Behavioural Analytics in Policing Data-Driven Offender Management: Assessing the Evidence Base

Babuta, Alexander (2023) Behavioural Analytics in Policing Data-Driven Offender Management: Assessing the Evidence Base. Doctoral thesis, University of West London.

https://doi.org/10.36828/xvqy0381

Behavioural Analytics in Policing

Data-Driven Offender Management: Assessing the Evidence Base

Alexander Babuta

Primary Supervisor: Prof Simon Harding

Professional Doctorate in Policing, Crime and Security
School of Humanities and Social Sciences
University of West London

Abstract

This doctoral thesis explores the use of behavioural analytics and data-driven offender management within UK policing. The findings are based on semi-structured interviews with criminal justice practitioners who have been directly involved in the development of such projects, and a process evaluation of a major data-driven offender management project delivered by one of the UK's largest police forces. The research explored the potential opportunities offered by new data-driven risk assessment tools, and sought to examine the barriers to successful implementation when the technology is deployed in an operational policing context. The process evaluation highlighted specific practical challenges associated with implementing a new data-driven system in an operational offender management context. The implications of these findings for future policing practice are explored, focussing on how such technology should be piloted and evaluated to assess its potential real-world benefits and limitations. The findings and discussion will be of interest to criminal justice practitioners and policymakers involved in developing or implementing new data-driven offender risk assessment tools in the UK criminal justice system.

Table of Contents

Acknowledgements

The author is very grateful to the police officers and staff who contributed their valuable time to participate in the study, without whom the research would not have been possible. The author would also like to thank Professor Simon Harding for his invaluable support and guidance throughout the research process, as well as Dr Allan Moore, Professor Shane Johnson, Professor Alison Wakefield, Dr Rashid Minhas, Dr Amy Stevens and Professor Caroline Lafarge for their helpful feedback on an earlier draft of this thesis.

Abbreviations

AI – Artificial Intelligence

APP – Authorised Professional Practice

ARAI – Actuarial Risk Assessment Instrument

AUC – Area Under the Curve

CJS – Criminal Justice System

ECHR – European Convention on Human Rights

FRAME – Framework for Risk Assessment, Management and Evaluation

GLM – Good Lives Model (of rehabilitation)

HMPPS – His Majesty's Prison and Probation Service

LFR – Live Facial Recognition

LOMU – Local Offender Management Unit

MAPPA – Multi-Agency Public Protection Arrangements

ML – Machine Learning

MoJ – Ministry of Justice

NPCC – National Police Chiefs' Council

OASys – Offender Assessment System

PSED – Public Sector Equality Duty

RMA – Risk Management Authority (Scotland)

RMP – Risk Management Plan

RNR – Risk, Need and Responsivity (framework)

SPJ – Structured Professional Judgement

SRA – Structured Risk Assessment

Chapter 1. Introduction

This doctoral study explored the use of behavioural analytics and data-driven offender risk assessment within the UK criminal justice system (CJS). The findings presented in this thesis are based on (i) detailed semi-structured interviews with police and criminal justice practitioners involved in developing or implementing offender management programmes, and (ii) a process evaluation of a pilot data-driven risk assessment project that was being conducted in one of the UK's largest police forces at the time the study was undertaken.

While often presented as novel and futuristic, the use of statistical methods to predict risk of future offending is now a well-established practice. However, recent years have seen an emergence of a new generation of risk assessment technology, particularly driven by advances in data science and machine learning (discussed further in Section 2.3). This new generation of data-driven technologies can be understood as a form of behavioural data science, or 'behavioural analytics' – defined here as the application of data science techniques to understanding or forecasting human behaviour. These new behavioural analytics capabilities have attracted considerable media and public attention, as part of a broader focus on the ethical challenges posed by the use of 'predictive policing' and related artificial intelligence technologies within the UK criminal justice system.

Despite this public and media attention, behavioural analytics systems in policing have been subject to very limited academic scrutiny. The little research that is available focuses almost exclusively on statistical validity, often overlooking the crucial question of whether the tools are *useful in practice* for criminal justice practitioners who are required to implement them in an operational context. This study seeks to address this gap, and assess the potential benefits and shortcomings offered by new data-driven offender risk assessment tools in an operational policing environment.

1.1 Background

This study is concerned with the use of statistical risk assessment within UK policing, which is often described as a form of 'predictive policing'. It is important to distinguish between two main categories of predictive policing technology: location-based and individual-based. Location-based predictive policing refers to *predictive* (or *prospective*) *crime mapping*, which has been the subject of extensive academic research (Bowers, Johnson and Pease, 2004; Johnson *et al.*, 2007; Pearsall, 2010; Perry, 2013). Predictive mapping uses past crime data to develop statistical models forecasting areas where future crime is likely to occur. This is based on the consistent observation that repeat victimisation accounts for a large volume of all crime (Farrell and Pease, 2001), and that crime is often contagious – with the risk of crimes such as burglary temporarily increasing for nearby properties in the immediate aftermath of the initial offence (Townsley, Homel and Chaseling, 2003; Ludwig and Kling, 2007).

The regularity of crime distribution in space allows us to make powerful predictions about where crime is likely to occur in the near future. Numerous field trials have demonstrated that predictive mapping is more effective at predicting the location of future crime than traditional intelligence-led techniques (Bachner, 2013). The technique has since been widely commercialised, for instance in the form of the computer software *PredPol*, first developed in 2011 by academics at UCLA and Santa Clara University, in partnership with the Los Angeles Police Department and Santa Cruz Police Department (Mohler *et al.*, 2011). In the UK, uptake of predictive mapping remains limited. This is despite numerous independent reviews urging UK police forces to make better use of predictive mapping to enable more targeted deployment of resources (Her Majesty's Inspectorate of Constabulary and Fire & Rescue Services (HMICFRS), 2017). The practice has also been subject to considerable criticism from privacy groups and academic commentators, as discussed later in this thesis. Although location-based predictive policing is not the focus of the current study, there are nevertheless important parallels to be drawn from the literature which are also relevant to individual-based predictive policing, particularly in terms of operational and ethical challenges.

One of the most consistent findings in the literature is that a relatively small number of offenders are responsible for a large proportion of all crime (Garside, 2004; Farrell, 2015), and that repeat victimisation accounts for a large number of all recordable offences (Ellingworth, Farrell and Pease, 1995; Farrell and Pease, 2001). Due to these concentration patterns, offender management programmes typically aim to target the highest intensity of treatment towards the most persistent and prolific offenders (Worrall and Mawby, 2004; Farrall, Mawby and Worrall, 2007). Accurately prioritising limited resources is crucial to ensuring that preventative interventions are focused on those most likely to offend again in future. While statistical risk assessment tools have been used throughout the CJS for many years to assist in such prioritisation (Craig and Beech, 2009), recent advances in data science and machine learning have enabled the development of more sophisticated modelling techniques, which are used to 'risk score' offenders based on analysis of historic data. The use of such technology raises numerous operational, policy and ethical considerations, which have been the subject of considerable public debate and commentary (Dencik *et al.*, 2018a; Couchman, 2019; Babuta and Oswald, 2020; Brayne and Christin, 2021).

This new generation of algorithmic risk assessment involves the use of data science and statistical modelling techniques to assign numerical risk scores to offenders, corresponding either to the level of harm generated by their *current* offending, or to their estimated risk of *future* offending. This type of data-driven risk assessment can be loosely characterised as a subset of behavioural analytics (Babuta, Oswald and Janjeva, 2020). Behavioural analytics can be understood as a new discipline at the intersection of behavioural psychology and data science, which aims to derive insights or make predictions regarding human behaviour based on the use of advanced data science techniques such as machine learning modelling.

The increased interest in algorithmic risk assessment across UK policing has been driven largely by resourcing pressures and the need to allocate limited resources more efficiently based on a data-driven assessment of risk and demand (Babuta and Oswald, 2020). The NPCC National Policing Digital Strategy 2020-2030 included a commitment to 'translate evolving definitions of threat, harm and risk into digital formats that complement human judgement' and to 'use digital tools to rapidly identify harm related behaviours in order to target interventions' (National Police Chiefs' Council, 2020, p. 7).

Alongside this, the police service is increasingly expected to adopt a preventative, rather than reactive, posture, with greater emphasis on targeting resources towards the areas of highest risk and vulnerability (Crawford and Evans, 2017). The *NPCC Policing Vision* 2025 describes how 'by 2025 the police service will have transformed the way it delivers its mission with a keen focus on prevention and vulnerability and the effective management of risk' (National Police Chiefs' Council, 2016). The purpose of algorithmic analysis is often framed as 'predictive', and aimed towards identifying criminogenic risk before it occurs to inform the use of preventative interventions (Brayne and Christin, 2021). This study is primarily concerned with policing interventions directed towards individual offenders (or suspects). However, it is important to note the wider shift towards preventative tactics witnessed across UK policing in recent years, for instance through greater uptake of situational crime prevention methods. Situational crime prevention is focused on reducing opportunities for crime by increasing perceived risks and reducing rewards, and dates back to at least 1980 (Clarke, 1980, 1983, 1995).

Several police forces have now deployed such risk scoring tools operationally, as detailed in the following sections. However, unlike traditional statistical risk assessment tools which have been subject to considerable academic scrutiny and evaluation research, there is a concerning lack of empirical evidence regarding the potential benefits and limitations of new and emerging behavioural analytics and data-driven risk assessment methods. The few evaluation studies that have been published focus almost exclusively on statistical validation (i.e., technical evaluation of the tools themselves); often neglecting to include an evaluation of the real-world *use* of the tool in its operational context. There is therefore a notable absence of research addressing the crucial question of whether these tools are useful *in practice* to assist CJS practitioners in making reliable risk decisions. There is now a pressing need to engage more closely with CJS practitioners to understand their perspectives on the strengths and limitations of these tools before they are deployed for enduring use.

This study seeks to address this gap, to inform future policy and practice regarding the use of data-driven risk assessment tools in the English CJS. The research comprised detailed semi-structured interviews with CJS practitioners who have been directly involved in the development of new data-driven risk assessment projects; and a process evaluation of a major data-driven risk assessment project, delivered by one of the UK's largest police forces. To the author's knowledge, this represents the most comprehensive study of the use of data-driven offender risk assessment tools in UK policing. The findings will be of relevance to criminal justice practitioners and policymakers involved in developing technology-supported approaches to offender management; to private sector companies supplying such technology to police forces; and to academic researchers with an interest in this field.

1.2 Research aims and objectives

This study aimed to explore the opportunities and risks presented by new behavioural analytics and data-driven risk assessment tools within the English CJS, through direct engagement with CJS practitioners involved in relevant projects. It is hoped that these findings will inform future approaches to data-driven risk assessment by the police and other criminal justice agencies, and identify the measures required to ensure that these tools are deployed in an effective and responsible way.

The following research questions are designed to critically assess offender risk assessment methods throughout the English CJS, with particular focus on novel data-driven approaches currently in development.

- **RQ1:** What are the strengths and limitations of existing offender risk assessment methods used throughout the English criminal justice system?
- **RQ2:** What opportunities do new data science techniques offer for assessing risk and targeting interventions most effectively?
- **RQ3:** How could advanced data science tools be effectively integrated within existing offender management processes?

This project addresses a notable gap in existing research and is intended to directly inform future approaches to data-driven offender risk assessment within the English CJS. The research questions above were formulated based on a comprehensive review of existing literature, which highlighted a concerning lack of empirical evidence regarding the potential benefits and limitations of novel data-driven risk assessment tools within the English CJS.

This study is concerned with the use of data-driven risk assessment tools for forecasting risk of future offending at the individual level. This includes both violent and non-violent offending. However, most academic literature on the topic of statistical risk assessment focuses on the specific practice of *violence* risk assessment. Therefore, much of the discussion in Section 2 relates specifically to violence prevention and risk assessment, although the case study evaluation presented in Section 4 is concerned with all forms of individual-level risk assessment – whether violence related or otherwise.

The research does not seek to statistically evaluate the data science tools under examination. Rather, the aim of the research is to identify key shortcomings and limitations of existing risk assessment approaches, and the opportunities presented by novel data-driven methods. In doing so, the research aims to identify potential future uses of data-driven risk assessment tools, and make practical recommendations as to how new and emerging risk assessment technologies could be effectively integrated into offender management processes.

1.3 Professional context

[Removed for public version]

1.4 Remaining chapters

The remainder of this thesis is structured as follows. Chapter 2 comprises a literature review of existing academic publications, policy documents and relevant grey literature on the practice of offender risk assessment in the UK criminal justice system, with particular focus on new and emerging data-driven approaches. It also explores the legal, ethical and operational challenges associated with new approaches to data-driven risk assessment. Chapter 3 describes the research design and methodology, summarising the sampling strategy, data collection and analysis methods used for both components of the research, including ethical considerations and methodological limitations. Chapter 4 presents the findings from the primary research, divided according to the strengths and limitations of data-driven risk assessment tools identified in the project, and the findings from the process evaluation component of the study. Chapter 5 comprises a discussion and analysis of the research findings presented in Chapter 4, including proposing 21 recommendations for police forces and others seeking to develop or deploy future data-driven risk assessment systems. Finally, Chapter 6 concludes by summarising the implications of these findings for future policing practice and policy.

Chapter 2. Review of existing literature

This literature review explores the development of behavioural analytics and data-driven offender risk assessment tools throughout the UK CJS. Section 2.1 summarises the literature search and selection strategy used for the review. Section 2.2 then provides an overview of the three main approaches to offender risk assessment discussed in the literature: unstructured clinical judgement, statistical prediction and Structured Professional Judgement. Section 2.3 critically assesses current approaches to offender risk assessment adopted within the UK CJS, including reviewing the development of 'next-generation' data scoring tools. Finally, Section 2.4 concludes by summarising key relevant legal, ethical and operational considerations that have been discussed in the literature.

2.1 Literature search and selection strategy

The primary information sources for this literature review comprised academic journal articles, books and book chapters on the topic of offender risk assessment and risk management in the UK and North America. While the primary research conducted for this project focuses specifically on the English criminal justice system, the geographic scope of the literature review was not restricted as such, to allow a full exposition of the current academic landscape. The literature review provides a retrospective summary of developments in offender risk assessment throughout the 20th century, as such the timeframe for inclusion was also not restricted.

Google Scholar was the main online database consulted for this literature review. The primary search terms were as follows: ("risk assessment" OR "risk management") AND (violen* OR offen*); "violence prediction"; "structured professional judgement". The search was limited to English language publications. Core readings were first identified, and reference lists were reviewed to identify further relevant studies using a snowballing method. In addition to academic publications, the literature review also included a number of relevant government reports published in the last 10 years, with focus on the publications produced by the National Police Chiefs' Council (NPCC), Ministry of Justice (MoJ) and Her Majesty's Prison and Probation Service (HMPPS). Finally, there is little or no formally published academic material in relation to police use of new data-driven risk assessment tools. For this reason, the review also included a small number of other relevant publicly available documents, for instance formal minutes from police digital ethics committee meetings.

It is important to note that the literature review focuses primarily on the practice of *violence* risk assessment, a subset of the wider field of offender risk assessment. Although the case study evaluation presented later in this report is broader in scope, most academic literature on this topic tends to focus on the specific task of violence risk assessment, which is generally considered the highest-priority component of most offender management programmes. While the literature presented here on violence risk assessment in the criminal justice system is largely generalisable to the broader practice of offender risk assessment, it should be borne in mind that some content discussed in the following sections relates specifically to the practice of forecasting risk of future violence.

2.2 Competing approaches to offender risk assessment

This section provides an overview of the main approaches to offender risk assessment discussed in the literature, with particular focus on how the practice of offender risk assessment has developed since the mid-20th century.

2.2.1 The Risk-Need-Responsivity Model

Risk assessment is an uncertain process, which involves estimating the probability of an undesirable event, assessing the likelihood of the event occurring and the nature of any potential harm (Denney, 2005). The prevailing approach to offender management throughout the UK CJS is based on the Risk, Need and Responsivity (RNR) principles (Andrews and Bonta, 2010). Central to the RNR framework is the assumption that crime is caused by quantifiable patterns of psychological and social factors which increase the likelihood of an individual breaking the law, and that offending rates will be reduced by identifying and targeting these individual 'risk factors' (Andrews and Bonta, 2010). This approach has been described in the literature as the 'Risk Model' (Visher, 2006; Ward and Maruna, 2007, p. 20). Practice centres around the use of structured assessment tools to identify individual-level risk factors, where the primary objective is to reduce identified risk factors in the most cost-efficient manner (Andrews and Bonta, 2010).

According to the 'Risk' principle, individuals assessed as posing a greater risk should receive a greater dosage or intensity of treatment (Andrews and Dowden, 2006). 'Need' is understood as a set of 'dynamic risk factors' causally linked to criminal behaviour (Gendreau and Andrews, 1990; McGuire, 2000; Hanson, 2001), and is closely related to the concept of 'Risk' in the sense that unmet criminogenic needs often contribute to an increased level of risk. The 'Responsivity' principle relates to the matching of interventions to individual characteristics and circumstances (Andrews, Bonta and Wormith, 2006). In rehabilitative terms, risk assessment may benefit the assessed individual if the assessment is used to identify beneficial treatments that reduce the likelihood of future offending (Douglas *et al.*, 2017). Through this lens, the risk assessment process simultaneously serves two ultimate goals – public protection, and individual rehabilitation.

The RNR framework and its practical implementation in the UK CJS have remained largely unchanged since the early 2000s. Some commentators have credited the RNR model with transforming contemporary approaches to rehabilitation (Cullen, 2005; Andrews, 2012; Polaschek, 2012). However, others maintain that the RNR model is fundamentally limited in several ways. Most notably, some have argued that RNR is essentially a 'risk framework', and the notion of offender "need" is at best treated as secondary in RNR-based offender assessment processes (Hannah-Moffat, 2005; Ward and Maruna, 2007). Some offenders may not be assessed as posing a significant *risk*, but nevertheless have various offending-related *needs* which should be accounted for as part of an individualised treatment plan. As such, some have argued that RNR alone does not constitute an adequate rehabilitation theory, but rather a set of loosely related principles (Polaschek and Collie, 2004). In practice RNR often equates to a standardised or 'one size fits all' approach to service delivery, failing to sufficiently capture how service plans can be most appropriately tailored to address the specific needs, circumstances and vulnerabilities of the service user (Ward and Maruna, 2007, p. 23; Mair, 2013; Fox *et al.*, 2018).

Another limitation is that RNR adopts an inherently 'individualist' perspective of risk: risk is reduced to a set of measurable individual characteristics related to offending, with individuals placed on a behavioural continuum ranging from low to high (Brown, 2000; Ward and Maruna, 2007, p. 79). This results in a reductionist approach to assessment, where psychometric analysis of pre-defined 'risk factors' is central to the development of service plans. This contrasts with the 'categorical' perspective, where risk is considered to reflect aspects of human character and values, in combination with cultural, social or environmental factors (Robinson, 1999; Sparks, 2001; Denney, 2005). According to the categorical perspective, risk cannot be quantified, but rather holistically assessed through structured and systematic judgement (McGuire, 2000). Some have further argued that an overly-individualist perspective on risk can in fact alienate service users and could even *increase* the likelihood of future offending (Beech and Ward, 2004; Hanson and Morton-Bourgon, 2005).

One alternative to the RNR framework that has been presented in the literature is Ward and Maruna's 'Good Lives Model' of rehabilitation (GLM). In contrast to RNR, GLM treats the concept of offender need as central to the delivery of effective treatment programmes. According to GLM, chief among these 'needs' is the innate desire to attain 'primary human goods'. As summarised by Ward and Maruna:

> 'What is required at the clinical level is some attention to helping offenders build a better life (not just a less harmful one) in ways that are personally meaningful and satisfying, and socially acceptable… Concentrating on criminogenic needs is arguably not that helpful to clinicians (and offenders) because it encourages them to focus largely on the elimination or modification of criminogenic needs rather than on how to attain primary human goods.' (Ward and Maruna, 2007, p. 83).

They argue further that the dominance of RNR throughout the UK CJS has resulted in a risk-centric, reductionist approach to offender assessment, that may fail to adequately capture individual motivational, behavioural, and environmental nuances of different offending patterns. The authors conclude that the RNR model alone does not constitute an adequate rehabilitation theory, and must be supplemented by additional, crime-specific theory and clinical models to result in effective treatment.

However, despite these limitations regarding the development of effective rehabilitation programmes, RNR nevertheless provides a practical framework for the police and other agencies to *prioritise* limited resources to the areas of greatest need. Before criminal justice agencies can develop tailored rehabilitation plans for individual service users, they must first 'triage' cases to identify those that require the most immediate attention. A risk-based prioritisation framework is arguably the most defensible approach in the public safety context. When considering the benefits and shortcomings of different approaches to offender management, it is therefore crucial to clearly distinguish between the process of prioritisation and triage on the one hand, and individual-level risk assessment on the other. These are two fundamentally different tasks, but are often conflated in the literature. Many of the 'risk assessment' tools and methods presented in the literature are not in fact intended to assess the nature and severity of risk at the *individual* level, but rather to filter and prioritise cases to identify individuals who should be subject to more in-depth, individualised assessment. This issue is discussed further in the subsequent chapters.

While RNR-based frameworks still dominate the approaches to offender management and rehabilitation adopted throughout the UK and North America, risk assessment methods have nevertheless evolved considerably since the mid-20th century (Campbell, French and Gendreau, 2009). Approaches to offender risk assessment can be broadly divided into three categories: 'unstructured clinical judgement', where conclusions are based solely on the professional opinion of the decision-maker; 'statistical (actuarial) forecasting', where the aim is to 'predict' future behaviour based on statistical, algorithmic analysis; and 'Structured Professional Judgement' (SPJ), a systematic but discretionary approach that does not rely on statistical calculations. These three approaches are summarised in turn below.

2.2.2 The shortcomings of clinical judgement

Historically, assessments of future offending risk relied entirely on the clinical judgement of the decision-maker, an approach that was found to be prone to error and bias (Campbell, French and Gendreau, 2009). This so-called 'first generation' of offender risk assessment has been characterised in the literature as 'the dark days of unstructured clinical opinion, where the basis for any opinion about future risk was at best opaque, the evidence base relied upon unclear, and the validity of the conclusions uncertain' (Cooke, 2012, p. 221). Of all potential approaches, unstructured clinical judgement has the least empirical support, and is now accepted as the least accurate approach to individual risk assessment (Andrews, Bonta and Wormith, 2006; Skeem and Monahan, 2011).

Numerous meta-analyses from the 1950s onwards have demonstrated that statistical (actuarial) forecasting typically yields more accurate predictions than unstructured clinical judgement, across many disciplines and in a wide range of decision-making contexts (Meehl, 1954; Dawes, Faust and Meehl, 1989; Grove *et al.*, 2000a; Ægisdóttir *et al.*, 2006; Kahneman and Klein, 2009). Of particular note is the early work of Meehl (1954), who analysed 20 studies on clinical and actuarial decision-making, and found that in all but one of these studies, actuarial forecasting was as accurate or more accurate than clinical judgement (Meehl, 1954). Another meta-analysis by Grove et al. (2000) of 136 studies covering a range of decision-making contexts found that on average, statistical prediction methods were around 10% more accurate than clinical predictions, and in only eight of the 136 studies did clinical prediction outperform statistical prediction. The authors concluded that 'these data indicate that mechanical predictions of human behaviours are equal or superior to clinical prediction methods for a wide range of circumstances' (Grove *et al.*, 2000b).

In the specific context of violent offending, research consistently shows that clinical decision-makers perform only marginally better than chance when predicting future violence risk. A study by Lidz et al. (1993) required 148 nurses and 67 psychiatrists to assess 714 psychiatric patients living in the community in terms of potential violence towards others during the next 6 months. Overall, the study found that clinicians perform better than chance in their predictions of violence, but also that 'the relatively low sensitivity and specificity of their predictions shows that there is substantial room for improvement.' The authors concluded that 'the low sensitivity and specificity of these judgements show that clinicians are relatively inaccurate predictors of violence'(Lidz, Mulvey and Gardner, 1993, p.1010). Another meta-analysis by Ægisdóttir et al. (2006) examined 67 studies specifically related to mental health practitioners (counselling psychologists), and found that 'in general, statistical prediction methods are somewhat more accurate than the clinical method'. Of particular relevance is the finding that 'one area in which the statistical method is most clearly superior to the clinical approach is the prediction of violence, *r* = -.09. Out of 1,000 predictions of violence, the statistical method should correctly identify 90 more violent clients than will the clinical method' (Ægisdóttir *et al.*, 2006, p.367).

A main argument against the unstructured clinical approach is that all human decision-making is inevitably influenced by heuristics in judgement and cognitive bias (Tversky and Kahneman, 1974). As summarised by Quinsey et al., when human decision-makers are required to make judgements based on probabilities, these simplifying judgement heuristics cloud the decision-making process and can lead to systematic and gross errors. The biased reasoning that results can lead decision-makers to 'ignore the profound effects that differing base rates have on probabilistic judgements' (Quinsey *et al.*, 2006, p.172). Of particular concern is the risk of confirmation bias. A 1979 study by Quinsey and Ambtman required nine schoolteachers and four forensic psychiatrists to make predictions of future offending risk based on psychiatric assessments, individual histories and offence descriptions of 20 offenders (Quinsey and Ambtman, 1979). When participants were provided with the entire case file, they displayed a greater tendency to assess the patient as posing a risk of committing an offence than when they were provided with a limited subset of information. The authors conclude that 'it appeared that all raters sought signs of dangerousness and were more likely to find them the more information they had' (Quinsey *et al.*, 2006, p.175). This finding points to a risk of over-confidence and confirmation bias among assessors when provided with larger quantities of information. However, it should be noted that the small sample size in the Quinsey and Ambtman study means that generalisability may be limited beyond the specific context in which this study was conducted.

In addition to its poor validity, unstructured clinical judgement is neither transparent nor evidence-based. The evaluator is not required to specify their reasons for reaching a decision, making it difficult for the judgement to be challenged by others (Hart, 1998). With this in mind, the lack of transparency, auditability and and accountability of unstructured clinical decision-making should preclude its use for offender risk assessment, because decisions made in relation to probation and parole can have significant consequences on the individual. As discussed by Logan and Lloyd, 'decision-making as regards risk should be transparent, accountable, and defensible, which should be the case even in security and intelligence agencies' (Logan and Lloyd, 2019a, p.7) As such, it is imperative to ensure clear and defensible reasoning for how assessors arrived at a decision, and the specific factors they took into account when making their assessment (Scottish Risk Management Authority, 2011; HM Prison & Probation Service, 2019). For these reasons, unstructured clinical judgement is no longer considered an acceptable approach to offender risk assessment within the UK CJS.

2.2.3 Statistical prediction

Recognising the limitations of the unstructured clinical approach to risk assessment, research efforts focussed on developing actuarial risk assessment instruments (ARAIs) for use in clinical and criminal justice settings. Throughout the 1980s and 90s, numerous 'second-generation' risk assessment tools were developed, which relied on statistical models that generate a predictive 'risk score' purporting to indicate an individual's likelihood to offend. They do so by assigning weighted numerical scores to input variables and combining the weighted scores to arrive at an overall risk score, typically using traditional statistical modelling techniques such as regression modelling.

Early ARAIs relied solely on 'static' risk factors (such as age and criminal history), meaning that the tools did not enable ongoing monitoring of changes over time and did not allow for the identification of specific areas for intervention (Andrews, Bonta and Hoge, 1990; Hoge and Andrews, 1996; Kraemer *et al.*, 1997; Wong and Gordon, 2006; Campbell, French and Gendreau, 2009). To address these shortcomings, the next generation of ARAIs emphasised the need to not only *predict* risk but also to identify specific needs and vulnerabilities that could be targeted to reduce the risk of future offending. With the inclusion of dynamic risk factors (such as employment, housing, relationships and mental state), the updated instruments allowed changes in individuals' personal situations to be monitored over time and across contexts, enabling the formulation of bespoke intervention plans (Gendreau, Little and Goggin, 1996; Bonta, 2002; Andrews, Bonta and Wormith, 2006).

Such structured, statistical assessment tools are now widely used to support offender risk management across the UK and North America (Craig and Beech, 2009). Some are purely actuarial methods that do not incorporate any clinical judgement, while others require the assessor to use the output of the algorithmic prediction in combination with their professional judgement to arrive at an overall risk assessment. The ARAIs in use today are specifically designed to be integrated into risk management processes, to directly inform the selection of intervention methods, and to assess rehabilitation progress over time (Wong and Gordon, 2006; Campbell, French and Gendreau, 2009; Andrews and Bonta, 2010).

While proponents of the actuarial approach argue that statistical methods consistently outperform unstructured judgement, research shows that violence risk assessment approaches that incorporate a degree of professional judgement yield more successful results than relying purely on actuarial methods (Douglas, Yeomans and Boer, 2005). Furthermore, critics have argued that actuarial methods have relatively weak validity and are of limited use to practitioners seeking to develop offender-specific risk management plans (RMPs), concluding that their potential utility has been significantly overstated (Hart, Michie and Cooke, 2007; Hart and Cooke, 2013; Cooke and Michie, 2014; Kewley and Blandford, 2017).

It has also been argued that 'prediction' is a fundamentally flawed concept in the context of violence risk assessment, as an intervention is typically delivered that *prevents* the predicted outcome from happening (Hart, 1998). There is no way of knowing what would have happened had we not intervened, and therefore no way of reliably measuring the accuracy of the initial risk prediction. In other words, if our interventions are effective, all our predictions will appear to be incorrect – because we have prevented the predicted outcome from happening.

In short, the academic debate in this regard is intense and ongoing (Logan and Lloyd, 2019b). However, despite these concerns, ARAIs remain widely used for offender assessment purposes throughout the UK CJS, and statistical risk scores are often instrumental in determining what interventions are available to service users, and the level of risk management required for each case (discussed further in the following sections).

2.2.4 Structured Professional Judgement

Structured Professional Judgement (SPJ), also referred to as the 'guided clinical approach', is an approach to risk assessment and management that aims to bridge the gap between unstructured clinical judgement and statistical prediction (Hanson, 1998; Douglas and Kropp, 2002). SPJ can be summarised as an analytical method that is discretionary, but relies on evidence-based guidelines to structure and systematize the exercise of discretion (Cooke, 2012; Logan and Johnstone, 2012). SPJ involves the use of risk assessment *guidelines* (rather than 'tools') which incorporate specifically defined risk factors, identified on the basis of empirical research in combination with the clinical experience of professional experts (Borum, 1996; Douglas, Blanchard and Hendry, 2012; Webster, Haque and Hucker, 2013; Logan, 2017). SPJ guides are updated and refined as research progresses and more relevant factors are identified through retrospective studies.

The fully operationalised SPJ approach involves the use of an assessment worksheet for documenting evidence of risk factors and justification for decision-making. Evaluators are required to systematically consider each factor listed in the worksheet and determine whether risk factors are present, possibly present or absent. Considering all evidence for the presence and relevance of risk factors, the evaluator is required to develop a case formulation, including hypothesising possible scenarios in which an adverse outcome (such as violent offending) may be realised. They then make recommendations for risk management planning, which may include direct intervention, supervision and victim-safety plans (Logan and Lloyd, 2019b). Some SPJ guides also take into account protective factors, i.e. those which may mitigate or reduce the potential risk of future violence (de Vries Robbé, de Vogel and de Spa, 2011).

In contrast to the actuarial approach, SPJ does not incorporate a scoring system for risk factors. Rather, the focus is on *case prioritisation*; identifying the level of risk management required for each individual. A core aspect of SPJ is that the risk assessment is continuously reviewed and the RMP modified as needed. Crucially, the central purpose of the SPJ approach is to *prevent* – rather than *predict* – violence.

Among behavioural experts, SPJ is widely considered the most evidence-based approach to violence risk assessment (Hart and Logan, 2011; Logan and Lloyd, 2019b). The structured and systematic nature of SPJ ensures logical and coherent reasoning, and a clear link between risk factors and the ultimate intervention (Douglas and Kropp, 2002). Its flexibility means the assessor can incorporate both general ('nomothetic') risk factors (those which prior research has shown to be linked to a higher likelihood for violence), but also individual-specific ('idiographic') factors, which may be pertinent to risk on a case-specific basis, but are not necessarily included on the worksheet as explicitly pre-defined risk factors (Borum, 2015).

Best practice in SPJ has been discussed at length elsewhere. In 2011 the Scottish Risk Management Authority (RMA) published a *Framework for Risk Assessment, Management and Evaluation* (FRAME), which included 13 principles to guide risk assessment and management practice (Scottish Risk Management Authority, 2011). In summary, the FRAME approach advocates a policy agenda focussed on supporting evidence-based practice in public protection to reduce reoffending while limiting the use of custody. To this end, the RMA lists four guiding principles for risk practice: balancing rights;

proportionality; collaboration; and evidence-based practice. The FRAME principles include (but are not limited to): explicitly acknowledging the uncertainty of risk (risk cannot be eliminated nor accurately predicted); adopting a collaborative, multi-agency approach to risk management; ensuring risk assessment directly informs decision-making; adopting a systematic, structured and timely approach; ensuring assessments are based on the best available information; ensuring risk management is sufficiently individualised; ensuring effective transparency and communication between stakeholders; and ensuring that risk management is dynamic, iterative and responsive to change. In addition to the FRAME principles, Webster et al. (2013) also listed 20 principles for effective risk management, many of which overlap with those listed above (Webster, Haque and Hucker, 2013).

A recurring theme in the SPJ literature is that the risk assessment process should incorporate a wide range of evidence from different sources. The interaction between risk assessment and risk management should also be iterative, with future risk judgements being updated on the basis of the outcomes from risk management interventions. Importantly, unlike other approaches to risk assessment, cases assessed through SPJ are not given a 'risk score' or quantitative estimate of risk. Rather, the focus is on *case prioritisation*; identifying the level of management required for each individual. Case prioritisation can change depending on individual circumstances, and should be subject to ongoing review. Therefore, a crucial aspect of the SPJ approach is that the risk assessment is continuously reviewed and the risk management plan modified as needed.

SPJ methods are now widely used in clinical settings. The fully operationalised SPJ approach is typically carried out by experienced and trained assessors, such as psychologists and behavioural scientists. SPJ is time and resource intensive, and therefore arguably not appropriate in situations where the aim is to prioritise or triage a large volume of cases for more detailed, manual review. An alternative version of SPJ – "SPJ lite" (Logan and Lloyd, 2019b) – can be deployed in situations where assessors do not have specialised behavioural expertise, but nevertheless have access to relevant information or sources, such as police officers (McEwan, Bateson and Strand, 2017). Rather than requiring the assessor to consider potential scenarios and make recommendations for RMPs, the SPJ-lite approach requires assessors to identify risk factors and produce a summary risk rating of high, medium, or low, based on an overall assessment of the pattern of risk factors present in the case (Logan and Lloyd, 2019b). As detailed in the following section, several discretionary risk assessment frameworks are used throughout the UK CJS, which can be loosely characterised as a form of 'SPJ-lite'.

2.3 Offender risk assessment in the UK CJS today

The previous sub-section tracked the evolution of offender risk assessment methods throughout the mid to late-20th century. This sub-section now focuses on the current practice of offender risk assessment within the UK CJS, with particular focus on the structured tools currently used to support the offender risk assessment process, and the so-called 'next generation' data scoring systems that have been enabled by recent development in data science and machine learning technology.

2.3.1 Structured risk assessment tools in the UK CJS

Existing risk assessment frameworks used throughout the UK CJS combine the use of ARAIs and 'SPJ-lite' guides to assess risk of harm posed by individual service users and to develop individualised risk management plans. These actuarial and SPJ-lite assessment tools are referred to collectively herein as 'structured risk assessment' (SRA) tools.

Consistent with the RNR model, SRA tools are central to service delivery within the UK CJS. For instance, all offenders subject to Multi-Agency Public Protection Arrangements (MAPPA) undergo a thorough risk assessment incorporating various SRAs, to assess their likelihood of reoffending and categorise them according to their risk of serious harm (HM Prison & Probation Service, 2019). All MAPPA offenders are assigned a 'risk level' (low, medium, high, very high) on the basis of this assessment. This risk level determines the level of risk management required for each service user and informs the development of the individualised RMP.

Numerous SRA tools are used to risk assess service users in the UK CJS, and it is beyond the scope of this study to provide an exhaustive analysis of each. However, a brief summary of the most commonly used SRA tools provides important context for the discussion that follows:

i. **The Offender Assessment System (OASys)** is the national risk and needs assessment tool for adult offenders, and is routinely used by the police and HMPPS to measure individuals' likelihood of reoffending and to develop individual risk management plans (Howard, Clark and Garnham, 2003). OASys incorporates both static risk factors (such as age and criminal history), and dynamic risk factors (such as accommodation, employment, relationships and substance use), which allow progress and changes in offender behaviour to be monitored over time (Moore, 2015). OASys includes two predictive models: one for general (i.e. all recordable) reoffending (OGP1), and one specifically for violent reoffending (OVP1). OASys also incorporates Risk of Serious Harm (RoSH) ratings, determined through structured professional judgement, which allow assessors to prioritise public protection issues and identify requirements, conditions and controls for managing specific risks.

ii. **The Offender Group Reconviction Scale (OGRS)** is an actuarial tool used by HMPPS to assess reoffending risk at pre-sentence court report stage, post-sentence, or for offenders who receive out of court disposals (Copas and Marshall, 1998; Howard *et al.*, 2009). As OGRS includes only a limited range of static risk factors (age, gender and criminal history), it can be used on a wider group of offenders than OASys, for instance in situations where it is not possible to complete a more thorough assessment. The fourth iteration of OGRS includes two predictive models: one for general (i.e. all recordable) reoffending (OGRS4/G), and one specifically for violent reoffending (OGRS4/V). OGRS is now incorporated within OASys, meaning that an OGRS score is calculated for all offenders assessed through OASys.

iii. **The Risk Matrix 2000 (RM2000)** is the NPCC accredited risk assessment tool used by the police to assess adult male MAPPA offenders (Thornton *et al.*, 2003). RM2000 comprises two actuarial tools for predicting sexual recidivism (RM2000/S) and nonsexual violent recidivism (RM2000/V). The scores of both tools can also be combined to give an overall risk of reconviction (RM2000/C). RM2000 uses a stepwise approach to scoring, incorporating only static risk factors (the majority of which relate to offending history). According to the authors, an advantage of using only static variables and not taking into account complex psychological factors is that the tool 'can appropriately be used by probation officers, police officers, and correctional personnel, so long as they are given specific training', and that 'the data it employs are of a kind that can easily be routinely collected and computerized' (Thornton *et al.*, 2003, p. 233).

iv. **The Active Risk Management System (ARMS)** is an SPJ framework developed to assist practitioners in assessing risk and strengths of offenders with sexual convictions (Kropp and Hart, 2000; Kropp and Gibas, 2010; Helmus and Bourgon, 2011). It has subsequently been implemented across all police areas in England and Wales. ARMS was based in part on the RM2000, but sought to provide a holistic assessment of dynamic factors related to both risk and protective factors that would be clinically helpful for practitioners. ARMS involves the assessor collating evidence from various sources and using this information to assign an overall case priority level, and hypothesise potential future scenarios in which offending might occur. Based on this assessment, the assessor develops an individualised risk management strategy and details the specific actions required to implement the strategy.

v. **ASSET**, first introduced by the Youth Justice Board in 2000 (Youth Justice Board, 2000) is an actuarial risk assessment tool for young people used by all youth offending teams (YOTs) in England and Wales. ASSET includes four static risk factors (offence type, age at first reprimand/caution/warning, age at first conviction, and number of previous convictions) and 12 dynamic risk factors (such as living arrangements, education and employment, lifestyle and mental health) (Wilson and Hinks, 2011). YOT workers are required to assign a rating from 0 to 4 to each of these risk factors to arrive at a score ranging between 0 and 64. Individuals are then grouped into three levels of interventions: standard (0–14), enhanced (15–32), and intensive (33–64) (Wilson and Hinks, 2011). A comprehensive 2004 study revealed two important findings of relevance to the current project. The research found that experienced practitioners were less likely to see ASSET as having value for their work, whereas new practitioners found that the tool provided a helpful structure for assessment. The research also found that a potential advantage of using a structured assessment tool was that it could lead to the discovery of information that may not otherwise have been identified (Burnett and Appleton, 2004)

vi. **The Violence Risk Appraisal Guide (VRAG),** first developed by Harris, Rice and Quinsey in 1993, is a purely actuarial method for predicting violent recidivism, which assigns weighted scores to different static risk factors and then combines item responses to produce an overall risk score (Harris, Rice and Quinsey, 1993). VRAG uses 12 static input predictors to assign individuals a score ranging from -27 to +35, with a higher score indicating greater likelihood of future violence. A related tool, the Sex Offender Risk Appraisal Guide (SORAG), was developed specifically to assess risk of recidivism for sexual offending, and more recently the Violence Risk Appraisal Guide–Revised (VRAG-R) has now combined and replaced VRAG and SORAG (Quinsey *et al.*, 2006). As noted by the Scottish Risk Management Authority, as VRAG is composed solely of static risk factors with the purpose of predicting recidivism, it does not have the capacity to inform treatment protocol or monitor offender progress. As discussed by Daffern (2006), owing to the tool's reliance on static risk factors, it 'does not provide staff with dynamic appraisals of risk level or assist in the identification of treatment targets that might remediate risk level. Nor, like many other schemes does it assist in the identification of important dimensions of risk including the nature, severity, frequency and imminence of future violence' (Daffern, 2007). Nevertheless, a 2009 study by Khiroya et al. found that SORAG and VRAG were widely used for violence risk assessment in medium secure forensic units in the UK (Khiroya, Weaver and Maden, 2009).

vii. **The Spousal Assault Risk Assessment Guide (SARA)** focusses specifically on spousal violence (intimate partner violence, 'IPV') (Kropp and Hart, 2000; Kropp and Gibas, 2010; Helmus and Bourgon, 2011). The SARA incudes 20 risk factors, 10 of which relate to violence risk in general and 10 of which relate specifically to risk of spousal violence. Static risk factors include those relating to criminal history, previous violence and personality disorder, while dynamic risk factors include considerations such as recent substance abuse, recent relationship problems, recent suicidal or homicidal ideation, and recent psychotic or manic symptoms. Earlier versions of the SARA incorporated a numerical system, but the most recent version simply requires assessors to determine whether each of the 20 factors are 'absent', 'possibly present' or 'present'. Assessors should also note the presence of any additional case-specific risk factors and evaluate the overall degree of risk posed by the individual, taking into account the nature, severity, likelihood, frequency, and imminence of any future violence.

As discussed by Campbell et al., the selection of which risk assessment tool to use should be informed by the specific context in which it is being delivered and the overall objective of the risk assessment process (Campbell, French and Gendreau, 2009). Andrews & Bonta stress the importance of adhering to the RNR principles; i.e. that the purpose of risk assessment should be to identify specific individual needs and context-specific interventions that can be delivered to effectively mitigate identified risks (Andrews and Bonta, 2010).

However, in many cases, the SRAs listed above are not used for risk management purposes, but for 'screening', i.e. to identify a smaller subset of a given population of offenders who require further, more detailed risk assessment. As discussed by Cooke, this is potentially problematic, as there is a lack of empirical evidence that these instruments are in fact effective 'screens', as this is not the purpose for which they were created (Cooke, 2010). Moreover, 'in practice this rarely happens; the social worker and the police officer do not have the time – and they probably do not have the training – to provide the systematic risk assessment required if the offender is caught in the screen. The decision maker in court is provided with the results of the actuarial scale without any consideration of certitude or risk formulation' (Cooke, 2010). This again emphasises the crucial distinction between group-level prioritisation and individual-level risk prediction, and the need to clearly distinguish between the two categories when implementing structured approaches to offender risk assessment.

Across all settings, assessors are strongly advised to exercise their discretion and professional judgement, rather than relying solely on the use of SRA tools. For instance, the College of Policing's *Authorised Professional Practice* (APP) notes that 'it is impossible to use the information derived from a formal risk instrument to predict with certainty the behaviour of an individual or the outcome of a particular situation. RI [risk identification], RA [risk assessment] and RM [risk management] tools should be regarded as an excellent but limited, means of improving the likelihood of identifying and preventing future offending or victimisation. They can enhance professional judgement but not replace it' (College of Policing, 2014a). Similarly, the official MAPPA guidance stresses that 'Risk assessment must never become formulaic. There must always be a place for discretion and professional judgment. Static and dynamic indicators and protective factors should be taken into account when determining the overall risk of reoffending and risk of serious harm and deciding upon the level of management' (HM Prison & Probation Service, 2019).

However, questions remain over the extent to which assessors engage their discretion and professional judgement *in practice* when completing offender assessments. In a 2015 questionnaire-based study exploring probation service and prison service assessors' views of OASys, 24% of respondents reported that the amount of professional judgement required to complete an OASys assessment was too little, with one respondent suggesting that '…sometimes the way things are worded and some of the tick boxy bits of OASys make it feel like professional judgement isn't being exercised' (Moore, 2015, p. 28). Moreover, the element of the OASys framework that was most commonly viewed as being useful to assessors was the Risk of Serious Harm (RoSH) component, which is based on SPJ rather than statistical prediction (Moore, 2015, p. 26). Interviewees highlighted the strength of the SPJ approach to RoSH assessment, suggesting that this was easier to understand than actuarial scoring. However, the same study also found that actuarial scoring performed significantly better than SPJ RoSH assessments when predicting 'grave reoffending', concluding that public protection could be improved by increasing the weighting of actuarial scores on RoSH ratings (Moore, 2015, p. 151). These findings suggest that it cannot be assumed that a professional judgement-based approach is preferable to a purely statistical approach in all risk assessment contexts.

In summary, further practitioner-focused research is required to explore the interaction between the risk assessment tool and the user, and specifically the role of discretion and professional judgement in the structured assessment process. The same risk score can be used for numerous purposes: for instance for intelligence purposes to identify individuals who should be subject to ongoing monitoring; for offender management prioritisation to 'triage' individuals most in need of more in-depth risk assessment; or at the point of sentencing or parole hearings to assist in determining what level of intervention is appropriate for any given offender. While numerous statistical risk assessment tools are now in use throughout the UK CJS, evidence is lacking on how these are used in practice, and specifically how they can be most effectively integrated into existing offender management practices.

2.3.2 Next-generation data scoring tools

While the use of actuarial and "SPJ-lite" risk assessment tools is now standard practice across the UK CJS, a more recent development is the use of complex algorithms and machine learning models to "risk score" offenders based on statistical analysis of police-recorded data. These next-generation data scoring tools are founded on the same theoretical principles as the traditional ARAIs discussed previously, but with several key differences. First, the statistical modelling used has become increasingly sophisticated, often incorporating the use of machine learning (ML). Discussed further in Section 2.3.3, ML differs from other forms of statistical modelling as traditional modelling aims to derive *inferences* by plotting a line of best fit across a single distribution, to characterise the relationship between the input data and the outcome variable. By contrast, ML aims to 'train' a model on a subset of historic data, to then make predictions on new, unfamiliar data. The contrast between statistical inference on the one hand, and ML predictions on the other, is an important distinction in the context of offender risk prediction. Second, improvements in computing power and exponential growth in data availability means these systems now have access to far more data from a wider range of sources. Finally, the resulting data scores are increasingly used by the police for *intelligence* purposes, rather than solely the development of offender management plans. These developments are discussed in detail in the following sections.

The growth in 'next-generation' data scoring tools is part of a wider shift towards 'predictive policing' experienced throughout the UK CJS in the past 10-15 years. Historically, predictive policing research focussed on location-based predictive analytics, i.e., the use of statistical modelling to analyse historic data and make forecasts about *where* crime is most likely to happen in the near future (Babuta and Oswald, 2020). The use of such technology dates back to at least 2004 (Bowers, Johnson and Pease, 2004), and recent research suggests that at least 12 (of 43) police forces in England and Wales are currently using or developing predictive crime mapping systems, or have done so in recent years (Couchman, 2019). Predictive mapping is based on the well-observed phenomenon of repeat and near-repeat victimisation. Crime is often 'contagious', and the risk of crime greatly increases in the immediate geographic vicinity in the aftermath of an initial offence, with this risk then decaying over time (Townsley, Homel and Chaseling, 2003; Ludwig and Kling, 2007). Research has repeatedly shown that the use of predictive mapping software consistently increases the likelihood of detecting future crime and results in net reductions in overall crime rates, and its use is now widely advocated by academic criminologists (Johnson *et al.*, 2007; Braga and Bond, 2008; Guerette and Bowers, 2009; College of Policing, 2013; Mohler *et al.*, 2015).

Despite the large body of empirical evidence in favour of predictive crime mapping, the practice has been subject to considerable criticism. Most notably, critics have highlighted the potential risk of bias and discrimination when using historic arrest data to make predictions regarding future criminal offending (Brayne, 2017; Ferguson, 2017; Selbst, 2017; Couchman, 2019; Richardson, Schultz and Crawford, 2019). For instance, a 2019 report from Liberty recommended that 'police forces in the UK should end their use of predictive policing "mapping" programs, which rely on problematic historical arrest data and encourage the over-policing of marginalised communities' (Couchman, 2019). However, it is important to note that in practice predictive mapping typically relies on recorded crime data. Arrest data is rarely (if ever) used for predictive purposes, as it is widely recognised that arrest data is not representative of the underlying distribution. In addition, the vast majority of analysis purporting to indicate bias in predictive policing technology has been conducted in the US, and there is insufficient evidence to conclude that these observations are transferable to the UK context (Babuta and Oswald, 2020). Very few studies have empirically examined the extent of bias in predictive crime mapping algorithms. The only randomised controlled trial that has been conducted found no significant differences in the proportion of arrests by racial-ethnic group between locations where mapping software was and was not deployed (Brantingham, Valasik and Mohler, 2018). It is beyond the scope of this thesis to elaborate further on the use of predictive crime mapping in UK policing. In summary, the practice remains the subject of considerable debate and controversy, particularly concerning the risk of biased or discriminatory outcomes – but there is no empirical evidence that predictive mapping does in fact lead to over-policing of minority groups or any other form of unfair discrimination.

Data-driven offender risk assessment can be understood as a different form of predictive policing technology, where the focus of analysis is not geographical locations – but individual offenders or suspects. Of the 43 police forces in England and Wales, only a small number (fewer than 10) are known to be currently using 'next-generation' offender risk assessment tools. While definitive data is lacking, the author's previous research has sought to map the data-driven policing landscape and identify the most significant data-driven risk assessment projects currently underway nationwide (Babuta and Oswald, 2020). Among others, this research has identified the following case studies as particularly notable examples:

i. Durham Constabulary's Harm Assessment Risk Tool (HART), a random forest forecasting model which uses 34 predictor variables to assign offenders into low-, medium- or high-risk groups corresponding to their predicted likelihood of re-offending over a 24-month period. The system is used to assess offenders' eligibility to participate in the *Checkpoint* programme, an out-of-court disposal scheme designed to reduce offending by addressing the root causes of criminal behaviour (Urwin, 2016). The risk score is intended to be used as one of many factors that officers should consider making an overall risk assessment, enabling more effective targeting of offender intervention programmes.

ii. Hampshire Constabulary's domestic violence risk forecasting algorithm, which aims to improve existing domestic violence risk assessment tools by including an additional perpetrator-based risk classification, calculated using machine learning (Babuta and Oswald, 2020).

iii. Avon and Somerset Constabulary's *Qlik Sense* analytics software, which uses predictive modelling to produce individual risk profiles and assist the force in triaging offenders according to their perceived level of risk in various categories, including likelihood of offending, risk of offending escalation and risk of becoming a victim of crime (Dencik *et al.*, 2018a). The Qlik dashboard synthesises data across the force's internal data systems and local authority datasets, and produces daily risk scores for all offenders on the force crime and intelligence database. The machine learning models used for the Qlik dashboard predictions are subject to ongoing re-validation to ensure their accuracy is maintained (Babuta and Oswald, 2020).

iv. The 'Most Serious Violence' (MSV) use case being developed as part of the National Data Analytics Solution (NDAS), led by West Midlands Police (WMP). The MSV use case applies machine learning modelling to historic data, to identify factors of individuals who escalate from low-level offending to serious violent offending. The model is then used predictively to identify a cohort of individuals who match a number of these key predictive factors but have not yet committed a serious violent offence (West Midlands Police, 2020).

v. The West Midlands Police 'Integrated Offender Management' (IOM) model, which uses machine learning prediction to assign a 'harm score' to every offender in the force intelligence database, and identify individuals who are escalating towards more serious offending (West Midlands Police, 2019a).

vi. The Essex Police 'Fearless Futures' knife crime and violence model, which uses machine learning modelling to identify individuals at heightened risk of becoming involved in serious violence, to target effective multiagency support services aimed at assisting an individual's exit from criminality (Essex Police, 2021).

As data science technology has developed and police forces face increasing pressure to 'do more with less', efforts have focussed on developing new approaches to allocate limited resources most efficiently based on a data-driven assessment of risk and demand. These developments are characteristic of a broader proliferation in the use of 'data scoring' tools to support service delivery across the UK public sector (Dencik *et al.*, 2018b). However, these novel data scoring tools have not been subject to nearly the same level of academic scrutiny or empirical validation as the traditional, rules-based ARAIs discussed previously. The author's previous research has concluded that 'the development of policing algorithms is often not underpinned by a robust empirical evidence base regarding their claimed benefits, predictive accuracy, scientific validity or cost effectiveness.' (Babuta and Oswald, 2020, p. 9). This is a particular concern in relation to individual-level assessment tools which are used to score offenders according to their perceived risk of (re-)offending.

The next-generation data scoring tools detailed above share certain commonalities, but there are important differences in the approach taken to development and implementation. For instance, some have been developed entirely in-house or with the support of academic partners, and do not rely on third-party commercial technology. Others are delivered by third-party commercial providers. A reliance on third-party providers to develop such risk scoring tools gives rise to several operational and policy considerations. While commercial and procurement considerations are largely outside the scope of the current study, they must nevertheless be accounted for when developing future policy for the use of new data science tools in the English CJS. A particular concern in this regard is the extent to which the agency procuring the technology can meaningfully evaluate the system's performance, or test the model for unacceptable bias.

Only a minority of UK police forces are currently using next-generation data scoring systems operationally, and the approach to implementation can be characterised as fragmented and uncoordinated. Nevertheless, the examples above indicate a clear trend towards increased use of advanced analytics and 'big data' technology to support risk assessment at the individual person level. A natural evolution of the actuarial approach to offender management, it seems the UK CJS is now entering a new era of 'data-driven offender risk assessment', where a core focus is on prioritising limited resources according to a statistical assessment of risk and demand. As technology evolves and such tools become increasingly embedded across the CJS, this will have far-reaching implications for the approach taken to offender risk management and rehabilitation, specifically when considering the extent to which risk management plans are tailored to the individual needs of offenders. The lack of criminological research evidence supporting the use of novel data-driven risk assessment tools is concerning, and should be addressed as a priority before such methods become embedded within professional practice.

2.3.3 Developments in machine learning

The next-generation data scoring tools detailed above have been enabled primarily by developments in machine learning (ML), one of the core components of artificial intelligence (AI). AI can be understood as a general purpose technology, which 'encompasses a huge variety of subfields, ranging from the general (learning and perception) to the specific, such as playing chess, proving mathematical theorems, writing poetry, driving a car on a crowded street, and diagnosing diseases.'(Russell and Norvig, 2002, p. 1). Modern AI is underpinned by machine learning algorithms. An algorithm can be defined as 'a set of mathematical instructions or rules that, especially if given to a computer, will help to calculate an answer to a problem.'[1] Non-learning algorithms are static, meaning the content of their mathematical instructions is pre-defined and will not change, regardless of the data they are exposed to. By contrast, ML algorithms build a 'model' based on their training data, which is then applied predictively on new, unfamiliar data.

There are three main types of machine learning: supervised; unsupervised; and reinforcement learning.

- **Supervised learning** involves the use of pre-labelled input-output pairs to teach the agent a function that maps from input to output. For example, for object classification, training data could include many photographs of different types of fruit, and labels defining which fruit is in each photo. The trained model 'generalises' well if it correctly identifies the type of fruit when presented with new, unfamiliar photos.
- **Unsupervised learning** involves the agent identifying and learning patterns in the input data in the absence of any explicit feedback. For image recognition, training data could include thousands of individual photographs of 5 types of animal, but no labels identifying the animals. The model performs well if it correctly divides the photographs into 5 piles, each containing the photos of one type of animal.
- **Reinforcement learning** is a goal-oriented form of learning, where the agent improves at a task over time based on exposure to positive and negative feedback. For recommender systems, a human listener may be recommended music based on their previous listening habits. The user provides feedback indicating whether they like the computer-recommended track. This feedback helps the model to learn the user's listening preferences, meaning that the recommendations become more accurate over time.
- **Semi-supervised learning** is a fourth category of ML, involving datasets where some input-output pairs are labelled but a large proportion are unlabelled. Returning to the fruit classification example, the model can be pre-trained on the entire training set (using unsupervised methods), before it is fine-tuned using the labelled subset.

For further discussion, see (Russell and Norvig, 2002).

[1] Definition of 'algorithm' from the Cambridge University Dictionary & Thesaurus, Cambridge University Press.

The use of machine learning can offer significant improvements when compared with traditional data science techniques. This is particularly the case for tasks that require processing very large volumes of data, such as machine translation or image recognition. Developments in deep learning (particularly deep neural networks) have driven many of the major AI breakthroughs witnessed in recent years. Notable examples include Google's *AlphaGo,* an AI system underpinned by deep neural networks that defeated several *Go* world champions*,* an ancient Chinese board game widely considered to be the most complex game to master (Chen *et al.*, 2018); large language models such as GPT-3, which uses generative pre-training to generate highly fluent prose that is difficult to distinguish from text written by humans (Dale, 2021); and facial recognition systems that rely on convolutional neural networks to produce highly accurate results for both retrospective and live facial matching tasks (Hu *et al.*, 2015).

In the context of offender risk assessment, supervised machine learning methods (most notably random forest forecasting, Bayesian trees and gradient boosting) have shown significant potential when compared with traditional statistical modelling. Traditional models used for risk assessment tools such as OASys and OGRS use input variables that are pre-programmed by the developer, with logistic regression being the statistical method of choice. Tests of correlation (such as Chi-Square tests) are run on a historic dataset to identify the most powerful 'predictors' of future offending. These predictor variables are then used as the input for a regression model, which is applied predictively on a new dataset of individuals. Based on the weighted value of each input variable, the model then calculates a probability of each individual (re)-offending, in order to generate a 'risk score' for each nominal in the dataset.

It is important to note that logistic regression modelling can still be considered a form of machine learning. However, traditional, logistic regression-based models are fundamentally limited in several ways. A regression model estimates the probability of an instance belonging to a single class (i.e. a single binary dependent variable). The use of linear decision boundaries can often lead to inaccurate forecasting, with research demonstrating that the predictions generated by such models can produce a false negative rate of up to 99.7%, rendering them unusable in a criminal justice setting (Berk *et al.*, 2009). Other forms of supervised learning such as random forest forecasting do not assume the model has a linear relationship, meaning the model can be adjusted depending on the distribution of the underlying data; features can be weighted depending on their potential influence on the outcome variable. Moreover, ML techniques such as random forests and gradient boosting are 'ensemble methods', meaning they use multiple algorithms to derive better predictive performance (Rokach, 2010) – in contrast to the single algorithm used to build traditional regression models.

Research has sought to compare the statistical accuracy of new machine learning-based models with traditional logistic regression models (Breiman, 2001; Berk, 2012; Berk and Hyatt, 2015) This research has consistently demonstrated the benefits of ensemble approaches, due to the ability to model for more than two outcome variables and the non-linearity of the modelling techniques, meaning the model can be adjusted and weighted according to different types of error and cost ratio (Berk and Bleich, 2013; Berk, Sorenson and Barnes, 2016; Urwin, 2016). For instance, in the criminal justice setting, a random forest model can be weighted more heavily to avoid false negatives (i.e. under-predicting the risk of future offending) rather than false positives (over-predicting risk), if false negatives are judged to be a more 'costly' error than false positives. This 'customisability' of classification boundaries for models such as random forests has made them the technique of choice for recent data-driven risk assessment tools, such as Durham Constabulary's HART model (Urwin, 2016). Classification thresholds are discussed further later in this thesis.

While ML performance continues to improve both in terms of speed and accuracy, the use of increasingly complex model architectures (particularly deep learning) has made it more difficult for human operators to understand how ML models arrive at their outputs (Gleaves, Schwartz and Broniatowski, 2020). Much discussion has focused on the so-called 'black box' problem associated with contemporary AI systems (McGovern *et al.*, 2019; Watson *et al.*, 2019; Azodi, Tang and Shiu, 2020), giving rise to a new sub-field of data science research focussed on developing 'explainable AI systems' (often referred to as 'XAI' research) (Gunning and Aha, 2019; Gunning *et al.*, 2019; Arrieta *et al.*, 2020). The computational complexity of deep learning systems means that only the inputs and outputs may be observed directly, and not the specific calculations made to arrive at any given prediction.

The black box nature of many machine learning systems is particularly problematic in high-stakes decision-making contexts such as criminal justice and law enforcement, where it is necessary to maintain a clear and auditable record of the factors that led to a certain decision being made (Rudin, 2019; Knack, Carter and Babuta, 2022). For this reason, it is important to assess whether the potential accuracy improvements that could be gained from using machine learning-based systems are worth the potential loss in explainability that could result. This is an important consideration in the context of the current study and is discussed further in the following sections.

2.4 Legal, ethical and operational considerations

The rise in the use of algorithmic risk assessment in policing raises various legal and ethical concerns, which have been discussed at length in the literature (Ferguson, 2017; Bayamlıoğlu and Leenes, 2018; Bennett Moses and Chan, 2018; Jansen, 2018; Oswald, 2018; Grace, 2019; Richardson, Schultz and Crawford, 2019). The author's own research has highlighted the limited evidence regarding the efficacy of different systems, their impact on individual rights and the extent to which they serve valid policing aims (Babuta, Oswald and Rinik, 2018; Babuta and Oswald, 2020, 2021). There is also a need for further research to explore the extent to which police decision-making is influenced by algorithmic risk assessment tools in practice, and the role of discretion and human judgement in this process (Lynskey, 2019).

Data-driven risk assessment offers significant opportunities to target preventative interventions to the areas of greatest need. If deployed effectively, the technology could enable more accurate decision-making, improving the police's ability to manage risk and vulnerability, thereby supporting core policing aims of protecting the public and rehabilitating offenders. However, these approaches introduce new operational and ethical challenges, and the potential benefits they present are yet to be evidenced in practice. Further research is needed to assess the relative benefits and risks of these new techniques in an operational policing context, before they are deployed in a way that could interfere with individuals' human rights and civil liberties.

2.4.1 Predictive validity

A fundamental challenge in the field of offender risk assessment lies in demonstrating the relative accuracy or effectiveness of different methods. This is crucial not just from an operational perspective, but also to ensure proportionate use of potentially intrusive tools and techniques. If the benefits resulting from the use of a particular data-driven method have not been clearly established, it becomes challenging to argue that any potential interference with rights or freedoms arising from the use of the tool is necessary, proportionate and in accordance with the law (Babuta and Oswald, 2021).

Randomised control trials are inappropriate in the context of public safety: it would be unethical *not* to intervene when an individual is expected to commit a criminal act, for the purpose of determining whether the prediction was accurate. For this reason, retrospective validation is the most commonly used method to evaluate offender risk assessment tools, alongside studies of inter-rater reliability. These evaluation methods are fundamentally limited as they offer little insight into whether a particular tool is useful *in practice* to inform the development of individual-specific risk management plans.

Experts continue to disagree over the predictive validity of structured risk assessment tools (Hart, Michie and Cooke, 2007; Hart and Cooke, 2013; College of Policing, 2014a). Predictive validity can be understood as 'the extent to which scores on an assessment tool are able to predict some outcome measure' (Debidin, 2009). The most commonly used measurement for predictive accuracy is the Area Under the Curve (AUC) of the Receiver Operating Statistic (ROC) (Douglas and Webster, 1999). The ROC plots the true positive rate against the false positive rate, and the AUC is a measurement of the total area within the resulting curve. An AUC value of 1 denotes perfect prediction (100% accuracy), while a random model achieves a value of 0.5 (chance prediction). Nicholls et al. suggest that an AUC of 0.75 or above can be considered as a moderate to large effect size (Nicholls, Ogloff and Douglas, 2004). An advantage of AUC is that it is unaffected by variations in selection ratio and base rate, allowing direct comparisons of accuracies of different tests used with different selection ratios and base rates (Rice and Harris, 2005).

Evidence suggests that the ARAIs currently in use in the UK CJS demonstrate good predictive validity. A comprehensive systematic review and meta-analysis conducted by Farrington et al. sought to compare the accuracy of the most widely used risk assessment tools based on AUC results. The review assessed 31 studies which included analyses of 9 different structured risk assessment methods. Based on AUC results, the authors concluded that 'all risk assessment instruments included in this review performed significantly better than chance in the prediction of future violence' (Farrington, Jolliffe and Johnstone, 2008a). Specifically for the UK context, analysis of OASys data suggests that both models have good predictive validity (for OGP1, AUC=0.79; for OVP1, AUC=0.74) (Moore, 2015). In relation to OGRS, AUC analysis of historic data suggests that both tools have good predictive validity (for OGRS4/G, AUC=0.77; for OGRS4/V, AUC=0.76). ASSET was also found to have good predictive validity (AUC=0.70), but better validity when the dynamic risk factors are calculated in conjunction with OGRS 3 (AUC=0.73) (Wilson and Hinks, 2011).

However, despite numerous validation studies purporting to indicate good predictive validity of the assessment tools currently in use, it has been argued that the AUC statistic is fundamentally misleading in the context of offender risk assessment due to the very high margins of error often involved (Cook and Paynter, 2011; Sutherland *et al.*, 2012; Cooke and Michie, 2014). Furthermore, the predictive validity of actuarial tools invariably 'shrinks' when they are used on samples different from those on which they were developed (Webster, Haque and Hucker, 2013). As summarised by Douglas et al., 'highly optimized actuarial estimates […] tend to vary as a function of a myriad of possible study and sample features', and 'to state that any given person has some precise probability of reoffending (say, 40%) is potentially highly sample-dependent. The same person could have a 60% (or 20%, or y%) probability of reoffending depending on sample characteristics, methodology, definition of violence, follow-up period, and so forth.' (Douglas, Yeomans and Boer, 2005, pp.501-502).

For these reasons, other authors have argued that 'existing data suggests that most risk assessment tools have poor to moderate accuracy in most applications' (Douglas *et al.*, 2017, p. 135). A 2012 systematic review and meta-analysis of the nine most commonly used SRA tools also found considerable variation in predictive accuracy depending on how the tool is used (Fazel *et al.*, 2012). The authors concluded that 'the view that violence, sexual, or criminal risk can be predicted in most cases is not evidence based', that 'these tools are not sufficient on their own for the purposes of risk assessment', and that 'actuarial instruments focusing on historical risk factors perform no better than tools based on clinical judgement' (Fazel *et al.*, 2012, p. 5). Research has also found that long-term external validation of SRA tools is uncommon, and when it does occur, predictive accuracy is generally reduced (Siontis *et al.*, 2015).

In summary, there remains a concerning lack of reliable and unbiased data regarding the 'predictive accuracy' of offender risk assessment tools. But beyond concerns regarding statistical validity, it has also been argued that "What Works" is probably the "wrong question" in the context of offender rehabilitation (Maruna, 2001; Lin, 2002; Farrall, 2004; Mair, 2013). "Works" implies some degree of predictable consistency (Ward and Maruna, 2007), which is paradoxical in the crime prevention context. When individuals are judged to pose a risk of offending, an intervention is typically delivered which prevents the predicted outcome from happening. If an intervention is delivered on the basis of the assessment, we cannot know what may have happened had we not intervened, and therefore there is no way to test the accuracy (or otherwise) of our prediction. For these reasons, rather than focussing on the 'predictive validity' of the assessment tool, it would be more constructive to address the question of "what helps" offenders to desist from (violent) crime, a question that does not lend itself to a statistical approach (Maruna *et al.*, 2004, p. 13).

Regardless of concerns over statistical validity, perhaps the most fundamental limitation with the actuarial approach is that a 'risk score' provides no insight into the nature of the specific risk to be prevented, the precipitating factors that may lead an individual to offend again in the future, nor measures that could be taken to mitigate this risk. Actuarial tools therefore offer little guidance to practitioners seeking to develop an offender-specific risk management plan aimed at reducing the influence of these individual and situational precipitators. As summarised by Hart (1998), 'to put it simply, the clinical task is violence *prevention*, not violence *prediction*' (Hart, 1998, p.123). The College of Policing's *Authorised Professional Practice* also recognises that actuarial prediction 'is recognised as more accurate than unstructured judgement, but is inflexible and blind to specific contexts' (College of Policing, 2014b).

Considering the issues outlined above, it could be argued that offender risk assessment research has become overly-focussed on issues concerning statistical accuracy and predictive validity, often neglecting to address the fundamental question of whether a particular tool is helpful *in practice* to enable interventions to be appropriately tailored to the specific risks, needs and circumstances of the individual. Of the studies examined for this literature review, few sought to include the perspectives and views of practitioners, and none incorporated the perspectives of service users themselves.

Of the few studies that have sought to assess the practical utility of SRA tools for risk management purposes, the most comprehensive findings are to be found in a 2018 systematic review conducted by Viljoen et al. The authors analysed the results of 73 studies, and found mixed results in the perceived utility of SRA tools for risk management; some professionals viewed the tools as useful whereas others did not. The authors also found that "slippage" often occurs between risk assessment and risk management, meaning SRA tools are not consistently used to guide the development and delivery of risk management plans. Notably, the review found limited adherence to the "need" principle following the use of SRA tools, meaning that the use of such tools does not consistently assist practitioners in developing strategies to target specific criminogenic needs. The authors conclude that there is 'insufficient evidence from non-randomized studies to conclude that risk assessment tools reduce violence and offending', and that 'use of risk assessment tools does not consistently improve assessees' outcomes, nor does it consistently improve professionals' risk management practices' (Viljoen, Cochrane and Jonnson, 2018, p. 23).

Of the few studies that have examined UK criminal justice practitioners' perceptions of SRA tools, the most relevant findings are reported in the 2015 Ministry of Justice Compendium of Research and Analysis on the Offender Assessment System (OASys). Pike and Smith-Yau conducted a comprehensive qualitative study to capture prison and probation assessors' views and experiences of OASys, with a particular focus on identifying potential improvements (Moore, 2015). This involved an online self-completion questionnaire (eliciting 1,093 responses) followed by individual structured interviews with twelve assessors. The findings were largely positive, with 89% of respondents reporting that the information recorded in an OASys assessment supported them very well or fairly well in managing offenders' risks and needs. As mentioned previously, the RoSH ratings were reported as being the most useful component of the OASys assessment, which is notable considering this is an SPJ rather than actuarial assessment.

Another questionnaire-based study by Farrington et al. (2008) sought to assess the extent to which professionals perceived existing SRA tools as adhering to RMA Scotland's best practice Standards and Guidelines for Risk Assessment (Scottish Risk Management Authority, 2011). On the basis of 29 completed questionnaires, the authors found that most instruments met most standards, but notably none of the SRA tools were assessed as adequately enabling the identification of protective factors (Farrington, Jolliffe and Johnstone, 2008b). However, the authors note that this study is inherently limited by the small sample size, limiting generalisability of the findings.

In summary, experts continue to disagree regarding the predictive validity of the offender risk assessment tools currently in use, and there remains a notable gap in research exploring the practical use of such tools from the perspective of criminal justice practitioners, in particular whether they are useful in practice to enable more effective identification of future offending risk. This is the gap that this study seeks to address.

2.4.2 The legal context

Much public discourse regarding police use of AI and data analytics has focussed on the legal framework governing its use, and perceived deficiencies regarding the safeguards currently in place (Couchman, 2019; Lynskey, 2019; Richardson, Schultz and Crawford, 2019). However, much of the concern raised in this regard tends to focus on US case studies, and it is unclear the extent to which these same concerns are applicable to the UK policing context, which operates within a very different legal framework (Babuta and Oswald, 2021). To contextualise the discussion regarding new data-driven risk assessment tools and their potential benefits and risks for UK policing, it is important to briefly assess the legal frameworks relevant to their use.

The legal frameworks governing policing powers in England and Wales are primarily principles-based. This means that a strong emphasis is placed on the role of discretion and professional judgement in making context-based decisions, rather than imposing prescriptive rules. The underlying legal authority of the police in England and Wales is not laid out in primary legislation or statute; rather it stems from common law principles (Bronitt and Stenning, 2011).

Beyond these common law powers, several legal frameworks are relevant to police use of artificial intelligence and data analytics more broadly. These include the European Convention on Human Rights (ECHR, transposed into UK law in the form of the Human Rights Act 1998), particularly Article 8 ECHR, the right to respect for one's private life, family life, home and correspondence. Article 8 is a qualified right, meaning the state can interfere with this right provided such interference is demonstrated as 'in accordance with the law', 'necessary in a democratic society', and 'proportionate' for the prevention of crime or preservation of public safety. The ECHR also imposes *positive* obligations on the state (and by extension, the police), most notably Article 2 and Article 3 ECHR (the right to life; and prohibition on torture and inhumane and degrading treatment respectively). It has been ruled that Article 2 entails a positive obligation on the police to take measures to protect a person whose life is at risk, emphasising the importance of risk assessment to identify individuals in need of safeguarding or protection.[2]

Nevertheless, in order for a measure to be 'in accordance with the law', it has been ruled that such a measure must be *accessible* to the individual concerned and *foreseeable* in terms of its effects.[3] This entails having clear, transparent and professionally approved policies in place regarding use of new police tactics, to maintain accessibility and foreseeability of the law. For instance, in the appeal case of *R (Bridges) v Chief Constable of South Wales Police* [2020], the Court of Appeal ruled that – in the absence of primary legislation specifically governing the use of live facial recognition (LFR) – local policies are needed to satisfy the requirement of "in accordance with the law", and these policies do

[2] *LXD v The Chief constable of Merseyside Police* [2019] EWHC 1685 (Admin)

[3] *M.M. v the United Kingdom* (Application no.24029/07).

not necessarily need to be at the national level.[4] In *Bridges,* it was ruled that South Wales Police's use of LFR had been unlawful as it did not comply with legal requirements set out in the Equality Act 2010, Human Rights Act 1998 and Data Protection Act 2018 (as discussed further below). In response to this ruling, the College of Policing has since developed Authorised Professional Practice for LFR, setting out official national guidance for how police forces should deploy LFR technology while ensuring compliance with all relevant legal requirements. While such guidance is not a statutory instrument, it nevertheless established legally enforceable standards in the absence of primary legislation explicitly governing police use of LFR.

In addition to ensuring foreseeability of the law through clear and accessible policies, any police force deploying data-driven risk assessment tools must comply with various other existing legal requirements. These include (among others) the data protection principles set out in Part 3 of the Data Protection Act 2018, human rights principles set out in the Human Rights Act 1998, and the requirements for non-discrimination set out in the Equality Act 2010 and accompanying Public Sector Equality Duty (PSED). As was established in the case of *Bridges*, the PSED places a positive obligation on public bodies such as police forces to take all reasonable steps to satisfactorily address the potential for bias or discrimination in any new activity, including the deployment of new technology.[5] Specifically, the PSED states that 'a public authority must, in the exercise of its functions, consider the need to eliminate discrimination, harassment, victimisation and any other conduct that is prohibited by or under the act'.[6] In *Bridges,* the Court of Appeal found that South Wales Police had not taken reasonable steps to establish whether their facial recognition software contained algorithmic bias related to race or sex – thereby failing to meet their obligations under the PSED. This has important implications for future use of data-driven tools in UK policing, as it implies that all forces must take reasonable steps to test their models for any potential bias on the grounds of protected characteristics prior to their operational deployment. It is not simply sufficient to rely on bias testing results provided by third-party suppliers; the force must be able to complete its own operational testing and make these results publicly available.

To address these various complex legal requirements, the author's previous research has argued that additional policy and official guidance (such as Authorised Professional Practice) is required to ensure legitimate and lawful use of data-driven risk assessment tools in UK policing (Babuta and Oswald, 2020, 2021). However, data-driven offender risk assessment is not necessarily comparable to police use of LFR (Babuta and Oswald, 2021). The calculation of 'risk scores' or 'harm scores' represents the creation of new contestable information regarding an individual, which may then directly influence their subsequent treatment (including, for example, interfering with their Article 8 rights). Therefore, in order to demonstrate that the creation of such scores is *necessary* and *proportionate*, it is essential to articulate the overall decision-making process that such scoring will be used to inform – for instance whether the interventions that could result are supportive or punitive. This is the focus of the following sub-section.

[4] *R (Bridges) v Chief Constable of South Wales Police* [2020] EWHC Civ 1058, para 61.

[5] *R (Bridges) v Chief Constable of South Wales Police* [2020] EWHC Civ 1058.

[6] 'Equality Act 2010' (UK).

2.4.3 Impact on individual rights

As has been demonstrated above, a fundamental consideration regarding the introduction of next-generation data scoring tools is what interventions or decision-making processes they will be used to inform. While existing risk assessment tools are specifically used to guide the development of individualised risk management plans (such as MAPPA interventions), next-generation tools used by the police and other agencies could feasibly be applied for a far wider range of purposes. For instance, the most recent problem statement on the NDAS MSV use case suggests that the model could be used to augment current decision-making processes used to classify and prioritise 'elevated risk' individuals who may be at risk of escalating to serious violent offending (West Midlands Police, 2020). Questions arise regarding what interventions such 'elevated risk' individuals may be subject to, and how the calculation may influence the development of individual-specific treatment plans.

Moreover, it has been argued that the term 'predictive policing' is potentially problematic, as these tools are essentially classifying individuals into different groups based on their similarity to a historic behavioural 'profile', rather than generating individual-level predictions about future behaviour (Babuta and Oswald, 2020). There is a risk that the categories that result may represent new targeted groups resulting from systematic profiling of individuals (Bennett Moses and Chan, 2018). In the case of the so-called 'Gangs Matrix' used by the Metropolitan Police Service, an Amnesty International investigation concluded that:

> Once on the matrix, they become *de facto* 'gang nominals', a label which carries the stigma and suspicion of involvement in violent crime… the person is often automatically treated as someone who poses a risk of violence – even if they should not be on the matrix, or are on the matrix only because they have been a victim of violence. (Amnesty International, 2018).

A particular risk in this regard relates to 'scope creep' – and specifically a potential blurring of boundaries between offender management and intelligence purposes. The same data scoring system could be used to inform the development of offender management plans for those on probation, or for intelligence purposes to identify potentially high-risk individuals not currently under supervision. These are two fundamentally different uses of the same system, with profoundly different implications for individual rights.

To avoid these risks, it is essential to have clear policies in place specifying the interventions that may result on the basis of any given risk assessment, and any other factors that should be taken into account when making such a judgement. When first assessing the proposed Integrated Offender Management Model risk scoring system developed by West Midlands Police, the WMP Digital Ethics Committee suggested that 'far more detail is required around what interventions might be applied to those individuals identified', and requested that the force clarify how the model is going to be used operationally and what the benefit would be for policing purposes (West Midlands Police, 2019b). These factors must be considered when developing any new data-driven risk assessment project, as whether the use of such a tool can be justified as a necessary and proportionate use of police powers will depend largely on the outcomes that could result from its use.

While not related specifically to data-driven risk assessment, the College of Policing's Authorised

Professional Practice on Risk describes risk assessment tools as ‘an excellent, but limited, means of improving the likelihood of identifying and preventing future offending or victimisation’ (College of Policing, 2014a). It advises further that decision-makers should ‘consider the value and likelihood of the possible benefits of a particular decision against the seriousness and likelihood of the possible harms’ (College of Policing, 2014a). This emphasises the context-specific nature of risk assessment, and the high degree of discretion placed on individual officers to consider all factors relevant to the process and all consequences that may foreseeably result from any subsequent intervention.

This context-specificity is challenging in relation to data-driven risk assessment tools, which are largely prescriptive and blind to individual context. Although contemporary machine learning systems can identify predictive indicators (‘features’) in historic data that may go unnoticed by traditional statistical techniques, they nevertheless rely on such indicators appearing frequently enough to represent statistically significant correlates of future offending. As a result, any data-driven risk assessment model (however sophisticated) will not be able to account for the full range of idiosyncratic contextual factors which may be relevant to an individual’s risk of offending. They also cannot identify specific individual-level needs and vulnerabilities that should be addressed as part of a bespoke risk management plan. In other words, even if the quantitative scoring is highly accurate such that the police can effectively prioritise the highest-risk offenders who should be subject to further intervention, such harm scores alone do not provide insight into the specific *action* that is required to reduce the level of risk in any given context. In short, ‘not everything that can be counted counts, and not everything that counts can be counted’ (Cameron, 1963).

Given the potential impact on individual rights and the critical importance of context in offender management processes, officers must take into account all factors that may be relevant to an individual’s circumstances when making any decision regarding further intervention. This is essential to maintain the appropriate degree of decision-making accountability, which is particularly important when relying on algorithmic outputs which have been generated by an opaque statistical model. As discussed in recent research, the use of machine learning models for individual risk assessment could introduce ambiguity regarding who should take accountability for decisions, in circumstances where the human operator is unable to fully comprehend the overall analytic process (Knack, Carter and Babuta, 2022). Accountability is directly linked to the potential consequences of a decision, and higher-impact outcomes are likely to require a more detailed explanation regarding the specific factors that were taken into account to arrive at any given decision. For these reasons, when making risk-based decisions at the individual person level, human officers must also take into account other context-specific factors that are not captured by the data-driven system, and these factors should be clearly recorded in a traceable and auditable format.

In summary, while police use of new data-driven techniques for offender risk assessment can likely be accounted for within existing legal frameworks, use of any new capability must be demonstrated as necessary and proportionate, and crucially – accessible to the individuals concerned and foreseeable in terms of its potential effects. This emphasises the importance of force-level policy documents detailing the circumstances in which a particular tool will be used, and the interventions that could result on the basis of any given risk score. This is a crucial issue that must be taken into account when considering the findings presented in the following section.

2.5 Gaps in existing research

Recent years have seen significant public and media attention regarding the rise of 'big data policing', 'predictive policing', and the broader proliferation of artificial intelligence and data scoring tools across the UK public sector. However, despite a significant increase in the use of algorithmic risk assessment tools in policing, this public attention has not been matched with the same degree of academic scrutiny one might expect. The lack of empirical research evidence regarding the benefits and limitations of new and emerging data-driven risk assessment tools is concerning, particularly as their use has the potential to significantly impact on individuals' human rights and civil liberties.

While statistical risk assessment in policing is by no means a new approach, the use of complex algorithmic systems to collate data from multiple sources and 'predict' risk of future offending at the individual level represents a significant development in the police's approach to offender risk management. The literature suggests that some forces are now investing significant resource into developing and deploying such systems, but with a lack of any agreed standards or national guidance as to how the technology will be used. This gap should be addressed as a priority before these systems are deployed for wider operational use.

When considering principles for effective evaluation of crime prevention interventions, several dimensions of analysis are required beyond simply measuring effect size. Johnson et al. (2015) note that many systematic reviews of crime prevention interventions focus only on effect size but neglect other crucial aspects of evaluation, creating a 'lacuna of knowledge' in relation to other important factors such as implementation challenges and economic costs of the program under evaluation (Johnson, Tilley and Bowers, 2015). To address this deficiency in existing approaches, the authors propose a new framework, "EMMIE", to ensure that evaluation research covers all five relevant dimensions of analysis: *effect* direction and size; *mechanisms* or *mediators* activated by the intervention; *moderators* and *contexts* relevant to the production or otherwise of both intended and unintended effects; *implementation* issues and how they contribute to the success or failure of the intervention; and the *economic* costs and benefits associated with the intervention. With regard to implementation challenges, the authors note that 'even simple interventions can be fraught with difficulties', and that 'it is important for the practitioner to know what was done, what was crucial to the intervention and what difficulties might be experienced if it were to be replicated elsewhere'. This is particularly relevant to the current study, due to the lack of existing research evaluating the implementation challenges associated with deploying risk assessment systems in an operational policing context.

User experience is one important aspect of implementation-related challenges that has received only limited attention in existing literature, with some notable exceptions. In relation to predictive mapping, a 2007 Home Office report from Johnson et al. examined implementation challenges encountered when deploying a prospective mapping system in one Basic Command Unit in the East Midlands (Johnson *et al.*, 2007). This study included a process evaluation comprising semi-structured interviews, survey methods, and direct observation, and elicited new insights regarding officers' perceptions of the usefulness of the maps and how they might be improved. This study generated important findings regarding the visual presentation of algorithmic outputs and user requirements of system interfaces. In relation to individual risk assessment, a 2015 questionnaire-based study by

Moore examined prison and probation risk assessors' views and experiences of OASys, including user experience considerations (Moore, 2015). This study also included questions related to the training and guidance available for practitioners. However, beyond OASys, no recent studies have been identified examining police officers' views and experiences of the so-called 'next-generation' data-driven risk assessment tools in Section 2.3.2.

In summary, the literature review has identified three specific implementation-related knowledge gaps that this study seeks to address. The first relates to *officers' perceptions* regarding the limitations of existing approaches to offender risk assessment, and where they believe new data-driven systems could add the most value. The second relates to *the user interface and design requirements* of new data-driven systems, and what guidance is required for software developers to ensure an intuitive and comfortable user experience for officers. The final gap relates to the *overall decision-making process* that algorithmic systems are used to inform. It remains unclear how algorithmically-generated risk scores should be incorporated within wider offender management processes, and what interventions or decisions may be made as a result. This study seeks to address these gaps, through direct engagement with officers involved in the development and use of data-driven risk assessment tools in UK policing.

Chapter 3. Design and Methodology

This chapter describes the research design and methodology used for this study. Section 3.1 discusses the overarching research philosophy guiding the practitioner-focused approach used for the research, before summarising the general design and methodology used to structure the primary research component of the study. Section 3.2 then discusses the sampling strategy used, as well as the approach to data collection and analysis adopted for both the interview stage and the process evaluation component of the project. Section 3.3 explores the ethical considerations arising from the research and what steps were taken to address these, while Section 3.4 summarises the main limitations associated with the methods used for the research.

3.1 Research Design

3.1.1 Research philosophy

This study adopts a constructivist research philosophy, underpinned by relativist ontological assumptions (Crotty, 1998). The relativist view recognises that reality is a subjective construct and differs according to individual experience (Guba and Lincoln, 1994). As such, the research focuses on exploring the subjective perceptions and individual experiences of practitioners. The study adopts an interpretivist epistemology (Scotland, 2012), acknowledging that meaning is constructed in different ways by different people and that the social milieu can only be understood from the perspective of individuals engaged in it (Cohen, Manion and Morrison, 2017). Consistent with the interpretivist approach, the research aims to not only examine the specific phenomena under investigation, but also to yield new insights regarding the wider social and cultural context, in this case the English criminal justice system.

Existing research into offender risk assessment has typically adopted a positivist paradigm. Most studies to date have focussed on statistical validation of assessment instruments; the goal is to establish objective, quantitative measurements of success for different assessment tools (Howard, Clark and Garnham, 2003; Howard *et al.*, 2009; Fazel *et al.*, 2012). These previous studies have often neglected to address the key question of whether different tools are useful *in practice* to enable practitioners to effectively prioritise cases and develop offender-specific risk management plans (two notable exceptions are a 2018 systematic review of the overall usefulness of risk assessment tools in forecasting future violence (Viljoen, Cochrane and Jonnson, 2018) and a 2015 qualitative study examining assessors' views and experiences of the Offender Assessment System, OASys (Moore, 2015)).

Moreover, given their recency, there are very few publicly available studies exploring criminal justice practitioners' perceptions and expectations of new data-driven approaches to violence risk assessment. Is such a development considered desirable by those who would be required to implement it? Do they perceive deficiencies in existing methods, which could be addressed by the use of more sophisticated technological approaches? If so, how could these tools provide the greatest benefit for offender managers and other practitioners? How can they be integrated within existing processes in a way that maximises their potential benefits while minimising risks? These are crucial factors to consider for the success of future data-driven approaches to offender management. The constructivist, interpretivist approach adopted here is intended to enrich this existing limited body of research, by exploring in detail criminal justice practitioners' perceptions and expectations of the data-driven tools under examination.

3.1.2 General design and methodology

This study adopted a qualitative, emic research design (Given, 2008), focused on collecting subjective perspectives of practitioners involved in offender risk assessment throughout the English CJS. The emic approach is person-centred and recognises the importance of the social context and circumstances of personal experiences. As such, participants' contributions have formed the basis of the ultimate research findings and conclusions. The overall research design could best be described as a general interpretive approach, incorporating elements of case study research (Merriam, 2002).

A qualitative approach is desired given the focus on collecting descriptions and interpretations of subjective experiences, and theory-building through discovering patterns in qualitative data (Tesch, 2013). Qualitative research is initially guided by relatively broad questions (rather than specific hypotheses), with more specific questions being generated iteratively as the study evolves (Rice and Ezzy, 1999). With this in mind, the methodology adopted here is intended to be sufficiently structured to address the main research questions listed above, while allowing flexibility to explore other lines of inquiry not initially anticipated in the research design. The practitioner-centred approach emphasises the 'privileging of lay knowledge' (Popay, Rogers and Williams, 1998, p. 345), recognising that previous research on this topic has often neglected to take into account the individual perspectives of criminal justice practitioners.

Bearing in mind the specialist and technically complex nature of the subject matter, a case study approach was adopted, whereby the author identified a small number of specific case study projects to examine in detail. The case study approach seeks to answer focused questions by providing detailed descriptions over a relatively short period of time (Hays, 2004). Case study research focuses on the complexity and particular nature of a single phenomenon (the case), where one or more programmes are selected because they are in some way typical or unique in relation to the overall topic of investigation (Merriam, 2002). In this instance, the author identified a small number of 'critical cases' (Yin, 2009) of data-driven offender risk assessment projects to examine in granular detail, with a view to eliciting rich data from participants directly involved in those case study projects.

Primary research for this study was conducted in two stages. The first stage involved semi-structured qualitative interviews with subject matter experts in offender risk assessment, to understand the strengths and limitations of existing risk assessment frameworks, and the opportunities and risks presented by new data-driven methods. The second stage comprised a process evaluation of one specific data-driven risk assessment project, currently being undertaken by a large UK police force. The process evaluation was a mixed-methods study incorporating interviews, focus groups and a written survey of practitioners. The specific data collection and analysis methods used for each of the two stages of the study are discussed further below.

3.2 Research Methods

3.2.1 Sampling strategy

Qualitative sampling is focussed on information richness, meaning it is important to identify appropriate participants who can best inform the study (Kuzel, 1992). For this reason, a purposive, non-probabilistic sampling strategy was used to identify suitable participants who could provide meaningful insights regarding the research questions under examination (Bryman, 2016). Purposive sampling is a non-random sampling technique whereby participants are deliberately selected based on their specific knowledge or experience (Bernard, 2017). In this instance, participants were selected based on their first-hand experience of developing, using or researching data-driven violence risk assessment projects within the English CJS. This targeted sampling approach is aimed to ensure the identification of information-rich sources to maximise the limited resources available for the project (Patton, 1990).

The purposive sampling approach used here could be described as a form of 'critical case sampling' (Palys, 2008; Etikan, Musa and Alkassim, 2016). Critical case sampling involves sampling key case studies that allow logical inferences to be drawn that are likely to be generalisable to other contexts (Bryman, 2016). A core focus was on identifying individuals involved in these case study projects who can provide detailed, specialist information regarding the tools that have been developed and how they may be used. Critical case sampling was appropriate in this context as it focuses on identifying individuals who can provide compelling insights regarding the phenomena under investigation (Onwuegbuzie and Collins, 2007).

In addition, a snowball sampling approach (also known as "chain referral sampling" (Biernacki and Waldorf, 1981)) was adopted whereby the initial cohort of interviewees were asked to propose other suitable participants who could provide an informed view regarding the issues under examination (Noy, 2008). The snowball approach is appropriate in this context given the specialist subject matter, the fact that there are a limited number of individuals with in-depth knowledge of data-driven risk assessment tools, and the difficulties in identifying such individuals whose contact details are not in the public domain (Bryman, 2016).

3.2.2 Data collection

Semi-structured interviews were the primary data collection method used for the initial stage of the project. Interviews are the most widely used method of data collection in qualitative social research (Cassell, 2005; Nunkoosing, 2005). Qualitative interviews are defined as a mode of inquiry focussed on individuals' stories and experiences because they are of worth (Seidman, 2006). Interviews were appropriate in this context given their inherent flexibility (Bryman, 2016), and their value not only in obtaining insight into social issues by exploring individuals' experience, but also in capturing important contextual information (Denzin, 2001).

Ensuring sound qualitative research requires developing a systematic and rigorous approach to the collection and analysis of data, and the interpretation and reporting of findings (Fossey *et al.*, 2002).

A semi-structured approach was adopted to ensure data collection remained focused on the target research questions, while allowing sufficient flexibility to explore other areas of interest not initially anticipated in the research design. Semi-structured interviews are conversational in tone, and allow for an open response rather than a binary "yes" or "no" answer (Longhurst, 2003). In contrast to either the fully structured or unstructured approach, the semi-structured format enables the researcher to adopt a broadly consistent line of questioning in each interview, while allowing space to probe specialised areas of knowledge and experience in respondents. Semi-structured interviews are well suited to studies such as this, where the focus is on exploring under-researched territory and allowing interviewees maximum latitude to pursue new angles of enquiry (Newcomer, Hatry and Wholey, 2015).

Interview request letters were sent in advance (by email), alongside a project information sheet so respondents had a clear understanding of the purpose of the project and were able to give their informed consent to the interview. Once participants had reviewed the information sheet and had an opportunity to ask any questions regarding the project, they were requested to return a signed consent form, which was stored securely and separate to any interview data.

A semi-structured interview guide was developed, focussed on exploring the three research questions outlined previously. A semi-structured interview guide is an outline of key questions to guide the discussion, but does not constitute an exhaustive list of all questions the researcher may ask in interview (Newcomer, Hatry and Wholey, 2015). The interview guide was structured to open with a brief introduction, asking the participant to provide some background information about their current role and relevant experience. The following section of the guide focusses on understanding the strengths and weaknesses of existing offender risk assessment tools used within the CJS. The next section then goes on to explore the potential opportunities and risks arising from the use of novel data-driven approaches to violence risk assessment. The guide concludes with a brief closing section providing participants the opportunity to ask any final questions about the study and reaffirming that all data gathered is anonymous and non-attributable.

All interviews were conducted remotely using a secure videoconferencing platform (Microsoft Teams). Given the ongoing restrictions on movement as a result of the COVID-19 pandemic, it was not possible to conduct any in-person interviews at the time this study was undertaken. Interviews were not video or audio recorded. Instead, interview notes were transcribed directly into a secure document at the time of interview. Bearing in mind the sensitive nature of the subject matter and the background of respondents, it is unlikely that participants would consent to a recording being made of their interview, and any who did consent would likely be inhibited from exploring sensitive or potentially contentious issues.

A total of seven subject matter experts were interviewed during this preliminary stage of the project, between February and March 2021. These included five police respondents and two non-police criminal justice practitioners. This figure does not include the process evaluation phase of research (discussed further below).

3.2.3 Data analysis

Interview data was analysed following a general inductive approach, whereby the aim is to derive theory from the data, rather than 'test' pre-defined hypotheses (Bryman, 2016). The inductive approach involves adopting a systematic procedure for analysing qualitative data, guided by specific research questions and objectives (Thomas, 2003). The inductive approach is appropriate in this context as it enables transparent and defensible links to be established between the research questions and resultant findings. Adopting a systematic approach to concept development helps to ensure analytical rigour, which leads to credible interpretations and ultimately plausible conclusions (Gioia, Corley and Hamilton, 2013).

A preliminary open coding process allowed recurring themes and categories to be identified, and then a more granular analysis allowed trends and patterns within these themes to be explored in further detail (Corbin and Strauss, 2014). Following close reading of the interview transcripts, an "in vivo" coding process allowed distinct categories to be derived from units of phrases (also referred to as "verbatim coding" or "literal coding") (Saldaña, 2014). Pertinent text segments were then copied and pasted into their relevant category, allowing interview data to be analysed thematically. Bearing in mind the limited sample size and manageable volume of text data, the use of specialised qualitative analysis software was deemed unnecessary for this purpose.

Once interview data had been organised according to these broad categories, a more granular analysis allowed subtopics to be explored within each category, providing detailed insights into the range of views and perspectives put forward by participants. This next stage of analysis could be described as a form of "axial coding" (Corbin and Strauss, 1990), where data is scrutinised to identify relationships between categories and sub-categories. As noted by Brown et al., there are four analytical processes involved in axial coding: continually relating categories to subcategories; comparing those categories with collected data; expanding the density of categories by exploring their properties and dimensions; and exploring variations in the phenomena (Brown *et al.*, 2002). This axial approach is intended to be iterative, whereby the analysis and coding of data will inform the questions to be posed in later interviews.

This multi-stage process of open and axial coding bears close resemblance to the systematic set of procedures involved in Grounded Theory (Corbin and Strauss, 1990, 2014; Charmaz and Belgrave, 2007; Glaser and Strauss, 2017). However, it is important to note that this study does not adopt a Grounded Theory methodology in its true sense. The overall purpose of Grounded Theory is to *explain* a substantive topic at a broad conceptual level (Creswell, 2002). The primary focus of the current study is rather to critically assess the strengths, limitations, opportunities and risks involved in the phenomena under investigation. Nevertheless, adopting elements of the Grounded Theory approach ensures a degree of systematic rigour to inductively derive key observations and findings. A main advantage of this inductive approach is that it enabled detailed analysis of the content of the interviews, including identifying divergences of views and contradictions between participants.

3.2.4 The process evaluation

The second stage of research comprised a process evaluation of a data-driven offender risk assessment project being undertaken within one of the UK's largest police forces. The evaluation focused specifically on the beta-testing phase of the project, which was being conducted within two of the force's Local Offender Management Units (LOMUs) at the time the study was undertaken. The purpose of beta-testing (or 'user testing') is to trial a piece of software with a group of target users in a real-world environment to evaluate its performance in an operational environment. This process evaluation focused specifically on the beta-testing phase of the project.

As discussed by the Education Endowment Foundation, when implementing a new intervention, it is important to first conduct a pilot trial in a small number of settings to 'develop and refine the approach and test [the intervention's] feasibility' (Foundaton, 2015, p. 3). The purpose of a pilot trial is to test whether a new intervention has potential, and qualitative research is the primary data collection method used for this purpose. Fox et al. have also argued that pilots in the CJS are often 'implemented prematurely with insufficient time and resource put into first developing a sound theory of change and then testing key elements prior to a larger pilot' (Fox *et al.*, 2018, p. 40).

With this in mind, the process evaluation sought to scrutinise the implementation of the beta-testing phase of the project, and assess whether the intervention has potential in the context in which it is to be implemented. The evaluation followed the approach set out in the College of Policing's Policing Evaluation Toolkit (College of Policing, 2018). The toolkit provides a framework to assist the police in evaluating the impact of tactics, projects or policies in their local area. Stage 2.3 of the toolkit (pp. 26 – 28) sets out the key factors to consider when conducting a process evaluation of a policing intervention. It encourages the use of interviews and survey methods to understand participants' perceptions of an intervention, and identify whether the intervention was delivered as intended, what worked well and what could be improved. The methodology used for this process evaluation is based directly on the guidance set out in Stage 2.3 of the Evaluation Toolkit.

Three data collection methods were used to conduct the process evaluation: semi-structured interviews; focus groups; and a practitioner survey. The purpose was to examine the real-world use of the tool in its operational policing context.

Interviews and focus groups were conducted using the same procedure detailed above (Section 3.2.2), but using an interview guide specifically tailored to the project being evaluated. A copy of this interview guide is included in Annex 1. The name of the project and the police force have been redacted throughout the thesis to preserve anonymity of the force and all research participants. Interviews and focus groups were conducted throughout January and February 2022, with 15 respondents within the force who were taking part in the beta-testing phase of the project. This included officers of varying ranks between Constable and Inspector, and a number of police staff such as police data scientists.

Interview and focus group request letters were sent in advance (by email), and the two Inspectors overseeing each LOMU were responsible for distributing these letters among all officers involved in the process evaluation. As before, signed consent forms were requested and were stored securely and separate to any interview data. The semi-structured interview guide used for the process evaluation (included in Annex 1) was divided into two sections: the first focusing on the harm scoring element of the dashboard; and the second focusing on the predictive modelling component. Respondents were required to explain how they currently use each component of the dashboard, whether they find it useful, what interventions could result on the basis of the calculations, the extent to which they understand how the scores have been calculated, whether they have confidence in the system's accuracy, and what modifications may improve their confidence in the system. Questions were also included to explore respondents' overall perspectives of benefits and limitations, as well as what additional guidance and training may be required. As before, all interviews were conducted remotely using a secure videoconferencing platform (Microsoft Teams). Two online focus groups were also conducted (one for each LOMU), which brought together a small number of PCs from each LOMU in a group environment. Although the questions asked in these focus groups were identical to those explored in interviews, the focus groups elicited rich insights regarding group dynamics between officers – as the conversational interactions between respondents generated new findings that were not elicited through interviews alone.

Finally, a closed-ended written survey was distributed to all officers involved in the beta-testing exercise, eliciting a total of 11 responses. The survey was distributed between March and May 2022, immediately following completion of the interview and focus group component of the evaluation. A copy of the survey questions is included in Annex 1. The survey explored the extent to which users find the harm score and predictive modelling useful, whether they use it regularly as part of their offender management responsibilities, and whether it has delivered operational benefit in their force area. As before, certain details have been redacted to preserve anonymity of the force and all respondents. Although the survey questions covered many of the same issues explored in the interviews and focus groups, this multi-methods approach ensured the full spectrum of perspectives was triangulated and validated through multiple data collection methods. By the inclusion of closed-ended questions with multiple choice answers, respondents were forced to make definitive value judgements regarding the usefulness of the system. This provided crucial data (presented in Chapter 5) that could not be collected through interviews and focus groups alone, and ensured an added layer of robustness in the findings and conclusions. Nevertheless, the survey findings must be interpreted with caution bearing in mind the small sample size.

3.3 Ethical considerations

When conducting research involving human participants, tensions may arise between the investigative aims of the project and participants' right to privacy (Fujii, 2012). Potential harms can be prevented or reduced through the application of appropriate ethical principles (Orb, Eisenhauer and Wynaden, 2001). Numerous examples of "research ethics frameworks" have been developed over the years, which aim to provide a set of guiding principles for researchers to follow (Pimple, 2002; Smith, 2003). The most relevant frameworks for the current study are the ESRC framework for research ethics[7] and the UWL Research Ethics Code of Practice 2018.[8] Both are guided by the over-arching

[7] https://esrc.ukri.org/funding/guidance-for-applicants/research-ethics/our-core-principles/

[8] https://www.uwl.ac.uk/sites/default/files/Departments/Research/Web/PDF/research_ethics_code_of_practice_january_2019f.pdf

principle of beneficence and non-maleficence: research should maximise beneficial outcomes while minimising potential risk and harms.

All stages of the research process were guided by the two ethics frameworks mentioned above. Informed consent was received from all participants as a pre-requisite to their participation. A project information sheet was provided to participants outlining the parameters of the project, the aims of the research, and contact details of a third-party they could contact for further information. Participants were able to follow up with any questions regarding the study before they agreed to participate. Once participants had the opportunity to review the information sheet and ask any questions, they were asked to provide a signed consent form, acknowledging that they understood the purpose of the project and what their data would be used for, and that they freely consented to taking part in an interview. These consent forms were stored securely in a separate location to any interview data.

Research for this study was conducted on an anonymised (non-attributable) basis. The academic literature and research ethics frameworks tend to conflate two distinct issues: information confidentiality; and participant anonymity (Kaiser, 2009). There is a crucial distinction between the two. Information that is treated as *confidential* is information that will not be disclosed to a third party or made publicly available. Confidential information may contain personally identifiable information (an example would be a doctor's medical records, which contains highly personal information about specific individuals, but is treated as strictly confidential). Information that is *anonymous* (but not confidential) is information that is stripped of any personally identifiable information, but may be disclosed to third parties or otherwise made publicly available. Qualitative research data typically falls into this latter category: the data is stripped of any personal information relating to participants, but it is not *confidential*, as it may ultimately be used for the purposes of publication or otherwise disseminated to third parties.

All data collected for the current study falls into this latter category: it is anonymous (non-attributable), but it is not confidential. (However, it is necessary to bear in mind the possibility of re-identification of research participants using their rank or job title, as discussed further below). The distinction between anonymisation and confidentiality is particularly important in the law enforcement and intelligence context because non-attributable information may still be highly sensitive if disclosed to the public. With this in mind, participants were requested not to discuss any classified or otherwise sensitive information that should not be disclosed in the public domain. Nevertheless, it was necessary for the evaluation report to undergo a review process by the force, to minimise the risk of publishing any sensitive information that should not be in the public domain. The College of Policing's Evaluation Toolkit notes that process evaluation should be as independent as possible (College of Policing, 2018, p. 27). In most circumstances, this should entail that the report drafting, review and publication process is undertaken independently of the organisation or project being evaluated. However, given the sensitive nature of the subject matter and the information the author was provided with, the requirement for the force to review the report was a pre-requisite to the study being conducted. It was clarified at the outset of this review process that information should only be redacted for reasons of classification or security sensitivity. This review process also ensured that participants had an opportunity to object to the use of job titles and/or police ranks if they believed there was any reasonable risk of re-identification. The inclusion of ranks and job titles was considered an important element of the analysis to explore differences or disagreements between groups. No participants objected to the data being reported in this way, and the risk of re-identification through the inclusion of job titles was assessed by the force to be negligible.

Another important ethical consideration in this context is the need for transparency of data collection, analysis and reporting (Fossey *et al.*, 2002). To ensure transparency, participants were offered the opportunity to request a copy of their interview transcript. One participant made such a request and their transcript was provided to them one day after the interview was conducted.

A final ethical issue taken into account when designing the study was the potential for the methods used to cause unintended distress, discomfort or otherwise adverse outcomes for participants. The risk of psychological harm or distress to participants arising from the interview process was assessed to be negligible. Participants were criminal justice practitioners and the interviews focussed on understanding their views and experiences of offender management projects they had been involved in developing. The questions were non-personal: at no stage were participants asked for information that did not relate directly to their professional role. Should any participants have any concerns or questions regarding the study, a project debrief sheet was provided that included contact details for two third-party organisations they could contact for further information regarding the issues under discussion. In addition, research activities were conducted with careful consideration of issues of diversity and inclusion in the selection of stakeholders and case studies.

The research design and associated materials were reviewed by the UWL School of Human and Social Sciences Research Ethics Panel, which initially granted provisional approval of the project and requested further clarification regarding points around data collection and participant recruitment. These points were clarified and the research materials updated accordingly. Full approval was then secured from the Panel before any data collection was undertaken.

3.4 Limitations

This study has several limitations, primarily relating to external validity. Validity in this context refers not to the data but rather the inferences drawn from the data (Aktinson and Hammersley, 1998). Validity standards in qualitative research are inherently challenging given the need to simultaneously incorporate empirical rigor and subjectivity into the research process (Johnson, 1999). Nevertheless, credibility has been described as a main objective of qualitative research, meaning the research findings should reflect the experiences of participants and their context in a believable way (Guba and Lincoln, 1994).

Internal validity in qualitative research can be understood as comprising two components: ‘descriptive validity’ and ‘interpretive validity’ (Maxwell, 1992). Descriptive validity relates to the factual accuracy of reporting: is the researcher recording a true and accurate representation of the things they saw and heard? Interpretive validity pertains to the researcher’s interpretation of participants’ perspectives on issues, and the extent to which the researcher is accurately inferring aspects of participants’ beliefs and feelings (Maxwell, 1992).

In the context of the current study, descriptive validity may be affected by the fact that all research activities were conducted remotely, and the method of live transcription used to record interview notes. It is possible that research participants conveyed certain non-verbal cues that were not apparent over a video call, and the interview notes inevitably did not record every word uttered by participants. The interview notes are likely to be less descriptively rich than a verbatim interview transcript derived from audio recordings. However, audio recording was not feasible nor desirable in this instance given the sensitivities discussed previously.

External validity (also referred to as generalisability) refers to the extent to which the research findings can be generalised beyond the specific time and place in which the study was conducted (Noble and Smith, 2015). Qualitative studies typically do not aim to make systematic generalisations to the wider population in the way that quantitative studies do (Maxwell, 1992). Nevertheless, representativeness is still a crucial aspect to consider to ensure the external validity of an interview-based study (Robinson, 2014). In the context of the current study, external validity is inherently limited due to the purposive sampling strategy and the limited sample size. As detailed previously, seven experts took part in a semi-structured interview during the initial phase of research, 15 officers took part in a process evaluation interview or focus group, and the closed-ended survey elicited a total of 11 responses. This sample size is below average for comparable policing process evaluations. For instance, the Fox et al. (2018) evaluation of personalised offender management interventions comprised a total of 58 interviews (Fox *et al.*, 2018), and the Johnson et al. study examining predictive mapping in operational context included interviews with 12 Sergeants, and a survey of 57 front-line officers (Johnson *et al.*, 2007). However, the critical case sampling approach used for the current study was deemed appropriate given the specialist nature of the subject matter and the fact that only a small number of individuals have in-depth knowledge of the case study projects under examination.
As research participants are generally subject matter experts with in-depth knowledge of the use of data-driven risk assessment, it is hoped that their perspectives and opinions will be generalisable to other comparable projects elsewhere in the English CJS.

There are several additional limitations associated with the process evaluation component of the project. The primary limitation relates to participant recruitment and sampling. Although efforts were made to ensure that all offender managers involved in the beta-testing phase of the project were engaged in the process evaluation, staffing changes within the force meant that some research participants had been using the system for longer than others, potentially influencing their perspectives on its strengths and limitations. Another challenge to internal validity is the risk of self-censorship: although all research activities were conducted on an anonymised basis, it is possible that some participants felt reluctant to express views that could reflect negatively on the force.

External validity of the process evaluation is also inevitably limited. As detailed previously, the evaluation focused specifically on the beta-testing phase of the project under examination. The findings resulting from this component of research may not be reflective of other data-driven offender management projects currently in development, either within the same police force or nationwide. Moreover, as noted by the College of Policing, process evaluation is not a substitute for good impact evaluation (College of Policing, 2018, p. 28). As such, further evaluation research will be needed to assess the overall *impact* of the project within the force in question, and whether it has resulted in improvements in the force's overall approach to offender risk management. It was beyond the scope of the current process evaluation to assess the overall impact of the project on the force’s offender management processes.

Despite these limitations, the practitioner-focused research design adopted for this study has elicited new insights regarding the use of data-driven risk assessment tools in UK policing, particularly concerning the practical implementation of such technology in an operational policing environment. As the following sections will demonstrate, the findings represent a novel contribution to the research landscape and provide an important initial evidence base to inform the future development and use of such technology within the UK criminal justice system.

Chapter 4. Results

As detailed in the previous section, research for this study was conducted in two stages. The first stage comprised semi-structured interviews with seven subject matter experts from across the UK CJS, focused on understanding the potential strengths and limitations of new data-driven approaches to offender risk assessment. The second stage comprised a process evaluation of a major data-driven offender risk assessment project, which was being piloted in one of England's largest police forces at the time the research was undertaken. This section reports the findings from each of these components in turn.

This chapter is structured in two parts. Section 4.1 reports the findings from interviews with subject matter experts regarding the overall strengths and limitations presented by new data-driven risk assessment tools. Section 4.2 then reports the findings from the process evaluation specifically, comprising interviews, focus groups, and survey data.

4.1 Interview findings

Interviews with CJS practitioners revealed new insights regarding the perceived strengths and limitations of novel data-driven risk assessment tools for offender management purposes. All participants interviewed in this phase of the project were either senior-ranking police officers (Chief Inspector or above), or experienced criminal justice practitioners with more than seven years of experience working in offender management roles. It is notable, however, that these *perceived* strengths and limitations identified by interviewees do not correspond to the real-world benefits and shortcomings of the system that was evaluated in the second phase of the project. In other words, it appears that the actual strengths and limitations of a data-driven risk assessment tool when implemented operationally are different to the factors that interviewees thought to be most important during the initial phase of research. These disparities are discussed further in the following section.

Before discussing the perceived strengths and limitations identified by interviewees, it is important to note that interviewees widely recognised the limited evidence base, and lack of professional standards or regulation for police use of data analytics. As explained by one senior officer:

> I don't think we in policing have identified the relative advantages precisely enough, or made the case of how we would use them… I don't think the case has been made yet nationally around what advantage these tools are adding, especially in particular use cases… We don't have a standard we operate to in terms of developing these tools... there's just a whole lack of regulation as to how the tools are built and used. (Chief Superintendent)

This concern was shared by the majority of interviewees. Therefore, when considering the findings presented in these sections, it is important to bear in mind that interviewees' perspectives are based on a nascent and limited body of knowledge regarding the real-world benefits this technology could offer. The lack of official guidance or regulation governing police use of data-driven risk assessment tools also highlights the critical importance of developing a stronger evidence base regarding their potential strengths and limitations.

Analysis of interview data revealed six distinct themes, categorised according to three ‘strengths’ and three ‘limitations’. The three strengths identified were: efficiency and ease of use; managing data volume; and predictive validity. The three limitations were: data quality and availability; loss of explainability; and fettering discretion. Each theme is considered in turn below.

4.1.1 Strengths of data-driven risk assessment tools

Efficiency and ease of use

The first strength identified in the interviews relates to improving the efficiency of the risk assessment process. Police respondents described traditional (non-data-driven) risk assessment tools as ‘bureaucratic’, describing existing processes as ‘really arduous’ (Detective Chief Inspector). By contrast, new data-driven tools could present information in a more accessible way, allowing users to more efficiently identify individuals who pose the highest level of risk:

> The basic idea is that we will provide that information in a dashboard to offender managers… it provides them with a filtered list of people that they can then go away and look at using their own normal processes… If it means that offender managers can start looking at the people who are committing the higher types of harm in a consistent manner, then that risk can be mitigated and managed, allowing them to do that more efficiently. (Police data scientist)
>
> Once we create that data, it will be visible to an IOM [integrated offender management] team and a frontline officer attending a domestic abuse incident. So if the officer asks, the result of that forecast should be available on our records management system. (Chief Superintendent)

This perceived ease of use and added efficiency of new data-driven systems was a recurring sentiment among interviewees. Rather than spending time manually completing risk assessment forms and adding up scores on a checklist, automated data-driven risk assessment tools could rapidly derive insights from multiple systems and present an algorithmically-generated ‘risk score’ on-demand to the user, in an accessible and consistent format. This was described as having the potential to significantly reduce the time and effort required to assess risk.

As was discussed in the previous sections, a core focus of new data-driven systems is to enable police forces to ‘do more with less’, i.e., to make more efficient use of limited resources. It is therefore unsurprising that improving the efficiency of time-consuming offender risk assessment processes was viewed as one of the main potential strengths of new data science-based techniques.

Managing data volume

The second strength identified by all interviewees relates to the ability to manage increasingly large volumes of police data across multiple computer systems:

> If there is a way of accessing larger volumes of data that are very difficult to weigh up in an individual case, then that could be helpful. Particularly around some of the more serious and rarer types of offence. That's where it becomes more difficult to develop robust predictors. If you can obtain large volumes of data to help with those particular predictions, that would be helpful. (Senior criminal justice practitioner)

One officer explained how new data-driven tools would enable data to be extracted from multiple policing systems, rather than needing to search each database individually:

> The strengths are that it gives us the data from those different policing systems. It uses all of the data available to us to give us a list of offenders who are suitable for offender management. (Chief Inspector)

It was further suggested that more advanced tools could extract data not just from police systems, but also from partner agency databases:

> The main opportunity is to make sense of large amounts of data. Potentially police and partner data. (Detective Chief Inspector)

This could enable more effective 'triage', by effectively screening out individuals who are not assessed as posing an immediate risk:

> If you look to what you think the benefits of data-driven risk assessment tools are, I think it's a mistake to say they'll get you a needle in the haystack. But what they can do is drive out your high volume of people who aren't going to offend again, which might be up to 50 and 60% of all cases… That's the reason we're exploring data-driven tools, to provide more effective triage and better identification. (Chief Superintendent)

The challenge of managing large volumes of data across multiple systems was frequently mentioned by officers, and new data-driven risk assessment tools were described as offering the potential to automate a large portion of this process. However, from a legal and ethical perspective, it is important to consider the proportionality of intrusion that could result from enabling officers to rapidly collate larger volumes of personal information from multiple disparate systems. Such intrusion is likely to impact on individuals' right to privacy (Article 8 ECHR), and any potential interference must be judged as necessary and proportionate in relation to the anticipated benefits.

Predictive accuracy

The final strength relates to the accuracy of the risk assessment. While most tools under discussion have only been implemented recently and quantitative data is therefore lacking, practitioners nevertheless recognised that new modelling techniques could offer improvements in reliability and validity when compared with traditional statistical algorithms:

> I think it will give better accuracy. Previously we used to have data on a monthly basis, but this can be done on an ongoing basis… It considers other elements that may not have been considered previously, such as weapons, firearms, violent crime etc… We can use that tool alongside the probation scoring, to create a better joined up approach. (Chief Inspector)
>
> Potentially, [new data science techniques] could further improve aspects of reliability and validity. From a technical perspective, that is the potential for improvement. (Senior criminal justice practitioner)
>
> First of all, with better data, with different data, I think it is possible to get a better precision rate and potentially predict more effectively who is going to commit violent crime in the future. (Detective Chief Inspector)

Increased accuracy was mentioned by all interviewees in this stage of research as a potential advantage of new data-driven risk assessment techniques. Indeed, a senior data scientist within a force that has recently implemented a predictive risk assessment model explained that 'so far it's proved very accurate' (police data scientist). However, it remains unclear whether these improvements in accuracy are due to the type of statistical modelling used, or simply because more recent tools have access to larger volumes of higher quality data.

While all interviewees believed that new data science techniques should improve the overall accuracy of individual risk assessment processes, it is important to reiterate the very limited evidence base regarding the real-world predictive accuracy of such systems when deployed operationally. As such, this potential strength highlighted by interviewees should be interpreted as aspirational rather than factual; it remains to be seen whether the use of advanced algorithmic systems for offender risk assessment will in fact result in more accurate risk judgements in the long term, when compared with traditional actuarial or SPJ-based assessment protocols.

4.1.2 Limitations of data-driven risk assessment tools

Data quality and availability

The first main limitation identified in interviews concerns data quality and reliability. Several interviewees emphasised that the performance of any statistical model is entirely dependent on the data used to build it:

> One of the limitations around all of these models is just the limitations of the data we collect or the data we've got available. If we're going to try to develop insights, that will shine a light on a threat type. The problem is, if you shine a light on something, you put everything else in the shade. In improving our view of a threat type, we should never get complacent. There will be elements of that threat that will not be present in that data. (Detective Chief Inspector)
>
> One of the great things that needs to happen is around the data. Data quality standards, integrated platforms, architecture. (Chief Superintendent)

A particular concern in this regard was the risk of 'algorithmic bias' in the development of complex statistical models. As articulated by one senior officer:

> We've got to understand the biases that go into the data before we think about NLP [natural language processing] on the data and the output that will come out of that model… We like to think there's no bias in the model but inevitably there are biases in the data that go into the model. We've got to treat our intelligence with caution, if we're going to use it to identify risk. (Detective Chief Inspector)

Several interviewees also highlighted that most predictive models are developed using local police data, but that other police forces or partner agencies may have access to other relevant information that is not included in the modelling:

> The main limitations for the model are the data available to it. We don't have data from other forces, or information from other sources. (Police data scientist)
>
> There is a vast amount of data in the IOM [Integrated Offender Management] world across other agencies, which we don't use… If we were to have more partnership data in an algorithm, that would give us a more balanced view of the offenders based on the external data from those other providers. (Chief Inspector)

However, as explained by one DCI, while it may be possible to improve predictive accuracy by incorporating more data from other sources, this could potentially entail a higher degree of privacy intrusion, which may not be proportionate in relation to the potential benefits:

> If you think about police data, even on somebody who is very well known to the police, we only hold a certain amount of data. We only have a small window into their lives. 90% precision would require access to a huge amount of personal data about that individual… with better data, with different data, I think it is possible to get a better precision rate and potentially predict more effectively who is going to commit violent crime in the future… but I wouldn't be comfortable with having personal health data on a police platform. So I think it's possible, but I don't think it's viable. (Detective Chief Inspector)

With these limitations in mind, while machine learning techniques such as natural language processing could allow relevant information to be rapidly extracted from multiple disparate data systems, if there are deficiencies in the quality or reliability of the underlying data, this could result in the modelling producing inaccurate or biased outputs. A second factor is to consider is the additional privacy intrusion that could result from providing readier access to the sensitive data held on partner agencies' systems.

Loss of explainability

The second perceived limitation of data-driven risk assessment tools is the loss of explainability associated with complex machine learning modelling. Traditional statistical algorithms provide a high level of explainability, meaning it is possible to calculate the influence of different input variables on the model's final prediction. Complex machine learning modelling techniques do not provide a clear indication regarding the relative weighting of different input variables ('features') on the model's prediction, meaning it is often impossible to understand what factors were most relevant when calculating an overall risk score. This was a concern among several interviewees:

> Some of this technology is so sophisticated that no-one can even explain it. If you can't pass that test in an accredited process, you can't use it in law. (Chief Superintendent)
>
> With the model, we can see the significance of each individual variable, but we don't know whether it has a positive or negative impact on the dependent variable… You get non-linearities in the way that variables interact... To identify which factors contributed to an individual's harm score, you would need to go back and run the model again with certain features included or excluded. (Police data scientist)

Two senior criminal justice practitioners working within agencies that have chosen *not* to use advanced machine learning for offender risk assessment argued that maintaining a high degree of explainability was crucial to ensure practitioner buy-in to the tools:

> The fundamental question is whether you want to go for that bit of extra accuracy versus having something that's easy to understand. In the arena that we are in, we are really cautious about making predictions that we can't back up… That seems a really important property to having an overall system of risk assessment that people can have confidence in, given the impact that any one risk decision can have. No risk assessment system is entirely accurate, but at least we can explain how we are doing it. (Senior criminal justice practitioner)
>
> If you're in the space of relatively small increases in predictive validity, but you also get some of the challenges from the practitioner perspective about how those predictors are working and how they've ended up with this score… It's more a black box approach, it's about working out whether those trade-offs are worth it. (Senior criminal justice practitioner)

A DCI within a force that has implemented a complex machine learning risk prediction model also expressed disappointment that the tool was not able to provide 'strategic insights' into the underlying factors contributing to violent crime across the force area:

> I was constantly asking the questions of *why* violent crime is being committed. There were a number of key predictive indicators, or "features" – factors that have the greatest impact on an individual's risk score. If you look at an individual, you could understand which KPIs led to the risk score, which would help you understand the context around why they were offending. My question was how can you use that for better strategic insights, but we looked at it a number of ways and there was no way of doing that. (Detective Chief Inspector)

Similarly, a senior data scientist explained that their force had attempted to develop a 'one-off explanatory model' to understand at a strategic level the factors contributing to youth violence. However, 'where it falls down is that the relationships between the variables and any particular pathways are actually so complicated that it could not easily identify a factor or pathway of factors that a human being could easily understand and do something with' (police data scientist). As such, while complex machine learning models could provide more *accurate* predictions of future offending, it will often be impossible to disaggregate the risk factors that contributed to this overall prediction, limiting their utility for criminal justice processes.

Fettering discretion

The third main limitation identified is the risk that advanced data scoring tools could undermine the discretion or professional judgement of the human decision-maker. Interviewees explained that traditional, manual risk assessment involves a high degree of professional judgement, and requires the assessor to consider all relevant contextual factors:

> The strengths [of manual assessment instruments] are that first, they bring in professional judgement. Secondly they force the professionals to think about everything that's relevant rather than just what is in front of them. To seek the answers to questions they don't know, so they can get a more holistic view. (Detective Chief Inspector)

By contrast, data-driven risk assessment tools were described by several interviewees as highly automated, providing little scope to incorporate professional judgement into the scoring process:

> It's all data-driven, there's nothing subjective there. That's the harm score. We end up with a quantitative score that is essentially grouped into five different groups, from low to very high… What it means is that when somebody gets put into a particular group, it's relative to everybody else in the dataset. (Police data scientist)

Two specific issues were identified in this regard. First, the risk that automated analysis could lead to punitive interventions with little human oversight. And second, the risk that human officers may feel reluctant to contradict or override the algorithmic output with their own professional judgement:

> The risk is that you have an unsupervised decision that arguably has a punitive effect… The higher stakes the decision, the more it matters that you are confident you're being fair. Obviously predictive accuracy matters as well, but you would want to have some oversight of matters that really make a difference to people. (Senior criminal justice practitioner)
>
> Cops were worried about what would happen if their decision differed from that of the algorithm, and how were they expected to resolve that. (Chief Superintendent)

For these reasons, several senior officers suggested that complex data scoring algorithms are useful tools for filtering and triaging from a large volume of offenders, but individual-level risk assessments would still need to be conducted by trained professionals, taking into account other relevant factors that are not coded into the statistical model:

> It would definitely be useful for overall triage. But even if you do overall triage, you've still got to filter down. So you've always got to have a professional conversation on the 50 or 100 high-risk

> cases you've identified this month… So I see this as more useful for triage than I would see it as the point of contact for a front-line officer. (Chief Superintendent)
>
> It's a way to produce a consistent approach to looking at people using normal procedures, before they create that harm… But they would still need to go off and do their own risk assessment, because at the end of the day the offender managers will have access to information that we don't. (Police data scientist)

In conclusion, while advanced statistical modelling was viewed as useful for filtering from a very large group of offenders down to a smaller proportion of higher-risk individuals, there was consensus that individual-level risk assessments will still need to be conducted manually, by trained offender managers with access to other relevant contextual information. This is consistent with the suggestions made in previous academic research, discussed in the previous sections.

4.2 Process evaluation findings

This section reports the findings from the process evaluation component of research. As detailed in Section 3.2.4, the process evaluation comprised semi-structured interviews, focus groups and a practitioner survey with offender managers based within two local offender management units (LOMUs) at a large UK police force, which were beta-testing a novel data-driven risk assessment dashboard at the time the study was undertaken. The identity of the police force and name of the project have been redacted to preserve the anonymity of respondents.

Participants in this phase of research fall into one of three sub-categories: Police Constables and Police Sergeants taking part in the beta-testing phase of the project in one of the two LOMUs under examination; Police Inspectors responsible for overseeing the pilot project in each of the two LOMUs; and the police data scientist responsible for developing the project. Due to the anonymised data collection methods used for the process evaluation, further demographic information (such as gender and years of service) was not requested. Although several research invitations were distributed to all officers who were beta-testing the dashboard, only 11 responded to the written survey. Therefore, throughout this section, results from the survey should be interpreted with caution, given this small sample size.

The dashboard was first developed in 2019 and has two distinct components. The first component (the 'harm score') is a statistical harm score calculated for nominals in the force database who have previously been charged with an offence, corresponding to the overall level of harm associated with their current offending (ranging from 'low' to 'super high'). The definition of the differing levels of harm was derived from the Cambridge Crime Harm Index (Sherman, Neyroud and Neyroud, 2016). The second component of the application (the 'harm model') is a predictive model which produces forecasts at the individual level calculating each offender's likelihood of escalating from low-level offending to more serious offending, or of offending at a scale that cumulatively leads to more harm generated (i.e. the probability of moving from the 'low' or 'medium' harm groups into the 'high' or 'super high' harm groups).

The harm score is a descriptive measure corresponding to an individual's *current* level of offending. By contrast, the output from the harm model is a predictive forecast corresponding to their expected risk of *future* offending. The model predictions are calculated using Xgboost (extreme gradient boosting), a type of supervised machine learning typically used in regression and classification tasks (Chen *et al.*, 2015). The harm score and the outputs of the predictive modelling are both available to offender managers on-demand via an interactive dashboard. Both components (the harm score and the predictive model) together constitute the 'application' or 'dashboard'.

As summarised by the police data scientist responsible for developing the project:

> The basic idea is that we will provide that information in a dashboard to offender managers. There are two ways in which they will potentially use it. First of all to look at the harm score and identify whether they are managing who they should be managing. And then in terms of the harm model, it provides them with a filtered list of people that they can then go away and look at using their own normal processes. (Police data scientist)

The overall objective is to enable the force to more effectively manage risk and target preventative interventions to the highest-risk offenders. As explained by one Inspector responsible for overseeing the project:

> Risk is a massive amount of what we do… we're coming away from enforcement more into rehabilitation areas through use of partners and pathways… we're trying to understand by way of engagement with individuals those crime-causing catalysts, and divert individuals away from further offending by addressing those needs... [we're moving] away from enforcement towards more of a prevention approach… morphing into the use of pathways, partners and policing skills to address and understand why people commit crime. (Inspector)

With this in mind, this project should be assessed within the context of the wider shift towards preventative and pre-emptive policing tactics discussed earlier. It is important to note that neither the harm score nor the predictive model are intended alone to constitute a full risk assessment; the purpose is to enable officers to more efficiently *prioritise* higher-harm nominals who require more detailed individual risk assessment. Any decisions related to further management or supervision will be taken by the officer, who will be expected to conduct a subsequent (manual) risk assessment in addition to the initial machine-generated forecast. As discussed previously, the numerical scores alone do not provide insight into the underlying factors related to an individual's offending, or supportive interventions that could address their individual criminogenic needs – hence the need for a more detailed, individualised risk assessment to develop a bespoke risk management plan for each nominal under supervision.

The harm score and accompanying predictive model were rolled out for beta-testing in October 2021, for a period of approximately 8 months until May 2022. Offender managers within the two pilot LOMUs were required to use the dashboard in conjunction with existing data systems to support the offender risk assessment process. All participants took part in an online training session to brief them on the purpose and function of the dashboard before being required to use it operationally (the author also attended one of these training sessions).

Those involved in the beta-testing were all invited by their respective senior officers to take part in the process evaluation. Of the 17 distinct users recorded as using the app between October 2021 and May 2022, 8 took part in the evaluation by way of a research interview or focus group, while a total of 11 completed questionnaire responses were received. Due to the anonymised data collection method, it is not possible to assess the degree of intersection between these two samples. It is possible that some users who took part in a research interview or focus group did not complete a written questionnaire, and vice-versa. Nevertheless, consistent themes and findings emerged across all research activities, which are discussed below.

4.2.1 Cross-cutting findings

Before discussing the specific strengths and limitations of the application identified in the process evaluation, it is important to reflect on three cross-cutting findings that emerged from the research.

First, the research highlighted a fundamental divergence in views between junior officers (PCs and Sergeants) on the one hand, and more senior officers (Inspectors) on the other. The latter group were markedly more positive and complimentary regarding the new system, as indicated by comments such as:

> I'm infinitely more happy with [the harm score] than [the previous application]. I was not convinced with the maths that sat behind the scoring. I'm much happier with the transparency around [the new dashboard], the filters that it's brought in. It's more interactive, that gives me a lot more confidence that it's a more precise tool. (Inspector)

This is in stark contrast to the comments returned by PCs and Sergeants:

> I have no idea how the [harm score] and crime predictor tool come up with the data. The training is an hour over Skype but I left as confused as when I started. The system only works on charges but this does not show intelligence so will only pick up people charged, not arrested or those with lots of intelligence to suggest offending. I don't use it. (Sergeant)

These findings were also reflected in survey data. In response to the prompt 'I find the [harm score] useful', not a single PC or Sergeant 'agreed' or 'strongly agreed' with this statement. 33% (*n* = 9) neither agreed nor disagreed; 33% (*n* = 9) disagreed; while 33% (*n* = 9) strongly disagreed. In relation to the crime prediction modelling, again no PC or Sergeant respondents reported that they find the modelling useful: 33% (*n* = 9) neither agreed nor disagreed; 22% (*n* = 9) disagreed; while 44% (*n* = 9) strongly disagreed. When asked whether the new application is an overall improvement over the previous system, no PC or Sergeant respondents 'agreed' or 'strongly agreed' with this statement: 22% (*n* = 9) neither agreed nor disagreed; 44% (*n* = 9) disagreed; while 33% (*n* =9) strongly disagreed. By contrast, both Inspectors who responded to the questionnaire 'agreed' that they find the harm score and the crime prediction model useful, and both 'agreed' that the application represents an improvement over the previous system. This further emphasises the divergence in perspectives between PCs and Sergeants on the one hand, and Inspectors on the other. This is an important finding of the research, and its implications are discussed further in the following section.

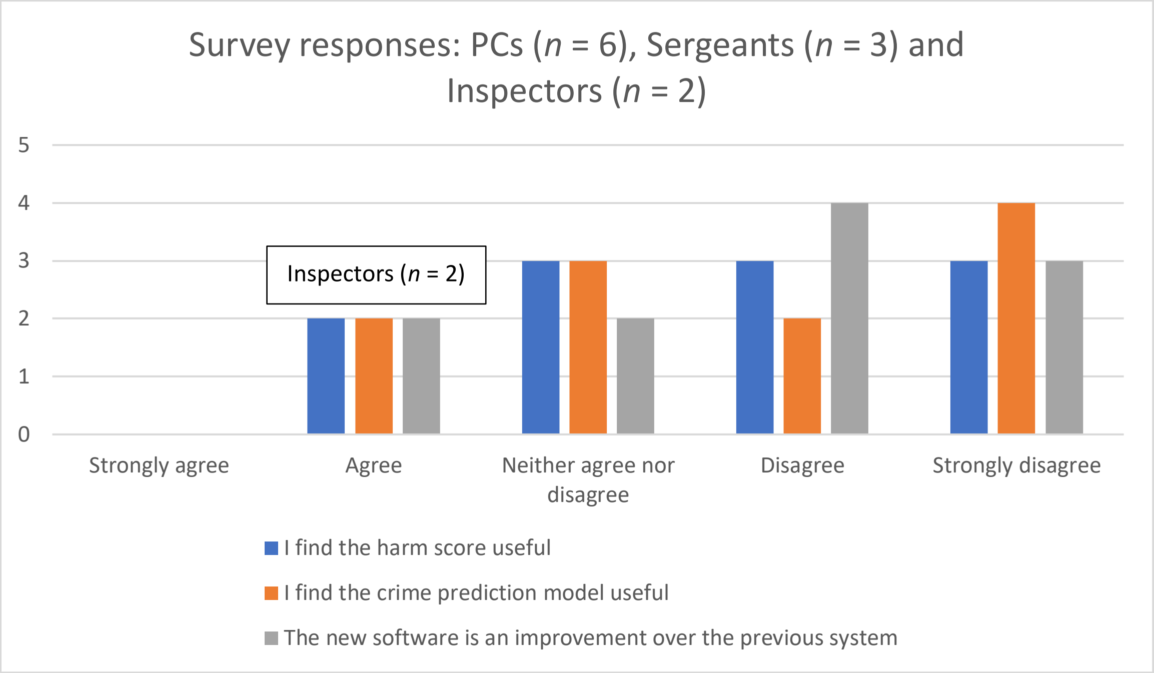

Figure 1. No PCs or Sergeants reported finding the dashboard useful.

The second key finding of the research is that the practical implementation of the application does not appear to have resulted in a better user experience for officers. On the one hand, officers believed that the new system should enable them to 'be a bit more strategic to recommend people… to management in neighbourhoods' (PC). However, in practice, PCs and Sergeants unanimously reported that the user experience and usability of the new application was not an improvement over the previous system, as indicated by comments such as:

> [The previous application] was an easier system to navigate. [The new dashboard] throws up a lot of names, but the interface isn't as user-friendly as [the previous application]. I think [the previous application] was better as people are able to manage better with visual aids. (Sergeant)
>
> I've had a play with it over the last few days. I've put a nominal's name in and it just takes forever to load. 3, 4, 5 minutes just to look at one person. (PC)

This poor user experience appears to be the main reason that respondents reported not regularly using the new system as part of their offender management responsibilities, nor finding the new harm score or predictive modelling useful, as reflected in the survey data.

Based on these observations, it appears that the practical implementation of the system is yet to result in the improvements for offender managers that were envisaged at the time it was rolled out for beta-testing. It remains unclear whether this is due to limitations in the underlying statistical application itself; deficiencies in the visual presentation of outputs and usability of the interrace; or a lack of sufficient training and guidance for users (or a combination of these factors). These issues are discussed further in the following section.

A third and final cross-cutting finding relates to the distinction between the two components of the application: the descriptive harm score on the one hand; and the predictive modelling on the other. These represent two very different components of the dashboard – the harm score is purely descriptive and does not rely on any machine learning modelling; while the predictive model produces a probabilistic forecast indicating an individual's likelihood to progress to higher-harm offending. For example, an individual's *current* harm score may be 'LOW', while their risk of escalating from low harm to higher-harm offending may be a high probability, such as 90%. However, this distinction was not clearly recognised by the majority of respondents. The large majority of interview quotes and survey data relate specifically to the descriptive harm score. Although prompted several times to provide an opinion on the probability modelling component specifically, interviewees had little insight into this aspect of the application. As such, all findings discussed in this section should be interpreted as relating primarily to the descriptive harm score element, unless specific reference is made to the probability modelling element of the application.

4.2.2 Strengths

As mentioned previously, there was a clear divergence of perspectives between Inspectors on the one hand, and PCs and Sergeants on the other, which must be taken into account when interpreting the findings presented in this section. At times the feedback provided by PCs and Sergeants directly contradicted the comments provided by Inspectors. The causes and implications of this finding are discussed further in the following section. As such, the perceived strengths of the system identified in this section of the report were derived primarily from interviews with Inspectors, as indicated throughout.

Inspectors responsible for overseeing the beta-testing of the application recognised that existing approaches to offender risk management were 'inconsistent', with one interviewee explaining that 'everyone's done it differently. Inconsistency with regard to risk is the big one for me.' (Inspector). It was suggested that the new application could 'give us a good push in the right direction to be more precise with our risk assessment.' This interviewee described the new application as 'more smart really', adding that 'I have a better understanding at where the scores are coming from.' (Inspector). In relation specifically to the offender escalation model, the Inspector reported that:

> I don't know much about this one. My understanding is that it's going to predict harm and predict who may re-offend… that could allow us to intervene sooner with someone on an upward trajectory. If that was really nice and clear, that would allow us to intervene really quickly by way of a more intensive offender management visit. At the moment, we're reliant on local knowledge but there's been no formal system to feed that into. (Inspector)

Based on interviews and questionnaire responses from the two *Inspectors* specifically, the research highlighted three main strengths of the new application: precision; transparency; and identification of 'hidden risk'. These are discussed in turn below.

Precision

The first perceived benefit of the new risk assessment application is improved precision of targeting. It is important to note that precision is distinct to accuracy: accuracy refers to the overall validity of the statistical predictions, i.e. the proportion of individuals that are assigned the 'correct' risk score; while precision refers to the ability to improve the granularity of these predictions. Precision is directly linked to efficiency; in a resource-constrained environment it is essential to efficiently prioritise limited resources to those areas of greatest need.

As explained by one interviewee, 'with reduced officer capacity, there is an absolute requirement for us to act with precision. [The new application] helps us to do that beautifully.' (Inspector). As a result, 'many individuals have been de-selected on the basis of [the scoring]', meaning the tool is enabling the force to more effectively 'screen out' lower-risk offenders who no longer require supervision. The Inspector described how 'we use it as a tool to justify de-selections.' An important feature in this regard was said to be the ability to monitor changes in an individual's harm score across specific offending behaviour, to track 'an increase or decrease in that risk, with numerous sub-filters.' (Inspector).

This is consistent with the anticipated benefits described by the senior data scientist responsible for overseeing the project, who explained that:

> [The harm score] would provide that filtering capability that they might not have had before… if it means that offender managers can start looking at the people who are committing the higher types of harm in a consistent manner, then that risk can be mitigated and managed within [the force], allowing them to do that more efficiently. (Police data scientist)

While Inspectors listed improved precision as one of the main motivations behind developing and implementing a data-driven risk assessment tool, it remains unclear whether these intended benefits have been realised in practice, as discussed further below.

Confidence

The second perceived benefit of the system is increased confidence in the risk assessment process. One interviewee suggested that the application has 'brought an element of transparency to the [risk assessment] process, so you can see where the scores are coming from.' (Inspector). This transparency has helped to build confidence in the validity of the outputs, as the factors that contributed to a particular harm score can be identified and triangulated across other data sources:

> I would say it is very accurate. When we've gone through [the harm scores] we've recognised names of individuals, and we cross-reference to our intel system, and you can see where the

> score has been generated from… I'm very confident that it's identifying the right people. (Inspector)

This degree of transparency was viewed as essential to maintain confidence in the output: 'if it was a black box, I think you'd naturally feel less confident about making that decision. I'd feel significantly less confident if I didn't know what sat behind that.' (Inspector).

It is notable that lack of transparency and loss of explainability is one of the most frequently mentioned risks of data-driven systems discussed in the literature, and emerged as one of the main potential limitations highlighted by subject matter experts in the interview phase of research. However, in reality, the Inspectors perceive that the harm scoring system has in fact *increased* the transparency and auditability of the overall risk assessment process (when compared with non-statistical methods). As the harm score is purely descriptive, it is possible to identify the specific factors that led to each individual harm score (in contrast to the predictive model). This issue is discussed further in the following section.

<u>Identification of 'hidden risk'</u>

A main advantage identified in the research is the ability to identify high-risk nominals who may otherwise not have reached the threshold for offender management or further scrutiny.

While not an original intended purpose of the system, Inspectors report using the harm score to verify and triangulate the risk assessments conducted by other agencies (most notably the probation service) using traditional risk assessment frameworks such as OASys and OGRS:

> We use it now to take to Day One selection meetings with Probation. To confirm that Probation are also selecting the right people with their OGRS scores. (Inspector)

However, a more significant development is the use of the harm score to identify individuals who may not have previously been subject to offender management orders:

> We've been using [the harm score] to identify [an Under-25 cohort] for either offender management or neighbourhood management, depending on the level of risk… the vast majority were not being managed or were being managed as part of a wider grouping… they were not given any proactive management plans… Previously, there would be no system that was flagging them to us. The neighbourhood officers would probably know them, but they would have no way of carrying that forward, managing that. The system to flag them. (Inspector)

Assuming this is an accurate observation, this could be interpreted as both a strength and limitation of the system. While the surfacing of 'hidden risk' could lead to pertinent individuals being identified who may otherwise have gone unnoticed, this also introduces the risk of false positives, and increased demand placed on already-stretched offender managers to scrutinise a larger number of cases. The same harm score could be used in multiple ways; either to verify and triangulate risk assessments conducted by others, or to effectively 'screen out' lower-risk individuals who do not meet the threshold for further scrutiny, or to identify new individuals not currently subject to offender management who warrant further scrutiny. Any force deploying such a system will need to establish and articulate which of these functions the tool is intended to support. This issue is discussed further in the following section.

4.2.3 Limitations

User Experience

As mentioned previously, a main limitation identified in the research relates to the user experience and usability of the software interface. As reported by almost all PCs and Sergeants interviewed:

> It's certainly not user-friendly if I'm honest. I'm clicking around, pressing buttons, I don't really know how to navigate around it if I'm honest… I don't think the tool itself is user friendly. (PC)

A common issue mentioned was that the dashboard is not integrated with other policing systems, meaning officers must access multiple systems to triangulate the information provided by the harm score:

> There was no way in the app of knowing why people are scored in a certain way. You have to then go into a different system to look at one individual. (PC)
>
> It needs to be a bit more user-friendly and available on the same system as everything else… it needs to be very simple, very clear. You really have to spend time on it, and time is of the essence, especially in policing… Customers wouldn't use it, but since it's a police system we have to use it and we have to make it work. (Sergeant)

This was also recognised by the senior data scientist responsible for developing the project, who explained that:

> They [offender managers] would still need to go off and do their own risk assessment, because at the end of the day the offender managers will have access to information that we don't… They would go away and apply their own processes and if they think that person should be managed then they will be managed in the normal fashion. (Police data scientist)

Nevertheless, this poor user experience appears to be one of the main reasons why no PCs or Sergeants who participated in the questionnaire reported using the harm score regularly as part of their offender management responsibilities, and none reported that the harm score has delivered operational benefit in their force area. This is an important finding of the research: while much attention has focussed on the *data science* elements of 'predictive policing systems' and the challenges of implementing these operationally, comparatively less attention has been paid to the *software engineering* challenges associated with integrating such tools into existing police computer systems. This is discussed further in the following section.

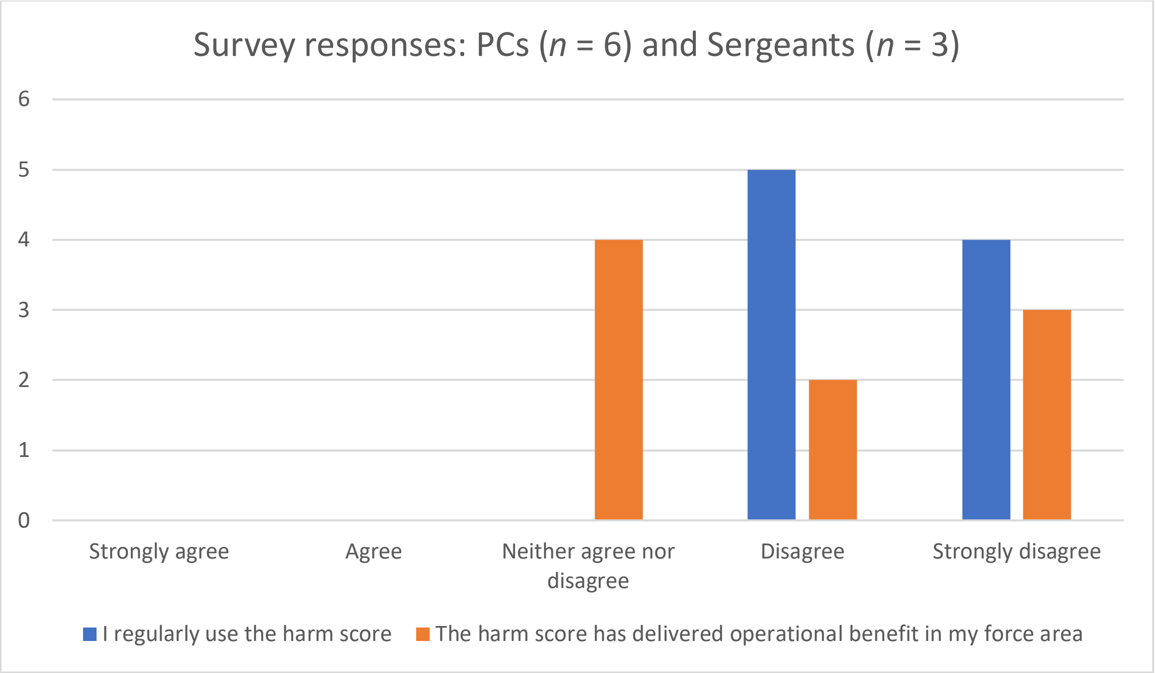

Figure 2. No PCs or Sergeants reported regularly using the harm score, or that it has delivered operational benefit in their force area.

<u>**Over-classification of risk**</u>

A second limitation identified in the research relates to the over-classification of risk: individuals who receive high harm scores but on further examination are assessed not to pose an immediate risk. As described by one interviewee, 'Sometimes it's incorrect, there are people in prison who have high [harm] scores.' (PC). As a result, officers report that the system is scoring a disproportionately large number of nominals as 'high' or 'super high', resulting in an unmanageably large list of individuals to review. This limitation was recognised by a large proportion of PCs and Sergeants interviewed:

> I've seen some inconsistencies in it. It wasn't showing any of our current youths that we would look at. I've had another look recently and a lot of ours are starting to feature in our "high" cohort, and there's now one who's scoring as "super high", but I have some doubt as to how that scoring is calculated… having looked at him I don't think he needs to be on that radar. (Sergeant)
>
> On the harm side of it, I've had a look at it and there were 900 offenders who scored high across my area. I'm not going to go through all of them, it's too time-consuming. And the breakdown of crimes is not detailed, it just says "acquisitive", it doesn't say burglary or robbery or whatever. So it's really difficult for me to work out who I'm meant to look at. (PC)

There is a risk that the large number of high harm scores generated by the system create additional demand for already-stretched offender manager teams. This risk was well recognised by the Inspector responsible for overseeing the delivery of the project:

> If we're going out looking for people to manage, that does create more demand. But we've made an agreement with the neighbourhoods that we're not going to overload them and we're only going to choose the highest risk individuals. The system will help us identify those who are causing the most risk. (Inspector)

Nevertheless, they did recognise that 'it does generate a lot of people, and you have to go in and double-check that you've got the right offending type.' (Inspector). This is an important limitation to consider when assessing whether the system has resulted in overall efficiency gains for offender management units. While Inspectors suggested that the system should result in improved 'precision' of targeting, this will not be achieved in practice if the system is producing an unmanageably large number of high-scoring nominals who require further scrutiny. This issue is closely related to the classification thresholds of the statistical model, as discussed in the following section.

Missing nominals

Conversely, as well as over-classification errors, officers also report that a large proportion of individuals who should be scored as high-risk are not currently being identified by the system. There appear to be two main reasons for this. First, the risk scoring only includes nominals who are classed as 'defendants', i.e., those who have been previously charged with an offence. It does not include 'suspects' (those who have been arrested but not yet charged). Second, the system only uses local data, meaning that data from other police forces or national databases will not be incorporated into the risk scoring. This problem was highlighted in particular by the Sergeants who took part in the process evaluation phase of the project:

> 'Individuals who are committing offences are not necessarily on the system if they are "suspects" and not "defendants"… If someone moves in from our area and they are a prolific burglar, they won't feature whatsoever because nothing is transferred.' (Sergeant)

As summarised by one Sergeant who responded to the written questionnaire:

> 'I have concerns over how up to date the info is that is then being used to give the scores. Having checked when the trial initially started, no offenders we manage were featuring in the high brackets… I'm managing youths of significant risk, which [the dashboard] isn't picking up. It is therefore not something I would use as part of my daily business nor on a regular basis. It would be something I would check probably on a monthly basis to see if there are any names on there which we are not aware of.' (Sergeant)

It is important to note that the discrepancy between the officer's assessment and the machine-generated harm score is based on the subjective judgement of the officer; it was not clear what factors led to the officer concluding that the youths in question are 'of significant risk', or why these factors had not been identified by the dashboard. This reported limitation should therefore be interpreted with caution, as it represents the subjective opinion of officers.

This perceived limitation appears to be one of the main reasons why no PCs or Sergeants report having confidence in the accuracy of the harm scoring, or assess that it has delivered operational benefit in their force area (0%, *n* = 9). In relation specifically to the predictive model (the machine learning component of the application), it was suggested that:

> 'The suspect thing would come in handy there. If they're a suspect for offending but not a defendant, it can show us the frequency even if they're not charged or convicted… Where the system could come in useful is if people are getting arrested, there's lots of intel coming in about them… The people in the "Pursue" cohort, but where there are no conditions on them but they are on our radar.' (Sergeant)

It was reported that the decision to exclude suspects from the dashboard was made on the basis of ethical considerations associated with 'risk scoring' individuals who have not yet been charged with an offence. However, it was conversely pointed out that the force will continue to actively monitor, risk assess and manage others who have not been charged with an offence, but without the use of the app. It is therefore unclear whether the choice to exclude suspects from the dashboard does in fact represent a more 'ethical' approach, particularly if this is damaging users' trust and confidence in the overall system. This is discussed further in the following section.

4.2.4 Priorities for further improvement

Beyond the specific limitations discussed in the previous section, respondents also identified several areas of focus for future improvement, if the application were to be deployed for enduring use.

<u>Training</u>

The first and most significant priority for the future use of the application relates to the training provided to officers. As indicated by survey responses (reported below), the majority of respondents (PCs, Sergeants and Inspectors alike) reported that they have not received sufficient training or written guidance on the application before being required to use it (*n* = 11).

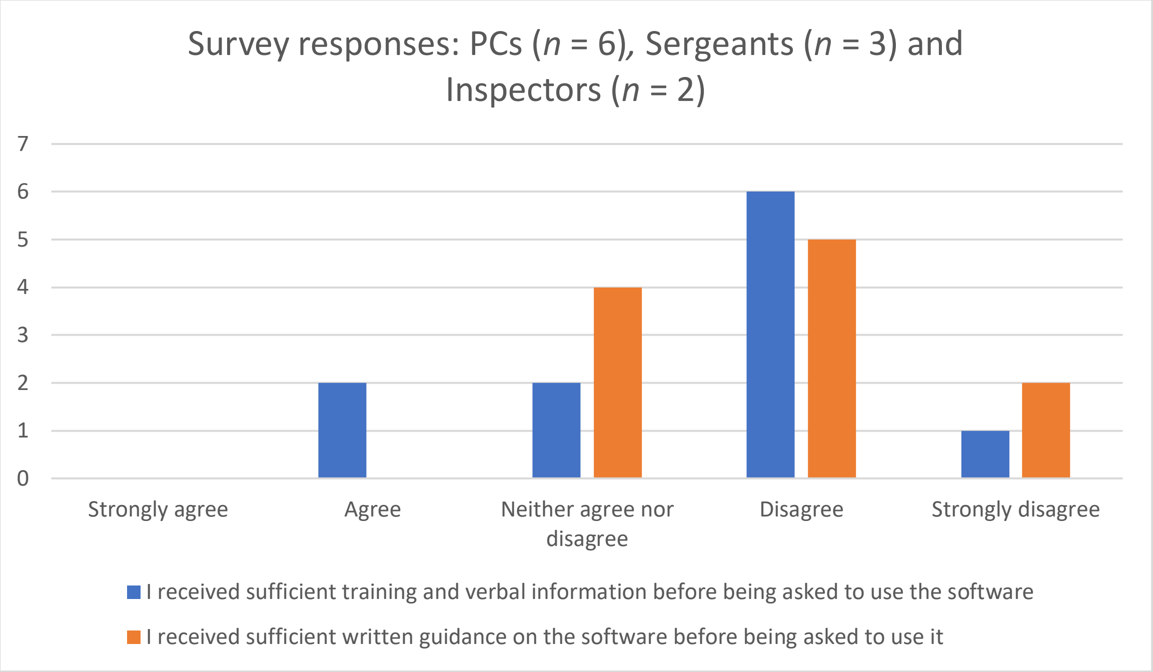

Figure 3. A lack of sufficient training and written guidance was widely recognised among survey respondents.

Inspectors and Sergeants particularly recognised the importance of further training, as indicated by comments such as:

> The training is an hour over Skype but I left as confused as when I started. (Sergeant)
>
> I think there should have been some written training guidance that we could refer to when using the system… The training given could have been more detailed. (Inspector)
>
> Some refreshed and specific training inputs would be beneficial prior to (or at the point of) further roll out of [the application] to ensure that all users are aware of the system and its benefits – maybe some written instructions that can be retained. (Inspector)

It is notable that two respondents did report receiving sufficient training and verbal information before being asked to use the software, but no respondents reported receiving sufficient *written* guidance before being asked to use the software. This suggests a need to focus on developing refreshed written guidance for officers regarding the new system, for instance on the organisational intranet. The perceived lack of sufficient training and guidance materials may have partially contributed to the divergence of views between Inspectors on the one hand, and PCs and Sergeants on the other. This is discussed further in the following section.

Filters

A particular missing feature that interviewees wanted to see added to a future version of the dashboard is the ability to filter by more specific crime types. PCs explained that 'because the selector criteria are not filtered, it's difficult to narrow down according to crime type.' (PC). To address this, they requested: 'can we have a top 20 or top 30 of offenders in a particular crime type, like burglary, knife crime etc.' (PC). This requirement was also recognised by Inspectors interviewed, who suggested it would be particularly helpful to filter for domestic-related incidents specifically:

> I'd like to see more specific filters for violence, domestic abuse… I'd like to see a "domestic" tab in there as well, that would be a really useful filter to have in so we can divide out anything with a domestic element… And just to retain that ability to have filters built into it depending on what we'd like to look at… Perhaps something just to filter through anything that's domestic. (Inspector)

Lack of this functionality may be one of the factors leading most respondents to report not regularly using the harm score as part of their offender management responsibilities.

This raises an important question regarding the scope of data-driven risk assessment tools more broadly. As has been discussed in the literature, numerical 'risk scores' often do not provide insight into the specific *nature* of the risk to be prevented, or specific steps that can be taken to mitigate that risk. It appears that a generic 'harm score' indicating the level of harm associated with an individual's *overall* offending does not provide sufficient granularity to be practically useful to officers who are trying to prioritise individuals at risk of committing specific types of offending. This is discussed further in the following section.

<u>Data integration</u>

The final missing feature that several respondents requested was the ability to view other intelligence related to an individual within the dashboard, most notably pictures of offenders:

> When we look at it there are no pictures of offenders, it's just names. On [a separate data system] you would see pictures and other Intel, [the dashboard] is just words… it's too data-driven. (PC)

This primarily appears to be an issue of data integration and cross-compatibility with other information management systems. As explained by one PC:

> It wouldn't be the sole basis of a decision. You'd go through other things, you'd speak to your neighbourhoods… [the dashboard] is more of a selection process tool. It might be better if [the dashboard] was part of [our main data system] in the future. (PC)

Indeed, this was also recognised by the project's lead data scientist as one of the main limitations of the current application:

> The main limitations for the model are the data available to it. We don't have data from other forces, or information from other sources. (Police data scientist)

However, it must be borne in mind that this is not unique to the tool being assessed but is a recurring issue across all police information management systems. The issue of data integration is complex: on the one hand, a lack of integration with other information management systems causes added inconvenience for officers, who are required to manually triangulate risk scores against other relevant data systems. On the other hand, the aggregation of multiple sources of sensitive information about an individual could increase the potential intrusion incurred during the risk assessment process, and this would need to be assessed as necessary and proportionate in relation to the anticipated efficiency gains.

Chapter 5. Discussion and Recommendations

As outlined in the previous section, the research generated new insights regarding the potential benefits and limitations of new data-driven risk assessment tools for UK policing. This section analyses the findings of this research and explores their implications for future policy and practice.

Interviews with practitioners highlighted the potential of new data science approaches to improve the efficiency of existing risk assessment processes, enabling the analysis of far greater volumes of data than was previously possible. These interviews also highlighted the practical limitations of new data science approaches, including the challenge of gaining access to high-quality data with which to build a reliable model, the potential loss of explainability when using complex machine learning methods, and the risk of fettering the discretion of the human decision-maker.

With regard specifically to the process evaluation component of research, it is notable that the real-world strengths and limitations of the application in question when deployed in an operational policing environment did not align with the *perceived* strengths and limitations of data-driven risk assessment tools identified in the interview phase of the project, nor in the literature review. In other words, the issues that interviewees were most concerned about regarding the *potential* consequences of implementing data-driven methods for offender risk assessment were not the same as the issues that materialised in practice for offender managers involved in testing such a system operationally. This is an important finding, as it demonstrates that existing perceptions of data-driven risk assessment methods – both among CJS practitioners and more widely in the academic literature – do not reflect the operational reality faced by offender managers when such a system is deployed in a real-world policing environment.

With this in mind, any future policy and practice regarding the use of data-driven risk assessment methods within the UK CJS should be grounded in a clear understanding of the highest-priority challenges faced by the end-users who will be expected to use the system (in this case, offender managers). This suggests a shift away from 'theory-led' approaches to policy development towards a more 'practitioner-led' approach, where the end-users are active participants in the integration of new approaches to the operationalisation of data-driven systems within their existing working processes. The following recommendations and discussion that follows aim to provide an initial guide to the most pertinent issues that should be considered in developing a more practitioner-led approach to policy and implementation in this context.

5.1 Recommendations

Recommendation 1: Future approaches to developing data-driven offender risk assessment tools for use in UK policing should start from a clear understanding of the operational challenges faced by end-users who will ultimately be expected to use the new system. This should be established through initial user research, for instance comprising interviews and surveys. Academic literature on this topic should be interpreted with caution, as many research publications relate to studies undertaken in the US, and it is unclear the extent to which these findings are directly transferable to the UK policing context.

Recommendation 2: The process evaluation has not established sufficient positive evidence in favour of deploying the harm dashboard and accompanying predictive model for long-term operational use. The application should be subject to further, detailed evaluation research to conclusively establish its benefits and limitations before it is deployed for wider operational use.

Recommendation 3: The force should establish a clear impact evaluation plan to measure the outcomes of the harm dashboard and predictive model on an ongoing basis. This should include developing a basic logic model or Theory of Change to describe the intended outputs and outcomes of the project, as outlined in the College of Policing's Evaluation Toolkit. It is important to define measurable evaluation criteria to assess the ongoing business case for the project and demonstrate that it is delivering its intended outcomes. The Maryland Scientific Methods Scale provides useful methodological guidance for ensuring the validity of policing evaluation research, and should be used as a guide to establish this evaluation plan and Theory of Change.

Recommendation 4: Offender Managers who are required to pilot or trial any new data-driven risk assessment tool should be consulted at an early stage in the project development, giving them an opportunity to directly contribute to the application development process. An initial survey of end-users should be distributed, requesting feedback on the limitations of existing processes, and the user interface and design requirements for any new system.

Recommendation 5: Any future development of the harm score should focus on identifying nominals not currently subject to offender management orders who should be subject to more in-depth risk assessment. To avoid the risk of false negatives (which could lead to high-priority nominals being erroneously de-selected), individuals already subject to offender management plans should be excluded from the harm scoring system.

Recommendation 6: The research was inconclusive regarding the potential benefits offered by the predictive modelling component of the application. The predictive model should be subject to dedicated, controlled evaluation research before it is deployed operationally. The criteria for success for the predictive modelling component should be more stringent than the harm scoring component, given the additional ethical considerations associated with making individual-level forecasts of future human behaviour.

Recommendation 7: If the predictive model is to be deployed for enduring use, model outputs should be more clearly distinguished from the descriptive harm scores. A caveat should be included alongside the model outputs, with the following 'health warning': *Prediction generated by statistical model. Accuracy and confidence may vary depending on context. Validate alongside other data sources before taking further action.*

Recommendation 8: Future practitioner research into data-driven offender risk assessment should assess the feasibility and desirability of establishing a two-stage approach to data-driven risk assessment – where the first stage involves automated harm scoring to prioritise individuals who should be subject to more in-depth risk assessment, with the subsequent stage involving predictive forecasting to identify individuals within this higher-risk cohort who are demonstrating behaviours which may indicate a potential escalation in their offending trajectory.

Recommendation 9: As part of the evaluation research detailed in Recommendation 2, further user testing is required to establish an appropriate classification threshold for the harm scoring and offender escalation model, specifically to identify the most appropriate balance to strike between false positives and false negatives. The model's threshold should then be updated accordingly and kept under regular review – for instance through surveys with end-users to establish whether the number of alerts generated by the system is creating an unmanageable volume of referrals for more in-depth assessment.

Recommendation 10: Further research should assess the feasibility of generating harm scores related to particular offending types, to assist offender managers in identifying the specific nature of the risk associated with each nominal. For instance, a more granular scoring system would provide an overall harm score for each individual, broken down into several 'dimensions' of risk (e.g. risk of future violence, risk of future acquisitive crime, risk of drug-related crime etc.).

Recommendation 11: Any police force deploying a data-driven risk assessment system should produce a list of potential interventions that the resulting harm scores may be used to inform, as well as a list of any external agencies or third-parties with whom the scores may be shared for intelligence or offender management purposes.

Recommendation 12: The level of explainability required of a data-driven system (and therefore the choice of which statistical method to use, machine learning or otherwise) should be directly informed by an assessment of the potential impact of subsequent decision-making on individual rights. If a risk-based decision has the potential to significantly impact an individual, there must be very strong justification for integrating algorithmic data scoring into this process.

Recommendation 13: Any future development of the application should focus on improving the front-end user experience, incorporating best practice in data visualisation and software accessibility. The dashboard should incorporate a 'Feedback' section, where users can provide written feedback on the application and submit suggestions for improvement. Monthly feedback meetings should be held for officers to provide verbal feedback to the development team. Where possible, behavioural scientists should be consulted to advise on the most effective visual presentation of outputs to support decision-making and minimise risk of cognitive bias.

Recommendation 14: The most pertinent data points from other information management systems (most notably custody images) should be included within the harm scoring dashboard. Integrating the dashboard within the force's existing information management system is likely to achieve this and should be a priority for any future development of the application.

Recommendation 15: The force should reconsider the decision to exclude suspects from the harm scoring dashboard. It is unclear if this represents a more ethical approach as such individuals will still be monitored and assessed through other means. The force should consult with its Ethics Committee to seek refreshed guidance on this issue, in light of the findings presented in this report.

Recommendation 16: Additional training should be delivered to all officers with access to the harm dashboard and predictive model. This should cover how the application is intended to be used, the input variables used to calculate the harm score and to build the predictive model, and an overview of the inherent limitations of the statistical techniques underpinning the system.

Recommendation 17: Written guidance (and frequently asked questions, FAQs) should be developed for all officers with access to the harm dashboard and predictive model. This guidance should include a summary of how the application generates the harm scores and statistical predictions, as well as a workflow diagram of how the algorithmically-generated insights should be integrated into existing offender management processes. Pre-recorded video training should also be produced and made available on demand through the force intranet.

Recommendation 18: The training and guidance described in Recommendations 16 and 17 should include a basic explanation of precision and recall at different classification thresholds, including a simple set of 'exam questions' to ensure that all officers understand the relationship between precision, recall, false positives and false negatives. This is essential to ensure accountability of the overall risk assessment process.

Recommendation 19: Training and guidance materials should be developed for offender managers specifically focused on managing risk of cognitive bias, due to the potentially prejudicial impact of risk labels on human decision-making. Behavioural science expertise should be consulted and involved in the development of these materials.

Recommendation 20: Any police force considering developing or deploying a data-driven risk assessment system for operational use should conduct a full integrated impact assessment prior to commencing the project. This should include four elements: 1) a data protection impact assessment; 2) a human rights impact assessment (with particular focus on Article 8 concerns and the proportionality of privacy intrusion); 3) an equality impact assessment (including testing the data and model for any unacceptable bias); and 4) a community impact assessment, to understand any potential disproportionate impact of the system on particular groups or communities. This integrated impact assessment should be made publicly available, and reviewed periodically throughout the project lifecycle.

Recommendation 21: Any police force considering developing or deploying a data-driven risk assessment system for operational use should consult with an external ethics committee during the design stage of the project. This committee should include multidisciplinary expertise in fields such as law, ethics, computer science and criminology. A focus should be on identifying potential ethical risks associated with the project design, which should be communicated in writing to the force and made publicly available on the force website. The police force should then formally respond to any ethical concerns identified in this review process, and provide a clear plan for how these risks will be mitigated.

5.2 Testing and evaluation

The first issue to consider when implementing any new data-driven system for offender risk assessment is how the system should be tested and evaluated. This is becoming more complex as such systems increasingly incorporate machine learning (ML) modelling, which entails additional testing and evaluation considerations to traditional data systems.

Recent research has highlighted the critical importance of testing ML platforms prior to their operational deployment (Braiek and Khomh, 2018). This testing process should incorporate three elements: testing for data issues; testing for model issues; and testing for implementation issues. In relation to implementation issues, recent research has demonstrated the importance of *user testing* ML classifier models with a group of target users prior to their operational deployment, especially if they are to be embedded within a decision-making process that could have a legally significant impact on individuals (Knack, Carter and Babuta, 2022). Controlled trials are the most empirically robust approach to user testing, for instance to compare the error rates of using ML tools for analysis in contrast to conducting the task without ML (Knack, Carter and Babuta, 2022). However, controlled trials are often resource-intensive and time-consuming to implement, and may be ethically unfeasible in the policing context (Babuta and Oswald, 2020)

In relation to the process evaluation component of this study, it is laudable that the police force in question established a controlled trial methodology to beta-test the application in a controlled environment with a small number of offender managers prior to wider operational deployment. This is particularly important as the system in question is to be integrated into decision-making processes that have the potential to directly impact on individuals. However, for reasons discussed below, the process evaluation was not able to establish sufficient evidence to support wider operational deployment of the harm score dashboard and accompanying predictive model. Further testing and evaluation are essential to assess the potential benefits of the system if it is to be deployed for enduring use.

First, due to a combination of staffing changes, absence and non-participation, not all officers involved in the Beta-testing of the application took part in the process evaluation. As such, there may be other important factors relevant to the development and use of the system that were not captured by this research. As detailed below, the evaluation findings demonstrate only limited positive evidence in favour of wider deployment of the application, and the little evidence that was established was provided primarily by Inspectors who took part in the project, rather than PCs or Sergeants.

Second, most officers engaged for the research did not have a clear understanding of the distinction between the harm score on the one hand, and the predictive model on the other. For this reason, it has not been possible to meaningfully assess the benefits and limitations of the *predictive model* specifically, and further evaluation research is needed to assess this component of the system (distinct to the harm score).

Finally (and perhaps most importantly), there were no pre-defined evaluation metrics in place by which the overall impact of the project will be assessed. Without such evaluation criteria, it remains unclear what the intended outcomes for the project are and how these will be measured. As part of the longitudinal evaluation plan specified above, it is essential to develop clear evaluation criteria against which the project will be assessed, alongside a theory of change or logic model articulating the overall intended outcomes of the project.

The impact evaluation plan described above should be time-bound, and the project should not proceed unless it can be demonstrated that it is delivering its intended outcomes at the end of a specified evaluation period. These results should be independently reviewed by an impact evaluation specialist following completion of the evaluation period.

5.3 Does the harm score help officers to manage risk more effectively?

The previous chapters have demonstrated that the accurate identification and prioritisation of risk is an essential pre-requisite to the delivery of effective and timely interventions. As explored at the outset of this thesis, the Risk, Need and Responsivity (RNR) framework is central to offender management approaches throughout the UK CJS. Experts continue to disagree over the relative benefits and limitations of statistical approaches to offender risk assessment. However – with resourcing constraints in mind – it is inarguable that police forces must adopt structured and systematic methods to prioritise limited preventative interventions to those who pose the greatest risk of harm. The crucial question then is whether the use of new and advanced data scoring systems effectively support offender managers *in practice* in prioritising limited resources to the areas of greatest need.

As discussed previously (Section 4.2), the overall purpose of the statistical harm score under evaluation is to enable offender managers to manage risk more effectively and target preventative interventions to the highest-risk offenders. Inspectors responsible for overseeing the project described a shift away from enforcement towards more preventative approaches, highlighting the importance of robust risk assessment for the early identification of those who should be prioritised for further intervention.

Perhaps the most significant finding of the evaluation is a fundamental divergence in views between PCs and Sergeants on the one hand, and Inspectors on the other. Both Inspectors who responded to the questionnaire reported that they find both the harm score and the predictive model useful, and that the application represents an improvement over the previous software. In stark contrast, no PC or Sergeant reported finding the harm score or the crime prediction model useful, and none reported that it represents an improvement over the previous software. Based on these findings, it appears that the harm score and predictive model are significantly more useful for Inspectors than they are for PCs and Sergeants. There are two likely reasons for this.

First, the Inspectors who participated in the evaluation have been more directly involved in the planning and development of the system and are therefore likely to have a more detailed understanding of its strengths and limitations. They are likely to be more familiar with the rationale for its deployment and have had longer to familiarise themselves with the application. This demonstrates the critical importance of ensuring wide consultation and engagement in the early planning stages of a new data-driven system, to ensure end-users have had sufficient opportunity to contribute to the early development process.

A second potential reason for this divergence in views relates to the need for more senior officers to understand strategic-level insights across their force area. At the individual offender level, PCs and Sergeants reported numerous perceived over-classification errors, requiring users to validate risk scores manually via other systems. Conversely, they also reported numerous perceived under-classification errors and false negatives (missing nominals), suggesting that a large proportion of individuals who should be scored as high risk are not being identified by the system. The perceived occurrence of both over-classification errors and false negatives appears to have significantly damaged PCs and Sergeant's trust in the overall validity of the system, leading none of them to report that the harm score has delivered operational benefit in their force area. One Sergeant explained that the occurrence of false negatives 'discredits the info for me personally', demonstrating that the experience of even a small number of false negatives could cause users to lose trust in the validity of the system as a whole. (It is important to note, however, that these perceived errors are based on the subjective judgement of officers – there is no way to validate that the purported 'false negatives' and 'missing nominals' are in fact genuine errors of the system).

By contrast, at the more strategic level, Inspectors report being very confident that the system is identifying the right people, explaining how the harm scoring has identified a new cohort of under-25 offenders, the majority of whom were not previously subject to proactive management plans. As such, despite individual-level over-classification and under-classification errors that appear to have weakened officers' confidence in the validity of individual outputs, at aggregate level, the system appears to be surfacing high-risk nominals who should be subject to further intervention, but may have otherwise gone unnoticed. This identification of nominals who may otherwise not be subject to further scrutiny appears to be the single greatest strength of the harm score (as reported by Inspectors) and should be the focus of any future development of the application.

A final potential reason for the divergence in views between Inspectors and more junior officers could relate to the command and reporting structures within the force: as the Inspectors interviewed report directly to the Senior Responsible Owner for the project, they are inevitably incentivised to ensure the project achieves its intended outcomes – and may therefore be more likely to emphasise potential benefits of the pilot project over limitations and risks. These dynamics should be considered as part of any future practitioner-focused research in this area.

5.4 Does the predictive model help officers to manage risk more effectively?

Data-driven risk assessment systems can be broadly divided into two categories: those that assign static 'harm scores' to offenders based on their *current and recent* offending behaviour; and those that make predictive forecasts calculating an individual's risk of *future* offending. These imply two fundamentally different uses of the same data – harm scoring enables offender managers to identify those individuals whose offending behaviour is causing the greatest level of *present* harm, while predictive modelling enables offender managers to identify those who may be on a trajectory towards more serious offending (but whose level of existing harm may be comparatively lower). Predictive modelling is associated with a much higher degree of uncertainty than statistical harm scoring, which must be taken into account when integrating such systems into human decision-making and risk assessment processes.

PCs and Sergeants interviewed for this research did not distinguish clearly between the descriptive harm score on the one hand, and the predictive modelling on the other. The majority of interview and survey responses related specifically to the harm score. When prompted to provide feedback on the predictive modelling component as distinct from the harm score, most interviewees reported that the predictive model is not yet routinely used. The evaluation was therefore inconclusive regarding the potential benefits offered by the predictive modelling component of the application. Further evaluation research is required to determine whether the predictive model is useful in practice to help officers manage risk more effectively.

Nevertheless, one important conclusion can be drawn in this regard. From a statistical perspective, the harm score and the offender escalation predictions are fundamentally different categories of output. The harm score is a descriptive score corresponding to an individual's *current* level of offending. By contrast, the outputs of the predictive model are based on machine learning forecasting of *future* risk, and therefore entail a degree of inherent uncertainty. However, this important distinction has not been clearly articulated to users, reducing their ability to critically assess the validity of model outputs.

It is essential that end users are made fully aware of this inherent uncertainty associated with machine learning predictions, and that model outputs are treated with a significantly greater degree of caution than descriptive harm scores. It is concerning that officers interviewed for this evaluation were not aware of this crucial distinction between the descriptive harm score and the predictive modelling. If the application were to be deployed for enduring use, it will be essential to clearly communicate to users that the outputs of the predictive model are inherently uncertain and probabilistic, and should be treated with a higher degree of caution and scrutiny than the descriptive harm score.

Another consideration for future research is whether the predictive model and harm scoring should be used to inform the same risk assessment processes, or whether they should be used for two different purposes. At present, both the harm scores and the model predictions are presented alongside each other within the same user interface or 'dashboard'. It is expected that officers will use both the static harm score and the predictive forecast – in conjunction with their own professional judgement – to prioritise individuals who should be subject to further scrutiny or more in-depth risk assessment. However, an alternative approach would be for the harm scores to be used to prioritise individuals who should be subject to more in-depth risk assessment, with the predictive model being used as an 'early warning system' to identify those individuals who do not currently require active management, but who should be kept under close observation to monitor for changes in their offending behaviour.

This would entail two separate processes – a harm scoring process to surface individuals who should be subject to more in-depth risk assessment *now*, and a predictive forecasting process to identify individuals who should be monitored for any potential changes in their offending behaviour *in the near future.* Future research in this area should assess whether this two-tiered approach to data-driven assessment would improve officers' ability to manage future offending risk.

5.5 Classification thresholds

When developing a machine learning model, it is necessary to determine the classification threshold above which a data item is categorised into a certain class. For example, a classification threshold set at 0.99 only generates an alert if the model calculates a 99% or higher probability that the target variable belongs to a certain class. The same model could instead be set to a 0.95 threshold, meaning a 95% or higher probability would trigger a positive alert (Knack, Carter and Babuta, 2022). This sensitivity is a crucial factor to consider when developing any machine learning classification model, as it has direct implications on the level of confidence users should place on probabilistic outputs.

The model's classification threshold determines its precision and recall rates, which in turn determine the expected level of false positives and false negatives. Precision describes what proportion of positive identifications were correct, while recall describes what proportion of actual positives were identified correctly. A higher precision entails a lower false positive rate, while a higher recall entails a lower false negative rate. In practice, there is often a trade-off between precision and recall, meaning the developer must make a judgement as to whether to favour false positives or false negatives when setting the model's threshold. This decision is highly context-specific, as in some decision-making contexts it may be more important to avoid the risk of potentially important information being missed (false negatives), while in other contexts it may be more important to minimise false positives – for example if there is only limited resource available to manually review the model's outputs.

As has been discussed above in Section 5.2, the primary benefit of the application appears to be in identifying individuals not currently subject to offender management orders who should be subject to a more in-depth risk assessment. For this reason, it follows that the model's classification threshold should be set lower, to favour higher recall and lower precision (i.e., to favour false positives over false negatives). This is also supported by the observation in the literature that tolerance for false negatives in the intelligence analysis context may be very low, given the risk of letting potentially high-risk individuals 'slip through the net' (Knack, Carter and Babuta, 2022).

However, although false negatives are arguably a more 'costly' error in the offender management context, there is also an important balance to be struck to avoid the risk of too many false positives creating an unmanageable volume of alerts for already over-stretched offender management teams. The research has shown that the system is already demonstrating over-classification errors, where officers perceive that individuals are erroneously being scored as high risk. There is therefore a risk that reducing the model's classification threshold may increase the perceived number of over-classification errors, which may discourage officers from using the system. For this reason, the classification threshold should be determined based on more in-depth engagement with end users, to understand from an offender manager's perspective the appropriate balance to strike between false positives and false negatives.

Moreover, it is crucial for users of the system to understand the relationship between precision, recall and classification thresholds in order to appropriately interpret model outputs. For instance, if a model had been weighted to favour higher precision (i.e., to minimise false positives), it is important for users to take into account the higher likelihood of encountering false negatives – which could lead to potentially important information being effectively 'screened out'. These additional training requirements for users are discussed further below.

5.6 Risk of what?

As was discussed in Section 2.4, the calculation of statistical 'harm scores' for individual offenders represents the creation of new information regarding an individual, which could then influence their subsequent treatment. As the interventions that may result on the basis of the harm scoring could significantly impact on individuals' human rights and civil liberties, it is essential to demonstrate that the creation of such scores is *necessary* and *proportionate*. This means clearly articulating the overall decision-making process that such scoring will be used to inform – for instance whether the interventions that could result are supportive or punitive.

The question of whether the use of a data-driven tool can be justified as a necessary and proportionate use of police powers will depend largely on the outcomes that could result from its use. In simple terms, what interventions could result on the basis of the harm scoring, and what is the risk that these interventions are designed to prevent? To use a crude example, if a system were used to harm score all prior violent offenders whose offences were alcohol-related, and identify those who should be offered additional support in the form of addiction therapy, counselling and other offender management programmes, this is more likely to be assessed as proportionate than if the same system were used to harm score all known gang-affiliated nominals in a particular area, and then inform stop-and-search targeting conducted by neighbourhood officers. This again illustrates that the same data-driven system can be used for numerous potential purposes, and it is crucial at the outset to articulate the overall decision-making process which the statistical outputs will be used to inform.

A common limitation of statistical approaches to offender risk assessment (which has been discussed at length in the academic literature) is that a harm score in isolation provides no insight into the specific nature of the risk that is to be prevented. This is why additional, individualised assessment is required to identify the risks and needs associated with each offender, to establish a bespoke risk management plan. Officers interviewed for the process evaluation identified the lack of 'filtering' functionality as a priority for future improvement, suggesting that it would be helpful to identify high-risk nominals according to specific offending categories. This is an important consideration as the system under examination currently produces 'generic' harm scores for all offenders, which provide no insight into the specific nature of the risk for each individual.

Beyond exploring new methods for calculating more granular risk predictions at the individual level, the force should also clearly articulate a list of potential interventions that could result on the basis of the harm scoring, as well as other agencies with whom the scores may ultimately be shared. This is essential to assess in advance the proportionality of any intrusion that could result on the basis of being included on the system.

5.7 Transparency and explainability

The impact of data-driven scoring systems on the transparency and explainability of the risk assessment process was a key theme identified in the research. On the one hand, interviewees suggested that a key limitation of new advanced data science-based approaches is their perceived lack of explainability. While traditional statistical risk assessment tools allow clear relationships to be identified between input variables ('features') and the resulting risk score, complex modelling techniques (particularly those incorporating machine learning) often do not allow the human observer to identify which input factors or combination of variables were most significant in contributing to an overall prediction or output. This potentially limits their use for offender management purposes, where decision-makers need to maintain a clear and defensible link between the various factors that may lead to certain outcomes.

These concerns raised by interviewees are consistent with recent discussion in the wider AI literature. While ML techniques have significantly improved in speed and accuracy, they are also becoming increasingly complex and difficult to understand (Knack, Carter and Babuta, 2022). The most popular model architectures – particularly those incorporating deep learning – typically involve many hidden computations which obfuscate the relationships between input features and the resulting output – a phenomenon which has been widely described in the literature as the 'black box' problem (Veale, Van Kleek and Binns, 2018; Holzinger *et al.*, 2022; Zhou, Chen and Holzinger, 2022). In the risk scoring context, a model may provide a certain classification (e.g. 'high risk' or 'low risk'), but it is impossible for the user to understand the logic of the model or which factors were assessed to be more important in predicting this outcome (Busuioc, 2021; Knack, Carter and Babuta, 2022). Recent research has explored the potential implications of the 'black box' problem in an intelligence and security context, which include the risk of hindering algorithmic assessments from being challenged or casting doubt on who should take accountability for decisions (Knack, Carter and Babuta, 2022). This is particularly concerning in the policing context where the actions taken on the basis of risk assessments may have significant consequences for the individuals targeted.

Despite these concerns raised in the literature and research interviews, it is notable that the process evaluation conducted for this study arrived at somewhat different conclusions. When asked to describe what they perceived to be the greatest strengths of the data scoring system under evaluation, Inspectors reported that the new application has in fact *increased* the transparency of the risk assessment process, as the factors that contributed to any given score can be readily identified and cross-referenced against other systems. This transparency was described as directly linked to officers' confidence in the outputs – with Inspectors reporting that they would feel less confident making decisions on the basis of the risk score if they were not able to understand how the scores had been calculated.

It is important here to reiterate the fundamental distinction between the harm scoring system on the one hand, and the predictive modelling on the other. The increased transparency described by Inspectors appears to relate specifically to the harm scoring element of the system. As discussed previously, as the harm score is *descriptive* rather than *predictive,* the calculations involved are not associated with the same degree of 'black box' opacity discussed above regarding complex ML modelling. It is therefore possible to identify which risk factors have contributed to each harm score (although it may not be possible to identify the specific weighting of each input variable).

This has important implications for future policy and practice. While much commentary has focussed on the 'black box' problem of complex machine learning systems, it appears that descriptive data scoring that is *not* derived using machine learning can in fact *improve* the perceived transparency of the initial filtering process. One likely reason for this is that offender management decision-making has historically depended to a large extent on the professional judgement and discretion of officers. While it is important to retain an appropriate degree of professional judgement throughout the risk management process, it is nevertheless important to systematise this exercise of discretion – to enable a defensible link to be drawn between risk factors and any resultant outcomes. At the individual level, it is essential for officers to be able to articulate clear and coherent reasoning as to why certain risk decisions have been made. At the aggregate level, it is not possible to assess each case in detail – meaning automated systems can help to highlight the most pertinent factors that have led to certain individuals being prioritised over others.

Recent research has emphasised that the level of explainability required from data-driven systems is highly context-dependent and linked directly to the potential impact of subsequent decision-making. For instance, if the decision(s) made on the basis of automated analytics are likely to have a significant impact on an individual, users will generally require a much higher granularity of explanation regarding the factors that led to a certain prediction (Knack, Carter and Babuta, 2022). In some decision-making contexts, the use of algorithmic risk scoring may be entirely inappropriate – if the decision has very significant consequences for the individual being assessed. For this reason, the suitability (or otherwise) of an algorithmic approach, the complexity of the modelling used, and the level of explainability required from the system, must all be assessed in the context of the wider decision-making process in which the system is being used. In their role as Chair of a policing ethics committee, the author has seen first-hand that ethical decisions regarding algorithmic systems are often made in isolation – with insufficient attention paid to the subsequent action that may be taken on the basis of a data-driven output. This is an important gap which must be accounted for in future policy and practice.

5.8 User experience

As mentioned at the outset of this thesis, existing academic analysis of data-driven policing systems has focused almost exclusively on issues of statistical validity and mathematical validation. While this research is important, it often fails to consider the fundamental question of whether the software is *useful in practice* for those who are required to implement it in an operational policing context. This is because academic analysis tends to approach the issue from a data science perspective, rather than a software engineering perspective.

In this context, data science refers to the back-end statistical models that make the complex calculations required to derive individual risk scores or offending predictions. Software engineering refers to the front-end user interface of how this information is presented to users, including how to communicate key information such as accuracy, precision and recall at different classification thresholds. Any holistic evaluation of data-driven policing systems must take into account both the data science elements of the system and software engineering considerations related to useability, accessibility and interpretability.

One of the main potential strengths of data-driven policing systems identified in the research interviews was the ability to improve the *efficiency* of the risk assessment process. Automated analytics systems – particularly those incorporating machine learning – can rapidly triage and prioritise information in massive volumes of data, at a volume and velocity that far surpasses manual human analysis. However, deficiencies in front-end system usability are likely to result in officers becoming frustrated or confused with the system, or result in additional time spent triangulating information across other systems – limiting adoption.

As outlined in the previous section, the process evaluation found that the implementation of the dashboard has resulted in a poor user experience for officers. There are two main reasons for this. The first relates to the user interface, with officers reporting that the graphical interface of the dashboard is not user friendly. Regardless of the technical performance of the system, a poor user interface is likely to deter officers from regularly accessing the dashboard, resulting in low adoption levels. Any future development of the dashboard should focus on continuously improving user interface and accessibility features, by requesting regular feedback from officers.

Recent research has highlighted the importance of including end-users in the development stage of data-driven intelligence systems, particularly to identify specific interface design requirements that may otherwise be overlooked (Knack, Carter and Babuta, 2022). As well as requiring software engineering (not just data science) expertise, there is also a strong behavioural science element to this process – it is important to carefully consider factors like the visual presentation of outputs, the language used to communicate complex concepts like 'risk', and the way that system performance and limitations (such as false positive and false negative rates) are communicated to officers.

5.9 Data integration

Beyond the graphical user interface issues identified above, the second reason for the poor user experience relates specifically to data integration challenges. As is common for police data systems, officers report needing to access multiple systems separately to triangulate the information provided by the harm score. A common request was for all relevant information to be available on the same system, rather than needing to access the dashboard in parallel to other existing police data systems. While this is symptomatic of a broader data integration challenge across UK policing (Babuta, 2017), future efforts should focus on extracting the most pertinent data points from other key databases to be presented alongside the harm score and predictive outputs.

There are important ethical considerations with regard to data integration. Much of the data in question is highly sensitive personal data, and access must be restricted only to those individuals with a genuine *need* to have access to that data. Although the fragmentation of information across police data systems is a long-standing problem that is the source of much exasperation (Babuta, 2017), there are many circumstances in which the siloing of particularly sensitive personal datasets is more ethically defensible than centralising large volumes of personal data on a single system that can be readily accessed by a large number of officers. So there is an important balance to be struck – between ensuring sufficient data integration so officers can readily access all the information they need to make a reliable and informed risk judgement, while not creating a highly intrusive database which contains a disproportionately large amount of detailed information about many individuals, which can be readily accessed by many officers with limited oversight.

As discussed previously, the harm scoring dashboard only includes information relating to those who have been charged with an offence. It was reported that the decision to exclude suspects from the dashboard was made on the basis of ethical considerations associated with risk scoring individuals who have not yet been charged with an offence. However, it was also pointed out that the force will continue to actively monitor, risk assess and manage those who have *not* been charged with an offence, but without the use of the app.

It is therefore unclear whether the choice to exclude suspects from the dashboard does in fact represent a more 'ethical' approach, if such individuals will still be actively assessed and monitored. It is clear from interview data that the choice to exclude suspects from the dashboard has weakened the utility of the system from the perspective of officers, who are still required to manually access various other intelligence systems to collate information relating to known suspects who have not yet been charge. This in turn has damaged users' trust and confidence in the overall system.

The research was inconclusive regarding whether suspects should be included in the harm scoring dashboard alongside convicted offenders. However, as a minimum, it is clear that pertinent data items from other systems (such as custody images, address history and other relevant personal data) should be embedded within the dashboard, rather than requiring officers to access multiple systems individually. This will improve the user experience of the dashboard as a whole, thereby increasing adoption and acceptance among officers.

5.10 Training and guidance

Finally, there was broad consensus that further training and written guidance would be required if the application were to be deployed for enduring use. In addition to the specific limitations of the system identified by interviewees, the research also highlighted a general lack of sufficient understanding among PCs and Sergeants of how the application works in practice, and how the insights derived from the system are expected to be integrated within existing offender management processes. The fact that one Sergeant reported leaving the training session 'as confused as when I started' is concerning, and suggests that further training should be a high priority if the application is to be deployed operationally.

This lack of sufficient training could also be a main reason why PCs and Sergeants reported not finding the application useful, and not routinely using it as part of their offender management responsibilities. It may also have partially contributed to the divergence of views between Inspectors on the one hand, and PCs and Sergeants on the other. If operational users do not sufficiently understand how the system works – including the strengths and limitations inherent in the application – they are unlikely to feel comfortable using it to inform operational decision-making. This is particularly important when the decisions informed by the application will have a direct impact on individuals being assessed, requiring a high degree of confidence and accountability throughout all stages of the decision-making process.

Another observation of the research is the high staff turnover during the time period that the study was undertaken. Many individuals who took part in the later stages of the project had not been present during the training sessions that were delivered at the outset of the beta-testing process. Given the relatively high turnover in many offender management teams, it is essential to also develop clear written guidance (such as Wiki pages on the organisational intranet) for users to refer to, in lieu of attending in-person training sessions. Pre-recorded video training would also be a useful measure to ensure that all officers using the dashboard have access to the same information that was provided at the time initial training is delivered.

Two specific areas of focus were identified as particular training priorities for officers to responsibly use data-driven insights to inform risk assessment processes.

The first relates to training on machine learning precision, recall and classification thresholds. As discussed previously, machine learning systems produce probabilistic outputs, which are associated with an inherent degree of uncertainty. At the development stage, a model must be weighted to either favour higher recall (fewer false *negatives* at the expense of more false *positives*), or higher precision (fewer false *positives* at the expense of more false *negatives*). The classification threshold for the model – and the precision and recall rate that this entails – has profound implications for the confidence and reliability of the system outputs, and must be sufficiently understood by all those who are using the predictions to inform subsequent decision-making. For example, if a model has been weighted to favour higher recall, it must be communicated to officers that this is likely to result in more false positives – or over-classification errors, where individuals are erroneously judged to pose a heightened risk of future offending. Conversely, if a model is weighted to favour higher precision, officers must be aware that this may lead to a higher number of false negatives – meaning the model may not identify all individuals who are at increased risk of future offending.

The second area of focus for future training relates to cognitive bias and heuristics in judgement. As has been discussed in the literature, while the risk of cognitive bias is often used as an argument in favour of the statistical approach, a ‘risk score’ is potentially highly prejudicial to the human decision-maker (Cooke, 2010). As Cooke and Michie (2012) discuss, the ‘anchoring bias’ is a well-established cognitive bias that influences human judgement: ‘It is difficult for the decision-maker to disregard the number and alter their evaluation even if presented with detailed, credible and contradictory information’ (Cooke and Michie, 2012, p. 10). Quantifications of risk are also liable to be misinterpreted by others who were not directly involved in the assessment, and can be misrepresented, either deliberately or otherwise (Cooke, 2010).

For example, a ‘low risk’ label could be automatically interpreted to mean that an individual requires no further monitoring or intervention. Such ‘low risk’ individuals may have specific needs and vulnerabilities that should be addressed as part of a bespoke risk management plan; needs and vulnerabilities which may not be detected by a statistical algorithm. Such individuals may then fail to receive the necessary support to prevent them from returning to problematic behaviour (Cooke, 2010). Similarly, a ‘high risk’ label could influence offender managers’ judgement in the other direction

– they may be reluctant to subsequently assess such individuals as *not* posing an immediate risk, due to the increased accountability risk of not acting on potentially important information. Future training should therefore focus on ensuring offender managers understand the potential impact of cognitive bias and judgement heuristics on the decision-making process, and can critically assess risk scores in conjunction with their own professional judgement.

5.11 Ethical deployment

As explored earlier in Section 3.3, police use of behavioural analytics and data-driven risk assessment raises numerous legal and ethical considerations. It is beyond the scope of this thesis to provide a comprehensive analysis of the full range of these, but it is important to highlight several ethical concerns that emerged specifically in relation to the process evaluation component of the study.

The first ethical challenge identified is the risk of undermining the professional judgement and discretion of the human decision-maker. Although data-driven systems such as the one being evaluated are presented as 'supporting' rather than replacing human judgement, in practice human officers may be reluctant to contradict the outputs or predictions provided by the system – due to the additional accountability risk this entails. This is a particular concern as officers reported encountering numerous instances of 'over-classification' of risk, where they perceived that an individual had been erroneously assigned a higher-risk classification. In practice, this could create a situation where a human officer would not have assessed a particular individual to pose a heightened risk, but because the algorithm has assigned that individual a 'high risk' label, the officer feels obliged to conduct further scrutiny and analysis – which may entail a degree of collateral intrusion that would not have occurred otherwise. Assuming some of these high-risk calculations will be the result of false positives, the use of a data-scoring risk assessment system could therefore lead to unnecessary and intrusive scrutiny of individuals who would not have otherwise been subject to such scrutiny.

Conversely, there is the inverse ethical risk of *not* identifying individuals who pose a heightened risk and are therefore in need of additional supportive interventions. As mentioned previously, the decision was made to only include on the dashboard individuals who have previously been charged with an offence. However, a risk of this approach is that potentially high-risk nominals will not appear on the scoring dashboard. Over time, if officers come to rely on the scoring dashboard as a primary information source for prioritising nominals for further risk assessment, there is a risk that individuals who do not appear on the dashboard will 'slip through the cracks' and become deprioritised in relation to subsequent risk management planning. This suggests the force should reconsider the decision to exclude suspects from the harm scoring dashboard, given the additional ethical concerns associated with *not* including such individuals in the system.

Finally, a broader (but highly contentious) ethical issue relates to proportionality of intrusion – and the question of whether it is more intrusive for sensitive personal data to be reviewed by a conscious human, or processed by an automated system. There has been much academic discussion on this topic. As noted by Babuta, Oswald and Janjeva (2020):

> 'The use of AI arguably has the potential to reduce intrusion, both in terms of minimising the volume of personal data that needs to be reviewed by a human operator, and by resulting in

> more precise and efficient targeting, thus minimising the risk of collateral intrusion. However, it has also been argued that the degree of intrusion is equivalent regardless of whether data is processed by an algorithm or a human operator. According to this view, the source of intrusion lies in the collection, storage and processing of data. The methods by which this is achieved – whether automated or manual – are immaterial.' (Babuta, Oswald and Janjeva, 2020)

While this issue remains a matter of open debate, it is important to consider two important implications of this. First, the possibility that the use of automated systems – such as a data-driven risk scoring algorithm – could potentially reduce privacy intrusion by minimising the number of individual cases that must be manually reviewed by human officers. From this perspective, behavioural analytics systems could present a more proportionate alternative to traditional risk assessment methods. And second, the volume of individuals affected and whether the use of automated systems could lead to the processing of far more data than was previously possible. One can conceive of a scenario in which a highly effective automated risk scoring system results in human officers only targeting their manual review processes at a smaller number of high-priority nominals, but simultaneously many more individuals are being passively processed by the automated system than were previously being reviewed manually.

The ethical concerns raised above must be formally assessed at the outset of any police behavioural analytics or data-driven risk assessment project, and pre-emptive measures implemented at the project design stage to mitigate against potential unintended consequences. It is recommended that any police force considering developing data-driven risk assessment systems for operational use conducts a full 'integrated impact assessment' prior to commencing the project (Babuta and Oswald, 2020). This impact assessment should include four central components, including a data protection impact assessment, human rights impact assessment, equality impact assessment, and community impact assessment.

While this initial impact assessment is essential to understand and mitigate against potential ethical risks, it must be accompanied by appropriate governance, review and external oversight functions to ensure transparency and accountability in the deployment of the new data-driven system. Different forces have different approaches to ethical governance of technology projects, and there is no 'one-size-fits-all' model that can be recommended for use nationwide. However, as a minimum it is recommended that new data-driven risk assessment projects that have the potential to impact on individual rights should be formally reviewed by an external ethics committee before they are authorised for operational deployment. This committee should include multidisciplinary expertise in fields such as law, ethics, computer science and criminology. The committee's advice should be made publicly available and the police force should formally respond to any ethical concerns identified at the review stage, with a clear plan in place as to how these risks will be mitigated.

Chapter 6. Conclusions

The research examined the use of behavioural analytics and data-driven risk assessment within UK policing. The project adopted a case study approach, including a process evaluation of a major behavioural analytics project currently underway in one of the UK's largest police forces. This section summarises the key conclusions from the research and their implications for future policy and practice, as well as priority avenues for future academic research.

The first key observation is that numerous police forces nationwide are now developing and deploying next-generation data scoring tools for offender risk assessment purposes, but with a lack of any formal policy or national-level guidance on how such systems should be developed or used. Given the significant operational, legal and ethical considerations highlighted in this research, this is a concerning gap which should be addressed as a matter of urgency. Future policy efforts should focus on developing official guidance – potentially in the form of Authorised Professional Practice – for police forces seeking to develop or deploy such systems operationally.

While existing academic research has focused primarily on statistical evaluation of data-driven systems, this study has highlighted that the most pressing concerns for those deploying such systems in an operational policing context are largely organisational – not technological – considerations. Issues such as the human-machine interaction, the impact of automated analytics on existing decision-making processes, user experience, training and data integration all emerged as consistent themes in the research. Future policy for the use of data-driven risk assessment in policing must account for this full range of organisational and operational considerations, beyond simply providing technical guidance on the design and development of systems.

The importance of testing and evaluation was highlighted consistently throughout the research. In relation to the process evaluation specifically, it is notable that the force in question did not have a formal evaluation plan in place – or clear success criteria by which the project would be assessed. The process evaluation presented here was inevitably limited in scope, and did not establish sufficient evidence in favour of deploying the harm scoring dashboard and accompanying predictive model for long-term operational use. Further, detailed evaluation research is required to conclusively establish the potential benefits and limitations of the system before it is deployed for wider operational use. This should include developing clear longitudinal evaluation metrics, and a theory of change describing the overall intended outcomes of the project. The research remains inconclusive on whether the application will ultimately provide the operational benefits for the force that were envisaged at the time of its development.

Perhaps the most notable finding of the process evaluation was the fundamental divergence in views between PCs and Sergeants on the one hand, and Inspectors on the other. Inspectors were significantly more positive regarding the new application, while the feedback returned by PCs and Sergeants was unanimously negative. One potential reason for this is that Inspectors involved in the research had been more directly involved in the development of the application, emphasising the critical importance of ensuring end-user engagement at an early stage in the design process. The beta-testing may have yielded a more positive outcome if PCs and Sergeants had been more closely involved in the early project development stage, for instance to advise on user interface requirements. This may have identified at an earlier stage the perceived deficiencies in user experience that emerged in the course of the evaluation. As part of the evaluation plan mentioned above, formalised reporting mechanisms should be established for end-users (PCs and Sergeants) to provide ongoing feedback on the user experience and design requirements of the system, including at the early project development stage.

At an operational level, PCs and Sergeants report encountering both over-classification errors (i.e., an over-estimation of risk) and under-classification errors (i.e., individuals not being flagged by the system despite posing a high level of risk). In the offender management context, false negatives are likely to be a more ‘costly’ error, as they could result in high-risk nominals being erroneously de-selected. For this reason, the research broadly concluded that the harm score should *not* be used to assess offenders who are currently subject to offender management orders; such individuals should be subject to detailed individual risk assessment (incorporating structured professional judgement) to assess whether they are eligible for de-selection. Instead, the harm score would provide greater value as a ‘risk identification’ tool, enabling the identification of high-risk nominals *not* currently subject to offender management orders, who should be prioritised for more in-depth (manual) risk assessment.

Several priority issues should be addressed if the application is to be deployed for enduring use. This includes ensuring a comprehensive training plan for users, including written guidance summarising how the harm score and model outputs are calculated. This is important not just to ensure officers understand how to use the system in conjunction with existing processes, but also to maintain accountability throughout the full decision-making chain. At the technical level, efforts should be made to include more selection criteria within the dashboard to enable users to filter according to specific crime types, and to extract relevant data points from other systems to be integrated within the dashboard itself. The ideal outcome described by interviewees would be for harm scores to be integrated within the force’s existing information management systems.

The findings presented here relate specifically to the beta-testing phase of the project under evaluation. Many findings are likely to be generalisable to other behavioural analytics and data-driven risk assessment projects nationwide, although this cannot be assumed. Nevertheless, the study highlighted several priority areas for future research. Specifically, future research should aim to:

- Examine in detail the organisational power dynamics between operational users and more senior officers overseeing data science projects, to explore whether the divergence in views identified in this research is consistent across other police technology projects nationwide.
- Explore the relative merits and shortcomings of using behavioural analytics systems for ‘risk identification’ purposes, rather than to assess individuals already subject to offender management orders.
- Examine the long-term impact of data-driven risk scoring on officer decision-making, for instance to understand over time the factors that lead to increased (or reduced) trust and confidence in the system.
- Assess whether there is real-world operational value to producing predictive forecasts indicating escalation towards *future* offending, or whether offender managers benefit more from *harm scores* indicating the level of harm associated with *current* offending behaviour.

- Establish acceptable classification thresholds for offender risk scoring algorithms, and specifically how to strike an appropriate balance between false positive and false negatives.
- Identify decision-making contexts in which the use of data-driven risk scoring may be unacceptable, for instance of the decisions being made have a direct and significant impact on individual rights.
- Critically assess different approaches to ethical review and oversight of police technology projects, to identify best practice and establish a set of guidelines for robust external scrutiny of new data-driven policing projects.

It is hoped that the findings presented here will offer a valuable reference point for police forces seeking to deploy behavioural analytics and data-driven risk scoring systems operationally, to learn from the lessons (both positive and negative) from previous projects and appropriately direct future resource allocation.

References

Ægisdóttir, S. *et al.* (2006) 'The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction', *The Counseling Psychologist*, 34(3), pp. 341–382. Available at: https://doi.org/10.1177/0011000005285875.

Aktinson, P. and Hammersley, M. (1998) 'Ethnography and participant observation', *Strategies of Qualitative Inquiry. Thousand Oaks: Sage*, pp. 248–261.

Amnesty International (2018) *Trapped in the Matrix: Secrecy, stigma, and bias in the Met's Gangs Database*.

Andrews, D.A. and Bonta, J. (2010) *The psychology of criminal conduct*. Routledge.

Andrews, D.A., Bonta, J. and Hoge, R.D. (1990) 'Classification for effective rehabilitation: Rediscovering psychology', *Criminal justice and Behavior*, 17(1), pp. 19–52.

Andrews, D.A., Bonta, J. and Wormith, J.S. (2006) 'The Recent Past and Near Future of Risk and/or Need Assessment', *Crime & Delinquency*, 52(1), pp. 7–27. Available at: https://doi.org/10.1177/0011128705281756.

Andrews, D.A. (Don) (2012) 'The risk-need-responsivity (RNR) model of correctional assessment and treatment', in *Using social science to reduce violent offending*. New York, NY, US: Oxford University Press (American psychology–law society series), pp. 127–156.

Andrews, D.A. and Dowden, C. (2006) 'Risk principle of case classification in correctional treatment: A meta-analytic investigation', *International journal of offender therapy and comparative criminology*, 50(1), pp. 88–100.

Arrieta, A.B. *et al.* (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information fusion*, 58, pp. 82–115.

Azodi, C.B., Tang, J. and Shiu, S.-H. (2020) 'Opening the black box: interpretable machine learning for geneticists', *Trends in genetics*, 36(6), pp. 442–455.

Babuta, A. (2017) 'Big Data and Policing', *RUSI Occasional Papers* [Preprint].

Babuta, A. and Oswald, M. (2020) 'Data Analytics and Algorithms in Policing in England and Wales', *RUSI Occasional Papers* [Preprint].

Babuta, A. and Oswald, M. (2021) 'Machine learning predictive algorithms and the policing of future crimes', in *Policing and Artificial Intelligence (Dr John LM McDaniel and Prof Ken Pease OBE, eds.)*. Routledge.

Babuta, A., Oswald, M. and Janjeva, A. (2020) 'Artificial intelligence and UK national security: policy considerations'.

Babuta, A., Oswald, M. and Rinik, C. (2018) 'Machine Learning Algorithms and Police Decision-Making', *RUSI Whitehall Report* [Preprint].

Bachner, J. (2013) *Predictive policing: preventing crime with data and analytics*. IBM Center for the Business of Government Washington, DC.

Bayamlıoğlu, E. and Leenes, R. (2018) 'The 'rule of law'implications of data-driven decision-making: a techno-regulatory perspective', *Law, Innovation and Technology*, 10(2), pp. 295–313.

Beech, A.R. and Ward, T. (2004) 'The integration of etiology and risk in sexual offenders: A theoretical framework', *Aggression and violent behavior*, 10(1), pp. 31–63.

Bennett Moses, L. and Chan, J. (2018) 'Algorithmic prediction in policing: assumptions, evaluation, and accountability', *Policing and society*, 28(7), pp. 806–822.

Berk, R. *et al.* (2009) 'Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1), pp. 191–211.

Berk, R. (2012) *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Berk, R. and Hyatt, J. (2015) 'Machine learning forecasts of risk to inform sentencing decisions', *Federal Sentencing Reporter*, 27(4), pp. 222–228.

Berk, R.A. and Bleich, J. (2013) 'Statistical procedures for forecasting criminal behavior: A comparative assessment', *Criminology & Pub. Pol'y*, 12, p. 513.

Berk, R.A., Sorenson, S.B. and Barnes, G. (2016) 'Forecasting domestic violence: A machine learning approach to help inform arraignment decisions', *Journal of Empirical Legal Studies*, 13(1), pp. 94–115.

Bernard, H.R. (2017) *Research methods in anthropology: Qualitative and quantitative approaches*. Rowman & Littlefield.

Biernacki, P. and Waldorf, D. (1981) 'Snowball sampling: Problems and techniques of chain referral sampling', *Sociological methods & research*, 10(2), pp. 141–163.

Bonta, J. (2002) 'Offender risk assessment: Guidelines for selection and use', *Criminal justice and behavior*, 29(4), pp. 355–379.

Borum, R. (1996) 'Improving the clinical practice of violence risk assessment: Technology, guidelines, and training.', *American Psychologist*, 51(9), p. 945.

Borum, R. (2015) 'Assessing risk for terrorism involvement.', *Journal of Threat Assessment and Management*, 2(2), pp. 63–87. Available at: https://doi.org/10.1037/tam0000043.

Bowers, K.J., Johnson, S.D. and Pease, K. (2004) 'Prospective hot-spotting: the future of crime mapping?', *British journal of criminology*, 44(5), pp. 641–658.

Braga, A.A. and Bond, B.J. (2008) 'Policing crime and disorder hot spots: A randomized controlled trial', *Criminology*, 46(3), pp. 577–607.

Braiek, H.B. and Khomh, F. (2018) ‘On Testing Machine Learning Programs’. arXiv. Available at: http://arxiv.org/abs/1812.02257 (Accessed: 5 December 2022).

Brantingham, P.J., Valasik, M. and Mohler, G.O. (2018) ‘Does predictive policing lead to biased arrests? Results from a randomized controlled trial’, *Statistics and public policy*, 5(1), pp. 1–6.

Brayne, S. (2017) ‘Big data surveillance: The case of policing’, *American sociological review*, 82(5), pp. 977–1008.

Brayne, S. and Christin, A. (2021) ‘Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts’, *Social Problems*, 68(3), pp. 608–624. Available at: https://doi.org/10.1093/socpro/spaa004.

Breiman, L. (2001) ‘Statistical modeling: The two cultures (with comments and a rejoinder by the author)’, *Statistical science*, 16(3), pp. 199–231.

Bronitt, S.H. and Stenning, P. (2011) ‘Understanding discretion in modern policing’, *Criminal law journal*, 35(6), pp. 319–332.

Brown, M. (2000) ‘Calculations of risk in contemporary penal practice’, *Dangerous offenders: Punishment and social order*, pp. 93–108.

Brown, S.C. *et al.* (2002) ‘Exploring complex phenomena: Grounded theory in student affairs research’, *Journal of college student development*, 43(2), pp. 173–183.

Bryman, A. (2016) *Social research methods*. Oxford university press.

Burnett, R. and Appleton, C. (2004) *Joined-up youth justice: Tackling youth crime in partnership*. Russell House.

Busuioc, M. (2021) ‘Accountable artificial intelligence: Holding algorithms to account’, *Public Administration Review*, 81(5), pp. 825–836.

Cameron, W.B. (1963) *Informal sociology: A casual introduction to sociological thinking*. Random house.

Campbell, M.A., French, S. and Gendreau, P. (2009) ‘The Prediction of Violence in Adult Offenders: A Meta-Analytic Comparison of Instruments and Methods of Assessment’, *Criminal Justice and Behavior*, 36(6), pp. 567–590. Available at: https://doi.org/10.1177/0093854809333610.

Cassell, C. (2005) ‘Creating the interviewer: identity work in the management research process’, *Qualitative research*, 5(2), pp. 167–179.

Charmaz, K. and Belgrave, L.L. (2007) ‘Grounded theory’, *The Blackwell encyclopedia of sociology* [Preprint].

Chen, T. *et al.* (2015) ‘Xgboost: extreme gradient boosting’, *R package version 0.4-2*, 1(4), pp. 1–4.

Chen, Y. *et al.* (2018) ‘Bayesian optimization in alphago’, *arXiv preprint arXiv:1812.06855* [Preprint].

Clarke, R.V. (1980) ‘Situational crime prevention: Theory and practice’, *Brit. J. Criminology*, 20, p. 136.

Clarke, R.V. (1983) ‘Situational crime prevention: Its theoretical basis and practical scope’, *Crime and justice*, 4, pp. 225–256.

Clarke, R.V. (1995) ‘Situational crime prevention’, *Crime and justice*, 19, pp. 91–150.

Cohen, L., Manion, L. and Morrison, K. (2017) *Research methods in education*. routledge.

College of Policing (2013) *The Effects of Hot-Spot Policing on Crime: What Works Briefing*.

College of Policing (2014a) ‘Authorised Professional Practice, “Risk”’.

College of Policing (2014b) ‘Authorised Professional Practice: Understanding risk and vulnerability in the context of domestic abuse’.

College of Policing (2018) *The Policing Evaluation Toolkit*.

Cook, N.R. and Paynter, N.P. (2011) ‘Performance of reclassification statistics in comparing risk prediction models’, *Biometrical Journal*, 53(2), pp. 237–258.

Cooke, D.J. (2010) ‘More prejudicial than probative’, *The Journal of the Law Society of Scotland*, 55(1), pp. 20–23.

Cooke, D.J. (2012) ‘Violence risk assessment’, *Antisocial Behavior and Crime: Contributions of Developmental and Evaluation Research to Prevention and Intervention*, p. 221.

Cooke, D.J. and Michie, C. (2012) ‘Violence risk assessment: from prediction to understanding–or from what? To why?’, in *Managing Clinical Risk*. Willan, pp. 22–44.

Cooke, D.J. and Michie, C. (2014) ‘The generalizability of the Risk Matrix 2000: On model shrinkage and the misinterpretation of the area under the curve.’, *Journal of Threat Assessment and Management*, 1(1), pp. 42–55. Available at: https://doi.org/10.1037/tam0000004.

Copas, J. and Marshall, P. (1998) ‘The offender group reconviction scale: a statistical reconviction score for use by probation officers’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(1), pp. 159–171.

Corbin, J. and Strauss, A. (2014) *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

Corbin, J.M. and Strauss, A. (1990) ‘Grounded theory research: Procedures, canons, and evaluative criteria’, *Qualitative sociology*, 13(1), pp. 3–21.

Couchman, H. (2019) ‘Policing by machine: predictive policing and the threat to our rights’, *Liberty, January*, 15.

Craig, L.A. and Beech, A. (2009) ‘Best practice in conducting actuarial risk assessments with adult sexual offenders’, *Journal of sexual aggression*, 15(2), pp. 193–211.

Crawford, A. and Evans, K. (2017) ‘Crime prevention and community safety’.

Creswell, J.W. (2002) *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall Upper Saddle River, NJ.

Crotty, M. (1998) *The foundations of social research: Meaning and perspective in the research process*. Sage.

Cullen, F.T. (2005) ‘The twelve people who saved rehabilitation: How the science of criminology made a difference: The American Society of Criminology 2004 Presidential Address’, *Criminology*, 43(1), pp. 1–42.

Daffern, M. (2007) ‘The predictive validity and practical utility of structured schemes used to assess risk for aggression in psychiatric inpatient settings’, *Aggression and Violent Behavior*, 12(1), pp. 116–130. Available at: https://doi.org/10.1016/j.avb.2006.03.005.

Dale, R. (2021) ‘GPT-3: What’s it good for?’, *Natural Language Engineering*, 27(1), pp. 113–118.

Dawes, R.M., Faust, D. and Meehl, P.E. (1989) ‘Clinical versus actuarial judgment’, *Science*, 243(4899), pp. 1668–1674.

Debidin, M. (2009) *A compendium of research and analysis on the Offender Assessment System (OASys) 2006-2009*. Ministry of Justice.

Dencik, L. *et al.* (2018a) ‘Data Scores as Governance: Investigating uses of citizen scoring in public services’, p. 144.

Dencik, L. *et al.* (2018b) ‘Data Scores as Governance: Investigating uses of citizen scoring in public services’, p. 144.

Denney, D. (2005) *Risk and society*. Sage.

Denzin, N.K. (2001) ‘The reflexive interview and a performative social science’, *Qualitative research*, 1(1), pp. 23–46.

Douglas, K.S., Blanchard, A.J. and Hendry, M.C. (2012) ‘Violence risk assessment and management: Putting structured professional judgment into practice’, in *Managing Clinical Risk*. Willan, pp. 46–72.

Douglas, K.S. and Kropp, P.R. (2002) ‘A Prevention-Based Paradigm for Violence Risk Assessment: Clinical and Research Applications’, *Criminal Justice and Behavior*, 29(5), pp. 617–658. Available at: https://doi.org/10.1177/009385402236735.

Douglas, K.S. and Webster, C.D. (1999) 'Predicting violence in mentally and personality disordered individuals', in *Psychology and law*. Springer, pp. 175–239.

Douglas, K.S., Yeomans, M. and Boer, D.P. (2005) 'Comparative validity analysis of multiple measures of violence risk in a sample of criminal offenders', *Criminal Justice and Behavior*, 32(5), pp. 479–510.

Douglas, T. *et al.* (2017) 'Risk assessment tools in criminal justice and forensic psychiatry: The need for better data', *European Psychiatry*, 42, pp. 134–137. Available at: https://doi.org/10.1016/j.eurpsy.2016.12.009.

Education Endowment Foundation (2015) 'EEF evaluation: A cumulative approach', *London: EEF* [Preprint].

Ellingworth, D., Farrell, G. and Pease, K. (1995) 'A victim is a victim is a victim-chronic victimization in four sweeps of the British Crime Survey', *Brit. J. Criminology*, 35, p. 360.

Essex Police (2021) *Knife Crime and Violence Model – Fearless Futures | Essex Police*. Available at: https://www.essex.police.uk/police-forces/essex-police/areas/essex-police/au/about-us/privacy-notices/knife-crime-and-violence-model--fearless-futures/ (Accessed: 10 May 2022).

Etikan, I., Musa, S.A. and Alkassim, R.S. (2016) 'Comparison of convenience sampling and purposive sampling', *American journal of theoretical and applied statistics*, 5(1), pp. 1–4.

Farrall, S. (2004) 'Social capital and offender reintegration: Making probation desistance focused', *After crime and punishment: Pathways to offender reintegration*, pp. 57–82.

Farrall, S., Mawby, R. and Worrall, A. (2007) 'Prolific/persistent offenders and desistance', *Handbook of probation*, pp. 352–380.

Farrell, G. (2015) 'Crime concentration theory', *Crime prevention and community Safety*, 17, pp. 233–248.

Farrell, G. and Pease, K. (2001) *Repeat victimization*. Criminal Justice Press.

Farrington, D.P., Jolliffe, D. and Johnstone, L. (2008a) 'Assessing violence risk: A framework for practice', *Institute of Criminology, Cambridge University, Cambridge* [Preprint].

Farrington, D.P., Jolliffe, D. and Johnstone, L. (2008b) 'Assessing violence risk: A framework for practice', *Institute of Criminology, Cambridge University, Cambridge* [Preprint].

Fazel, S. *et al.* (2012) 'Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis', *BMJ*, 345. Available at: https://doi.org/10.1136/bmj.e4692.

Ferguson, A.G. (2017) *The rise of big data policing*. New York University Press.

Fossey, E. *et al.* (2002) 'Understanding and evaluating qualitative research', *Australian & New Zealand Journal of Psychiatry*, 36(6), pp. 717–732.

Fox, C. *et al.* (2018) ‘Piloting different approaches to personalised offender management in the English criminal justice system’, *International Review of Sociology*, 28(1), pp. 35–61. Available at: https://doi.org/10.1080/03906701.2017.1422886.

Fujii, L.A. (2012) ‘Research ethics 101: Dilemmas and responsibilities’, *PS: Political Science & Politics*, 45(4), pp. 717–723.

Garside, R. (2004) *Crime, persistent offenders and the justice gap*. Crime and Society Foundation London.

Gendreau, P. and Andrews, D.A. (1990) ‘Tertiary prevention: What the meta-analyses of the offender treatment literature tell us about what works’, *Canadian J. Criminology*, 32, p. 173.

Gendreau, P., Little, T. and Goggin, C. (1996) ‘A meta-analysis of the predictors of adult offender recidivism: What works!’, *Criminology*, 34(4), pp. 575–608.

Gioia, D.A., Corley, K.G. and Hamilton, A.L. (2013) ‘Seeking qualitative rigor in inductive research: Notes on the Gioia methodology’, *Organizational research methods*, 16(1), pp. 15–31.

Given, L.M. (2008) *The SAGE Encyclopedia of Qualitative Research Methods*. SAGE Publications.

Glaser, B.G. and Strauss, A.L. (2017) *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Gleaves, L.P., Schwartz, R. and Broniatowski, D.A. (2020) ‘The Role of Individual User Differences in Interpretable and Explainable Machine Learning Systems’. arXiv. Available at: https://doi.org/10.48550/arXiv.2009.06675.

Grace, J. (2019) ‘’Algorithmic impropriety’in UK policing?’, *Journal of Information Rights, Policy and Practice* [Preprint].

Grove, W.M. *et al.* (2000a) ‘Clinical versus mechanical prediction: A meta-analysis.’, *Psychological Assessment*, 12(1), pp. 19–30. Available at: https://doi.org/10.1037/1040-3590.12.1.19.

Grove, W.M. *et al.* (2000b) ‘Clinical versus mechanical prediction: A meta-analysis.’, *Psychological Assessment*, 12(1), pp. 19–30. Available at: https://doi.org/10.1037/1040-3590.12.1.19.

Guba, E.G. and Lincoln, Y.S. (1994) ‘Competing paradigms in qualitative research’, *Handbook of qualitative research*, 2(163–194), p. 105.

Guerette, R.T. and Bowers, K.J. (2009) ‘Assessing the extent of crime displacement and diffusion of benefits: A review of situational crime prevention evaluations’, *Criminology*, 47(4), pp. 1331–1368.

Gunning, D. *et al.* (2019) ‘XAI—Explainable artificial intelligence’, *Science robotics*, 4(37), p. eaay7120.

Gunning, D. and Aha, D. (2019) 'DARPA's Explainable Artificial Intelligence (XAI) Program', *AI Magazine*, 40(2), pp. 44–58. Available at: https://doi.org/10.1609/aimag.v40i2.2850.

Hannah-Moffat, K. (2005) 'Criminogenic needs and the transformative risk subject: Hybridizations of risk/need in penality', *Punishment & society*, 7(1), pp. 29–51.

Hanson, R.K. (1998) 'What do we know about sex offender risk assessment?', *Psychology, Public Policy, and Law*, 4(1–2), p. 50.

Hanson, R.K. (2001) *Age and sexual recidivism: A comparison of rapists and child molesters*. Solicitor General Canada Ottawa, ON, Canada.

Hanson, R.K. and Morton-Bourgon, K.E. (2005) 'The characteristics of persistent sexual offenders: a meta-analysis of recidivism studies.', *Journal of consulting and clinical psychology*, 73(6), p. 1154.

Harris, G.T., Rice, M.E. and Quinsey, V.L. (1993) 'Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument', *Crim. Just. & Behavior*, 20, p. 315.

Hart, S.D. (1998) 'The role of psychopathy in assessing risk for violence: Conceptual and methodological issues', *Legal and criminological psychology*, 3(1), pp. 121–137.

Hart, S.D. and Cooke, D.J. (2013) 'Another Look at the (Im-)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments: Imprecision of individual risk estimates', *Behavioral Sciences & the Law*, 31(1), pp. 81–102. Available at: https://doi.org/10.1002/bsl.2049.

Hart, S.D. and Logan, C. (2011) 'Formulation of Violence Risk Using Evidence-Based Assessments: The Structured Professional Judgment Approach', in P. Sturmey and M. McMurran (eds) *Forensic Case Formulation*. Chichester, UK: John Wiley & Sons, Ltd, pp. 81–106. Available at: https://doi.org/10.1002/9781119977018.ch4.

Hart, S.D., Michie, C. and Cooke, D.J. (2007) 'Precision of actuarial risk assessment instruments: Evaluating the "margins of error" of group *v.* individual predictions of violence', *British Journal of Psychiatry*, 190(S49), pp. s60–s65. Available at: https://doi.org/10.1192/bjp.190.5.s60.

Hays, P.A. (2004) 'Case study research', *Foundations for research: Methods of inquiry in education and the social sciences*, pp. 217–234.

Helmus, L. and Bourgon, G. (2011) 'Taking Stock of 15 Years of Research on the Spousal Assault Risk Assessment Guide (SARA): A Critical Review', *International Journal of Forensic Mental Health*, 10(1), pp. 64–75. Available at: https://doi.org/10.1080/14999013.2010.551709.

Her Majesty's Inspectorate of Constabulary and Fire & Rescue Services (HMICFRS) (2017) *PEEL: Police Effectiveness 2016 – A National Overview*.

HM Prison & Probation Service (2019) 'MAPPA Guidance 2012, Version 4.5 [Updated Jul 2019]'.

Hoge, R.D. and Andrews, D.A. (1996) *Assessing the youthful offender: Issues and techniques*. Springer Science & Business Media.

Holzinger, A. *et al.* (2022) 'Explainable AI methods-a brief overview', in *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Springer, pp. 13–38.

Howard, P. *et al.* (2009) *OGRS 3: The revised offender group reconviction scale*. Ministry of Justice.

Howard, P., Clark, D. and Garnham, N. (2003) 'Evaluation and validation of the Offender Assessment System (OASys)', *OASys Central Research Unit. Report to HM Prison Service and National Probation Service* [Preprint].

Hu, G. *et al.* (2015) 'When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition', in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 142–150.

Jansen, F. (2018) 'Data Driven Policing in the Context of Europe', *Data Justice Lab* [Preprint].

Johnson, M. (1999) 'Observations on positivism and pseudoscience in qualitative nursing research', *Journal of Advanced Nursing*, 30(1), pp. 67–73.

Johnson, S.D. *et al.* (2007) 'Prospective crime mapping in operational context: Final report', *UCL, Jill Dando Institute of Crime Science: London, UK* [Preprint].

Johnson, S.D., Tilley, N. and Bowers, K.J. (2015) 'Introducing EMMIE: An evidence rating scale to encourage mixed-method crime prevention synthesis reviews', *Journal of experimental criminology*, 11, pp. 459–473.

Kahneman, D. and Klein, G. (2009) 'Conditions for intuitive expertise: A failure to disagree.', *American Psychologist*, 64(6), pp. 515–526. Available at: https://doi.org/10.1037/a0016755.

Kaiser, K. (2009) 'Protecting respondent confidentiality in qualitative research', *Qualitative health research*, 19(11), pp. 1632–1641.

Keenan, B. (2021) 'Automatic facial recognition and the intensification of police surveillance', *The Modern Law Review*, 84(4), pp. 886–897.

Kewley, S. and Blandford, M. (2017) 'The development of the active risk management system', *Journal of Criminal Psychology*, 7(3), pp. 155–167. Available at: https://doi.org/10.1108/JCP-10-2016-0034.

Khiroya, R., Weaver, T. and Maden, T. (2009) 'Use and perceived utility of structured violence risk assessments in English medium secure forensic units', *Psychiatric Bulletin*, 33(4), pp. 129–132.

Knack, A., Carter, R.J. and Babuta, A. (2022) *Human-Machine Teaming in Intelligence Analysis: Requirements for developing trust in machine learning systems*.

Kraemer, H.C. *et al.* (1997) ‘Coming to terms with the terms of risk’, *Archives of general psychiatry*, 54(4), pp. 337–343.

Kropp, P.R. and Gibas, A. (2010) ‘The Spousal Assault Risk Assessment Guide (SARA)’, *Handbook of violence risk assessment*, pp. 227–250.

Kropp, P.R. and Hart, S.D. (2000) ‘The Spousal Assault Risk Assessment (SARA) guide: Reliability and validity in adult male offenders’, *Law and human behavior*, 24(1), pp. 101–118.

Kuzel, A.J. (1992) ‘Sampling in qualitative inquiry.’

Lidz, C.W., Mulvey, E.P. and Gardner, W. (1993) ‘The accuracy of predictions of violence to others’, *Jama*, 269(8), pp. 1007–1011.

Lin, A.C. (2002) *Reform in the making: The implementation of social policy in prison*. Princeton University Press.

Logan, C. (2017) ‘Reporting structured professional judgement’, in *The forensic psychologist’s report writing guide*. Routledge, pp. 82–93.

Logan, C. and Johnstone, L. (2012) *Managing clinical risk: A guide to effective practice*. Routledge.

Logan, C. and Lloyd, M. (2019a) ‘Violent extremism: A comparison of approaches to assessing and managing risk’, *Legal and criminological psychology*, 24(1), pp. 141–161.

Logan, C. and Lloyd, M. (2019b) ‘Violent extremism: A comparison of approaches to assessing and managing risk’, *Legal and Criminological Psychology*, 24(1), pp. 141–161. Available at: https://doi.org/10.1111/lcrp.12140.

Longhurst, R. (2003) ‘Semi-structured interviews and focus groups’, *Key methods in geography*, 3(2), pp. 143–156.

Ludwig, J. and Kling, J.R. (2007) ‘Is crime contagious?’, *The Journal of Law and Economics*, 50(3), pp. 491–518.

Lynskey, O. (2019) ‘Criminal justice profiling and EU data protection law: precarious protection from predictive policing’, *International Journal of Law in Context*, 15(2), pp. 162–176.

Mair, G. (2013) *What matters in probation*. Routledge.

Maruna, S. (2001) *Making good*. Washington, DC: American Psychological Association.

Maruna, S. *et al.* (2004) ‘Pygmalion in the reintegration process: Desistance from crime through the looking glass’, *Psychology, Crime & Law*, 10(3), pp. 271–281. Available at: https://doi.org/10.1080/10683160410001662762.

Maxwell, J. (1992) ‘Understanding and validity in qualitative research’, *Harvard educational review*, 62(3), pp. 279–301.

McEwan, T.E., Bateson, S. and Strand, S. (2017) ‘Improving police risk assessment and management of family violence through a collaboration between law enforcement, forensic mental health and academia’, *Journal of criminological research, policy and practice* [Preprint].

McGovern, A. *et al.* (2019) ‘Making the black box more transparent: Understanding the physical implications of machine learning’, *Bulletin of the American Meteorological Society*, 100(11), pp. 2175–2199.

McGuire, J. (2000) ‘Explanations of criminal behaviour’, *Behavior, crime and legal processes. A guide for forensic practitioners*, pp. 135–159.

Meehl, P.E. (1954) ‘Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.’

Merriam, S.B. (2002) ‘Introduction to qualitative research’, *Qualitative research in practice: Examples for discussion and analysis*, 1(1), pp. 1–17.

Mohler, G.O. *et al.* (2011) ‘Self-exciting point process modeling of crime’, *Journal of the American Statistical Association*, 106(493), pp. 100–108.

Mohler, G.O. *et al.* (2015) ‘Randomized controlled field trials of predictive policing’, *Journal of the American statistical association*, 110(512), pp. 1399–1411.

Moore, R. (2015) ‘A compendium of research and analysis on the Offender Assessment System (OASys), 2009–2013’, *Ministry of Justice Analytical Series*, p. 367.

National Police Chiefs’ Council (2016) *Policing Vision 2025*.

National Police Chiefs’ Council (2020) *National Policing Digital Strategy: Digital, Data and Technology Strategy 2020–2030*.

Newcomer, K.E., Hatry, H.P. and Wholey, J.S. (2015) ‘Conducting semi-structured interviews’, *Handbook of practical program evaluation*, 492.

Nicholls, T.L., Ogloff, J.R.P. and Douglas, K.S. (2004) ‘Assessing risk for violence among male and female civil psychiatric patients: the HCR-20, PCL:SV, and VSC’, *Behavioral Sciences & the Law*, 22(1), pp. 127–158. Available at: https://doi.org/10.1002/bsl.579.

Noble, H. and Smith, J. (2015) ‘Issues of validity and reliability in qualitative research’, *Evidence-based nursing*, 18(2), pp. 34–35.

Noy, C. (2008) ‘Sampling knowledge: The hermeneutics of snowball sampling in qualitative research’, *International Journal of social research methodology*, 11(4), pp. 327–344.

Nunkoosing, K. (2005) ‘The problems with interviews’, *Qualitative health research*, 15(5), pp. 698–706.

Onwuegbuzie, A.J. and Collins, K.M. (2007) ‘A typology of mixed methods sampling designs in social science research.’, *Qualitative Report*, 12(2), pp. 281–316.

Orb, A., Eisenhauer, L. and Wynaden, D. (2001) ‘Ethics in qualitative research’, *Journal of nursing scholarship*, 33(1), pp. 93–96.

Oswald, M. (2018) ‘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: Mathematical, Physical and Engineering Sciences*, 376(2128), p. 20170359.

Palys, T. (2008) ‘Purposive Sampling’, in *The Sage Encyclopedia of Qualitative Research Methods*. Thousand Oaks, CA: Sage.

Patton, M.Q. (1990) *Qualitative evaluation and research methods*. SAGE Publications, inc.

Pearsall, B. (2010) ‘Predictive policing: The future of law enforcement’, *National Institute of Justice Journal*, 266(1), pp. 16–19.

Perry, W.L. (2013) *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Pimple, K.D. (2002) ‘Six domains of research ethics’, *Science and engineering ethics*, 8(2), pp. 191–205.

Polaschek, D.L.L. (2012) ‘An appraisal of the risk–need–responsivity (RNR) model of offender rehabilitation and its application in correctional treatment’, *Legal and Criminological Psychology*, 17(1), pp. 1–17. Available at: https://doi.org/10.1111/j.2044-8333.2011.02038.x.

Polaschek, D.L.L. and Collie, R.M. (2004) ‘REHABILITATING SERIOUS VIOLENT ADULT OFFENDERS: AN EMPIRICAL AND THEORETICAL STOCKTAKE’, *Psychology, Crime and Law*, 10(3), pp. 321–334. Available at: https://doi.org/10.1080/0683160410001662807.

Popay, J., Rogers, A. and Williams, G. (1998) ‘Rationale and standards for the systematic review of qualitative literature in health services research’, *Qualitative health research*, 8(3), pp. 341–351.

Quinsey, V.L. *et al.* (2006) *Violent offenders: Appraising and managing risk*. American Psychological Association.

Quinsey, V.L. and Ambtman, R. (1979) ‘Variables affecting psychiatrists’ and teachers’ assessments of the dangerousness of mentally ill offenders.’, *Journal of Consulting and Clinical Psychology*, 47(2), p. 353.

Rice, M.E. and Harris, G.T. (2005) ‘Comparing effect sizes in follow-up studies: ROC Area, Cohen’s d, and r’, *Law and human behavior*, 29(5), pp. 615–620.

Rice, P.L. and Ezzy, D. (1999) ‘Qualitative research methods: A health focus’, *Melbourne, Australia* [Preprint].

Richardson, R., Schultz, J.M. and Crawford, K. (2019) ‘Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice’, *NYUL Rev. Online*, 94, p. 15.

Robinson, G. (1999) ‘Risk management and rehabilitation in the probation service: Collision and collusion’, *The Howard Journal of Criminal Justice*, 38(4), pp. 421–433.

Robinson, O.C. (2014) ‘Sampling in interview-based qualitative research: A theoretical and practical guide’, *Qualitative research in psychology*, 11(1), pp. 25–41.

Rokach, L. (2010) ‘Ensemble Methods in Supervised Learning’, in O. Maimon and L. Rokach (eds) *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, pp. 959–979. Available at: https://doi.org/10.1007/978-0-387-09823-4_50.

Rudin, C. (2019) ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, *Nature Machine Intelligence*, 1(5), pp. 206–215.

Russell, S. and Norvig, P. (2002) ‘Artificial intelligence: a modern approach’.

Saldaña, J. (2014) ‘Coding and analysis strategies’, in *The Oxford handbook of qualitative research*.

Scotland, J. (2012) ‘Exploring the Philosophical Underpinnings of Research: Relating Ontology and Epistemology to the Methodology and Methods of the Scientific, Interpretive, and Critical Research Paradigms’, *English Language Teaching*, 5(9), p. p9. Available at: https://doi.org/10.5539/elt.v5n9p9.

Scottish Risk Management Authority (2011) ‘Framework for Risk Assessment, Management and Evaluation: FRAME’. Risk Management Authority.

Seidman, I. (2006) *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. Teachers college press.

Selbst, A.D. (2017) ‘Disparate impact in big data policing’, *Ga. L. Rev.*, 52, p. 109.

Sherman, L., Neyroud, P.W. and Neyroud, E. (2016) ‘The Cambridge crime harm index: Measuring total harm from crime based on sentencing guidelines’, *Policing: a journal of policy and practice*, 10(3), pp. 171–183.

Siontis, G.C. *et al.* (2015) ‘External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination’, *Journal of clinical epidemiology*, 68(1), pp. 25–34.

Skeem, J.L. and Monahan, J. (2011) ‘Current Directions in Violence Risk Assessment’, *Current Directions in Psychological Science*, 20(1), pp. 38–42. Available at: https://doi.org/10.1177/0963721410397271.

Smith, D. (2003) ‘Five principles for research ethics’, *Monitor on psychology*, 34(1), p. 56.

Sparks, R. (2001) ‘Degrees of estrangement: The cultural theory of risk and comparative penology’, *Theoretical Criminology*, 5(2), pp. 159–176.

Sutherland, A.A. *et al.* (2012) ‘Sexual violence risk assessment: An investigation of the interrater reliability of professional judgments made using the Risk for Sexual Violence Protocol’, *International Journal of Forensic Mental Health*, 11(2), pp. 119–133.

Tesch, R. (2013) *Qualitative research: Analysis types and software*. Routledge.

Thomas, D.R. (2003) ‘A general inductive approach for qualitative data analysis’.

Thornton, D. *et al.* (2003) ‘Distinguishing and combining risks for sexual and violent recidivism’, *Annals of the New York academy of sciences*, 989(1), pp. 225–235.

Townsley, M., Homel, R. and Chaseling, J. (2003) ‘Infectious burglaries. A test of the near repeat hypothesis’, *British Journal of Criminology*, 43(3), pp. 615–633.

Tversky, A. and Kahneman, D. (1974) ‘Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.’, *science*, 185(4157), pp. 1124–1131.

Urwin, S. (2016) ‘Algorithmic forecasting of offender dangerousness for police custody officers: An assessment of accuracy for the Durham Constabulary model’, unpublished thesis, University of Cambridge. Available at: https:// www.crim.cam.ac.uk/global/docs/theses/sheena-urwin-thesis-12-12-2016.pdf’, p. 136.

Veale, M., Van Kleek, M. and Binns, R. (2018) ‘Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making’, in *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14.

Viljoen, J.L., Cochrane, D.M. and Jonnson, M.R. (2018) ‘Do risk assessment tools help manage and reduce risk of violence and reoffending? A systematic review.’, *Law and Human Behavior*, 42(3), pp. 181–214. Available at: https://doi.org/10.1037/lhb0000280.

Visher, C.A. (2006) ‘Effective reentry programs’, *Criminology & Public Policy*, 5(2), pp. 299–302.

de Vries Robbé, M., de Vogel, V. and de Spa, E. (2011) ‘Protective factors for violence risk in forensic psychiatric patients: A retrospective validation study of the SAPROF’, *International journal of forensic mental health*, 10(3), pp. 178–186.

Ward, T. and Maruna, S. (2007) *Rehabilitation*. Routledge.

Watson, D.S. *et al.* (2019) ‘Clinical applications of machine learning algorithms: beyond the black box’, *Bmj*, 364.

Webster, C.D., Haque, Q. and Hucker, S.J. (2013) *Violence risk-assessment and management: Advances through structured professional judgement and sequential redirections*. John Wiley & Sons.

West Midlands Police (2019a) *Minutes, Ethics Committee Meeting April 2019.* Available at: https://www.westmidlands-pcc.gov.uk/archive/april-2019/.

West Midlands Police (2019b) *Minutes, Ethics Committee Meeting July 2019.* Available at: https://www.westmidlands-pcc.gov.uk/archive/april-2019/.

West Midlands Police (2020) ‘National Data Analytics Solution: Submission to the WMP Ethics Committee July 2020’. Available at: file:///Users/Citizen/Downloads/07072020-EC-Agenda-Item-9-NDAS-Update.pdf.

Wilson, E. and Hinks, S. (2011) ‘Assessing the predictive validity of the Asset youth risk assessment tool using the Juvenile Cohort Study (JCS)’, *Ministry of Justice Research Series*, 10(11).

Wong, S.C. and Gordon, A. (2006) ‘The validity and reliability of the Violence Risk Scale: A treatment-friendly violence risk assessment tool.’, *Psychology, Public Policy, and Law*, 12(3), p. 279.

Worrall, A. and Mawby, R.C. (2004) ‘Intensive projects for prolific/persistent offenders’, *Alternatives to Prison*, p. 268.

Yin, R. (2009) ‘Yin, RK (2009). Case study research: Design and methods . Thousand Oaks, CA: Sage.’, *The Canadian Journal of Action Research*, 14(1), pp. 69–71.

Youth Justice Board (2000) *ASSET: Explanatory Notes*. London: Youth Justice Board.

Zhou, J., Chen, F. and Holzinger, A. (2022) ‘Towards explainability for AI fairness’, in *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Springer, pp. 375–386.

Annex 1: Research Materials

[Removed for public version]