. (2016) "Deep Learning for Identifying Metastatic Breast Cancer", Preprint, available at: <https://arxiv.org/abs/1606.05718>.

Wang, X., Mears, D.P. and Bales, W.D. (2010) "Race-specific employment contexts and recidivism," *Criminology*, vol. 48, no. 4, pp. 1171-1211.

Washington, D.M. (2018) "Mass Incarceration: Overview of Its Effects on Black and Brown Individuals, with Policy Recommendations Using Family Engagement to Address Recidivism", *Columbia social work review*, vol. 9, pp. 34-44.

Watt, B., Howells, K. and Delfabbro, P. (2004) "Juvenile Recidivism: Criminal Propensity, Social Control and Social Learning Theories," *Psychiatry, psychology, and law*, vol. 11, no. 1, pp. 141-153.

Weinstock, M. and Cronin, M.A. (2003) "The Everyday Production of Knowledge: Individual Differences in Epistemological Understanding and Juror-Reasoning Skill," *Applied Cognitive Psychology*, vol. 17, pp. 161-181.

Western, B., Davis, J., Ganter, F., Smith, N. (2021) "The cumulative risk of jail incarceration," *Proceedings of the National Academy of Sciences*, vol. 118, no. 16, pp. 1.

Western, B. and Wildeman, C. (2009) "The Black Family and Mass Incarceration," *The Annals of the American Academy of Political and Social Science*, vol. 621, no. 1, pp. 221-242.

Wexler, R. 2017 "When a Computer Program Keeps You in Jail", *The New York Times*, 13 June. Available at: <https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html> (Accessed: 30 May 2021).

White, M.D., and Marsh, E.E. (2006) "Content Analysis: A Flexible Methodology," *Library Trends*, vol. 55, no. 1, pp. 22-45.

Williams, W.S., Short, E.L., and Ghiraj, D. (2016) 'Ethnoviolence as Structural Inequality: Media Representations of Black/African Descent Women,' In, Short, E.L. and Wilton, L. (eds.) *Talking About Structural Inequalities in Everyday Life: New Politics of Race in Groups, Organizations, and Social Systems*, Charlotte, NC: Information Age Publishing, pp. 121-138.

Williamson, G.L. (1992) "Education and Incarceration: An Examination of the Relationship Between Educational Achievement and Criminal Behavior," *Journal of correctional education*, vol. 43, no. 1, pp. 14-22.

Wilson, J.Q., and Herrnstein, R. (1985) *Crime and Human Nature*, New York: Simon and Schuster.

Wilson, T.B. (1965) *The Black Codes of the South*, Tuscaloosa, AL: University of Alabama Press.

Wilton, L. (2016) 'Race, Sexuality, AIDS, and Activism in Black Same-Gender Practicing Men's Communities in Post-Apartheid South Africa,'

In, Short, E.L. and Wilton, L. (eds.) *Talking About Structural Inequalities in Everyday Life: New Politics of Race in Groups, Organizations, and Social Systems*, Charlotte, NC: Information Age Publishing, pp. 165-184.

Yildirim, B.O. and Derksen, J.J.L. (2015) "Clarifying the heterogeneity in psychopathic samples: Towards a new continuum of primary and secondary psychopathy," *Aggression and violent behavior*, vol. 24, pp. 9-41.

Yosso, T.J., Smith, W.A., Ceja, M., Solorzano, D.G. (2009) "Critical race theory, racial microaggressions, and campus racial climate for Latina/o undergraduates," *Harvard Educational Review*, vol. 79, no. 4, pp. 659-690.

Young, H.P. (1995) *Equity*, Princeton: Princeton University Press.

Zafar M.B., Valera I., Gomez Rodriguez M., Gummadi K.P. (2017) "Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment." In: Proceedings of the 26th international conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 1171–1180.

Zemel R., Wu Y., Swersky K., Pitassi T., Dwork C. (2013) "Learning fair representations," In: International conference on machine learning, pp. 325–333.

Zenou, Y. and Boccoard, N. (2000) "Racial Discrimination and Redlining in Cities," *Journal of Urban Economics*, vol. 48, no. 2, pp. 260-285.

Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W. (2018) "Learning Gender-Neutral Word Embeddings." Preprint, available at: <https://arxiv.org/abs/1809.01496>.

Zinn, H. (2005) *A People's History of the United States*. New York: HarperCollins.

Žliobaite, I. (2017) "Measuring discrimination in algorithmic decision making," *Data Mining Knowledge Discovery*, vol. 31, pp. 1060-1089.

Appendix A

COMPAS Risk Assessment Questionnaire¹⁴

PERSON			
Name:	Offender #:		DOB:
R	Gender:	Marital Status:	Agency: DAI

ASSESSMENT INFORMATION:			
Case Identifier:	Scale Set: Wisconsin Core – Community Language	Screener:	Screening Date:

Current Charges

- Homicide Weapons Assault
 Arson
 Robbery Burglary Property/Larceny
 Fraud
 Drug Trafficking/Sales Drug Possession/Use DUI/OUIL
 Other Sex Offense with Force Sex Offense w/o Force

- Do any current offenses involve family violence?
 No Yes
- Which offense category represents the most serious current offense?
 Misdemeanor Non-violent Felony Violent Felony

¹⁴ (Angwin et al., 2016).

3. Was this person on probation or parole at the time of the current offense?
 Probation Parole Both Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
 No Yes
5. Number of pending chargers or holds?
 0 1 2 3 4+
6. Is the current top charge felony property or fraud?
 No Yes

Criminal History

Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
 8. How many prior juvenile felony offense arrests?
 0 1 2 3 4 5+
 9. How many prior juvenile violent felony offense arrests?
 0 1 2+
 10. How many prior commitments to a juvenile institution?
 0 1 2+
-

Note to Screener: The following Criminal History Summary questions require you to add up the total number of specific types of offenses in the person's criminal history. Count an offense type if it was among the charges or count within an arrest event. Exclude the current case for the following questions

11. How many times has this person been arrested for a felony property offense that included an element of violence?
 0 1 2 3 4 5+
12. How many prior murder/voluntary manslaughter offense arrests as an adult?
 0 1 2 3+

13. How many prior felony assault offense arrests (not murder, sex, or domestic violence) as an adult?
 0 1 2 3+
14. How many prior misdemeanor assault offense arrests (not sex or domestic violence) as an adult?
 0 1 2 3+
15. How many prior family violence offense arrests as an adult?
 0 1 2 3+
16. How many prior sex offense arrests (with force) as an adult?
 0 1 2 3+
17. How many prior weapons offense arrests as an adult?
 0 1 2 3+
18. How many prior drug trafficking/sales offense arrests as an adult?
 0 1 2 3+
19. How many prior drug possession/use offense arrests as an adult?
 0 1 2 3+
20. How many times has this person been sentenced to jail for 30 days or more?
 0 1 2 3 4 5+
21. How many times has this person been sentenced (new commitment) to state or federal prison?
 0 1 2 3 4 5+
22. How many times has this person been sentenced to probation as an adult?
 0 1 2 3 4 5+

Include the current case for the following question(s).

23. Has this person, while incarcerated in jail or prison, ever received serious or administrative disciplinary infractions for fighting/threatening other inmates or staff?
 No Yes
24. What was the age of this person when he or she was first arrested as an adult or juvenile (criminal arrests only)?

Non-Compliance

Include the current case for these questions.

25. How many times has this person violated his or her parole?
 0 1 2 3 4 5+
26. How many times has this person been returned to custody while on parole?
 0 1 2 3 4 5+
27. How many times has this person had a new charge/arrest while on probation?
 0 1 2 3 4 5+
28. How many times has this person's probation been violated or revoked?
 0 1 2 3 4 5+
29. How many times has this person failed to appear for a scheduled criminal court hearing?
 0 1 2 3 4 5+
30. How many times has the person been arrested/charged w/new crime while on pretrial release (includes current)?
 0 1 2 3+

Family Criminality

The next few questions are about the family or caretakers that mainly raised you when growing up.

31. Which of the following best describes who principally raised you?
 Both Natural Parents
 Natural Mother Only
 Natural Father Only
 Relative(s)
 Adoptive Parent(s)
 Foster Parent(s)
 Other arrangement
32. If you lived with both parents and they later separated, how old were you at the time?

Less than 5 5 to 10 11 to 14 15 or older Does Not Apply

33. Was your father (or father figure who principally raised you) ever arrested, that you know of?

No Yes

34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?

No Yes

35. Were your brothers or sisters ever arrested, that you know of?

No Yes

36. Was your wife/husband/partner ever arrested, that you know of?

No Yes

37. Did a parent or parent figure who raised you ever have a drug or alcohol problem?

No Yes

38. Was one of your parents (or parent figure who raised you) ever sent to jail or prison?

No Yes

Peers

Please think of your friends and the people you hung out with in the past few (3-6) months.

39. How many of your friends/acquaintances have ever been arrested?

None Few Half Most

40. How many of your friends/acquaintances served time in jail or prison?

None Few Half Most

41. How many of your friends/acquaintances are gang members?

None Few Half Most

42. How many of your friends/acquaintances are taking illegal drugs regularly (more than a couple times a month)?

None Few Half Most

43. Have you ever been a gang member?

No Yes

44. Are you now a gang member?

No Yes

Substance Abuse

What are your usual habits in using alcohol and drugs?

45. Do you think your current/past legal problems are partly because of alcohol or drugs?
 No Yes
46. Were you using alcohol under the influence when arrested for your current offense?
 No Yes
47. Were you using drugs or under the influence when arrested for your current offense?
 No Yes
48. Are you currently in formal treatment for alcohol or drugs such as counselling, outpatient, inpatient, residential?
 No Yes
49. Have you ever been in formal treatment for alcohol such as counselling, outpatient, inpatient, residential?
 No Yes
50. Have you ever been in formal treatment for drugs such as counselling, outpatient, inpatient, residential?
 No Yes
51. Do you think you would benefit from getting treatment for alcohol?
 No Yes
52. Do you think you would benefit from getting treatment for drugs?
 No Yes
53. Did you use heroin, cocaine, crack or methamphetamines as a juvenile?
 No Yes

Residence/Stability

54. How often do you have contact with your family (may be in person, phone, mail)?
 No family Never Less than once/month Once per week Daily

55. How often have you moved in the last twelve months?
 Never 1 2 3 4 5+
56. Do you have a regular living situation (an address where you usually stay and can be reached)?
 No Yes
57. How long have you been living at your current address?
 0-5 mo. 6-11 mo. 1-3 yrs. 4-5 yrs. 6+ yrs.
58. Is there a telephone at this residence (a cell phone is an appropriate alternative)?
 No Yes
59. Can you provide a verifiable residential address?
 No Yes
60. How long have you been living in that community or neighborhood?
 0-2 mo. 3-5 mo. 6-11 mo. 1+ yrs.
61. Do you live with family—natural parents, primary person who raised you, blood relative, spouse, children, or boy/girlfriend if living together for more than 1 year?
 No Yes
62. Do you live with friends?
 No Yes
63. Do you live alone?
 No Yes
64. Do you have an alias (do you sometimes call yourself by another name)?
 No Yes

Social Environment

Think of the neighborhood where you lived during the past few (3-6) months.

65. Is there much crime in your neighborhood?
 No Yes
66. Do some of your friends or family feel they must carry a weapon to protect themselves in your neighborhood?
 No Yes

67. In your neighborhood, have some of your friends or family been crime victims?
 No Yes
68. Do some of the people in your neighborhood feel they need to carry a weapon for protection?
 No Yes
69. Is it easy to get drugs in your neighborhood?
 No Yes
70. Are there gangs in your neighborhood?
 No Yes

Education

Think of your school experiences when you were growing up.

71. Did you complete your high school diploma or GED?
 No Yes
72. What was your final grade completed in school?
73. What were your usual grades in high school?
 A B C D E/F Did Not Attend
74. Were you ever suspended or expelled from school?
 No Yes
75. Did you fail or repeat a grade level?
 No Yes
76. How often did you have conflicts with teachers at school?
 Never Sometimes Often
77. How many times did you skip classes while in school?
 Never Sometimes Often
78. How strongly do you agree or disagree with the following: I always behaved myself in school?
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
79. How often did you get in fights while at school?
 Never Sometimes Often

Vocation (Work)

Please think of your past work experiences, job experiences, and financial situation.

80. Do you have a job?
 No Yes
81. Do you currently have a skill, trade or profession at which you usually find work?
 No Yes
82. Can you verify your employer or school (if attending)?
 No Yes
83. How much have you worked or been enrolled in school in the last 12 months?
 12 Months Full-time 12 Months Part-time 6+ Months Full-time 0 to 6 Months PT/FT
84. Have you ever been fired from a job?
 No Yes
85. About how many times have you been fired from a job?
86. Right now, do you feel you need more training in a new job or career skill?
 No Yes
87. Right now, if you were to get (or have) a good job how would you rate your chance of being successful?
 Good Fair Poor
88. How often do you have conflicts with friends/family over money?
 Often Sometimes Never
89. How hard is it for you to find a job ABOVE minimum wage compared to others?
 Easier Same Harder Much Harder
90. How often do you have barely enough money to get by?
 Often Sometimes Never
91. Has anyone accused you of not paying child support?
 No Yes
92. How often do you have trouble paying bills?
 Often Sometimes Never

93. Do you frequently get jobs that don't pay more than minimum wage?

Often Sometimes Never

94. How often do you worry about financial survival?

Often Sometimes Never

Leisure/Recreation

Thinking of your leisure time in the past few (3-6) months, how often did you have the following feelings?

95. How often did you feel bored?

Never Several times/mo Several times/wk Daily

96. How often did you feel you have nothing to do you in your spare time?

Never Several times/mo Several times/wk Daily

97. How much do you agree or disagree with the following – You feel unhappy at times?

Strongly Disagree Disagree Not Sure Agree Strongly Agree

98. Do you feel discouraged at times?

Strongly Disagree Disagree Not Sure Agree Strongly Agree

99. How much do you agree or disagree with the following – You are often restless and bored?

Strongly Disagree Disagree Not Sure Agree Strongly Agree

100. Do you often become bored with your usual activities?

No Yes Unsure

101. Do you feel that the things you do are boring or dull?

No Yes Unsure

102. Is it difficult for you to keep your mind on one thing for a long time?

No Yes Unsure

Social Isolation

Think of your social situation with friends, family, and other people in the past few (3-6) months. Did you have many friends or were you more of a loner? How much do you agree or disagree with these statements?

103. "I have friends who help me when I have troubles."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
104. "I feel lonely."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
105. "I have friends who enjoy doing things with me."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
106. "No one really knows me very well."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
107. "I feel very close to some of my friends."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
108. "I often feel left out of things."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
109. "I can find companionship when I want."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
110. "I have a best friend I can talk with about everything."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
111. "I have never felt sad about things in my life."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree

Criminal Personality

The next few statements are about what you are like as a person, what your thoughts are, and how other people see you. There are no 'right or wrong' answers. Just indicate how much you agree or disagree with each statement.

112. "I am seen by others as cold or unfeeling."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
113. "I always practice what I preach."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
114. "The trouble with getting close to people is that they start making demands on you."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
115. "I have the ability to "sweet talk" people to get what I want."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
116. "I have played sick to get out of something."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
117. "I'm really good at talking my way out of problems."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
118. "I have gotten involved in things I later wished I could have gotten out of."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
119. "I feel bad if I break a promise I have made to someone."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
120. "To get ahead in life you must always put yourself first."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree

Anger

121. "Some people see me as a violent person,"
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
122. "I get into trouble because I do things without thinking."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
123. "I almost never lose my temper."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
124. "If people make me angry or lose my temper, I can be dangerous."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
125. "I have never intensely disliked anyone."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
126. "I have a short temper and can get angry quickly."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree

Criminal Attitudes

The next statements are about your feelings and beliefs about various things. Again, there are no 'right or wrong' answers. Just indicate how much you agree or disagree with each statement.

127. "A hungry person has a right to steal."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
128. "When people get into trouble with the law it's because they have not chance to get a decent job."
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
129. "When people do minor offenses or use drugs they don't hurt anyone except themselves,"

- Strongly Disagree Disagree Not Sure Agree Strongly Agree
130. "If someone insults my friends, family or group they are asking for trouble."
- Strongly Disagree Disagree Not Sure Agree Strongly Agree
131. "When things are stolen from rich people they won't miss the stuff because insurance will cover the loss."
- Strongly Disagree Disagree Not Sure Agree Strongly Agree
132. "I have felt very angry at someone or at something."
- Strongly Disagree Disagree Not Sure Agree Strongly Agree
133. "Some people must be treated roughly or beaten up just to send them a clear message."
- Strongly Disagree Disagree Not Sure Agree Strongly Agree
134. "I won't hesitate to hit or threaten people if they have done something to hurt my friends or family."
- Strongly Disagree Disagree Not Sure Agree Strongly Agree
135. "The law doesn't help average people."
- Strongly Disagree Disagree Not Sure Agree Strongly Agree
136. "Many people get into trouble or use drugs because society has given them no education, jobs or future."
- Strongly Disagree Disagree Not Sure Agree Strongly Agree
137. "Some people just don't deserve any respect and should be treated like animals."
- Strongly Disagree Disagree Not Sure Agree Strongly Agree

Appendix B

IMSIS INDEPENDENT STUDY

Dissertation Archive Permission Form




**International Master in Security, Intelligence and Strategic Studies
2019/2021**

Dissertation Archive Permission Form

I give the University of Glasgow and Charles University permission to archive an e-copy of my Master dissertation in a publicly available folder and to use it for educational purposes in the future.

Student Name MARÍA PATRICIA BEJARANO CARBÓ

Student Number: 2487073B

Student Signature:  **Date:** 02/08/2021