Amicus Curiae, Series 2, Vol 1, No 1, 84-96

Fairness in Algorithmic Decision-making: Trade-offs, Policy Choices, and Procedural Protections

Adam Harkens

Birmingham Law School, University of Birmingham

Abstract

This article discusses conceptions of fairness in algorithmic decision-making, within the context of the UK's legal system. Using practical operational examples of algorithmic tools, it argues that such practices involve inherent technical trade-offs over multiple, competing notions of fairness, which are further exacerbated by policy choices made by those public authorities who use them. This raises major concerns regarding the ability of such choices to affect *legal* issues in decision-making, and transform *legal* protections, without adequate legal oversight, or a clear legal framework. This is not to say that the law does not have the capacity to regulate and ensure fairness, but that a more expansive idea of its function is required.

Keywords: algorithmic decision-making, machine learning, fairness, criminal justice, administrative law

[A] INTRODUCTION

In June 2019, the Law Society of England and Wales launched a report on the use of algorithms in the criminal justice system of England and Wales, examining current and potential use cases in the United Kingdom, as well as related legal challenges and consequences (Law Society 2019). The report raised a number of concerns related to algorithmic decision-making, including the potential to produce biased and discriminatory decisions, the oversimplification of complex issues because of the methods of quantification employed, the loss of autonomy for those individuals who must enter the processes of the legal system and be decided about, a lack of transparency as to the reasons why a particular decision has been made, and the hindering of legal scrutiny of decisions, among others.

What each of these concerns share is a general common theme related to questions of fairness and justice in the legal system. Were any of these concerns to be realised in an operational algorithmic (or part-algorithmic) decision-making process, this would hinder the quality of justice being provided and ultimately undermine trust in the protections and capacities of the English legal system. This in itself is concerning because it demonstrates how each *design* question—whether minor or not—in the construction of an algorithmic decision-making process, is fraught with danger and can have significant consequences for the perception and practices of the wider legal system. This involves more than explicit technical design decisions and trade-offs, however, and any effort to analyse these tools should also incorporate the policy choices of public authorities which choose to incorporate them into decision-making processes.

This article aims at beginning to explore these design questions, including how they are made and how they are justified, in order to ask how they may affect the function of the legal system, and whether they can transform longstanding principles and concepts of legal protection in England and Wales, including procedural rights. It does so firstly by briefly discussing the context of algorithmic decision-making in the UK, before moving on to discuss technical trade-offs around fairness, inherent to machine learning and algorithmic tools. Next, it contextualises these trade-offs within the context of public policy and operational decisions made by public authorities, before finally analysing how this may affect and transform procedural rights in the English legal system, where the majority of clearly applicable current protections focus on issues relating to data protection.

[B] PUBLIC AUTHORITIES AND ALGORITHMIC DECISIONS IN ENGLAND AND WALES

Algorithmic decision-making is most notably being used in two specific areas of the English legal system: to automate, supplement and support *administrative* decision-making at the national and local council level; and to do the same for *criminal justice* in the context of resource management, surveillance, and risk assessment for policing, as well as more general offender management at varying post-arrest and pre-trial stages. Algorithmic decision-making for the purposes of immigration management spans both areas, depending on the type of decision being made.

Administrative Use

The types of administrative decision which have been targeted for automation include those relating to welfare and tax, immigration and residence checks, and social care. The extent to which these programmes have been developed and implemented varies, with some remaining in pilot status, while others have been fully implemented. For example, at the national level, Her Majesty's Revenue and Customs (HMRC) and the Department for Work and Pensions (DWP) have both attempted to improve 'service delivery' through the automation of decision-making processes. Through the 'Making Tax Digital' programme, and the 'Connect' database, HMRC intends to have fully digitalised services in the UK through the integration of real-time data streams to correct tax-code errors, automatically provide tax rebates, ensure debt collection, and to increase fraud detection capacities (Government Digital Service 2017).

The DWP, building on HMRC's real-time system, has been making use of available data to automatically assess individual Universal Credit claims, and to detect and pursue fraud (Government Digital Service 2017). This information is also central to the European Union EU settlement scheme, where EU citizens and their families must apply for 'settled status' following the UK's exit from the European Union. A system of initially automated checks makes the decision as to whether these citizens receive 'indefinite leave to remain' (settled status), 'limited leave to remain' (pre-settled status)—lasting for a period of five years—or are rejected and must provide further evidence to a human caseworker (Tomlinson 2019).[1] Simultaneously, local councils in England have begun to make use of algorithmic modelling for the purposes of predicting risk and the need for interventions in the home care of children. These are all separate systems, developed in-house and by private companies, and are not part of a wider policy (Dencik and Others 2018).

The Criminal Justice System

The picture in the criminal justice system is largely similar, in that current projects are still in development, and many make use of the same style of algorithmic modelling and risk prediction. However, algorithmic tools are also being used in the criminal justice system to enable other kinds of technologies, including live facial recognition and hotspot mapping.

For example, Durham Constabulary's Harm Assessment Risk Tool (HART) is designed to aid the decision-making of a custody officer

[1] EU Settlement Scheme guidance booklets for caseworkers.

immediately following the arrest of an individual within County Durham. If a charge is to be brought forward, the custody officer on duty is required to decide whether to 'bail (conditionally or unconditionally), hold in custody, prosecute, or divert [the suspect] from the Criminal Justice System (CJS) with an out of court disposal' and HART helps in doing so by sorting said suspects into three risk groups: high, medium and low (Urwin 2016).

A similar system is being developed by West Midland's Police—the 'Data Driven Insights' Programme (DDI)—alternatively attempts to identify high-risk individuals for intervention before they have committed any crimes (Alan Turing Institute and Independent Digital Ethics Panel for Policing 2017). West Midlands Police are also helping to develop the National Data Analytics Solution (NDAS), which is using algorithmic technologies with the goal of moving law enforcement 'away from its traditional crime related role and into wider and deeper aspects of social and public policy' by facilitating interventions for individuals at risk of harm (Alan Turing Institute and Independent Digital Ethics Panel for Policing 2017: 3).

London Metropolitan Police's Gangs Matrix is currently used to identify individuals who are members of a 'gang', at risk of becoming recruited as a gang member, or at risk of becoming a victim of gang violence. This is achieved through data analysis, which gives individuals a 'gang score' based upon their activities, interests and friendship groups (ICO 2018). London Metropolitan Police are also responsible for trialling Live Facial Recognition (LFR) systems within the capital and have come under significant scrutiny alongside South Wales Police for similar practices (Fussey and Murray 2019).

[C] TRADE-OFFS IN MACHINE LEARNING

A substantial body of literature exists on the ways in which fairness is defined, redefined, and operationalized within machine-learning systems; so much so that it would be beyond the scope of this article to cover sufficiently and in full detail. What is important for this article to acknowledge though is that within this literature there is a general recognition of the existence and requirement of trade-offs when defining fairness because of the limitations of algorithmic analysis (Berk and Others 2017).

One operational algorithmic decision-making in the United States, COMPAS, has received significant public attention because of these very issues. In 2016, journalists from ProPublica took a sample of 11,757 people that had been processed through the COMPAS system in Broward County, Florida, between 2013 and 2014. Their COMPAS scores were then compared with the county's records and analysed for their accuracy in predicting actual recidivism rates within two years of the initial risk assessment. The two-year standard is used by Northpointe, the developers, for its own validation studies (Northpointe 2015).

Following the study, ProPublica concluded that black defendants were 77% more likely to be classified as higher risk (medium to high), compared to white defendants, for violent recidivism risk, and 63% more likely for general recidivism risk. This discrepancy remains when looking at 'misclassifications' or errors in risk calculation. Here, black defendants who did not commit crimes within the next two years were almost twice as likely to be misclassified as higher risk (45% compared with 23% of white defendants), and white defendants who did commit crimes within the next two years were almost twice as likely to be labelled low risk (48% compared with 28% of black defendants). Figure 1 shows a visualization of the risk scores in this sample, including the more even distribution of scores for black defendants, contrasted with the low-risk heavy bars for white scores (Angwin and Others 2016a).

Later in the same year, COMPAS developers responded with a validity study claiming 'predictive parity' between black and white defendants in Broward County. The main counterpoint to ProPublica's findings, they argue, is that by splitting defendants into separate groupings of black and white, and therefore analysing the accuracy of this tool on a different basis, this shifted white defendants to a lower base risk 'relative to the norm' (Dieterich and Others 2016: 4-5).

Northpointe entirely rejected any accusations of bias by stating this information 'does *not* show evidence of bias, but rather is a natural consequence of using unbiased scoring rules for groups that happen to have different distributions of scores' (Dieterich and Others 2016: 8; original emphasis). 'Natural consequence' is used here because it is the algorithms within COMPAS that attempt to roughly split up the norming group into ten 'decile scores' of risk, from one to ten (Northpointe 2015: 8). The company makes the argument that once fed back into the system as a single grouping, the risk scores of white defendants will demonstrate a more even distribution.

While this is statistically justified, ProPublica still made the case that 'when you compare black and white defendants with similar characteristics, black defendants tend to get higher scores', which would suggest that the COMPAS algorithm could be setting in stone systematic

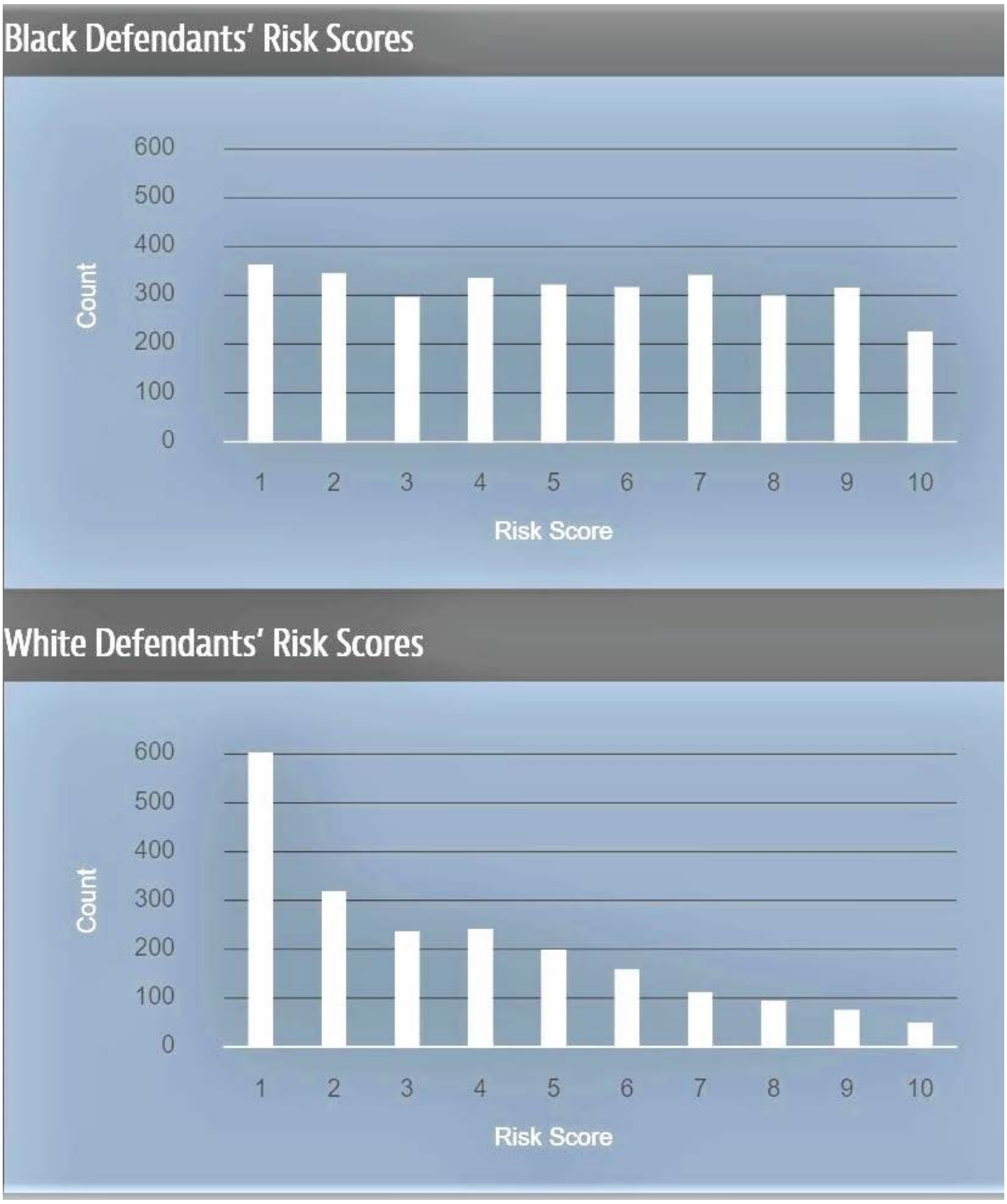


Figure 1: Distribution of risk by race in ProPublica's validation study of COMPAS. Source: Angwin and Others 2016a.

racial basis, already pre-existing within the Broward County criminal justice system (Angwin and Others 2016b). While COMPAS developers were making an argument of statistical fairness, ProPublica's represents one of political and legal fairness. Incorporating the two into one system is no mean feat. Often, constructing a 'fair' system requires decisions to be made regarding competing notions of fairness which are potentially incompatible.

[D] THE COMPLEXITY AND CONSEQUENCES OF POLICY CHOICES

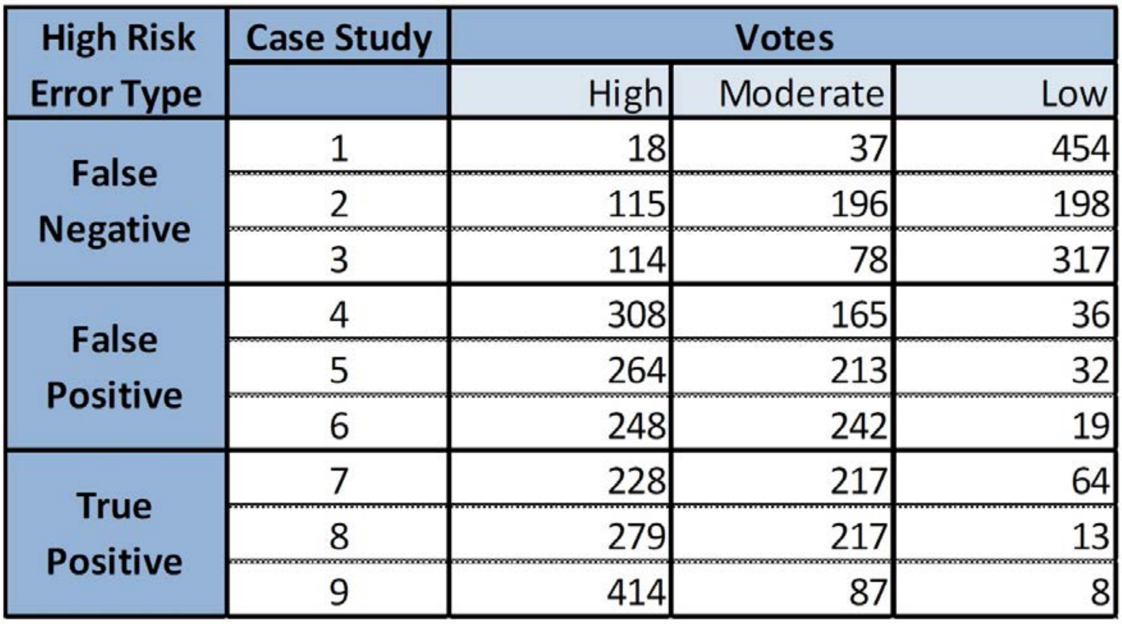
Harm Assessment Risk Tool

In the UK, similar trade-offs have been made which go beyond specific technical concerns, incorporating, and being justified by, public policy decisions. The best example of this is Durham Constabulary's HART. Because in the belief that 'not all errors are equal', HART is designed to be overly cautious and prioritise false positives over false negatives (Urwin 2016: 53, 75). In practice, this means that, if a decision between two risk groups is borderline, HART's algorithms are designed to predict a higher level of risk for the given individual. This is done in order to avoid the most dangerous errors, which would be when a person is classified as low risk, but goes on to commit a serious offence and therefore actually represented high risk to the community. Similar design choices may have been made with COMPAS, but this cannot be said with any certainty as the tool is proprietary, and available literature does not discuss choices like this (Brennan and Dieterich 2018).

The crucial point here is that these choices of what is considered 'fair' and 'safe' have been *designed* into the tool. While there is a level of complexity in terms of how the algorithms analyse available data itself, there is also a level of control available to policymakers as to what direction the analysis should take and how individuals are treated. In this situation, the decision has been made to treat potentially innocent individuals more harshly, on the basis of an algorithmic prediction.

HART in contrast uses 34 risk predictors—taking information on the defendant at time of arrest and combining this with Durham Constabulary's pre-existing records. The majority of these relate to the individual's criminal history, along with age, gender, two forms of postcode data, and the number of police intelligence reports collected on that person, for example (Oswald and Others 2018). The data from these parameters is combined to construct 509 different decision trees, including 'classification' and 'regression' trees, or CARTs (Oswald and Others 2018). Each of these 'trees' essentially represents a separate algorithm that analyses a random sample, or 'case profile' of an individual's data to categorize (classification) and make predictions (regression).

When each decision tree is completed and has come to a conclusion on how risky someone is, the tool draws from the 'wisdom of the crowd' by



High Risk Error Type	Case Study	Votes		
		High	Moderate	Low
False Negative	1	18	37	454
	2	115	196	198
	3	114	78	317
False Positive	4	308	165	36
	5	264	213	32
	6	248	242	19
True Positive	7	228	217	64
	8	279	217	13
	9	414	87	8

Figure 2: Case Study of 'majority voting' in HART. Source: Unwin 2016: 10 (Fig 10).

combining the trees, so that each one casts a 'vote' on whether the person at hand is low risk, medium risk, or high risk (Gollapudi 2016). Through an example of this voting process, it is possible to see how borderline decisions are treated. Figure 2 demonstrates a number of test case studies (see Urwin 2016). Case studies 5 and 6 are both borderline examples. In either of these, the individual in question could be either high or moderate risk. Moderate risk would enable them to be processed through an 'out-of-court disposal', however, in this example, the individual was required to continue through the courts process, along with the more extensive socio-legal consequences that this would cause.

Other Tools

As with HART, operational choices and non-technical definitions are crucial in other examples of decision-making systems, including the Gangs Matrix. For example, the following questions must be decided before a tool like this could be used: what level of association does one need to have in order to be considered a gang member? And how is the meaning of gang defined in this instance? A recent enforcement notice from the Information Commissioner's Office (ICO) demonstrates this exact problem. This showed that victims of gang crime were often assigned a risk score and included within the main database of the matrix itself as a potential risk, either because of the belief that this demonstrated gang associations, or it was registered as part of their crime history (ICO 2018: 9-11).

Concerns have also been raised regarding the NDAS and one of its designated purposes and justifications being to reduce 'harm' (West Midlands Police 2019). Harm is not defined in this situation, yet the different ways this term can be used could have a significant impact both on the decision being made and the capacities of the machine learning involved. This is particularly the case given West Midlands Police's determination to move into areas of 'social and public policy', beyond traditional policing, and thus expanding their powers of intervention.

These are certainly questions about fairness because they clearly impact upon the types of decision being made, as well as the treatment which an individual will receive as they are faced with various arms of the legal system. They are not technical issues—in that they are not questions surrounding the efficacy or efficiency of the algorithmic tools—yet they add layers of complexity through their interaction with technical choices that must be made during the design process, and can change the technical parameters of a decision based on how a given term is defined. Ensuring fairness therefore, requires more than an analysis of the ways in which the machine-learning model of a particular tool produces a risk score, but must also incorporate how this is interpreted by the public authority, how this authority has influenced the design process, how it has decided to make the decision-making system operational, and how its use has been justified, both legally, and in the language of 'social and public policy'.

[E] PROCEDURAL CONFLICTS

Much has been spoken so far regarding the types of technical and policy choices and trade-offs that must be made during the design process of an algorithmic decision-making tool and which affect fairness in the legal system, but less attention has been paid towards a crucial aspect of this: the law. Depending on how algorithmic decision-making is implemented, these tools sit at an important intersection of a number of different bodies of law.

As a result, legal frameworks from administrative law, data protection law, criminal law and criminal justice can all have an effect on how these systems are treated legally, and on their legality more crucially and generally. Much work has already been carried out from the perspective of data protection, particularly related to automated decision-making and concepts such as the right to an explanation.[2] Administrative law is also

[2] See, for example, among many others: Edwards and Veale (2018).

becoming a focus for this area, with research beginning to tackle algorithmic decision-making more seriously, using traditional concepts and principles (Oswald 2018; Cobbe 2019). This is the case for human rights law too, including international human rights law frameworks (McGregor and Others 2019). Much less has been written regarding the perspective of criminal law and criminal justice in the UK, though at least one project is working on this specific issue at the moment (Yeung 2019).

Tackling algorithmic decision-making, and analysing it through the perspective of each of these strands of law, is certainly crucial. This provides insights as to how traditional legal principles can be affected by this style of decision-making. Currently, the lack of combined and targeted legal frameworks (outside of the UK Data Protection Act 2018 and associated Data Protection Impact Assessments), means that the actions of public authorities in designing and implementing algorithmic tools are ensuring that it is largely public policy which is dictating what can be considered 'fair' in these circumstances, including the work of government bodies like the Centre for Data Ethics and Innovation.

This leaves questions over fairness in somewhat of a legal vacuum—in that the legal basis for these technologies has not been adequately identified or confirmed (Fussey and Murray 2019) when it should be the law which is primarily defining and protecting this concept. Data protection plays an important role in attempting to prevent such developments, as seen above in relation to the Gangs Matrix inclusion of victim's data and associate implications. However, this was primarily raised as a data retention issue, and one regarding the fair processing of data, rather than the specific question of what can be considered to be a fair decision (ICO 2018: 9-11).

To conceive of what is legally fair in this situation, requires an understanding of how legal concepts are affected, such as discretion, the duty to provide reasons, the right to liberty and the right to a fair trial—which are brought into question by the relative lack of transparency in the tools involved, as well as their predictive capabilities, and the ways in which they may constrain human decision-makers. This means that they are not approached simply from the perspective of the 'rules' of each individual legal area, but considered through the frame of fairness for the entire legal system as a whole. Given that these are legal concepts, they should also only change through legal methods, whether through the courts or legislation. Allowing such concepts to be transformed through policy actions may produce unwelcome shifts, without due care.

To achieve this, we must consider the design of the decision-making process from beginning to end and understand where trade-offs exist, where policy choices can apply, and decisions must be made as to how to secure a clear legal basis for algorithmic decision-making, where it is central to high stakes tasks. The creation of algorithmic decision-making tools is ultimately an extremely flexible process, as even where the limitations of machine learning have been reached in a given situation, further non-technical choices can be made which increase technical and legal complexity.

[F] CONCLUSION

Algorithmic decision-making is becoming more widespread in the United Kingdom, affecting an increasing number of procedures of the English system, including risk assessment by law enforcement for targeted intervention in crime prevention, the management of individual defendants, and the identification of *potential* offenders. It is also becoming important for administrative procedures, such as its central role in the EU Settlement Scheme.

This article has demonstrated that the design of these algorithmic tools involves a number of technical trade-offs and policy choices that can have a severe effect on the form of 'fairness' which a given tool can provide. These must be considered as being on a par, as their combination produces further complexity within the legal system and its associated procedures. Further, it has argued that these choices and trade-offs potentially result in legal fairness being treated as a public policy, where these tools exist within a certain degree of a legal vacuum. This should be prevented, and it should be the structures of the law which set out and define the kinds of choices that can be legally made in this area and which are legally fair. For example, decisions such as whether it is fair to treat an individual more harshly, based on a higher risk score which may be a false positive, should only be decided through the perspective of the law.

References

Alan Turing Institute & Independent Digital Ethics Panel for Policing (2017) 'Ethics Advisory Report for West Midlands Police' London: Alan Turing Institute.

Angwin, J, & Others (2016a) 'Machine Bias: There's Software Used across the Country to Predict Future Criminals. And it's Biased against Blacks' (Pro Publica, 23 May 2016)

Angwin, J and Others (2016b) 'Technical Response to Northpointe' (*Pro Publica*, 29 July 2016)

Berk, R & Others (2017, 28 May) 'Fairness in Criminal Justice Risk Assessments: The State of the Art' arXiv Working Paper 1703.09207 Ithaca NY: Cornell University .

Brennan, T & W Dieterich (2018) 'Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)' in Jay P Singh & Others (eds) *Handbook of Recidivism Risk/Needs Assessment Tools* Malden, MA: Wiley Blackwell.

Cobbe, Jennifer (2019) 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making' *Legal Studies* Cambridge: Cambridge University Press Cambridge Core .

Dencik, L & Others (2018) 'Data Scores as Governance: Investigating Uses of Citizen Scoring in Public Services' Cardiff: Data Justice Lab.

Dieterich, W & Others (2016, 8 July) 'COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity' Traverse City MI: Northpointe Inc.

Edwards L & M Veale (2018) 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?' 16(3) *IEEE Security and Privacy* 46-54.

Fussey, P & D Murray (2019) 'Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology' Colchester: Human Rights, Big Data and Technology Project, University of Essex.

Gollapudi, S (2016) *Practical Machine Learning: Tackle the Real-world Complexities of Modern Machine Learning with Innovative and Cutting-edge Techniques* Birmingham-Mumbai: Packt Publishing.

Government Digital Service (2017, 9 February) 'Government Transformation Strategy'.

ICO (Information Commissioner's Office) (2018) 'ICO Enforcement Notice for the Gangs Matrix'.

Law Society of England and Wales (2019) 'Algorithms in the Criminal Justice System: A Report by the Law Society Commission on the Use of Algorithms in the Justice System' London: Law Society.

McGregor, Lorna & Others (2019) 'International Human Rights Law as a Framework for Algorithmic Accountability' 68(2) *International and Comparative Law Quarterly* 309-43.

Northpointe (2015) 'Practitioner's Guide to COMPAS Core' Traverse City MI: Northpointe Inc.

Oswald, Marion (2018) 'Algorithm-assisted Decision-making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power' 376 (2128) *Philosophical Transactions of the Royal Society A* 1-20.

Oswald, Marion & Others (2018) 'Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental" Proportionality' 27(2) *Information and Communications Technology Law* 223-50.

Tomlinson, Joseph (2019) *Quick and Uneasy Justice: An Administrative justice analysis of the EU Settlement Scheme* London: Public Law Project: London.

Urwin, S (2016) 'Algorithmic Forecasting of Offender Dangerousness for Police Custody Officers: An Assessment of Accuracy for the Durham Constabulary Model' Research Presented for the purposes of gaining a Master's Degree in Applied Criminology and Police Management at Cambridge University.

West Midlands Police (2019, 3 April) 'Minutes of West Midland's Police and Crime Commissioner's Ethics Board Meeting 3 April 2019'.

Yeung, Karen (2019) 'Machine Decision-making in the Criminal Justice System: The FATAL4JUSTICE? Project' (*Oxford Human Rights Hub*, 8 April 2019) .

Legislation Cited

Data Protection Act 2018