Predictive Algorithms in Justice Systems and the Limits of Tech-Reformism

Pamela Ugwudike
University of Southampton, United Kingdom

Abstract

Data-driven digital technologies are playing a pivotal role in shaping the global landscape of criminal justice across several jurisdictions. Predictive algorithms, in particular, now inform decision making at almost all levels of the criminal justice process. As the algorithms continue to proliferate, a fast-growing multidisciplinary scholarship has emerged to challenge their logics and highlight their capacity to perpetuate historical biases. Drawing on insights distilled from critical algorithm studies and the digital sociology scholarship, this paper outlines the limits of prevailing tech-reformist remedies. The paper also builds on the interstices between the two scholarships to make a case for a broader structural framework for understanding the conduits of algorithmic bias.

Keywords

Algorithmic bias; algorithmic injustice; critical algorithm studies; design justice; digital criminology; digital sociology.

Please cite this article as:
Ugwudike P (2022) Predictive algorithms in justice systems and the limits of tech-reformism. *International Journal for Crime, Justice and Social Democracy.* 11(1): 85-99. https://doi.org/10.5204/ijcjsd.2189

Introduction

Digital technological innovations such as predictive algorithms are transforming the infrastructure of private and public sector services across several jurisdictions. This paper focuses on the predictive algorithms[1] that are increasingly applied in justice systems for crime risk prediction. The algorithms are mostly used to determine locational or recidivism risks in justice systems. Although the models, uses and modes of implementation vary, a common denominator that connects these technologies is that they are data-driven tools for predicting crime risks. The algorithms are increasingly informing high stakes decisions across several areas, from policing and sentencing to parole. They are also determining the intensity of probation and post-release supervision. Indeed, the technologies are proliferating across justice systems and could revolutionise the systems.

Nevertheless, studies have found that the algorithms can artificially inflate the risks associated with Black people (e.g., Angwin et al. 2016) and the areas in which they reside (Lum and Isaac 2016). In response, other studies have emerged to proffer remedies that can solve or at least mitigate such algorithmic bias[2] (Johndrow and Lum 2019; Skeem and Lowenkamp 2020). Yet, the competing principles inherent in this multidisciplinary body of work have received insufficient criminological attention. In particular, there are tensions between the tech-reformism inherent in some of the proffered remedies and broader structural transformations proposed by critical scholars. This paper utilises conceptual tools distilled from two scholarships—critical algorithm studies (CAS) and digital sociology (DS)—to address the paucity of insights and unravel the underpinning tensions. It makes an original contribution by drawing on the tools to, (1) outline the limits of prevailing tech-reformist remedies, and (2) demonstrate how poorly suited they are to the task of mitigating or remedying algorithmic bias.

Building on the interstices between the two scholarships, the paper makes a case for a structural framework for understanding the conduits of algorithmic bias. While this framework looks beyond the algorithms to address the wider structural roots of bias, tech-reformism focuses on the algorithms and seeks to improve their neutrality, objectivity, accuracy and other dimensions of technical efficiency. It is an approach that is rooted in what Benjamin (2019) describes as a soft determinism that views technology as neutral, unaffected by political, economic and social structures, and amenable to human intervention and governance. As such, it is a position that overlooks the interactive relationship between technology and society. Tech-reformism features mainly in the activities of organisations, institutions and researchers working on technical solutions. Examples of such solutions include debiasing datasets, performing algorithm audits, developing explainability techniques and establishing technical standards.

To achieve its objectives, this paper begins with an overview of how predictive algorithms in justice systems operate and some of the key challenges, focusing on predictive policing and risk assessment algorithms. It proceeds with a description of conceptual tools from the CAS and DS scholarships, which are useful for unravelling the broad structural roots and implications of the challenges. The paper then moves on to analyse the problem of algorithmic bias and the proffered tech-reformist remedies. Following this, the paper, inspired by the aforementioned conceptual tools, provides a structural framework for addressing such bias.

Predictive Algorithms in Justice Systems

Police services across Western and non-Western jurisdictions now deploy predictive algorithms to forecast crime risks. Examples include PredPol algorithm, now rebranded as Geolitica (PredPol 2021), which has been used in the United States (US) and the United Kingdom (UK), risk assessments via the Suspect Targeting Management Plan in Australia (Yeong 2020), and the GeoDASH Algorithmic Policing System in Canada (Kenyon 2020). The systems draw on inferences from patterns in collected data to forecast risks and inform police practice. They are varied systems, with substantive differences between locational (i.e., 'where to police?') and individualised (i.e., 'whom to police?') predictive policing models (Kaufmann et al. 2019). But a key feature of the algorithms is that they process large volumes of

administrative data, such as arrest records, and 'big data' (a by-product of increasing human interaction with data-driven technologies), to analyse crime patterns and forecast crime risks (Kauffman et al. 2018).

While predictive policing algorithms comprise several variants, with some focusing on locational crime risks, other commonly used predictive algorithms seek to determine individual recidivism risk. These are risk assessment algorithms that are applied to most individuals coming into contact with justice systems (see for example, Her Majesty's Inspectorate of Probation 2020). Risk assessment algorithms comprise statistical regression algorithms and, more recently, machine learning algorithms capable of detecting patterns in large-scale data to predict risk (Berk and Bleich 2013; Brennan and Oliver 2013). The algorithms are data-driven technologies in the sense that they are computerised systems with integrated algorithms trained on datasets, such as administrative data, for example, arrest and reconviction data, and formulated to compute risk of reconviction scores.

A fast-growing corpus of research has been highlighting several challenges associated with predictive algorithms applied in justice systems, outlining how the technologies can generate biased outcomes (Angwin et al. 2016; Ensign et al. 2018; Hao and Stray 2019; Lowder et al. 2019; Lum and Isaac 2016; Richardson, Schultz and Crawford 2019). In response, tech-reformism has arisen as a major approach to dealing with algorithmic biases while relatively limited attention has been paid to the broader structural approach proposed by this paper and elucidated later.

Some studies have, for example, found that risk of recidivism algorithms can over-predict risks in cases involving Black people (e.g., Angwin et al. 2016; Hao and Stray 2019; Lowder et al. 2019). Predictive policing algorithms can also expose Black communities in the US to over-policing (e.g., Ensign et al. 2018; Lum and Isaac 2016; Richardson, Schultz and Crawford 2019). Studies suggest that algorithmic bias can affect Indigenous populations in Australia (Allan et al. 2019; Shepherd et al. 2014) and Canada (Cardoso 2020). A reason for this is that predictor data based on Western values (such as those relating to family circumstances[3]) can be culturally insensitive and, as such, capable of over-predicting the risks posed by these populations (see also, Maurutto and Hannah-Moffatt 2006).

Biases can also stem from the reliance of predictive algorithms on data such as criminal history variables as proxies for crime risk. But the variables (e.g., arrest history) can be colour-coded since they derive from criminal justice data, which tend to show higher rates of negative outcomes (e.g., arrests) for ethnic minorities, in part reflecting racially-biased decision making (Hao and Stray 2019). Indeed, official statistics consistently reveal racial disparities in criminal justice outcomes across jurisdictions where the predictive algorithms are deployed. Examples include the UK (Ministry of Justice 2019), the US (Bureau of Justice Statistics 2018), Canada (Malakieh 2019) and Australia (Australian Bureau of Statistics 2018). Since biased decision making can partly explain the disparities, depicting criminal justice outcomes as risk predictors can disadvantage racial minorities. This is because predictive algorithms can perpetuate bias if they cannot mitigate the effects of racial bias. Indeed, as noted earlier, studies have found evidence of racial disparities in algorithmic predictions. The studies, which usefully highlight the technical problem of biased data, are currently fuelling the rise of technical remedies that focus primarily on 'fixing' the algorithms. But with conceptual tools from CAS and DS, this paper will go beyond the technical to unravel structural conduits of bias, understand the limits of technical remedies and provide a structural remedial approach.

Insights from Digital Sociology (DS) and Critical Algorithm Studies (CAS): Key Conceptual Tools

As already noted, this paper draws on conceptual tools to unpack tensions between the tech-reformism inherent in some of the remedies proffered for algorithmic bias and the broader structural transformations proposed by critical scholars. A concept emerging from DS and applied in this paper to unpack the tensions is the concept of digital capital. From CAS, the paper draws on the concepts of design justice, the digital racialisation of risk and algorithmic injustice. By connecting the two scholarships in its analysis of algorithmic bias and potential remedies, the paper also advances the extant criminological literature on

the prospects and challenges of the predictive algorithms currently proliferating across justice systems, from the UK and the US to Australia and Canada.

As I have argued elsewhere (Ugwudike 2020), a conceptual tool from DS that is useful for exploring the structural basis of artificial intelligence (AI) bias and the limits of narrow technical solutions is digital capital (van Dijk 2005). The concept alerts us to a key structural conduit of bias: the uneven distribution of digital resources with which predictive algorithms are created. Such inequality excludes affected populations from creational processes, conferring on others the power to inject their choices and preferences into algorithm design.

From a sociological perspective, digital capital refers to the ability to acquire resources required for creating and exploiting the full benefits of technologies such as predictive algorithms. Examples of these resources include the knowledge, skills or capital that confer the power and ability to create the technologies (van Dijk 2005). Therefore, in criminal justice contexts, the state and non-state creators of predictive algorithms are the ones equipped with digital capital. They have the power to infuse algorithms with their values, choices and preferences. Actors with limited digital capital lack such power. Consequently, they can exert no influence over the processes of creating technologies that can affect their lives. This power inequality means that their values, preferences and even their circumstances or the potential impact of the technologies on their lives may not be taken into account. This could potentially explain their status as people who, as studies now show, are often adversely affected by predictive algorithms (Angwin et al. 2016; Hao and Stray 2019; Lowder et al. 2019).

It is also worth noting that the predictive technologies are typically proprietary, and this insulates their algorithmic components from inspection and rebuttal (Kehl, Guo and Kessler 2017), masking subjective choices and decisions that can introduce bias and further empowering the creators by eliminating transparency and accountability imperatives. Further, algorithmic decision-making can become too complex and opaque to unravel. This paper will demonstrate that technical fixes that focus on algorithms themselves elide these power-laden dynamics that can sustain algorithmic bias (Ugwudike 2020).

Meanwhile, what the foregoing shows is that digital capital is a concept from DS that can illuminate structural conduits of algorithmic bias. The CAS scholarship, which is situated within the multidisciplinary field of science and technology studies, also known as science, technology and society studies, provides additional conceptual tools for understanding structural conduits of algorithmic bias and the limits of technical remedies. Design justice (Costanza-Chock 2020) is an example, and it refers to the need to democratise creational dynamics and even empower marginal groups to lead design processes. It is, therefore, a concept that emphasises the importance of reversing the unequal distribution of digital capital that excludes marginal groups from creational processes.

The CAS scholarship also draws attention to the digital racialisation of risk (Ugwudike 2020). This concept offers insights into a key implication of algorithmic bias, such as the over-prediction of risk in cases involving some minorities. An implication of such over-prediction is that it can entrench longstanding racial ideologies about supposed links between race, risk and crime. It can, therefore, expose affected groups to unwarranted profiling and penal intervention. In this way, biased systems can replicate existing structures of disadvantage and produce new forms of social control (see also, Benjamin 2019).

Alongside design justice and the digital racialisation of risk, algorithmic injustice is another concept from the CAS scholarship that also conceptualises structural implications of algorithmic bias. It refers to the ways in which data-driven technologies across several sectors can systematically reproduce systemic discrimination and broader structural inequalities that disadvantage already vulnerable and disempowered groups (Birhane 2021).

Together, the above concepts from CAS and DS are useful for understanding key structural bases and implications of algorithmic bias. The concepts will inform the structural approach to addressing such bias

(proposed by this paper). Unlike tech-reformism, the proposed approach, which will be elucidated later, is human-centric in that it does not prioritise technical imperatives over fair social outcomes and it pays attention to the unequal societal conditions in which algorithms are designed and deployed. Indeed, the conceptual tools outlined above highlight the nature of predictive algorithms as sociotechnical artifacts that impact on society just as society shapes and impacts them. Therefore, problems such as algorithmic bias are products of both technical and societal or structural dynamics.

Remedying Algorithmic Bias

Studies revealing the problem of algorithmic bias point to the need for remedial strategies (e.g., Angwin et al. 2016; Hao and Stray 2019; Lowder et al. 2019). Indeed, algorithmic bias in justice systems and beyond is a problem that is currently attracting significant attention nationally and internationally from academics, researchers, governments, civil society organisations and others (AI Now 2018; Angwin et al. 2016; Australian Government: Department of Industry, Science, Energy and Resources 2019; Ugwudike 2020; Benjamin 2019; Centre for Data Ethics and Innovation 2019, 2020; Government of Canada 2020; Hannah-Moffat 2019; Hao and Stray 2019; Lowder et al. 2019; The Law Society 2019; Lum and Isaac 2016; Meijer and Wessels 2019; Rovastos, Mittelsdat and Koene 2020). It is therefore not surprising that the field of AI ethics has emerged as a multidisciplinary scholarship concerned with remedying algorithmic bias (see generally, Raji et al. 2020).

Within this scholarship, two broad domains can be identified. One is a more prominent approach that responds mainly to technical conduits of bias such as data-related problems, while the second emphasises structural conditions that also foment bias. Although both can crosscut in the sense that some technical remedies may have broader structural objectives beyond fixing AI systems themselves, there are often clear points of departure between them. In the following sections, this paper unravels the two domains and delineates their differences. Ultimately, the paper makes the case that, as the aforementioned conceptual tools from the CAS and DS scholarships suggest, algorithmic bias is partly a structural problem that requires structural solutions. Technical remedies are useful only in so far as they take into account structural conduits of bias.

Bias in Commonly Used Predictive Algorithms

Before we analyse the tech-reformist solutions being proffered to address algorithmic bias, it is necessary to provide a summarisation of such bias and how it manifests. On this, a recent study published by ProPublica, which illustrates how bias can permeate algorithmic outputs, is quite insightful (see Angwin et al. 2016). The study focused on a generic risk of recidivism algorithm that is used in parts of the US and shares similar attributes (e.g., risk predictors and recidivism variables) with other generic algorithms such as the Level of Service Inventory–Revised, which is deployed in Australia (Gower et al. 2020), and some parts of the UK (Risk Management Authority 2019), the US (Lowder et al. 2019) and Canada (Bonta and Andrews 2017) (where it was created). The study found evidence of racial disparities in the form of higher rates of false positives for Black people (48%) and higher rates of false negatives for White people (28%) (Angwin et al. 2016). It follows that the algorithm over-predicted the recidivism rates of Black defendants and under-predicted that of White defendants, placing the Black defendants at higher risk of more punitive outcomes such as long prison sentences (see also Hao and Stray 2019). On the basis of this imbalance in error or misclassification rates, the study concluded that the algorithm did not produce fair outcomes; it was biased against Black people.

In response, the creators presented a validation study and argued that the results showed evidence of fairness (Dieterich et al. 2016). But they offered a different definition of fairness. Their metric was different from the definition used by Angwin et al. (2016), which emphasised 'equalised odds' or, in other words, an equal balance in error or misclassification rates (false positives and false negatives) across racial groups. In contrast, the creators defined fairness in technical terms, emphasising, inter alia, predictive parity in the sense that the tool predicts risk of recidivism equally well (at a similar level of accuracy) across different racial groups (Dietrich et al. 2016). Thus, technical fairness was emphasised over broader social justice,

highlighting the importance of the structural approach proposed by this paper, which focuses more on the wider societal impact of predictive algorithms, for example, unfair discrimination.

The creators also maintained that they used the same 'unbiased' rules to score all defendants, but the equalised odds problem occurred because the algorithm was using arrest data as a proxy for crime and was incapable of recognising structural issues such as the politics and nuances of criminal justice data, including the reality that some arrests and other criminal justice outcomes can correspond more to racial bias than actual crime rates (see also Hao and Stray 2019). In this case, the data presented Black defendants as having higher offending rates, which meant that their odds of receiving a false positive (and, therefore, being disadvantaged) were higher—although the algorithm attained predictive parity, which is a measure of technical accuracy.

These findings draw attention to how ostensibly neutral data can produce discriminatory outcomes. In the book *Race After Technology: Abolitionist Tools for the New Jim Code*, Benjamin (2019: 10) remarks that criminal justice and other widely used algorithms rely on flawed 'data that have been produced through histories of exclusion and discrimination'. Such data can foment the digital racialisation of risk by increasing the likelihood that Black people will attain high risk scores and be profiled as chronically criminogenic. Notwithstanding this problem, the algorithms that rely on such data appear to be scientific, 'value neutral' tools and credible sources of criteria for categorisations and governance.

Dietrich and colleagues' (2016) findings also appear to suggest that commonly used risk of recidivism algorithms can be biased against a group but considered 'fair' if fairness is defined in terms of technical accuracy rather than equity of outcomes for all. Therefore, a challenge for those genuinely interested in equitable algorithms is to determine how to reconcile the competing values and principles inherent in various definitions of fairness.

With the commonly used algorithms, a trade-off is currently being made between a definition that focuses on the technical issue of predictive parity and one that emphasises the broader structural aim of equalised odds. In justice systems, this trade-off is being resolved in favour of the former, meaning that some groups will continue to be disadvantaged by higher rates of false positives that expose them to worse criminal justice outcomes than others or, in other words, algorithmic injustice. Black people are likely to be disproportionately affected (Angwin et al. 2016). Therefore, the trade-off can foment the digital racialisation of risk (Ugwudike 2020) and algorithmic injustice (Birhane 2021); hence, the need for the structural framework recommended by this paper.

Unlike techno-reformism, the proposed framework draws attention to adverse social outcomes. Commenting on the adverse outcome documented by the ProPublica study, Humerick (2020: 234) notes in his analysis of algorithmic bias and potential remedies that 'to be "fair" to some defendants', commonly used predictive algorithms, such as the one analysed in that study, 'must be "unfair" to others'. Potentially, therefore, 'it is impossible to be both accurate and equal at the same time' (Humerick 2020: 234). Whether or not this situation can be considered 'fair' depends on what is prioritised by the creators and procurers: ostensibly equal treatment (via predictive accuracy) or unbiased/equitable treatment (via equalised odds). The former focuses on technical issues and can inspire tech-reformist remedies that pay greater attention to improving the accuracy of algorithms. In contrast, the latter prioritises broader social concerns, particularly equitable outcomes.

Here, we witness the importance of design justice principles (Costanza-Chock 2020), which hold that people most affected by algorithmic injustice should be empowered to contribute to key fairness decisions. As we have observed, when technical fairness metrics such as those based on predictive parity are selected, they can disadvantage Black people by over-predicting their risk. This can foment the aforementioned problems of algorithmic injustice and the digital racialisation of risk. But when the measure of fairness is equalised odds (equal classification errors/equal balance of error rates), it can benefit Black people by reducing their exposure to unfair discrimination via higher rates of false positives.

Humerick reinforces this with the comment that:

> the failure of the [Correctional Offender Management Profiling for Alternative Sanctions] COMPAS algorithm to equalize odds along racial lines led to 10% of [B]lack defendants (384 of the defendants in the sample) being unfairly disadvantaged in their risk assessment score. Thus, up to 10% of [B]lack defendants could be benefited—reclassified as Low or Medium Risk—by an algorithm operating under equalized odds. (2020: 231)

These debates about fairness also reveal that, in the absence of a legal framework, the digital capital (van Dijk 2015) to determine whether or not algorithms should rely on potentially biased administrative data and how fairness metrics should be established currently resides with the creators and the authorities that procure the tools, signalling a lack of design justice.

Data-driven bias also plagues predictive policing algorithms, and studies have highlighted the problem of racial bias, again triggering quite limited tech-reformist remedies. An example is Lum and Isaac's (2016) simulation study of the decision-making processes and outcomes of an algorithm that has been used by police services in the US and the UK. The study combined a synthetic sample of Oakland city residents and data from a national drug use and health survey. It found that the spread of drug use was fairly even across the city. But, when the study analysed police recorded drug crimes, it found that most of the drug crimes recorded by Oakland Police Department were unrepresentative and located in the areas populated mainly by 'non-white and low-income populations' (2016: 17).

Lum and Isaac (2016) also investigated the effects of using such police data for algorithmic predictions. They studied predictions produced by the PredPol algorithm when it used Oakland Police Department's drug crime data to forecast crime risks. The study found that the locations designated as high crime risk areas by the algorithm were the minority ethnic and low income areas that already featured the most in the police data on which the algorithm relied for predictions.

Reflecting on the findings, Lum and Isaac (2016: 18) noted that 'allowing a predictive policing algorithm to allocate police resources would result in the disproportionate policing of low-income communities and communities of colour'. Ensign et al. (2018) also arrived at similar conclusions about the algorithmic feedback loops they uncovered in their analysis of PredPol, which relied on policing data from Lum and Isaac's (2016) study (see also, Richardson, Schultz and Crawford 2019). As we shall see, findings such as these, which highlight data-related bias, along with concerns about the opacity of complex algorithmic data processing, have prompted the development of remedial strategies, most of which are tech-reformist and, as such, limited.

Apart from predictive algorithms, other data-driven technologies deployed by criminal justice systems have been found to be plagued by data-driven racial bias and have also prompted technical remedies. Facial recognition technologies, for example, have misidentified Black people, wrongly ascribing criminal labels to them, resulting in wrongful convictions (e.g., General and Sarlin 2021). These adverse outcomes are perhaps unsurprising given the findings of studies which show that facial recognition technologies can misidentify Black people mainly because this subgroup is under-represented in datasets used to train the technologies (Buolamwini and Gebru 2018).

Bias Beyond Justice Systems

Studies that have analysed the racial dynamics of data-driven decision-making technologies beyond justice systems have similarly found that the technologies can reproduce societal biases, of which racial bias represents an example. Evidence of such bias has been found in various algorithms, including those deployed by healthcare services, social welfare services, search engines, employers and finance companies (Ajunwa et al. 2016; Eubanks 2017; O'Neill 2016; Price 2019). Noble (2018), for instance, notes that search engines relying on biased data can produce results that demonise Black faces and Black lives. These

scholars and others contend that the algorithms are perpetuating existing societal biases and inequities (see also, Ugwudike 2020).

Together, these and other studies point to structural issues (e.g., racial discrimination) that disadvantage Black people. Data choices seem to be central to these problems; hence, the growing interest in data fixes such as debiasing data and developing auditing and explainability frameworks for unravelling how algorithms process data to produce outputs. Where these remedial approaches ignore broader structural conduits and implications of algorithmic bias, and focus on solving technical problems, they can be described as tech-reformist. Below, I provide key examples of tech-reformism.

Tech-Reformism and Its Limitations

As noted above, tech-reformist strategies include debiasing datasets such as the arrest data on which risk of recidivism algorithms rely for prediction. Skeem and Lowenkamp's (2020) debiasing technique, for example, focused on a risk assessment algorithm and assessed the effects of: (1) excising or reducing the influence of race-correlated variables; or (2) adding them and managing their impact on prediction variables. They found that option (2) produced a better outcome: it 'achieved the greatest racial balance in error rates' without significantly undermining predictive accuracy (Skeem and Lowenkamp 2020: 275).

It is however, worth noting that the aim of debiasing is to attain data neutrality and, as such, it is tech-reformist given the utopic assumption that technical components of an algorithm, in this case, underpinning data, can be neutral and unaffected by data collection and processing choices and general interpretations (see also Ugwudike 2021). It ignores structural conditions such as the digital capital with which the choices, preferences, views and values of those involved in creating algorithms and/or debiasing data permeate datasets and introduce often hidden biases, even when the data appear ostensibly neutral. As Gitlleman and Jackson (2013) rightly note, data loses its neutrality once it is decontextualised. Besides, debiasing techniques, focused as they are on technical considerations of improving the algorithm itself, may, unlike the structural approach proposed here, overlook structural issues such as the power to make subjective choices and choose theoretical standpoints that shape the design of other components of an algorithmic model. Such techniques can therefore operate as less visible conduits of bias. Another structural issue overlooked by debiasing techniques concerns the nuances of the criminal justice data on which some algorithms rely. For example, even when attempts are made to debias predictor data by excising race-related variables, historical biases in criminal history data (e.g., arrest and conviction data), which are often used as recidivism variables, can still bias predictions. Similarly, debiasing techniques may ignore commonly used predictors that appear ostensibly neutral (e.g., performance in employment and education) but continue to operate as proxies for race. In the UK, for example, the problems of racial inequality in these areas have long been documented in the education (Office for National Statistics 2020) and employment (Powell and Francis-Devine 2021) sectors.

Apart from debiasing techniques to enhance the efficiency of predictive algorithms, tech-reformism proposes additional technical strategies such as algorithm audits. Indeed, within justice systems and beyond, there have been calls for audits to, inter alia, identify conduits of algorithmic bias and embed ethical principles in algorithmic design. In their analysis of ethical audits, Brown, Davidovic and Hasan (2021: 1) note that 'nearly every research organization that deals with the ethics of AI has called for ethical auditing of algorithms'.

Audits are indeed useful in that they attempt to convert ethical principles into practical strategies for avoiding bias and attaining algorithmic fairness. Audits can also enhance transparency and help build public trust in AI systems. But when they focus on efficiency, commercial and other imperatives without due consideration of broader real-world harms (e.g., the capacity of algorithms to reproduce and entrench historical forms of discrimination) and how to avoid them, they can be described as tech-reformist. Later, when I outline the alternative structural approach, I will discuss some of the implications. Meanwhile, as yet, there are no official standards for auditing criminal justice algorithms. In addition, only few

independent/external evaluations have been conducted. Examples include the study of risk assessment algorithms by Angwin et al. (2016), which found evidence of data-driven bias (see also Hao and Stray 2019), and the evaluation of predictive policing algorithms (e.g., Hunt et al. 2014), which did not find significant evidence of effectiveness (see also, Meijer and Wessels 2019).

Alongside audits, techniques for improving the explainability of algorithmic outputs to attain transparency and accountability are also being developed. For example, Zeng, Ustun and Rudin (2015) used an approach known as Supersparse Linear Integer Models to demonstrate how to develop transparent, interpretable risk assessment algorithms. Parent and colleagues (2020: 52) have also developed an approach to creating an explainable predictive policing algorithm whose predictions can be explained using 'past event information, weather, and socio-demographic information'. Explainability techniques and audits are expected to address the problem of opacity since algorithmic decision-making can become very complex, making it difficult to investigate conduits of bias such as those that emerge during data processing. Nevertheless, the techniques are tech-reformist and limited when they focus more on the technical performance of algorithms.

Tech-reformism runs the risk of endorsing a 'technochauvinist's unique worldview' (Broussard 2018: 156). This worldview reflects a belief in the supremacy of computational and mathematical applications as scientifically neutral and fairer tools for solving societal problems. In criminal justice contexts, it manifests as a technocratic mindset that emphasises the role of such tools in fostering a rational and efficient criminal justice system. Systemic efficiency is thus accorded greater weight than broader structural aims such as equity and justice.

Beyond the Technical: A Broader Structural Framework for Understanding and Mitigating Algorithmic Bias

In this section, I provide a multifaceted, structural approach to unravelling conduits and implications of algorithmic bias. In terms of unravelling conduits of data-related bias, a structural approach would move beyond the technical conduits, such as those related to data, to highlight two structural conduits. One is the problem of data choices, and another is the problem of systemic bias, which yields biased data.

Regarding the issue of data choices, the proposed structural approach recognises that the datasets that inform algorithmic predictions are selected on the basis of subjective choices and principles and endorses a robust legal framework for regulating data selection. As Crawford (2013: xx) notes in a discussion about biased datasets, 'data and data sets are not objective; they are creations of human design'. Indeed, Gitelman and Jackson (2013) stress that the notion of 'raw data' or, in other words, neutral data, is an oxymoron, arguing that processed data is not neutral or objective; it is not a proxy for social reality. Instead, 'data are always already "cooked" and never entirely "raw"' (Gitelman and Jackson 2013: 2; Bowker 2005). They are imbued with assumptions and interpretations of human behaviour and attitudes made during collection and deployment, all of which can introduce bias. As Bennett-Moses and Chan (2016) observe in their problematisation of predictive policing algorithms and the underpinning presumptions, when the production and collection of data are imbued with racially-biased decision-making, though ostensibly objective, both processes can entrench historical biases.

Creators of predictive algorithms are currently equipped with the power to infuse algorithms with their values, choices and preferences. Examples include the selection of foundational or training data, the construction of prediction parameters and the choice of fairness metrics, all of which can inject subjective values into algorithmic data processing and outputs (Ugwudike 2020; Eaglin 2017). It is argued that access to digital capital, particularly the power to use digital technologies to influence or even dominate knowledge production, can be linked to intersecting social categories of race, gender and class (e.g., Benjamin 2019; Noble 2018; van Dijk 2005). That is, access to digital capital can be connected to positionality (social status) (see also, Ugwudike and Fleming 2021). A structural approach to addressing algorithmic bias would encompass a legal framework to regulate algorithmic design, creating standards

for ethical data choices and model design. Such regulation should balance bias elimination and other fairness imperatives with promoting human-centric technological innovation.

Unlike the tech-reformist strategy of debasing data, the structural approach takes systemic problems that yield biased data into account. Therefore, it emphasises that, to address the systemic roots of algorithmic bias, steps should be taken to address racially disparate decisions (e.g., racially-biased arrests), which generate biased data that then go on to trigger biased predictions. Failing this, such data (primarily criminal justice datasets such as arrest data) should not underpin risk predictions. Hao and Stray (2019) have provided a toolkit that demonstrates how racially-biased data can limit the ability of commonly used risk of recidivism algorithms to generate equalised odds (equal classification errors/equal balance of error rates) across different racial groups. This means that the algorithms render some groups more vulnerable to false positives (over-prediction) than others, and there is evidence that Black people are more disadvantaged (see Angwin et al. 2016). Therefore, until the problem of biased criminal justice processing is resolved, consideration should be given to eliminating criminal history variables as predictors and indicators of recidivism (cf. Johndrow and Lum 2019 and Skeem and Lowenkamp 2020). The same goes for socio-economic variables such as education and employment status, which can also operate as proxies for race and socio-economic disadvantage (see generally, Ugwudike 2020; Benjamin 2019; van Eijk 2017).

A structural approach will require algorithm audits that do not focus solely on technical efficiency but also crosscut with a broader structural framework. Such audits would investigate underpinning design logics (e.g., theories, assumptions and viewpoints that inform algorithm design) as a key step towards demystifying the technologies (Ugwudike 2021). A structural approach requires that steps be taken to unravel and question the typically less visible values, ideologies and theories that inform algorithm design. As Selbst and Barocas (2018: 1086) observe in their paper on model explanability, 'one must seek explanations of the process behind a model's development, not just explanations of the model itself'. This approach to algorithm audits should contribute to efforts to ensure that the logics are open, transparent, documented and amenable to robust critique.

A structural audit should also comprehensively evaluate the algorithm from its design logics, data inputs and processing to its outputs and potential impact on the target population. Raji and colleagues (2020) provide an example of such a structural auditing system. Described as Scoping, Mapping, Artifact Collection, Testing and Reflection, it is a structural auditing framework in the sense that it evaluates not only upstream choices and processes of algorithmic design but also downstream effects (e.g., unfair discrimination and other adverse social outcomes), providing an end-to-end internal auditing framework. This approach is consistent with what Benjamin (2019) conceptualises as 'equity audits'. Structural systems such as this are required for holistic audits, ensuring that algorithms are optimised for positive structural outcomes, not solely for technical imperatives such as validity and accuracy. But internal audits should not override the need for independent third-party ethical audits for external accountability (Brown, Davidovic and Hasan 2021; Raji et al. 2020).

Since there have been few audits of criminal justice algorithms and because the audits in other settings are currently conducted to meet ethical guidelines and standards that are not legally enforceable, lack of implementation of audit recommendations can be a problem. Indeed, in the absence of enforceable auditing standards, audit integrity issues such as the efficacy of the process itself can also become problematic (Mittelstadt 2019). The approach proposed here recognises that lack of adequate legislation is a structural problem that undermines algorithmic accountability and would require that a robust legal framework is instituted to encourage compliance with audit recommendations. A legal framework can set appropriate standards for maintaining audit integrity and integrating audit outcomes into practice.

A structural approach to addressing algorithmic bias also recognises that public engagement in key aspects of algorithm design can enhance trust. Therefore, the approach endorses the participation of historically disadvantaged groups who currently lack the digital capital required for influencing algorithmic design although, as studies have shown, they are most affected by unfair algorithmic predictions in justice

systems. To ensure the meaningful participation of disadvantaged groups, resources have to be made available to reverse the unequal distribution of digital capital, particularly the skills required for technology design. Resource investment can also advance the principles of design justice because such investment can help ensure that the processes of AI design and implementation are democratised. This can be achieved where the resource investment is channelled towards enabling historically marginalised groups who currently bear the burden of harms, such as the digital racialisation of risk and algorithmic injustice, to participate in their design.

Regarding the tech-reformist effort to enhance algorithmic explainability, a structural approach views this as laudable but limited. The reality is that even the most transparent and facilely accurate AI can still be discriminatory if consideration is not given to structural issues. The issues include power-laden creational dynamics such as the ability of some to select the data and rules on which algorithms rely. In other words, AI tools can be both explainable and unjust. For example, when a choice is made to use potentially biased data, algorithmic predictions can still fuel algorithmic injustice by disadvantaging groups affected by systemic bias and structural inequities even if the tools are themselves perfectly explainable (Ugwudike 2020).

The same applies when a decision is made to formulate a ruleset (algorithm) that is inspired by astructural theories of crime, which emphasise individual deficiencies, ignoring problems such as biased criminalisation that is fuelled by systemic and structural inequities. A problem with algorithmic opacity and the proprietary rights that protect creators from releasing their code for scrutiny is that such tools cannot be rebutted or challenged by defendants. This poses implications for due process principles and rights under Article 6 of the *European Convention on Human Rights* (European Court of Human Rights, Council of Europe 2021). Individuals have the right to contest or challenge accusations laid against them in court. But some believe that the tools have the capacity to eradicate due process rights if values based on those rights are not inscribed in their design (e.g., Završnik 2020).

In sum, as debates about algorithmic bias continue to gather momentum, insights from CAS and DS continue to remind us that a remedial framework that is cognisant of structural transformations is required to combat such bias.

Conclusion

This paper has used conceptual tools from CAS and DS to explore the problem of algorithmic bias and unravel the limitations of prevailing tech-reformist solutions. It has also drawn on insights from both scholarships to proffer a broader framework for understanding structural conduits of bias. The structural approach recognises that algorithmic bias is a societal problem with structural roots requiring structural solutions. In contrast, tech-reformism mostly depicts algorithmic bias as a technical problem. Thus, it obfuscates systemic biases such as racially discriminatory decision-making that can inject bias into the data on which the algorithms rely for prediction. Further, tech-reformism obscures structural inequities, such as the problem of unequal access to digital capital (van Dijck 2005). These problems impede design justice (Costanza-Chock 2020) by excluding affected groups from key creational processes, including the selection of data, theories and fairness metrics that should guide algorithmic design and implementation. Structural inequities may also fuel unequal distribution of algorithmic harms such as biased predictions, which, as studies now demonstrate, disproportionately affect Black people. Meanwhile, the consistent exposure of Black people to algorithmic harms such as artificially inflated risk scores can breed social harm. An example is negative labelling that can entrench the digital racialisation of risk (Ugwudike 2020) and broader algorithmic injustice (Birhane 2021).

In sum, digital capital, exclusion and divide, as well as algorithmic injustice, design justice and the digital racialisation of risk are all conceptual tools from DS and CAS that help illuminate structural sources and implications of algorithmic bias. The conceptual tools also draw attention to the fact that such bias can originate from sources beyond the technologies themselves. Its provenance is embedded in the unequal

societal structures that the technologies replicate. Therefore, in its focus on 'fixing' the technologies, tech-reformism can only address the symptoms of broader structural inequities. A structural approach, on the other hand, recognises that what is needed is a framework that can be responsive to both technical and structural problems.

Correspondence: Dr Pamela Ugwudike, Associate Professor of Criminology, University of Southampton, United Kingdom. p.ugwudike@soton.ac.uk

[1] An algorithm is a set of coded rules or instructions that can be followed (e.g., by computers) to perform functions or tasks such as making inferences from patterns in data to predict crime risks.
[2] Algorithmic bias is defined as 'the systematic, repeatable behaviour of an algorithm that leads to the unfair treatment of a certain group' (Rovastos, Mittelsdat and Koene 2020: 69).
[3] For a full list of commonly used predictor variables, see Hamilton (2015).

References

AI Now (2018) *Algorithmic accountability policy toolkit*. https://ainowinstitute.org/aap-toolkit.pdf
Ajunwa I, Friedler S, Scheidegger C and Venkatasubramanian S (2016) *Hiring by algorithm: Predicting and preventing disparate impact*. http://sorelle.friedler.net/papers/SSRN-id2746078.pdf
Allan A, Parry CL, Ferrante A, Gillies C, Griffiths CS, Morgan F, Spiranovic C, Smallbone S, Tubex H and Wong SCP (2019) Assessing the risk of Australian Indigenous sexual offenders reoffending: A review of the research literature and court decisions. *Psychiatry, Psychology and Law* 26(2): 274-294. https://doi.org/10.1080/13218719.2018.1504242
Angwin J, Larson J, Mattu S and Kirchner L (2016) Machine bias. *ProPublica*, 23 May. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
Australian Bureau of Statistics (2018) *4512.0 - Corrective services, Australia, June quarter 2018*. www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4512.0Main+Features1June%20quarter%202018?OpenDocument
Australian Government: Department of Industry, Science, Energy and Resources (2019) *Australia's artificial intelligence ethics framework*. https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles
Benjamin R (2019) *Race after technology: Abolitionist tools for the new Jim code*. Cambridge: Polity Press.
Bennett Moses L and Chan J (2016) Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *Policing and Society* 28(7): 806-822. https://doi.org/10.1080/10439463.2016.1253695
Berk RA and Bleich J (2013) Statistical procedures for forecasting criminal behaviour: A comparative assessment. *Criminology & Public Policy* 12(3): 513–544. https://doi.org/10.1111/1745-9133.12047
Birhane A (2021) Algorithmic injustice: A relational ethics approach. *Patterns* 2(2): 1-9. https://doi.org/10.1016/j.patter.2021.100205
Bonta J and Andrews DA (2017) *The psychology of criminal conduct*. 6th ed. New York: Routledge.
Bowker GC (2005) *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.
Brennan T and Oliver WL (2013) The emergence of machine learning techniques in criminology: Implications of complexity in our data and in research questions. *Criminology & Public Policy* 12(3): 551–562. https://doi.org/10.1111/1745-9133.12055
Broussard M (2018) Artificial unintelligence: How computers misunderstand the world. Cambridge, MA: MIT Press.
Brown S, Davidovic J and Hasan A (2021) The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* 8(1): 1-8. https://doi.org/10.1177%2F2053951720983865

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler SA and Wilson C (eds) *Conferences on Fairness, Accountability and Transparency*, vol 81 of *Proceedings of Machine Learning Research*: 77-91. http://proceedings.mlr.press/v81/buolamwini18a.html

Bureau of Justice Statistics (2018) *Prisoners in 2016*. www.bjs.gov/content/pub/pdf/p16_sum.pdf

Cardoso T (2020) Bias behind bars: A globe investigation finds a prison system stacked against Black and Indigenous inmates. *The Globe and Mail*, 24 October. https://www.theglobeandmail.com/canada/article-investigation-racial-bias-in-canadian-prison-risk-assessments/

Centre for Data Ethics and Innovation (2019) The Centre for Data Ethics and Innovation's approach to the governance of data-driven technology. *Gov.UK*, 19 July. https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovations-approach-to-the-governance-of-data-driven-technology/the-centre-for-data-ethics-and-innovations-approach-to-the-governance-of-data-driven-technology

Centre for Data Ethics and Innovation (2020) Review into bias in algorithmic decision-making. *Gov.UK*, 27 November. https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making

Costanza-Chock S (2020) *Design justice: Community-led practices to build the worlds we need*. Cambridge, MA: MIT Press.

Crawford K (2013) The hidden biases in big data. *Harvard Business Review*, 1 April. https://hbr.org/2013/04/the-hidden-biases-in-big-data

Dieterich W, Mendoza C and Brennan T (2016). COMPAS Risk Scales: Demonstrating
Accuracy Equity and Predictive Parity. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Eaglin JM (2017) Constructing recidivism risk. *Emory Law Journal* 67(1): 59-122. https://scholarlycommons.law.emory.edu/elj/vol67/iss1/2/

Ensign D, Friedler SA, Neville S, Scheidegger C and Venkatasubramanian S (2018) Runaway feedback loops in predictive policing. In *Conferences on Fairness, Accountability, and Transparency*, vol 81 of *Proceedings of Machine Learning Research*: 160-171. http://proceedings.mlr.press/v81/ensign18a.html

Eubanks V (2017) *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St Martin's Press.

European Court of Human Rights, Council of Europe (2021) European Convention on Human Rights as amended by Protocols Nos 11, 14 and 15, Supplemented by Protocols Nos, 1, 4, 6,7, 12, 13 and 16. https://www.echr.coe.int/documents/convention_eng.pdf

Powell A and Francis-Devine B (2021) Unemployment by ethnic background. *House of Commons Library*, 23 November. https://commonslibrary.parliament.uk/research-briefings/sn06385/

General J and Sarlin J (2021) A false facial recognition match sent this innocent Black man to jail. *CNN Business*, 29 April. https://edition.cnn.com/2021/04/29/tech/nijeer-parks-facial-recognition-police-arrest/index.html

Gitlleman L and Jackson V (2013) Introduction. In Gitlleman L (ed.) *"Raw data" is an oxymoron*: 1-14. Cambridge, MA: MIT Press.

Gower M, Spiranovic C, Morgan F and Saunders J (2020) The predictive validity of risk assessment tools used in Australia for female offenders: A systematic review. *Aggression and Violent Behavior* 53: 101428. https://doi.org/10.1016/j.avb.2020.101428

Government of Canada (2020) *Algorithmic impact assessment*. Version 0.8. https://canada-ca.github.io/aia-eia-js/

Hamilton M (2015) Risk-needs assessment: Constitutional and ethical challenges. *American Criminal Law Review* 52(2): 231-291. http://dx.doi.org/10.2139/ssrn.2506397

Hannah-Moffat K (2019) Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates. *Theoretical Criminology* 23(4): 453-470. https://doi.org/10.1177%2F1362480618763582

Hao K and Stray J (2019) Can you make AI fairer than a judge? Play our courtroom algorithm game. *MIT Technology Review*, 17 October. https://www.technologyreview.com/s/613508/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

Her Majesty's Inspectorate of Probation (2020) Assessment.
https://www.justiceinspectorates.gov.uk/hmiprobation/research/the-evidence-base-probation/supervision-of-service-users/assessment/

Humerick J (2020) Reprogramming fairness: Affirmative action in algorithmic criminal sentencing. *Columbia Human Rights Law Review Online*, April 15. http://hrlr.law.columbia.edu/hrlr-online/reprogramming-fairness-affirmative-action-in-algorithmic-criminal-sentencing/

Hunt P, Saunders J, and Hollywood JS (2014) *Evaluation of the Shreveport Predictive*
Policing Experiment. Santa Monica: RAND Corporation.

Johndrow JE and Lum K (2019) An algorithm for removing sensitive information: application to race-independent recidivism prediction. *Annals of Applied Statistics*. 13(1): 189-220.

Kaufmann M, Egbert S and Leese M (2019) Predictive Policing and the Politics of Patterns. *British Journal of Criminology*. 59, 674–692. https://doi.org/10.1093/bjc/azy060

Kehl DL, Guo P and Kessler S (2017) *Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing*. Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041

Kenyon M (2020) Algorithmic policing in Canada explained. *The Citizen Lab*, 1 September. https://citizenlab.ca/2020/09/algorithmic-policing-in-canada-explained/

Lowder EM, Morrison MM, Kroner DG and Desmarais SL (2019) Racial bias and LSI-R assessments in probation sentencing and outcomes. *Criminal Justice and Behaviour* 46(2): 210-233. https://psycnet.apa.org/doi/10.1177/0093854818789977

Lum K and Isaac W (2016) To Predict and Serve? *Significance* 13(5): 14-19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Malakieh J (2019) Adult and youth correctional statistics in Canada, 2017/2018. *The Canadian Centre for Justice Statistics*, 9 May. https://www150.statcan.gc.ca/n1/pub/85-002-x/2019001/article/00010-eng.htm

Maurutto P and Hannah-Moffat K (2006) Assembling risk and the restructuring of penal control. *British Journal of Criminology* 46(3): 438-454. https://www.jstor.org/stable/23639357

Meijer A and Wessels M (2019) Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration* 42(12): 1031-1039. https://doi.org/10.1080/01900692.2019.1575664

Ministry of Justice (2019) *Statistics on race and the criminal justice system 2018: A Ministry of Justice publication under Section 95 of the* Criminal Justice Act 1991. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/849200/statistics-on-race-and-the-cjs-2018.pdf

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1: 501-507. https://doi.org/10.1038/s42256-019-0114-4

Noble S (2018) *Algorithms of Oppression*. New York, New York University Press.

Office for National Statistics (2020) *Child poverty and education outcomes by ethnicity*. https://www.ons.gov.uk/economy/nationalaccounts/uksectoraccounts/compendium/economicreview/february2020/childpovertyandeducationoutcomesbyethnicity

O'Neil C (2016) *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishing.

Parent M, Roy A, Gagnon C, Lemaire N, Deslauriers-Varin N, Falk TH and Tremblay S (2020) Designing an explainable predictive policing model to forecast police workforce distribution in cities. *Canadian Journal of Criminology and Criminal Justice* 62(4): 52-76. https://doi.org/10.3138/cjccj.2020-0011

Pasquale F (2015) *The black box society: The secret algorithms that control money and information*. Cambridge: Harvard University Press.

PredPol (2021) Geolitica: A new name, a new focus. *PredPol*, 2 March. https://blog.predpol.com/geolitica-a-new-name-a-new-focus

Price M (2019) Hospital 'risk scores' prioritize white patients. *Science*, 24 October. https://www.science.org/news/2019/10/hospital-risk-scores-prioritize-white-patients

Raji DI, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D and Barnes P (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 33-44. New York: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Richardson R, Schultz JM and Crawford K (2019) Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review* 94: 192-233.

Rovastos M, Mittelsdat B and Koene A (2020) *Landscape summary: Bias in algorithmic decision-making: What is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?* Centre for Data Ethics and Innovation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819055/Landscape_Summary_-_Bias_in_Algorithmic_Decision-Making.pdf

Selbst AD and Barocas S (2018) The intuitive appeal of explainable machines. *Fordham Law Review*. 87(3) 1085-1139. https://ir.lawnet.fordham.edu/flr/vol87/iss3/11/

Shepherd SM, Adams Y, McEntyre E and Walker R (2014) Violence risk assessment in Australian Aboriginal offender populations: A review of the literature. *Psychology, Public Policy and Law* 20(3): 281-293. https://doi.apa.org/doi/10.1037/law0000017

Skeem J and Lowenkamp C (2020) Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & The Law* 38(3): 259-278. https://doi.org/10.1002/bsl.2465

The Law Society (2019) *Algorithm use in the criminal justice system report*. https://www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report

Ugwudike P (2020) Digital prediction technologies in the justice system: The implications of a 'race-neutral' agenda. *Theoretical Criminology*. Advance online publication. https://doi.org/10.1177/1362480619896006

Ugwudike P (2021) AI audits for assessing design logics and building ethical systems: The case of predictive policing algorithms. *AI and Ethics* https://doi.org/10.1007/s43681-021-00117-5

Ugwudike P and Fleming J (2021) Artificial Intelligence, digital capital, and epistemic domination on Twitter: A study of families affected by imprisonment. *Punishment and Society*. Advance online publication. https://doi.org/10.1177/14624745211014391

van Dijk J (2005) *The deepening divide: Inequality in the information society*. Thousand Oaks: SAGE Publications.

van Eijk G (2017) Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. *Punishment & Society* 19(4): 463-481. https://doi.org/10.1177%2F1462474516666282

Yeong S (2021) An evaluation of the Suspect Target Management Plan. *Crime and Justice Bulletin* 233. Sydney: NSW Bureau of Crime Statics and Research. https://www.bocsar.nsw.gov.au/Publications/CJB/2020-Evaluation-of-the-Suspect-Target-Management-Plan-CJB233.pdf

Završnik A (2020) Criminal justice, artificial intelligence systems, and human rights. *ERA Forum* 20: 567-583. https://doi.org/10.1007/s12027-020-00602-0

Zeng J, Ustun B and Rudin C (2015) Interpretable classification models for recidivism prediction. *arXiv*. https://arxiv.org/abs/1503.07810