Whitehall Report 3-18

Machine Learning Algorithms and Police Decision-Making

Legal, Ethical and Regulatory Challenges

Alexander Babuta, Marion Oswald and Christine Rinik

RUSI
www.rusi.org

Royal United Services Institute
for Defence and Security Studies

UNIVERSITY OF
WINCHESTER
CENTRE FOR INFORMATION RIGHTS

Machine Learning Algorithms and Police Decision-Making

Legal, Ethical and Regulatory Challenges

Alexander Babuta, Marion Oswald and Christine Rinik

RUSI Whitehall Report 3-18, September 2018

Royal United Services Institute
for Defence and Security Studies

UNIVERSITY OF
WINCHESTER
CENTRE FOR INFORMATION RIGHTS

187 years of independent thinking on defence and security

The Royal United Services Institute (RUSI) is the world's oldest and the UK's leading defence and security think tank. Its mission is to inform, influence and enhance public debate on a safer and more stable world. RUSI is a research-led institute, producing independent, practical and innovative analysis to address today's complex challenges.

Since its foundation in 1831, RUSI has relied on its members to support its activities. Together with revenue from research, publications and conferences, RUSI has sustained its political independence for 187 years.

Published in 2018 by the Royal United Services Institute for Defence and Security Studies.

Whitehall Report 3-18, September 2018. ISSN 1750-9432

Printed in the UK by Stephen Austin and Sons, Ltd.

Cover image: Courtesy of Pexels/Florian Weihmann

Royal United Services Institute
for Defence and Security Studies
Whitehall
London SW1A 2ET
United Kingdom
+44 (0)20 7747 2600
www.rusi.org
RUSI is a registered charity (No. 210639)

Contents

Acknowledgements v

Executive Summary vii

Introduction **1**
Definitions and Terminology 2

I. Predictive Policing and Offender Risk Assessment **5**

II. Discretion and Accountability **11**

III. Transparency and Intelligibility **17**
Technical Transparency 17
Transparency of Process 20

IV. Fairness and Bias **23**

V. Towards a Formal System of Regulation and Oversight **29**

Recommendations **33**

About the Authors 35

Acknowledgements

The authors of this report would like to thank the numerous experts who took part in the conference and series of focus groups held at the Royal United Services Institute (RUSI) on 2 May 2018. Thanks are due in particular to the panel speakers for their constructive and thought-provoking contributions, and to Lord Hogan-Howe for chairing the event and providing valuable analysis and insights. We are grateful to PA Consulting for their generous financial support and practical assistance in organising the event.

The authors would also like to thank a number of individuals who dedicated their time and expertise to reviewing earlier drafts of this paper, in particular Chief Superintendent David Powell, Dr Kieron O'Hara, Dr Nóra Ní Loideain and Professor Peter Fussey.

Our thanks are also extended to RUSI colleagues for their analytical and editorial input, including Professor Malcolm Chalmers, Dr Andrew Glazzard, Dr Emma De Angelis, Melanie Bell, Edward Mortimer and Stephen Reimer.

Executive Summary

This report explores the applications of machine learning algorithms to police decision-making, specifically in relation to predictions of individuals' proclivity for future crime. In particular, it examines legal, ethical and regulatory challenges posed by the deployment of such tools within an operational policing environment.

In the UK, the use of machine learning algorithms to support police decision-making is in its infancy, and there is a lack of research examining how the use of an algorithm influences officers' decision-making in practice. Moreover, there is a limited evidence base on the efficacy and efficiency of different systems, their cost-effectiveness, their impact on individual rights and the extent to which they serve valid policing aims. Limited, localised trials should be conducted and comprehensively evaluated to build such an evidence base before moving ahead with large-scale deployment of such tools.

There is a lack of clear guidance and codes of practice outlining appropriate constraints governing how police forces should trial predictive algorithmic tools. This should be addressed as a matter of urgency to enable police forces to trial new technologies in accordance with data protection legislation, respect for human rights and administrative law principles.

While machine learning algorithms are currently being used for limited policing purposes, there is potential for the technology to do much more, and the lack of a regulatory and governance framework for its use is concerning. A new regulatory framework is needed, one which establishes minimum standards around issues such as transparency and intelligibility, the potential effects of the incorporation of an algorithm into a decision-making process, and relevant ethical issues. A formalised system of scrutiny and oversight, including an inspection role for Her Majesty's Inspectorate of Constabulary and Fire and Rescue Services, is necessary to ensure adherence to this new framework.

There are various issues concerning procurement contracts between the police and private sector suppliers of predictive policing technology. It is suggested that all relevant public procurement agreements for machine learning algorithms should explicitly require that it be possible to retroactively deconstruct the algorithm in order to assess which factors influenced the model's predictions, along with a requirement for the supplier to be able to provide an expert witness who can provide details concerning the algorithm's operation if needed, for instance in an evidential context.

The legal and ethical issues concerning the use of machine learning algorithms for policing are complex and highly context-dependent. Machine learning algorithms require constant attention and vigilance to ensure that the predictions they provide are as accurate and as unbiased as possible, and that any irregularities are addressed as soon as they arise. For this

reason, multidisciplinary local ethics boards should be established to scrutinise and assess each case of algorithmic implementation for policing. Such boards should consist of a combination of practitioners and academics, and should provide recommendations to individual forces for practice, strategy and policy decisions relating to the use of algorithms.

A collaborative, multidisciplinary approach is needed to address the complex issues raised by the use of machine learning algorithms for decision-making. At the national level, a working group consisting of members from the fields of policing, computer science, law and ethics should be tasked with sharing 'real-world' innovations and challenges, examining operational requirements for new algorithms within policing, with a view to setting out the relevant parameters and requirements, and considering the appropriate selection of training and test data.

Officers may need to be equipped with a new skill set to effectively understand, deploy and interpret algorithmic tools in combination with their professional expertise, and to make assessments of risk using an algorithmically generated forecast. It is essential that the officers using the technology are sufficiently trained to do so in a fair and responsible way and are able to act upon algorithmic predictions in a way that maintains their discretion and professional judgement.

Introduction

This Whitehall Report explores the potential uses of machine learning algorithms for police decision-making, particularly in relation to individuals' proclivity for future crime. Much of the content is based on the proceedings of a half-day conference and a series of focus groups held at the Royal United Services Institute (RUSI) on 2 May 2018, organised in partnership with the Centre for Information Rights at the University of Winchester and made possible with the generous support of PA Consulting. Attendees included policymakers and practitioners from across government, law enforcement and the wider criminal justice system, academic and legal experts, and private sector representatives.

A participatory research approach was chosen due to the value that can be gained from the experience of stakeholders in the assessment of the real-world context, diagnosis of the issues and consideration of policy requirements.[1] A focus group method facilitated the exploration of deeper perspectives from those implementing and enforcing the law, and the identification of significant issues to the practitioner.[2] The report is also based on the authors' own interdisciplinary problem-oriented research undertaken over the past two years, which has involved working closely with several police forces in the UK, researching the integration of algorithmic decision-support tools into the policing process.

Despite a growing body of research examining the use of algorithms for policing and associated legal and ethical challenges,[3] there remains considerable uncertainty concerning how a future

1. Andrea Cornwall and Rachel Jewkes, 'What is Participatory Research?', *Social Science & Medicine* (Vol. 41, No. 12, 1995), pp. 1667–76.
2. Fiona de Londras, 'Participatory Research: Some Provocations for Doctoral Students in Law', in Laura Cahillane and Jennifer Schweppe (eds), *Legal Research Methods* (Dublin: Clarus, 2016), p. 150.
3. See, for example, Richard Berk and Jordan Hyatt, 'Machine Learning Forecasts of Risk to Inform Sentencing Decisions', *Federal Sentencing Reporter* (Vol. 27, No. 4, 2015); Richard A Berk, Susan B Sorenson and Geoffrey Barnes, 'Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions', *Journal of Empirical Legal Studies* (Vol. 13, No. 1, 2016), pp. 94–115; Angèle Christin, 'Algorithms in Practice: Comparing Web Journalism and Criminal Justice', *Big Data and Society* (Vol. 4, No. 2, 2017); Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" is Probably Not the Remedy You Are Looking For', *Duke Law and Technology Review* (Vol. 16, No. 18, 2017); Andrew Guthrie Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* (New York: New York University Press, 2017); Mireille Hildebrandt, 'New Animism in Policing: Re-Animating the Rule of Law?' in Ben Bradford et al. (eds), *The Sage Handbook of Global Policing* (London: Sage, 2016); Elizabeth E Joh, 'Policing by Numbers: Big Data and the Fourth Amendment', *Washington Law Review* (Vol. 89, No. 35, March 2014); Karen Yeung, 'Algorithmic Regulation: A Critical Interrogation', Regulation & Governance (31 July 2018), doi: 10.1111/rego.12158.

policy framework ought to operate in practice. The purpose of this report is to work towards bridging this gap between research and policy. The report seeks to critically assess specific multidisciplinary issues relevant to the use of algorithms for policing, and provide practical recommendations designed to contribute to the fast-moving debate over policy and governance in this area. Specifically, the report aims to provide an overview of legal, ethical, regulatory and practical challenges surrounding the authorities' use of algorithmic decision-support tools within policing, and to provide recommendations focused upon enabling controlled testing and appropriate innovation with such tools within a regulated framework.

The report is structured as follows. Chapter I provides an overview of the current situation in the UK regarding the police's use of machine learning algorithms, outlining the main issues posed by the introduction of such technology. Chapter II explores the topics of discretion, professional judgement and accountability within policing, and how new methods of data analysis may support or threaten this. Chapter III focuses on the transparency and intelligibility of the decision-making process, and the challenges of implementing statistical tools that are not entirely transparent in how they process data. Chapter IV discusses how law enforcement agencies could implement these tools in accordance with human rights principles of fairness and proportionality. Finally, Chapter V concludes by discussing whether there is now a need for a new formalised system of regulation and oversight of the authorities' use of machine learning algorithms, and if so, how such a framework might be implemented in practice.

Definitions and Terminology

This report is concerned with machine learning algorithms. An algorithm is defined here as a technologically automated mathematical formula, 'a sequence of instructions that are carried out to transform the input to the output'.[4] The focus of this report is on algorithmic tools that incorporate machine learning ('ML algorithms'), rather than those that do not ('non-ML algorithms'). In machine learning, the instructions are not directly provided by the programmer; instead, 'the aim is to construct a program that fits the given data'.[5] A machine learning program 'is a *general template with modifiable parameters*, and by assigning different values to these parameters the program can do different things'.[6] ML algorithms rely heavily on pattern recognition; they are provided with large volumes of data from different categories, and identify patterns that distinguish one category from another.[7] While ML algorithms are frequently referred to as a form of artificial intelligence (AI), this terminology is poorly defined and potentially misleading, and for the purposes of this report the technology under discussion is referred to as 'machine learning'.

4. Ethem Alpaydin, *Machine Learning* (Cambridge, MA: MIT Press, 2016), p. 16.
5. *Ibid.*, p. 24.
6. *Ibid.,* p. 24.
7. Nick Polson and James Scott, *AIQ: How Artificial Intelligence Works and How We Can Harness its Power for a Better World* (London: Bantam Press, 2018), p. 4.

The theories underlying machine learning are statistical, and therefore ML algorithms deal with probabilistic classifications or predictions, not certainties, and generalisations from particular observations. As Alpaydin (2016) points out, 'there is no guarantee that a machine learning model generalizes correctly – it depends on how suitable the model is for the task, how much training data there is, and how well the model parameters are optimized'.[8] ML algorithms are inherently limited in a number of ways, particularly in their inability to establish causation within any given correlation. The probabilistic nature of the statistical methods involved makes the use of ML algorithms for policing particularly complicated from a legal and regulatory perspective.

The particular application of ML algorithms under discussion can be loosely defined as a form of 'predictive policing'. Predictive policing is defined as 'taking data from disparate sources, analyzing them and then using results to anticipate, prevent and respond more effectively to future crime'.[9] Predictive policing is based on the notion that analytic techniques used by retailers to predict consumer behaviour can be adapted and applied to policing to predict criminal behaviour.[10] To date, predictive policing technology has primarily been used to make predictions of geospatial locations where crime is likely to happen in the near future.[11] Various field trials (both in the UK and the US) have demonstrated that predictive mapping software is in most cases significantly more effective at predicting the location of future crime than existing intelligence-led techniques.[12]

A more recent development in the field of predictive policing is the use of algorithmic risk assessment tools to make predictions related to *individuals*, for instance to identify high-risk offenders whose past behaviour indicates they may be at increased risk of reoffending in the near future. This paper focuses specifically on this latter type of crime prediction, referred to here as 'algorithmic decision-support'. The many other uses of algorithms for law enforcement purposes, such as automated facial recognition, biometric identification and predictive crime mapping, have been discussed at length elsewhere and therefore are not the focus of this report.[13] This report is written from an interdisciplinary perspective. It includes consideration

8. Alpaydin, *Machine Learning*, p. 42.
9. Beth Pearsall, 'Predictive Policing: The Future of Law Enforcement?', *National Institute of Justice Journal* (No. 266, May 2010), p. 16.
10. Jennifer Bachner, 'Predictive Policing: Preventing Crime with Data and Analytics', IBM Center for the Business of Government, Improving Performance Series, 2013, p. 4.
11. Kate J Bowers, Shane D Johnson and Ken Pease, 'Prospective Hot-Spotting: The Future of Crime Mapping?', *British Journal of Criminology* (Vol. 44, No. 5, September 2004), pp. 641–58; Shane D Johnson, 'Repeat Burglary Victimisation: A Tale of Two Theories', *Journal of Experimental Criminology* (Vol. 4, No. 3, 2008), pp. 215–40.
12. Shane D Johnson et al., *Prospective Crime Mapping in Operational Context,* Final Report (London: The Stationery Office, 2007); Pearsall, 'Predictive Policing'; Bachner, 'Predictive Policing'; Walter L Perry et al., *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (Santa Monica, CA: RAND Corporation, 2013).
13. For a broader overview of the UK police's use of data, see Alexander Babuta, 'Big Data and Policing: An Assessment of Law Enforcement Requirements, Expectations and Priorities', *RUSI*

of the implications of algorithmic decision-support tools within policing for compliance with key principles of UK law, although it is not intended to be an exhaustive review.[14] This Whitehall Report focuses in particular on how data protection,[15] human rights law[16] and administrative law principles[17] are engaged by the use of algorithmic decision-support, and the processing of personal data by such systems.

Occasional Papers (September 2017). For a more detailed discussion of the legal constraints governing the authorities' interception and collection of digital data, see Michael Clarke et al., 'A Democratic Licence to Operate: Report of the Independent Surveillance Review', *Whitehall Report*, 2-15 (July 2015).

14. The European Charter of Fundamental Rights is not assessed in this report.
15. 'Data Protection Act 2018 (UK)'.
16. 'Convention for the Protection of Human Rights and Fundamental Freedoms' (European Convention on Human Rights, as amended); 'Human Rights Act 1998 (UK)'.
17. See, for example, Marion Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power', Philosophical Transactions of the Royal Society A (6 August 2018), doi:10.1098/rsta.2017.0359; Andrew Le Sueur, 'Robot Government: Automated Decision-Making and its Implications for Parliament', in Alexander Horne and Andrew Le Sueur (eds), *Parliament: Legislation and Accountability* (Oxford: Hart, 2016).

I. Predictive Policing and Offender Risk Assessment

Previous research examining UK police forces' use of technology has consistently found that the police collect a vast amount of digital data on a daily basis, but often lack the technological capability to interrogate that data with the speed, precision and reliability required to take full advantage of the insights it could provide.[18] To address this problem of 'information overload', police forces are under increasing pressure to develop more effective ways of managing the data they collect: to use analytical tools to identify connections and patterns, predict future risk, and develop proactive policing strategies that direct resources to where they are most needed. While a large proportion of police data will not lend itself to this type of analysis, it is widely accepted that more sophisticated analysis of certain categories of police data (such as crime location data and offending history) would provide police forces with a richer understanding of different crime problems, allowing them to target limited resources more efficiently.[19]

In the UK, predictive policing algorithms have been in use for more than ten years, to identify geospatial locations that are most at risk of experiencing crime and to then pre-emptively deploy resources to where they are most needed – 'predictive crime mapping'.[20] Such technology employs a self-learning algorithm based on an Epidemic Type Aftershock Sequence model – the same type of statistical model that is used to predict earthquake aftershocks.[21] The software uses just three data points as the input for forecasting: crime type; crime location; and crime date and time. Despite the fact that field trials in the UK have found the system to be around twice as likely to predict the location of future crime as traditional intelligence-led techniques,[22] to date few forces in the UK have implemented the practice into daily policing activities. Its

18. See Babuta, 'Big Data and Policing'.
19. See, for example, Her Majesty's Inspectorate of Constabulary and Fire and Rescue Services (HMIC), *PEEL: Police Effectiveness 2016: A National Overview* (London: HMIC, 2017), p. 33; London Assembly Budget and Performance Committee, 'Smart Policing: How the Metropolitan Police Service Can Make Better Use of Technology', 2013; Europol, *European Union Serious and Organised Crime Threat Assessment 2017: Crime in the Age of Technology* (The Hague: Europol, 2017), p. 25.
20. See, for example, Bowers, Johnson and Pease, 'Prospective Hot-Spotting'; Johnson, 'Repeat Burglary Victimisation'.
21. For a mathematical explanation of how this type of model works, see G O Mohler et al., 'Self-Exciting Point Process Modelling of Crime', *Journal of the American Statistical Association* (Vol. 106, No. 493, 2011), pp. 100–08.
22. Kent Police Corporate Services Analysis Department, 'PredPol Operational Review [Restricted and Heavily Redacted]', <http://www.statewatch.org/docbin/uk-2014-kent-police-predpol-op-review.pdf>, accessed 12 August 2018.

use is more widespread in the US, however, where it has been attributed to significant crime reductions in various jurisdictions across California and Georgia, among others.[23]

More recently, a major development in the police's use of data is the implementation of algorithmic tools underpinned by machine learning to aid decision-making relating to individuals, and it is this latter type of analysis that is the focus of this report. In May 2017 it was reported that Durham Constabulary would become the first police force in the UK to implement a ML algorithm to assess the risk of individuals reoffending, in order to support custody decision-making.[24] Durham's Harm Assessment Risk Tool (HART) uses random forest forecasting (a form of supervised machine learning) to classify individuals in terms of their likelihood of committing a violent or non-violent offence over the next two years. The system bases this prediction on 34 input variables, 29 of which relate to past criminal history, and the remainder of which relate to background characteristics, such as the individual's age, gender and postcode.[25] The resulting risk score (high, medium or low) does not solely determine what further action will be taken with respect to the data subject: the risk score is one of a number of factors for the human officer to take into account when making their overall risk assessment.

While this is the first time a UK police force has implemented such a system, non-machine learning offender assessment tools have been in use elsewhere in the criminal justice system for many years. In England and Wales, the Offender Assessment System (OASys) is the national risk and needs assessment tool for adult offenders, and was initially developed between 1999 and 2001, with a new single electronic system being implemented in 2013.[26] OASys incorporates a number of statistical analysis methods to measure an offender's likelihood of reoffending, to identify any risk of serious harm, to develop individual risk management plans, and to measure progress and change over time.[27] Various other actuarial offender assessment tools exist, such as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool, which is used by criminal justice practitioners in several states in the US.[28]

23. See, for example, Bachner, 'Predictive Policing'; Predpol, 'Proven Crime Reduction Results', 2018, <http://www.predpol.com/results/>, accessed 12 August 2018.
24. Chris Baraniuk, 'Durham Police AI to Help with Custody Decisions', *BBC News*, 10 May 2017.
25. Sheena Urwin, 'Algorithmic Forecasting of Offender Dangerousness for Police Custody Officers: An Assessment of Accuracy for the Durham Constabulary Model', unpublished thesis, University of Cambridge, 2016, <http://www.crim.cam.ac.uk/alumni/theses/Sheena%20Urwin%20Thesis%2012-12-2016.pdf>, accessed 15 August 2018.
26. Robin Moore, 'A Compendium of Research and Analysis on the Offender Assessment System (OASys) 2009–2013', National Offender Management Service, Ministry of Justice Analytical Series, July 2015, p. 3.
27. *Ibid.*, p. 4.
28. Tim Brennan, William Dieterich and Beate Ehret, 'Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System', *Criminal Justice and Behavior* (Vol. 36, No. 1, January 2009), pp. 21–40.

The HART machine learning forecasting tool is currently being used for one specific purpose: to aid the assessment of offenders' eligibility to participate in Durham Constabulary's 'Checkpoint Programme', a voluntary out of court disposal designed to reduce reoffending by addressing the underlying factors causing individuals to engage in crime.[29] Only the 'medium'-risk cohort are eligible to participate in the Checkpoint scheme. In practice, this means that the difference between being assessed as 'high' or 'medium' risk could mean the difference between being charged with an offence and processed through the normal court system, and avoiding this altogether if accepted into the Checkpoint scheme. Other forces are in the process of developing similar systems, and it is anticipated that the use of ML algorithms for policing will become more commonplace in the coming years and be applied to a wider range of decision-making functions.[30] While such tools are currently being used for limited policing purposes, there is potential for the technology to do much more, and the lack of a regulatory and governance framework for its use is concerning.

Serious concerns have been raised over the use of algorithms for criminal justice purposes. For instance, there has been much discussion of the issue of indirect racial bias in models that predict offending.[31] While models such as HART and COMPAS do not include a variable for race, they do include postcode. In segregated areas, postcode can function as a proxy variable for race or community deprivation, thereby having an indirect and undue influence on the outcome prediction.[32] A ProPublica investigation conducted in 2016 into the COMPAS system found that black defendants were almost twice as likely to be deemed at risk of offending than white defendants, though this analysis is disputed by the system's developer, Northpointe (now 'Equivant').[33]

However, where it can be demonstrated that a prediction has been unduly influenced by some form of bias, in many cases the algorithm itself may not be the source of the bias. As mentioned previously, an algorithm is merely a calculation, whereby data has some operation performed on it according to a sequence of instructions in order to produce an outcome. If the input data is biased, the algorithm may replicate (and in some cases amplify) the existing biases inherent

29. For more information, see Durham Constabulary, 'Checkpoint', <https://www.durham.police.uk/information-and-advice/pages/checkpoint.aspx>, accessed 12 August 2018.
30. A recent example is Norfolk Constabulary's ongoing trial of an algorithm designed to assess the 'solvability' of burglary cases using 29 predictor variables in order to inform prioritisation decisions regarding which cases should be referred for further investigation. See Norfolk Constabulary, 'Response to Inaccurate Story Published', 2 September 2018, <https://www.norfolk.police.uk/news/latest-news/04-09-2018/response-inaccurate-story-published>, accessed 7 September 2018.
31. See, for example, Julia Angwin et al., 'Machine Bias', *ProPublica*, 23 May 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>, accessed 15 August 2018.
32. Marion Oswald et al., 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental" Proportionality', *Information & Communications Technology Law* (Vol. 27, No. 2, 2018), pp. 223–50.
33. Angwin et al., 'Machine Bias'.

in the dataset.[34] Any risk assessment or forecasting method – whether or not it incorporates an algorithm – can produce biased outcomes. For instance, the Metropolitan Police Service's (MPS) 'Gangs Matrix', a forecasting tool used to identify priority individuals at risk of engaging in gang violence, has been criticised by Amnesty International for being 'racially biased' and 'unfit for purpose'.[35] Police analysts interviewed for previous RUSI research described that the calculation of risk scores for the Gangs Matrix was almost entirely manual, explaining that it was an arduous and time-consuming process that 'could take weeks'.[36] Information provided by the Metropolitan Police Service to the authors under a Freedom of Information request also confirmed that the Matrix does not make use of an algorithm, stating 'the automated allocation of harm scores for the MPS gang violence matrix uses a scoring methodology developed internally in the MPS. It is not an algorithm as such and does not use any third-party software to calculate these scores'.[37]

Examples such as this exemplify the difficulties of using police data to make predictions (algorithmic or otherwise), data which by its nature is incomplete and often unreliable. Perhaps the most significant factor in this regard is the problem of under-reporting. The vast majority of crimes go unreported, meaning that police data provides only an incomplete snapshot of the phenomenon under investigation. This is particularly problematic when using arrest data to predict future crime, because arrest data is not necessarily representative of how crime is distributed in time and space. For instance, if a particular neighbourhood has been disproportionately targeted by police action in the past, an algorithm may incorrectly assess that neighbourhood as being at increased risk of experiencing crime in the future. Acting on the predictions may then cause that neighbourhood to again be disproportionately targeted by police action, creating a feedback loop whereby the predicted outcome simply becomes a self-fulfilling prophecy. Similarly, if a particular minority or social group has been disproportionately targeted by police action, the police may predict individual members of that group to be at increased risk of offending in the future. In reality, they may be no more likely to *offend* – just more likely to be *arrested*.

Despite the many difficulties associated with implementing ML algorithms for policing, these tools offer significant potential benefits. In particular, developing methods to more accurately predict reoffending risk has the potential to result in tangible reductions in overall crime. It has long been observed in the crime analysis literature that a small number of individuals are responsible for the majority of all crime.[38] According to the most recent statistics released by

34. For further discussion, see Alexander Babuta, 'Innocent Until Predicted Guilty? Artificial Intelligence and Police Decision-Making', *RUSI Newsbrief* (Vol. 38, No. 2, March 2018).
35. Amnesty International, 'Trapped in the Matrix: Secrecy, Stigma, and Bias in the Met's Gangs Database', May 2018.
36. Author interviews with four police analysts, London, 18 April 2017. See Babuta, 'Big Data and Policing', p. 23.
37. Information provided by the Metropolitan Police Service to Alexander Babuta under Freedom of Information Request 2018070000993, email, 20 August 2018.
38. See, for example, Graham Farrell, 'Crime Concentration Theory', *Crime Prevention and Community Safety* (Vol. 17, No. 4, November 2015), pp. 233–48; David Weisburd, 'The Law of

the Ministry of Justice, for individuals processed through the justice system in England and Wales during the time period July–September 2016, the overall proven reoffending rate was 29.5% (41.7% for juveniles). This means that of all offenders who were cautioned, received a non-custodial conviction at court or were released from custody between July and September 2016, 29.5% committed a proven reoffence within a year.[39] As repeat offending accounts for a large proportion of all crime, developing a system to more accurately identify those most likely to offend again would allow police forces to consistently make appropriate custody decisions and target preventive interventions to those who need them most. Incorporating statistical methods into the offender risk assessment process has the potential to improve the accuracy of these judgements: research across a wide range of fields has consistently shown that statistical forecasting in combination with professional judgement is typically more accurate in predicting future risk than clinical judgement alone.[40]

The police exist to prevent crime, arrest offenders and protect victims. If statistical forecasting tools have the potential to improve citizens' safety and security, the argument could be raised that authorities have a social obligation to trial these new methods, and to innovate and use technology in a way that is proportionate to the potential impact on human rights. The existence of a strict legal duty in this regard is likely to be highly context-specific, and may depend upon whether algorithmic methods can be demonstrated to result in a more effective investigation or policing method (and one that does not disproportionally infringe on individual rights), and importantly whether *not* using this method could be considered an obvious and significant failing resulting in the infringement of individual rights.[41]

Crime Concentration and the Criminology of Place', *Criminology* (Vol. 53, No. 2, May 2015), pp. 133–57; Dan Ellingworth, Graham Farrell and Ken Pease, 'A Victim is a Victim is a Victim?: Chronic Victimization in Four Sweeps of the British Crime Survey', *British Journal of Criminology* (Vol. 35, No. 3, 1995), pp. 360–65.

39. Ministry of Justice, 'Proven Reoffending Statistics Quarterly Bulletin, July 2016 to September 2016', 26 July 2018.
40. See, for example, Paul Meehl, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (Minneapolis, MN: University of Minnesota Press, 1954); Stefanía Ægisdóttir et al., 'The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction', *The Counseling Psychologist* (Vol. 34, No. 3, 2006), pp. 341–82; Daniel Kahneman, *Thinking, Fast and Slow* (New York, NY: Macmillan, 2011).
41. The court held in Commissioner of Police of the Metropolis v DSD & Anor [2018] UKSC 11 that, under Article 3 of the European Convention on Human Rights (which prohibits torture, or inhuman or degrading treatment), the state was obliged to conduct an effective investigation into crimes involving serious violence to persons, and that this should lead to 'the enhancement of standards and the saving of resources' [Lord Kerr]; the decision in R (DSD & Others) v The Parole Board of England and Wales [2018] EWHC 694 (Admin) might suggest that, should algorithmic assessments be demonstrated in particular circumstances to be more accurate or effective than other forms of assessment, then they should be proactively sought out.

While security technology is often perceived in dystopian terms, it is important to bear in mind that data-driven policing technologies have the potential to uphold rights as well as challenge them.[42] As mentioned previously, existing manual risk assessment methods used by the police have been criticised for being discriminatory and unfair. ML algorithms have the potential to remove certain types of bias from these systems, providing a more objective assessment that is unaffected by extraneous factors such as the decision-maker's mood, the weather, the time of day, and the multitude of additional factors that cloud a human decision-maker's judgement.[43] Similarly, computer-based tools may be able to provide clearer and more comprehensive reasoning for why they arrived at a certain prediction, whereas 'clinical' risk forecasting methods could be criticised for relying too heavily on the professional judgement of the human decision-maker (this is discussed further in Chapter III).

The use of ML algorithms to support police decision-making is in its infancy, and the potential outcomes of these tools are still poorly understood. The incorporation of machine learning into the criminal justice system may have unintended or indirect consequences that are difficult to anticipate.[44] The potential benefits of these tools are likewise yet to be fully established. It is clear that new technologies must be trialled in a controlled way in order to assess their effectiveness, before being rolled out in an operational environment where they are likely to have a significant impact on citizens' lives. However, there is currently no clear framework in place for how the police should conduct such trials. What is needed going forward are clear codes of practice to enable police forces to trial new algorithmic tools in a controlled way in order to establish whether or not a certain tool is likely to improve effectiveness of a certain policing function.

It is essential that such experimental innovation is conducted within the bounds of a clear policy framework, and that there are sufficient regulatory and oversight mechanisms in place to ensure fair and legal use of technologies within a live policing environment. Numerous recent reports have expressed concerns over the lack of policy and regulation governing the authorities' use of ML algorithms. A recent inquiry conducted by the House of Lords Select Committee on Artificial Intelligence 'encountered several serious issues associated with the use of artificial intelligence that require careful thought, and deliberate policy, from the Government', stressing the need for 'clear principles for accountability and intelligibility'.[45] This is of particular importance in the context of policing and criminal justice.

The chapters that follow seek to address the most pressing legal and ethical concerns of implementing algorithmic tools in a police setting, in order to inform the development of a future policy framework for the use of ML algorithms in policing.

42. Le Sueur, 'Robot Government'.
43. For further discussion, see Daniel Kahneman et al., 'Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making', *Harvard Business Review*, October 2016, pp. 36–43.
44. See, for example, Edwards and Veale, 'Slave to the Algorithm?'.
45. House of Lords Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?', HL Paper 100, Report of Session 2017–19, 16 April 2018, pp. 95, 98.

II. Discretion and Accountability

Those working in the criminal justice field are required to deploy discretion on a daily basis when exercising powers granted by statute or common law, such as decisions requiring 'reasonable' opinion or judgements of public interest. Appropriate application of discretion within policing and criminal justice ensures that the merits of individual cases are considered, individual rights are protected and rules are not applied unbendingly. Furthermore, discretion leads to accountability; the exerciser of the discretion can be held responsible for any perceived injustice or failure to deal properly with a case.

Exercise of discretion is constrained by law, codes of practice, guidance and policy. There are already many elements of a police officer's decision-making toolkit which constrain their discretion, such as the National Decision Model,[46] the principles of Management of Police Information (MoPI),[47] the framework for community resolutions under the Legal Aid, Sentencing & Punishment of Offenders Act 2012,[48] and codes of practice linked to the Police & Criminal Evidence Act 1984[49] and other legislation. In addition to these constraints on officer discretion, human officers do not make decisions in perfect conditions, but under various environmental pressures associated with responding to crime. They may also, for instance in confrontational operational situations, have limited time to take all relevant factors into account when deciding whether or not to take a particular action, or any action at all. Algorithms, appropriately deployed, can 'package' certain factors (or groups of factors) in a way that has the potential to improve the efficiency and efficacy of human decision-making, without the need for fully automated decisions.

The introduction of ML algorithms into decision-making processes involving an element of discretion may have a number of intended or unintended consequences. Where algorithms produce an output consisting of a classification or prediction, for instance in relation to risk, consideration has to be given to the relevancy of such classification or prediction to the overall

46. College of Policing Authorised Professional Practice, 'National Decision Model', last modified 15 December 2014, <https://www.app.college.police.uk/app-content/national-decision-model>, accessed 17 August 2018.
47. College of Policing Authorised Professional Practice, 'Management of Police Information', last modified 24 May 2018, <https://www.app.college.police.uk/app-content/information-management/management-of-police-information/>, accessed 17 August 2018.
48. College of Policing Authorised Professional Practice, 'Out of Court Disposals Framework', 2 July 2018, <https://www.app.college.police.uk/app-content/prosecution-and-case-management/justice-outcomes/#out-of-court-disposals-framework>, accessed 6 September 2018.
49. UK Government, 'Home Office Police and Criminal Evidence Act 1984 Codes of Practice', 21 August 2018, <https://www.gov.uk/guidance/police-and-criminal-evidence-act-1984-pace-codes-of-practice>, accessed 6 September 2018.

decision in question, and how an algorithmic output might affect the decision-making process *in practice*. Focus groups conducted with a range of practitioners from various agencies involved in policing and criminal justice revealed differing perspectives surrounding the implementation of forecasting algorithms in the field and their potential impact on discretion and professional judgement. Opinions on this ranged from 'our professional judgement is being taken out of the equation altogether' to the use of such algorithms being 'a very supportive friend to discretion'.[50]

In a policing environment, discretion is informed by experience and knowledge of complex contextual factors, and it is challenging to accurately 'datafy' all relevant information to be processed by an algorithm without losing that very context.[51] Fundamental principles of administrative law require public sector decision-makers to take all relevant factors into account – whether algorithmically generated or otherwise – and to assess whether the question or decision is one for which the algorithm was designed.[52] The above-mentioned research highlighted the perceived importance of an officer's 'gut instinct' and 'experience' in judging the 'best outcome',[53] comments that reflect an importance given to considering individual circumstances as they arise. However, there is a risk that individuals may consider an algorithmic assessment as relevant to their decision or action only when it aligns with their professional judgement or intuition, potentially leading to a confirmation bias effect.[54] Further research is needed to explore how practitioners' judgement is influenced by the implementation of an algorithmic tool in practice. It would seem premature to proceed with large-scale deployment of such technology without having established such an evidence base.

The impetus for the introduction of algorithmic tools tends to come from senior management or those with a policy role, who wish to take advantage of innovations in data analytics, improve on existing imperfect manual risk assessments and even constrain possibly cloudy exercise of discretion within current processes. These objectives have a high likelihood of failure if disconnected from the views of the officer on the ground, or where theoretical benefits do not take into account the difficulties of implementation. The influence of the algorithmic output – whether forecast, prediction, classification or otherwise – cannot be assessed in isolation, but

50. Research presented by Chief Superintendent David Powell, Hampshire Constabulary, to private audience at RUSI, London, 2 May 2018; David Powell, Christine Rinik and Marion Oswald, 'Policing, Algorithms and Discretion: A Critical Commentary on the Proposed Expansion of Algorithm-Assisted Decision-Making Within Policing Underpinned by Comments of Prospective Users', Institute of Advanced Legal Studies Information Law and Policy Centre Annual Conference 2018 (forthcoming).
51. Min Kyung Lee, 'Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management', *Big Data & Society* (Vol. 5, No. 1, January–June 2018), pp. 1–16.
52. Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector'.
53. Research presented by Chief Superintendent David Powell, Hampshire Constabulary, to private audience at RUSI, London, 2 May 2018.
54. Christin, 'Algorithms in Practice'.

should be considered in the 'messy, socio-technical context'[55] in which it will be implemented; this is referred to by Yeung as the 'sociotechnical assemblage' in which the algorithm operates.[56] Police forces implementing algorithmic tools should maintain a clear distinction between the algorithmic decision-support element and the human discretion element of the overall process being considered. The purpose of the algorithmic output should be to assist and complement officer discretion, rather than dictating the overall decision or outcome. It is appreciated that this may be challenging to achieve in environments that are highly time and resource constrained. However, this further suggests the need for robust research and testing to ascertain the implications of this immature technology before further significant deployments are undertaken.

Algorithmic tools in the policing and criminal justice field are commonly described as 'decision-support', the output becoming one of the factors that a human officer is required to consider when making a decision. This output may replace an assessment that was previously made by other means, for instance a manual risk assessment, or it may represent a new, altered or previously un-assessed output in the process. It can be argued that the human officer remains in control of the final assessment and can accept or reject the algorithmic output/recommendation as appropriate, after consideration of other relevant factors. There remains a risk, however, of officers coming to rely unthinkingly on an algorithmic result – and thereby illegally 'fettering their discretion' in practice. Furthermore, many decisions in policing depend on concepts such as reasonableness and opinions of necessity. Relying only upon a probabilistic algorithmic output could inadvertently be changing the question to be answered.[57]

Within policing, there is little research into how the introduction of an algorithmic decision-support tool will affect police decision-making in practice, and thus what frameworks are required to ensure both that the required exercise of discretion – and thus accountability – is preserved, and that relevant algorithmic outputs are in fact taken into account. In addition, Part 3 of the Data Protection Act 2018[58] provides safeguards for individuals in relation to significant decisions based solely on automated processing.[59] Although, as the Information Commissioner's Office comments, 'solely automated decision-making that leads to an adverse outcome is rarely used in the law enforcement context',[60] law enforcement bodies will now need to show that a human has provided meaningful review of the decision in order to demonstrate that the decision is *not* the product of automated processing.

55. Michael Veale, Max Van Kleek and Reuben Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making', paper presented at the CHI Conference on Human Factors in Computing Systems, Montréal, QC, Canada, 21–26 April 2018.
56. Yeung, 'Algorithmic Regulation'.
57. Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector'.
58. 'Data Protection Act 2018 (UK)'.
59. *Ibid*., ss 49–50.
60. Information Commissioner's Office, 'Guide to Law Enforcement Processing', <https://ico.org.uk/for-organisations/guide-to-law-enforcement-processing-part-3-of-the-dp-act-2018/individual-rights/right-not-to-be-subject-to-automated-decision-making/>, accessed 17 August 2018.

The practical implications of these issues may be highly contextual. We could find, for instance, that where high levels of accuracy are claimed (based on the algorithm's performance using training data), few officers will be willing to challenge or contradict the course of action that would follow from the algorithmic prediction, even if this is only one factor that should be considered. In some circumstances, such as where there are high levels of demand for the intervention and resources and/or time are limited, this may be a desirable or necessary outcome; in others – such as where the tool is subject to some extrinsic limit on its potential accuracy or relevance because of external factors – much less so.[61] On the other hand, if the tools are only being used in an 'advisory' capacity, individuals may accept the tool's deployment only when it aligns with their personal worldview and pre-existing assumptions; otherwise they may find ways in practice of minimising the impact of the tool's output (such as foot-dragging, gaming and open critique).[62] A clear process – and one which is cognisant of wider time and resource constraints – for resolving disagreements when professional judgement and the algorithm come to different conclusions would therefore appear to be essential.

All 'big data' and algorithmic analysis relies upon a human process of interpretation and inference, in respect of both inputs and outputs.[63] The above-mentioned focus groups conducted with a range of criminal justice practitioners documented concerns among officers over the legal status of algorithms within policing, affecting their confidence in use for policing purposes.[64] The research also revealed a desire among officers to understand what they are working with and how an algorithmic tool operates in order to build confidence. A tool which provided explanations, choices or options to the user of suitable granularity for the context, and related to the question or issue under consideration, could contribute to ensuring that the human officer becomes an active participant in the tool's use and development – a 'constructive critic' rather than an uninformed and potentially unengaged user.[65] Linked with this is the use of language such as 'AI' and 'algorithm', potentially giving such tools an aura of mystery and infallibility. If such tools were instead described in terms of what they actually did, and how they fit into the overall decision-making process, such reframing might help to improve practitioners' reactions to, and perceptions of, these tools.

The difficult question of whether algorithmic assessments are 'better' than current risk assessment frameworks is a recurring one. Comparisons will hardly ever be like-for-like. The aim of the introduction of the algorithm may have been to improve efficiency or decrease the burden

61. Jake M Hofman, Amit Sharma and Duncan J Watts, 'Prediction and Explanation in Social Systems', *Science* (Vol. 355, No. 6324, 3 February 2017), pp. 486–88.
62. Christin, 'Algorithms in Practice'.
63. Babuta, 'Big Data and Policing', pp. 4–5.
64. Research presented by Chief Superintendent David Powell, Hampshire Constabulary, to private audience at RUSI, London, 2 May 2018; Focus group of criminal justice practitioners and policymakers, RUSI, London, 2 May 2018.
65. See also Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector'.

of manual assessments, rather than to 'bring in shocking new insights'.[66] In addition, although there will inevitably be a focus on the term 'accuracy', true accuracy may be impossible to assess. Concerns may not materialise due to interventions, precautionary measures or other random factors. In the context of crime prevention, it is particularly difficult to demonstrate 'accuracy' of a forecasted prediction when that prediction leads to an intervention that prevents the forecasted outcome from happening. Some scenarios do not lend themselves to double-blind testing, for instance those involving serious offenders. As the Authorised Professional Practice risk principles state, 'By definition, [operational] decisions involve uncertainty, ie, the likelihood and impact of possible outcomes cannot be totally predicted, and no particular outcome can be guaranteed'.[67] The principles state further that 'assessments of decisions should concentrate on whether they were reasonable and appropriate for the circumstances existing at the time. If they were, the decision maker should not be blamed for a poor outcome'. It is likely that criminal justice policy frameworks will need to be adapted to support such a reasonableness assessment in an algorithmically-assisted environment.

There is yet little evidence to address how the introduction of an algorithmic-support tool might affect the discretionary elements of decision-making within policing and criminal justice. In the design of algorithmic tools and their implementation into decision-making processes, consideration must be given to the risk of impeding that discretion: 'either because all of the factors relevant to the decision cannot be included, or required elements of the decision itself cannot be appropriately codified into, or by, the algorithm'.[68]

66. Veale, Kleek and Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making'.
67. College of Policing Authorised Professional Practice, 'Risk', last modified 23 October 2013, <https://www.app.college.police.uk/app-content/risk-2/risk/>, accessed 17 August 2018.
68. Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector'.

III. Transparency and Intelligibility

A common criticism of certain types of ML algorithms is the opacity of their data processing methods. Such tools are often referred to as 'black boxes' that digest large volumes of data, far beyond the capacity of a human analyst, and produce an outcome without being able to show their working. Moreover, subjects may not be aware that a decision related to them has been made with the help of an algorithm, undermining the transparency of the overall justice process. When deploying ML algorithms in a criminal justice setting, clear processes should be established to ensure that an acceptable level of transparency and intelligibility is designed into the system from the outset. This is essential for external observers and the data subject themself to be able to challenge the process by which an outcome was reached, and to ensure that such tools are being used in accordance with the requirements of the relevant data protection legislation and principles of accessibility and natural justice under the Human Rights Act 1998.[69]

Technical Transparency

Transparency, intelligibility and auditability are three distinct but closely related issues and are often conflated as 'algorithmic transparency'. Making a system transparent does not necessarily entail that it is intelligible or auditable, and merely publishing an algorithm's source code does not mean that non-experts will be able to decipher and interpret that code. Transparency can be understood as the visibility and accessibility of the algorithm's source code and operating parameters. Intelligibility ('explainability') refers to the degree to which the code or disclosed information sufficiently explains how the model operates in practice appropriate to the context in which the algorithm is deployed. Auditability refers to the ability of a human observer to retroactively examine exactly how the tool arrived at a certain decision. It is important to distinguish further between transparency of the algorithmic tool itself, and transparency of the wider organisational decision-making process that incorporates the use of an algorithm.

On the issue of technical transparency, the report from the House of Lords Select Committee on Artificial Intelligence concluded that:

> Based on the evidence we have received, we believe that achieving full technical transparency is difficult, and possibly even impossible, for certain kinds of AI systems in use today, and would in any case not be appropriate or helpful in many cases … We believe it is not acceptable to deploy any artificial

69. See 'Data Protection Act 2018 (UK)' and 'Human Rights Act 1998 (UK)'.

> intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decisions it will take.[70]

Given technological limitations, it is perhaps unrealistic in all circumstances to expect machine learning decision-support systems to be able to generate full explanations for the predictions they make. The requirement for an algorithm to generate an explanation will be highly dependent on the context in which the algorithm is used and the type of decision to which it contributes. When making decisions that are likely to have a significant impact on the individual in question, it remains the responsibility of the police practitioner, rather than the machine, to be able to provide a satisfactory explanation of the various factors that led to a certain decision being made, of which the algorithmic prediction may be but one component. Furthermore, in cases where the human decision-maker's judgement contradicts that of the algorithm, they must be able to provide coherent alternative reasoning for why they arrived at a conclusion different from the predicted outcome.

While the term 'black box' is frequently used to refer to machine learning, machine learning methods vary considerably in the level of transparency they are able to provide. Durham Constabulary's HART system uses random forest forecasting, a form of supervised machine learning which uses 509 classification and regression trees to produce 509 individual predictions, with the prediction that occurs most often being used as the overall risk score.[71] Supervised machine learning systems are 'trained' using historic datasets, allowing the algorithm to create a model of how to map inputs to outputs, which can then be deployed predictively and applied to unfamiliar, new data. It is possible to retroactively deconstruct the model, and to examine the structure of each individual tree in the forest to understand how each tree has chosen to model the data. In addition, the output of the random forest forecast computes various measures of variable importance, indicating for each variable the weighted bearing it has on the model's prediction, and thereby the impact expected on such output if the variable were removed from the model.

In contrast to random forest forecasting, other forms of machine learning provide far less information in terms of how the model computes an outcome. For instance, neural networks, another common machine learning technique, frequently employ 'unsupervised' algorithms using deep learning methods to make predictions when presented with unfamiliar data. In contrast to supervised methods such as random forests, unsupervised algorithms are not trained using historic data. They are provided with a large volume of unfamiliar input data and instructed to detect patterns and trends in the data with minimal additional guidance. Rather than the classification and regression method employed by supervised algorithms, unsupervised algorithms use a method known as 'clustering', whereby input variables are grouped in terms

70. House of Lords Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?', p. 128.
71. Berk and Hyatt, 'Machine Learning Forecasts of Risk to Inform Sentencing Decisions'; Berk et al., 'Forecasting Domestic Violence'.

of their common features. This typically means that it is impossible to identify which individual variables have influenced the output's predictions.

It would seem, then, that the 'black box' predicament applies to certain types of algorithms but not others. In the case of machine learning techniques such as random forest forecasting, it is possible to identify which factors led to a certain risk prediction being made, thereby offering a reasonable degree of auditability, provided the information is given on actual results and not just during training. While it may prove technologically impossible to develop machine learning criminal justice algorithms that are entirely 'transparent' in their data processing methods, as a minimum it should be possible to retroactively deconstruct the model in this way in order to identify which factors led to a certain decision being made (counterfactuals and causal reasoning appear to also hold potential here).[72] This is essential in order to verify that certain factors did not indirectly or unduly affect the decision.

Given the wide variation in the levels of transparency different machine learning methods are able to provide, a future regulatory framework should establish minimum standards of technical transparency for algorithms used to support police decision-making relating to individuals.

To date, auditing of algorithms has been restricted somewhat by a lack of technical capability and computational resources.[73] However, this field is rapidly advancing, with recent studies demonstrating that it is possible to audit 'black box' algorithms for indirect influence 'from the outside', or in other words, without requiring access to the model's internal code.[74] There is now a growing industry of consultancy firms that claim to provide algorithmic auditing 'as a service' in order to assess existing algorithms for accuracy, bias, consistency, transparency, fairness and timeliness[75] (although it seems unclear how such services would apply to particular legal or practical contexts). As noted in a recent report from the Information Commissioner's Office, methods of data analytics such as Natural Language Generation (NLG) could be used to create coherent narratives in plain English explaining how an algorithm arrived at its prediction, and visualisation systems could be implemented to provide visual explanations of the decision-making process.[76]

72. Judea Pearl and Dana Mackenzie, *The Book of Why* (London: Allen Lane, 2018).
73. Brent Mittelstadt, 'Automation, Algorithms, and Politics: Auditing for Transparency in Content Personalization Systems', *International Journal of Communication* (Vol. 10, 2016), pp. 4991–5002.
74. Philip Adler et al., 'Auditing Black-Box Models for Indirect Influence', *Knowledge and Information Systems* (Vol. 54, No. 1, January 2018), pp. 95–122.
75. See, for example, O'Neil Risk Consulting and Algorithmic Auditing, <http://www.oneilrisk.com/>, accessed 11 June 2018; Natasha Lomas, 'Accenture Wants to Beat Unfair AI with a Professional Toolkit', *TechCrunch*, 9 June 2018, <https://techcrunch.com/2018/06/09/accenture-wants-to-beat-unfair-ai-with-a-professional-toolkit/?guccounter=1>, accessed 7 July 2018.
76. Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection', Version 2.2, March 2017, p. 87.

A limitation of the 'auditability' approach is that transparency may be retrofitted into the system rather than being built into the design from the outset. There is a risk that the fairness of the algorithm will only be scrutinised once a subject has put forward a challenge or initiated an appeals process. However, as new data is added to the system, the algorithm will change and develop, so there will always be a requirement to regularly review and challenge the fairness of the model. A preferred approach that should be prioritised in future research would be to incorporate built-in NLG functionality into the algorithm itself, so that a plain English explanation is provided alongside every risk prediction in a format that is appropriate to the context, the particular decision-maker and the application of the decision. However, such explanations can only go so far, and a range of other factors will need to be taken into account when interpreting the reliability and validity of the algorithmic prediction, such as whether the data on which the algorithm was trained matches the circumstances of the current situation, and the probabilistic margin of error used in calculating the likelihood of the predicted outcome.

Transparency of Process

Technical transparency alone is not sufficient to ensure intelligibility of the decision-making process to the data subject. If a subject wishes to challenge a decision that has been made with the help of an algorithm, it must be possible for the subject to scrutinise the algorithm's prediction and be provided with an intelligible summary, in plain English, of the factors the model took into account, and how these factors influenced the prediction. As noted in a recent ACM survey examining 'black box' machine learning models, 'there is no work that seriously addresses the problem of quantifying the grade of comprehensibility of an explanation for humans, although it is of fundamental importance'.[77] In addition to ensuring a sufficient degree of transparency concerning the input data and how it is processed, it is essential to also ensure the explanation provided is comprehensible to non-experts.

The Council of Europe study on algorithms and human rights states that 'the mere potential of [algorithms'] use raises serious concerns with regard to Article 6 of the ECHR [European Convention on Human Rights] and the principle of equality of arms and adversarial proceedings as established by the European Court of Human Rights'.[78] Furthermore, the duty to give reasons within administrative law at least requires a person to be informed of the gist of the case against them so that they can present their views in a more informed way to the decision-maker.[79] Similarly, an accused person must be given reasons post-decision where important rights are at stake. This constitutes a critical challenge if one of the factors in the decision has been produced by an algorithm that the decision-maker themself does not fully understand.

With this in mind, in addition to technical transparency, the wider organisational processes that incorporate the use of an algorithm must also be sufficiently transparent. While Section 49 of

77. Riccardo Guidotti et al., 'A Survey of Methods for Explaining Black Box Models', *ACM Computing Surveys* (Vol. 51, No 5, 2018), p. 36.
78. Council of Europe, 'Algorithms and Human Rights', Council of Europe Study DGI (2017) 12, p. 11.
79. R v Secretary of State for the Home Department ex p. *Fayed* [1997] 1 All ER 763.

the Data Protection Act 2018 ensures that individuals must not be subjected to solely automated decision-making in a law enforcement context,[80] this requirement alone may not represent sufficient protection. While decision systems incorporating an algorithmic tool may not fall into the category of fully 'automated' decision-making, the potential impact of the algorithm's prediction could be significant, and in some cases could lead to certain decisions being made that would not have been made without the use of an algorithm. For this reason, there is a need for clear legal and professional guidelines on how to present algorithmic predictions to those about whom the prediction is made.

Furthermore, in addition to providing assurances to the data subject that they have not been unfairly treated, ensuring a degree of algorithmic transparency is also essential to ensure that practitioners using the tool have sufficient confidence in the reliability of the system's predictions and are able to consider the relevance of the input factors and of the output itself. Focus groups with practitioners revealed a strong desire to understand what data was being used by the algorithm.[81] This would build officers' confidence in the predictions, making them more likely to want to use it.

In addition to knowing what data is used by the algorithm, it is essential that those using the system sufficiently understand the inherent limitations of the technology. Most significantly, ML algorithms are designed to detect correlation, but are unable to assess whether a correlation is associated with any type of causation. For this reason, human interpretation and review is essential to ensure that bias is not inadvertently introduced into the system, or that irrelevant considerations – in the form of untrustworthy outputs or input factors that have little connection to the purpose at hand – are not taken into account in the decision-making process.

Furthermore, methods such as random forest forecasting (the form of machine learning employed by Durham Constabulary's HART system) are able to make decisions about what types of error are more 'acceptable' than others. The model considers the cost ratios of different types of error, and then accounts for these calculations when the model is built. Durham's HART model is predisposed to favour false positives over false negatives, as it is deemed by the force more harmful to underestimate the risk posed by high-risk offenders than to overestimate the risk posed by low-risk offenders.[82] However, for the data subject, this could mean being labelled as high-risk by an algorithm when a human officer may have judged them as lower-risk. The fact that the model is predisposed to overestimate risk must therefore be made clear to the decision-makers required to use the algorithm, as factors such as these must be taken into account when interpreting the system's predictions.

80. See 'Data Protection Act 2018 (UK)'.
81. Research presented by Chief Superintendent David Powell, Hampshire Constabulary, to private audience at RUSI, London, 2 May 2018; Focus group of criminal justice practitioners and policymakers, RUSI, London, 2 May 2018.
82. Urwin, 'Algorithmic Forecasting of Offender Dangerousness for Police Custody Officers', pp. 30–31.

Finally, there may be situations where access to an algorithm's source code is required in an evidential context, and therefore this may need to be expressly stated in private sector procurement contracts. Given the transparency requirements outlined above, it is suggested that procurement contracts with private sector suppliers of police technology should be explicit in requiring access to an algorithm's source code. The supplier should also be required to provide an expert witness who can give evidence concerning the algorithm's operation if needed, in a court setting for instance. Moreover, previous cases have demonstrated the potential controversies arising from partnerships between police forces and private technology companies, particularly when new technological initiatives are pursued without public consultation or debate. The recent case of the piloting of predictive policing software developed by Palantir Technologies in New Orleans without the knowledge of public officials demonstrates the reputational risk to police forces of developing partnerships with technology companies without first assessing the views of the wider public.[83] The creation of interdisciplinary local ethics boards that scrutinise and assess the implementation of algorithmic tools at the local force level may go some way towards mitigating this risk.

Principles of transparency and intelligibility should form a core component of any future policy framework governing the use of algorithms for policing, and there is a need for clear governance and oversight mechanisms to ensure adherence to these principles.

83. Ali Winston, 'Palantir has Secretly Been Using New Orleans to Test its Predictive Policing Technology', *The Verge*, 27 February 2018.

IV. Fairness and Bias

Use of machine learning will need to not only advance efficiency and accuracy but also be seen to be fair in its recommendations in order to be of value to the police and wider criminal justice system. As Her Majesty's Inspectorate of Constabulary and Fire and Rescue Services articulated in its most recent police inspection (PEEL) report: 'Justice is fairness. Public confidence in the police depends on their being conspicuously fair to everyone'.[84] This section focuses on bias, defined here as the use or choice of data or algorithmic output that results, directly or indirectly, in an unlawful or unethical discriminatory effect on an individual, or where the data selected is unrepresentative.[85]

It should be noted that issues of bias, the use of personal data within algorithmic systems, and the predictions or classifications about an individual delivered by an algorithm may engage Article 14 of the European Convention on Human Rights (ECHR), the right to freedom from discrimination,[86] in conjunction with Article 8 of the ECHR, the right to respect for private life.[87] Furthermore, the new Data Protection Act 2018[88] sets out six principles which govern the processing of personal data when such activity is carried out in connection with law enforcement purposes. These principles address such issues as the requirement that the processing be lawful and fair,[89] that the purpose of the processing be specified and legitimate,[90] that the data be kept no longer than necessary[91] and that the data be processed in a secure manner.[92] In the context of the current discussion it is particularly important to note that the probabilistic nature of algorithmic outputs could give rise to questions of 'accuracy' and 'relevance' which

84. Her Majesty's Inspectorate of Constabulary and Fire and Rescue Services, 'State of Policing: The Annual Assessment of Policing in England and Wales 2017', p. 22.
85. European Union Agency for Fundamental Rights (FRA), *#BigData: Discrimination in Data-Supported Decision Making* (Vienna: FRA, 2018).
86. European Convention on Human Rights, Article 14, Prohibition of discrimination.
87. European Convention on Human Rights, Article 8, Right to respect for private and family life. Case law supports the position that Article 8 is likely to be engaged in respect to systematic storage of personal information or classification of an individual within a police database (*R (on the application of Catt)* v. *Commissioner of Police of the Metropolis* and *R (on the application of T)* v *Commissioner of Police of the Metropolis* [2015] UKSC 9; *PG and JH* v. *UK* (App no. 44787/98), ECHR 2001 IX); *S and Marper* (App no 30562/04), judgment 4 December 2008 (Grand Chamber) [2008] ECHR 1581.
88. 'Data Protection Act 2018 (UK)', c. 12.
89. 'Data Protection Act 2018 (UK)', c. 12, ss. 35.
90. 'Data Protection Act 2018 (UK)', c. 12, ss. 36.
91. 'Data Protection Act 2018 (UK)', c. 12, ss. 39.
92. 'Data Protection Act 2018 (UK)', c. 12, ss. 40.

are requirements under the third and fourth data protection principles.[93] These issues are highly context-specific, and for this reason cases must be considered on an individual basis when it is suggested that a decision made by the police may have violated any of these principles.

ML algorithms typically look for pattern recognition. The use of training data gives an opportunity for the algorithm to 'learn' from feedback and refine its future predictions based on past performance. The algorithm will look at the data provided and take 'decisions' based on prioritising, classifying, associating and filtering this information.[94] Brent Mittelstadt and others explain that in a classification task, the ML algorithm 'typically consists of two components, a *learner* which produces a *classifier,* with the intention to develop classes that can generalise beyond the training data'.[95] As an example, the algorithm may have learned that certain combinations of characteristics in an individual may correlate with an increased propensity to reoffend.[96] The question may then arise as to whether the use of this algorithm to classify an individual as a potential reoffender may be seen to be discriminatory under the law – particularly given the fact that ML algorithms merely identify correlations, rather than causation.

In a policing context, the learning process is inevitably based purely on historic data. It would be problematic for an algorithm to assess its performance over time in an operational policing environment, because the interventions that are delivered to subjects on the basis of the algorithm's predictions are likely to alter the future course of events that would otherwise have unfolded had the intervention not been delivered; the algorithm's prediction itself prevents the predicted outcome from happening. Caution must be exercised when using historic data to produce forecasts based on new, unfamiliar data, particularly to ensure the dataset used to train the algorithm is representative of the data on which the new predictions are being made.

Indeed, how this 'self-learning' process of the ML algorithm works in practice is not always clear to a human observer, and the level of sensitivity of the feedback mechanism may vary from algorithm to algorithm. As has been noted, 'Development is not a neutral, linear path; there is no objectively correct choice at any given stage of development, but many possible choices'.[97] Given the flexible and malleable way in which ML algorithms learn and develop, training the algorithm can be fraught with hazards as regards the characterisation and weighting given to the input data. Supervised machine learning models such as random forests rely to a large degree on the judgement of the individual who designs the model, and efforts must be made to avoid partiality or unintended bias creeping into the system. Furthermore, the determination of

93. 'Data Protection Act 2018 (UK)', c. 12, ss. 37, 38.
94. Nicholas Diakopoulos, 'Accountability in Algorithmic Decision Making', *Communications of the ACM* (Vol. 59, No 2, 2016).
95. Brent Mittelstadt et al., 'The Ethics of Algorithms: Mapping the Debate', *Big Data & Society* (July–December 2016), pp. 1–21.
96. See Oswald et al., 'Algorithmic Risk Assessment Policing Models'.
97. Mittelstadt et al., 'The Ethics of Algorithms'.

whether an algorithm may be considered to be biased, to involve illegal discrimination, or to be unfair can be complex and will involve a great deal of subjectivity.[98]

Conversely, it is possible to argue that the use of ML algorithms could assist the police in overcoming some human biases which could otherwise influence their decision-making. Brauneis and Goodman suggest that:

> Algorithmically informed decision making can also help government officials avoid the biases, explicit or implicit, that may creep into less formal, 'hunch'-based decision making … [T]he systematic use of data analytics can identify characteristics that have a significant correlation with recidivism and evaluate the strength of those correlations, either separately or in combination. Those correlations can then be encoded into an algorithm that estimates the risk of recidivism when fed input information about the prisoner.[99]

It can be argued that the algorithm is neutral in that it has no self-interest, and to that extent it can be expected to treat all similar data in a similar fashion providing a consistently fair result.

As mentioned previously, one difficulty encountered with ML algorithms is that although certain protected data (such as race or religion) can be expressly excluded from the dataset provided to the algorithm, the algorithm may identify a correlation between protected data and combinations of other data available, resulting in a response which is biased just as it would be if based on the excluded protected data.[100] This creates the possibility of generating a decision derived in part by reliance on probabilistically inferred protected information.[101]

Even apart from the algorithm, a database itself may be biased.[102] As mentioned previously, police-recorded data is not necessarily an accurate representation of how crime is patterned in time and space, but rather a representation of the places and people that have previously been targeted by police action. Osoba and Welser argue that '[a]pplying procedurally correct algorithms to biased data is a good way to teach artificial agents to imitate whatever bias the

98. For a discussion of these issues and a potential approach to dealing with them, see Michael Feldman et al., 'Certifying and Removing Disparate Impact', paper presented to the 2015 Association for Computing Machinery's (ACM) Conference of the Special Interest Group on Knowledge, Discovery and Data Mining (SIGKDD), Sydney, Australia, 10–13 August 2015.
99. Robert Brauneis and Ellen P Goodman, 'Algorithmic Transparency for the Smart City', *Yale Journal of Law & Technology* (Vol. 20, No. 103, 2018), p. 116.
100. See, for example, Simon DeDeo, 'Wrong Side of the Tracks: Big Data and Protected Categories' in Cassidy R Sugimoto, Hamid R Ekbia and Michael Mattioli (eds), *Big Data is Not a Monolith* (Cambridge, MA and London: MIT Press, 2016); Mittlestadt et al., 'The Ethics of Algorithms'; Edwards and Veale, 'Slave to the Algorithm?'.
101. Osonde Osoba and William Welser, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence* (Santa Monica, CA: RAND Corporation, 2017).
102. Amnesty International, 'Trapped in the Matrix'.

data contains'.[103] Negative reinforcement resulting from the repeated action of the feedback loop occurring over multiple uses of the ML algorithm can then magnify these undesirable results if the data and outputs are not kept under critical review. Appropriate safeguards must therefore be put in place to ensure that algorithms are frequently reviewed as new data is added, and that potential biases are not inadvertently reinforced or amplified.

The issue of algorithmic bias and discrimination is further complicated by the fact that crime data is inherently 'biased' in a number of ways, because certain classes of people commit more crimes than others. For instance, men commit crime at significantly higher rates than women, are more likely to be involved in violent offences, and are more likely to reoffend. This gender imbalance has been described as 'one of the few undisputed "facts" of criminology'.[104] Therefore, a crime prediction system that is operating correctly will assign many more male offenders to the 'high-risk' category than female offenders. This can be described as a 'fair bias', an imbalance in the dataset that reflects real-world disparities in how a phenomenon is distributed across different demographics. As mentioned in Google's AI Ethics Principles, 'distinguishing fair from unfair biases is not always simple, and differs across cultures and societies'.[105] The challenge of distinguishing 'fair' from 'unfair' bias in the policing domain is likely to pose significant problems to those seeking to develop 'fair' algorithmic systems, not least because such a judgement involves a significant degree of subjectivity. The protected characteristics laid out in the Equality Act 2010[106] are a reasonable starting point for ensuring algorithmic tools do not unduly discriminate against certain classes of individuals, but certain protected characteristics such as gender and age must clearly be factored into crime prediction tools in order to reflect the way that crime is patterned across different demographics.

Algorithmic processing of criminal justice data may also be criticised for treating a person as a member of a particular group, class or category, without taking into account additional contextual factors relevant to the decision. While the police officer dealing with the person before them can assess the specific characteristics and circumstances of the individual in question, the algorithm is merely comparing that person to a group of other people in order to assess whether they are likely to behave as the group. Thus, this method is a simplification which does not account for all of the relevant individual characteristics and idiosyncrasies of the person but treats them as a member of a predefined group,[107] in a way that human decision-makers do not.

103. Osoba and Welser, *An Intelligence in Our Image*, p. 17.
104. See Jennifer Schwartz et al., 'Trends in the Gender Gap in Violence: Reevaluating NCVS and Other Evidence', *Criminology* (Vol. 47, No. 2, 2009), pp. 401–25; Frances Heidensohn, 'Gender and Crime' in Mike Maguire, Rod Morgan and Robert Reiner (eds), *The Oxford Handbook of Criminology,* Third Edition (Oxford: Oxford University Press, 2002), pp. 491–530.
105. Google, 'Artificial Intelligence at Google: Our Principles', <https://ai.google/principles>, accessed 14 August 2018.
106. 'Equality Act 2010 (UK)', c. 15.
107. Brauneis and Goodman, 'Algorithmic Transparency for the Smart City'.

Given the technical, organisational and legal complexities of developing and implementing fair and transparent criminal justice algorithms, such development would benefit from the close cooperation of the police, mathematicians, computer scientists, data scientists and legal experts. Such an interdisciplinary group would be able to discuss the requirements for the tool in the particular context in which it is needed, and ensure that the tool would adequately meet the requirements of the commissioning body, in this case the police. Ensuring algorithmic fairness and eliminating bias often comes at the cost of predictive accuracy.[108] Achieving the correct balance between predictive power and algorithmic fairness would require understanding and input from all members of the team. The experience and skills gained by such an interdisciplinary group would also allow the formulation of algorithmic systems which could then be trained on other datasets held by the police to address different issues.

In terms of the supply of algorithmic systems, while the importance of managed services to policing is appreciated, caution should be exercised in relation to 'off-the-shelf' solutions. In the interest of developing expertise, and pursuing the multiple goals of uniformity, accuracy and fairness, control of both the software and relevant datasets should remain with the public sector body, and care should be taken to ensure that appropriate intellectual property rights are granted to the public sector body to both facilitate any modifications or updates which may need to be made in the future, and also to fulfil requirements of transparency and accountability. Developing and maintaining proprietary systems requires that there are individuals with the relevant skills and expertise working within the public authority. Efforts should be made to create attractive job opportunities within policing and criminal justice for software developers with the skills needed to develop proprietary algorithmic tools.

The potential for algorithmic bias has led some to question whether existing regulation is sufficient to ensure the fair use of machine learning. Certainly, the private sector has acknowledged the need for objectivity and fairness in the development of future machine learning systems. The Partnership on AI, composed of companies such as Facebook, Apple, Amazon, IBM and Microsoft along with numerous others, is seeking to bring 'together diverse, global voices to realize the promise of artificial intelligence'.[109] The Institute of Electrical and Electronics Engineers (IEEE), one of the world's largest technical professional organisations for the advancement of technology, has launched an international endeavour called The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,[110] emphasising their desire to ensure that ethics and human well-being are prioritised in the development of AI.[111] Critics,

108. Osoba and Welser, *An Intelligence in Our Image*, p. 18.
109. Partnership on AI, <https://www.partnershiponai.org/>, accessed 16 August 2018.
110. Institute of Electrical and Electronics Engineers (IEEE) Standards Association, 'About', <http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html>, accessed 16 August 2018.
111. See also Association for Computing Machinery, 'World's Largest Computing Association Affirms Obligation of Computing Professionals to Use Skills for Benefit of Society: Updated ACM Code of Ethics Adds New Principle on Adopting Standards of Care as Computing Systems Become Integrated into Society's Infrastructure', 17 July 2018, <https://www.acm.org/media-center/2018/july/acm-updates-code-of-ethics>, accessed 10 August 2018.

however, have suggested that allowing commercial entities with stakes in the result to propose ethical standards may be unwise,[112] and the specific legal grounding of these various technology initiatives remains uncertain.

Use of algorithms also raises the question of whether such tools are being used in a proportionate way – whether there is a fair balance between the rights of the individual and the public purpose being pursued.[113] In a society where the demands on the police are increasing while resources are more constrained, it could be argued that a tool to assist in decision making where the police officer still has authority to overrule the recommendation of the algorithm would be proportionate, but this assessment would need to be reviewed depending on the context. Clearly, if the tool produced erroneous or biased information which was not challenged by the police officer, that would be cause for serious concern, demonstrating why each case of algorithmic implementation must be handled individually and kept under vigilant review.

In a policing context, as the dataset utilised by the algorithm will be continually updated and revised, continued attention and vigilance is needed to ensure fairness of the system. In addition, the machine learning practiced by the algorithm may enhance predictive accuracy but must be monitored for bias and other unacceptable operations. Another potential difficulty is that a prediction, by its nature, is not neutral, but has the ability to exert an influence on the measured result. A report from the French data protection authority (Commission nationale de l'informatique et des libertés, or CNIL), has considered these challenges:

> One of the challenges identified has to do with the changeable, scalable nature of machine learning algorithms. This characteristic is compounded by the unprecedented scale of the potential impact of algorithms run by computer programs … How, then, should we tackle and regulate an unstable object, which is likely to engender new effects as it grows and learns – effects that could not be foreseen at the outset? Promoting a principle of 'required continued attention and vigilance' could be a way to address this challenge.[114]

It would seem clear that ML algorithms are not something that can be initialised, implemented and then left alone to process data. The ML algorithm will require constant 'attention and vigilance' to ensure that the predictive assistance provided is as accurate and unbiased as possible, and any irregularities are addressed as soon as they arise.

---

112. Jacob Turner, 'Letting Facebook Control AI Regulation is Like Letting the NRA Control Gun Laws', *Quartz,* 6 December 2017.
113. Article 8(2) ECHR: interference with the right must be in accordance with the law, necessary in a democratic society in the interests of a legitimate aim (meets a 'pressing social need' and proportionate to the aim pursued): *Handyside* v *UK* (App no 5493/72), judgment 7 December 1976, (1976) 1 EHRR 737, [1976] ECHR 5.
114. Commission nationale de l'informatique et des libertés, 'How Can Humans Keep the Upper Hand? The Ethical Matters Raised by Algorithms and Artificial Intelligence', December 2017, <https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf>, accessed 7 July 2018, p. 50.

V. Towards a Formal System of Regulation and Oversight

This report comes at a time when considerable activity has been focused on the governance and regulation of machine learning, algorithms and data analytics. The Information Commissioner has an established regulatory role in relation to personal data underpinning machine learning, including so-called 'Big Data'[115] and automated decision-making and profiling.[116] The Data Protection Act 2018 contains a number of accountability requirements applicable to law enforcement bodies, including requirements to:[117]

- keep internal records of processing activities;
- where relevant and as far as possible, keep clear distinctions between categories of individuals;
- integrate data principles into processing activities (data protection by design);
- carry out a data protection impact statement where processing is likely to result in a high risk to the rights and freedoms of individuals; and
- keep logs for audit and monitoring purposes of processing within IT systems.[118]

Deployment of ethical principles for machine learning is gaining momentum, both in the private sector, such as Google's recently announced principles for AI,[119] and in the public sector, including the Department for Digital, Culture, Media and Sport's (DCMS) Data Ethics Framework, revised from the 2016 Data Science Ethical Framework.[120] Despite this, there appears to be increasing concern that existing structures are not fit-for-purpose in terms of the regulation and oversight of new algorithmic deployments.

The British Academy and Royal Society report on data governance recommended a set of high-level principles to 'visibly sit behind all attempts at data governance across sectors' – with the overarching principle being the promotion of human flourishing – reinforced by

115. Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'.
116. Information Commissioner's Office, 'Guide to Law Enforcement Processing (Part 3 of the DP Act 2018)', 5 April 2018.
117. 'Data Protection Act 2018 (UK)'.
118. These requirements apply to at least the following processing actions: collection, alteration, consultation, disclosure (including transfers), combination and erasure.
119. Sundar Pichai, 'AI at Google: Our Principles', Google Blog, 7 June 2018, <https://www.blog.google/topics/ai/ai-principles>, accessed 16 August 2018.
120. Department for Digital, Culture, Media & Sport, 'Guidance: Data Ethics Framework', 13 June 2018.

the creation of a data stewardship body.[121] The House of Commons Science and Technology Committee advised that initiatives such as '[s]etting principles and "codes", establishing audits of algorithms, introducing certification of algorithms,[122] and charging ethics boards with oversight of algorithmic decisions' are all urgently needed.[123]

The Royal Society's 2017 report on machine learning, however, advised against governance structures for machine learning per se; instead, sector-specific mechanisms were preferred.[124] The House of Lords Select Committee on Artificial Intelligence's report also advised against blanket AI-specific regulation, commenting 'that existing sector-specific regulators are best placed to consider the impact on their sectors of any subsequent regulation which may be needed'.[125]

With the recent establishment of the Centre for Data Ethics and Innovation, the UK government has responded to a number of the above concerns. It does not appear, however, that the Centre will have a regulatory role. Its focus is likely to be on identifying best practice and advising on potential gaps in regulation, the proposed terms of reference suggesting a dual role consisting of maximising the benefits of data and AI for society and the economy, while advising on responsible and ethical use.[126]

Future policy and guidelines should address the challenge raised by the perceived obligation to innovate in a controlled way that gives the public reassurance that individual rights are being respected.[127] Uncertainty over acceptable uses of data and algorithms may mean policing organisations are reluctant to experiment and innovate. Although a number of legal frameworks apply to the use of algorithms within criminal justice – data protection, administrative law, human rights and policing legislation – these frameworks are often complex and inaccessible to practitioners working at the operational level. There is a pressing need for new guidelines that translate these legal frameworks into practical standards and organisational policy for the deployment of algorithms within policing and criminal justice.

---

121. The British Academy and the Royal Society, 'Data Management and Use: Governance in the 21st Century', June 2017.
122. An example of which is the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and the IEEE P7000 Standards Projects, <https://ethicsinaction.ieee.org/>, accessed 17 September 2018.
123. House of Commons Science and Technology Committee, 'Algorithms in Decision-Making', Fourth Report of Session 2017–19, HC 351, 23 May 2018.
124. The Royal Society, *Machine Learning: The Power and Promise of Computers that Learn by Example* (London: The Royal Society, April 2017).
125. House of Lords Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?', p. 116.
126. Department for Digital, Culture, Media & Sport, 'Centre for Data Ethics and Innovation: Consultation', June 2018.
127. See the 'experimental proportionality' model in Oswald et al., 'Algorithmic Risk Assessment Policing Models'.

Furthermore, a formalised system of scrutiny and oversight – specifically for policing and criminal justice – now appears necessary to ensure fairness is maintained. There is often a conflation between authorisation and inspection in oversight frameworks, and it would be important to separate the two when considering a suitable framework for the use of algorithmic technology. There is a legitimate concern, however, that the level of oversight required to use these tools may cancel out any efficiency savings. Therefore, the requirements of such a framework must not be so restrictive that they stifle appropriate innovation. Regulatory frameworks must also be flexible enough to account for unforeseen developments in technology in the coming years; regulation should not need to be retrofitted in response to technological advances.

It is likely that the proposed regulatory and oversight framework will take time to construct and implement, and there is a risk that policy development will be outpaced by technological advancements. Therefore, the immediate priority should be to establish clear codes of practice setting out how police forces should trial algorithmic predictive policing tools within an operational environment.

The recent controversy raised by the trials of facial recognition technology by the Metropolitan and South Wales police forces illustrates the challenges raised by experimentation and innovation in live policing environments.[128] As outlined in the London Policing Ethics Panel's Interim Report on Live Facial Recognition:

> Limited trials are of value to test whether a technology can effectively serve valid policing aims, whether expenditure on it is likely to be a good use of public funds, and whether it supports economical use of limited policing resources. However, trials should take place within clear and appropriate constraints and without any prior assumption that testing the technology justifies future deployment.[129]

Similarly, in the case of policing algorithms, it is essential that future development is informed by a reliable evidence base of the efficacy and efficiency of different systems, their cost-effectiveness, their impact on individual rights and the extent to which the implementation of the tool serves valid policing aims. In order to build such an evidence base, there is a need for clear codes of practice outlining clear and appropriate constraints governing how police forces should trial 'predictive policing' tools. Such limited trials must then be comprehensively and independently evaluated before moving ahead with large-scale deployment. Large-scale deployment of such tools will only be justified if the local trial stage has clearly demonstrated that deployment of the tool (i) serves valid policing aims, (ii) is effective in carrying out the function(s) it is intended to, (iii) is cost-effective and represents an efficient use of limited resources, and (iv) conforms to the requirements of data protection legislation and does not infringe on human rights or administrative law principles.

128. *Liberty*, 'Cardiff Resident Launches First UK Legal Challenge to Police Use of Facial Recognition Technology in Public Spaces', 13 June 2018.
129. London Policing Ethics Panel, 'Interim Report on Live Facial Recognition', July 2018, p. 8.

Regardless of whether these requirements are met, the issues discussed in this report highlight the urgent need to clarify appropriate regulatory and oversight mechanisms for the deployment of such tools for policing purposes. This framework should operate in such a way that it respects the autonomy of forces and the value of experimental innovation within operational environments, while setting minimum standards around issues such as transparency and intelligibility, the potential effects of the incorporation of an algorithm into a decision-making process and how to assess these, and the full consideration of all relevant ethical issues.

Recommendations

- The Home Office should develop codes of practice outlining clear and appropriate constraints governing how police forces should trial predictive policing tools, including instructions concerning 'experimentation' in live operational environments focused upon fairness and proportionality. Such limited trials must then be comprehensively and independently evaluated before moving ahead with large-scale deployment.
- The College of Policing should develop guidance within the Authorised Professional Practice with respect to the deployment of a ML algorithm within a decision-making process. This should include guidance on how police forces should present algorithmic predictions to those about whom the prediction is made. A clear process for resolving disagreements when professional judgement and the algorithm come to different conclusions should also be established within this guidance.
- The inspection role of Her Majesty's Inspectorate of Constabulary and Fire and Rescue Services should be expanded to include assessment of forces' compliance with the above-mentioned new guidance. This will provide an accountability mechanism to ensure police forces are developing new tools in accordance with relevant legislation and ethical principles.
- Officers may need to be equipped with a new, different skill set to effectively understand, deploy and interpret algorithmic tools (including the potential for bias), in combination with their professional expertise, and to make assessments of risk using an algorithmically generated forecast. The College of Policing should consider developing a course to equip officers with such a skill set, akin to the Digital Media Investigator training programme.
- A future regulatory framework should establish minimum standards of technical transparency for algorithms used to support police decision-making relating to individuals, which can be adapted for particular contexts and decision-making environments. As a minimum, ML algorithms should only be permitted for criminal justice purposes if it is possible to retroactively deconstruct the algorithm in order to assess which factors influenced the model's predictions and how the prediction has been generated. This requirement should be included in all relevant public procurement agreements, along with a requirement for the supplier to be able to provide an expert witness who can give evidence concerning the algorithm's operation if needed.
- When an algorithmic prediction is utilised in an evidential context, or is involved in judgements concerning criminal justice outcomes, in order for the data subject to be able to scrutinise the specific factors that led to a certain decision being made, the list of variables included in the model must be made available. The data subject should also have the option to review the algorithmic tool itself, and be provided with an intelligible output of the prediction, indicating in plain English the influence that each variable had on the system's overall prediction.
- A national working group consisting of members from the fields of policing, computer science, law and ethics should be tasked with sharing innovations and challenges, and

with looking at the requirements for new ML algorithms within policing, with a view to setting out the relevant parameters and requirements, and considering the appropriate selection of training and test data. Such a group should also work towards establishing principles concerning 'fair' and 'unfair' biases in predictive policing algorithms.

- The College of Policing and National Police Chiefs' Council should develop a standardised Terms of Reference and structure for local ethics boards, which will scrutinise and assess the implementation of algorithms for policing and criminal justice. Such ethics boards should consist of a combination of practitioners and academics, and should provide recommendations to individual forces for practice, strategy and policy decisions relating to the use of algorithms.
- ML algorithms must not be initialised, implemented and then left alone to process data. The ML algorithm will require constant 'attention and vigilance' to ensure that the predictive assistance provided is as accurate and unbiased as possible, and that any irregularities are addressed as soon as they arise. Each case of algorithmic implementation for policing should be regularly scrutinised in this way by the above-mentioned local ethics boards.
- Further research is needed to examine the real effect of the introduction of an algorithmic decision-support tool into a range of decision-making environments within UK policing and criminal justice. The impact of a variety of design and presentation options relating to both risk indicators and algorithmic output should be considered, in order to assess how these influence behaviour and acceptance.
- Future research into the use of transparent algorithms for policing and criminal justice purposes should prioritise developing NLG functionality that is incorporated into the system, so that individuals are always provided with an English language explanation outlining the factors the algorithm took into account when calculating its prediction, and any potential uncertainties or error margins present in the output.

About the Authors

Alexander Babuta is a Research Fellow in the National Security and Resilience Studies group at RUSI. His research focuses on policing and security in the digital age, transnational organised crime and counterterrorism.

Marion Oswald is a Senior Fellow in Law, Director of the Centre for Information Rights at the University of Winchester and a solicitor (non-practising). Her research focuses on the interaction between law and digital technology and on the deployment of innovative technologies particularly within the public sector.

Christine Rinik is a Senior Lecturer in Law at the University of Winchester and a solicitor (non-practising). Her research focuses on legal issues arising from the proliferation of digital technology, as well as the application of equitable principles and fiduciary duties to contemporary commercial relationships.