Policing: A Journal of Policy and Practice, 2026, 20, paag028
https://doi.org/10.1093/police/paag028
Published: 8 June 2026
Original article

OXFORD

Risk assessment tools in policing contexts: 10 key ethical challenges

Hope Kent [1]*, Seena Fazel[2,3], Tom Sorell [4], Rohan Borschmann [2,3,5,6], Lucia Zedner [7,8], Melissa Hamilton [9,10], Lewis Prescott-Mayling[11,12], Tori Pamela Anne Olphin [13,14], Thomas Douglas [15], Lee Hogarth [1], Stan Gilmour [16,17], William Huw Williams[1], Vittoria Porta[18], Timothy Lowe [11,18], George Leckie [19], Mackenzie Graham [18], James Hart [18], Mark Sheehan [18,20]

[1]Department of Psychology, University of Exeter, Exeter, United Kingdom
[2]Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[3]Oxford Health NHS Foundation Trust, Oxford, United Kingdom
[4]University of Warwick, Warwick, United Kingdom
[5]School of Social Sciences, Nottingham Trent University, Nottingham, United Kingdom
[6]Centre for Adolescent Health, Murdoch Children's Research Institute and Royal Children's Hospital, Melbourne, Australia
[7]All Souls College and Faculty of Law, University of Oxford, Oxford, United Kingdom
[8]Faculty of law, University of New South Wales, Sydney, Australia
[9]University of Surrey, Guildford, United Kingdom
[10]Surrey Institute for People-Centred Artificial Intelligence, Guildford, United Kingdom
[11]Thames Valley Police, United Kingdom
[12]Jill Dando Institute of Security and Crime Science, University College London, London, United Kingdom
[13]ResearchCore, Greater Manchester, United Kingdom
[14]Institute of Criminology, University of Cambridge, Cambridge, United Kingdom
[15]Uehiro Oxford Institute, Faculty of Philosophy, and Jesus College, University of Oxford, Oxford, United Kingdom
[16]University of Exeter, Exeter, United Kingdom
[17]Keele University, Newcastle, United Kingdom
[18]Ethox Centre, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
[19]School of Education, University of Bristol, Bristol, United Kingdom
[20]Oxford NIHR Biomedical Research Centre, Oxford, United Kingdom

*Corresponding author. Department of Psychology, University of Exeter, Washington Singer Laboratories, Perry Road, Exeter, Devon, EX4 4PY, UK. E-mail: h.kent@exeter.ac.uk

Abstract

Risk assessment tools are increasingly used in policing to enhance decision-making accuracy and objectivity; yet their implementation has raised significant ethical concerns regarding issues of bias, transparency, and governance. This paper examines the ethical complexities of risk assessment tools through an analysis of four instruments: the harm assessment risk tool, previously developed and used by Durham Constabulary; the Active Risk Management System (ARMS), used across all police forces in England and Wales; the Offender Assessment System, used to profile risk of reoffending by probation services in the UK; and the Correctional Offender Management Profiling for Alternative Sanctions, a widely researched tool deployed in US corrections contexts whose ethical challenges are directly relevant to tools now entering policing practice. A thematic framework identifies 10 key challenges for the field, including disparities in accuracy metrics, fairness trade-offs, bias linked to demographics and social identity, and retraining. The paper contextualizes these issues within influential roles, including tool developers, decision-makers, and oversight committees. The credible risk of ethical harms arising from the use of risk assessment tools underscores the need for rigorous validation, transparency, and adaptive governance to minimize these risks. This paper arises from a meeting of an interdisciplinary working group convened at Ethox, University of Oxford, comprising academics in philosophy, law, psychology, psychiatry, and criminology, as well as police stakeholders.

Introduction

Risk assessment in policing contexts aims to predict the likelihood of a behaviour or event occurring, the frequency with which it may occur, whom it may affect, and the extent to which the behaviour or event will cause harm (College of Policing 2025). In the UK, these assessments are underpinned by the College of Policing's core risk principles, which emphasize proportionality, defensibility, and a requirement for decisions to be informed, recorded, and regularly reviewed (College of Policing 2013).

Decisions based on risk assessment outcomes have serious consequences both for the individual and wider society. For example, assessments of an individual as being high risk may lead to decisions about pre-trial detention (i.e. being held 'on remand') that have a significant impact on the individual, including on their ability to prepare their defence effectively. By extension, these decisions also impact on their family, work, and social networks. Conversely, decisions to release an individual pre-trial due to being assessed as low risk could lead to serious harm in the community if made incorrectly. By their nature, risk assessment tools often end up determining resource allocation; if an individual is predicted to be of high risk, more resource will be diverted than to a low risk individual (Berk et al. 2019). Importantly, risk assessments—which are oriented towards preventing harm—should be distinguished from 'needs' assessments—which aim to identify interventions, which an individual may benefit from (Berk et al. 2019). This paper is concerned with risk assessments, rather than needs assessments.

Unassisted human assessment of risk is affected by cognitive biases relating to several extraneous factors including the political ideology or partisanship of the assessor (Harris and Sen 2019). The expression of such biases may systematically vary by demographic characteristics of both the assessor and person being assessed, such as gender and ethnicity (Harris and Sen 2019; Slobogin 2021; Viljoen et al. 2025). Human decision making is also inconsistent—in a notable study by Danziger et al. (2011), judges made less punitive decisions, which were more favourable to the defendant when they had recently had a meal break. In recognition of the bias and inaccuracy inherent in human decision making, actuarial[1] risk assessment tools have become embedded into decision-making processes in recent decades, with the objective of improving accuracy and enhancing objectivity (Noti and Chen 2022). In addition, these tools could help to control prison numbers, by providing empirical evidence to underpin a decision to release an individual pre-trial, where other evidence might not suffice (Slobogin 2021).

However, the use of actuarial risk assessment tools is highly contentious. In 2016, a Pro-Publica article published by Angwin et al. (2016) gained significant traction across public and academic media in discussing the racial bias inherent in the Correctional Offender Management Profiling for Alternative Sanction (COMPAS) algorithm (discussed in detail below). In 2020, more than 2,000 scholars from a range of disciplines campaigned for Springer (a major publisher in healthcare and behavioural science) to publish a statement 'condemning the use of criminal justice statistics to predict criminality', labelling it a 'tech to prison pipeline' (Coalition for Critical Technology 2020). They also advocated that Springer 'acknowledge their role in incentivizing such harmful scholarship in the past' (Coalition for Critical Technology 2020).

However, risk assessment is a necessity of the criminal justice system. It is essential to consider the least harmful option—and the objections to the use of algorithms frequently do not consider that the alternative is unassisted human decision making, with its inherent, inscrutable bias and inaccuracy. If algorithms offer even a minor improvement in reliability, fairness, and accuracy, there are strong arguments that they are the lesser of two evils. Moreover, algorithmic risk assessment tools enable some transparency around performance metrics such as consistency and accuracy, which are considerably more difficult to assess in the case of unstructured human judgement. Evidence generally suggests that risk assessment tools are more accurate than unstructured human judgement—see Viljoen et al. (2025) for a meta-analysis of 31 studies directly comparing risk assessment tools to unstructured human judgements of risk. Studies that have argued otherwise have typically been methodologically weak. For example, Dressel and Farid (2018) found COMPAS to be no more accurate than lay-person assessment, but this study was strongly criticized for methodological problems and was deemed largely invalid as discussed by Holsinger et al. (2018) and Lin et al. (2020).

The ability of the police to prevent harm relies on accurate risk assessment. Therefore, we argue that using imperfect risk assessment tools is often better than using no tool. However, genuine ethical issues arise when risk assessment tools are developed and implemented rapidly with little consideration of the central concerns and minimal oversight. Some estimates indicate that there are more than 200 such tools available for use in criminal justice contexts globally, and this number is increasing all the time (Singh et al. 2014; Fazel and Wolf 2018; Slobogin 2021). Developers vary—some are criminal justice professionals, data scientists, or academics, with varying vested interests in the outcomes. There is mixed reporting of accuracy statistics (Ogonah et al. 2023), and the quality of tools varies significantly (Fazel and Wolf 2018; Fazel et al. 2022a). Where tools are developed rapidly and implemented without proper consideration for potential harms, their use is delegitimized, and harms have been shown to disproportionately impact minoritized groups (Fiesler 2023). There is an acknowledged lack of consensus as to best practice, and a lack of transparency about the use of these tools across the justice system (Law Society of England and Wales 2019). This makes it difficult for decision-makers in the police to decide which tools are appropriate, how they should be used, and which tools are most accurate and fair. Consequently, developers are often not required to fully explore the ethical impacts of each decision made in the design of the tool. Tools are often then implemented with little oversight or monitoring, and deployed in settings where there are

[1]By 'actuarial' risk assessment tools, we refer to instruments that estimate the likelihood of an outcome such as violence or reoffending by assigning numerical scores to risk factors statistically associated with that outcome (Fazel and Wolf 2018).

few, if any, staff with the technical expertise to monitor performance, identify when a model is no longer functioning as intended, or retrain it appropriately (Myhill et al. 2023).

In this context, this paper aims to examine the ethical issues that arise when considering the application of several risk assessment tools in order to highlight key considerations that decision-makers may evaluate when selecting the most appropriate tool, implementing it, and monitoring its long-term use in policing. It is our intention here to make the use of risk assessment tools more transparent and considered, rather than to put up barriers to their use. These issues were discussed by a collaborative working group comprising academics from disciplines spanning philosophy, law, psychology, psychiatry, and criminology, as well as police stakeholders.

Influential bodies

We consider here three key influential bodies in the implementation of risk assessment tools. These include:

(a) Tool developers: The individual or organization who develops and configures the tool. This might be a commercial developer, academic, independent data scientist, or data analytics team within a police force, for example.
(b) Decision-makers: The individual or organization who makes the decision to use a tool and decides which to use. This could be, for example, a chief constable, or a police and crime commissioner.
(c) Oversight committees*:* Any individuals or groups who might oversee the implementation of, monitoring of, and commissioning of tools. For example, Data Ethics Committees such as those described by Sorell (2024).

There may be some overlap in these three categories; however, we find this a helpful way to frame the influential bodies in the process.

Aims and scope of this paper

This paper aims to achieve the following:

1. Identify and explain key ethical concerns, to support influential bodies in making informed decisions about the design and use of risk assessment tools in policing contexts.
2. Provide decision-makers with examples of a balanced approach to weighing up the relevant ethical risks when making decisions about the design and use of risk assessment tools in policing contexts.
3. Indicate how decision-makers could be more transparent in their decision making, meaning that they are better equipped to justify their decisions about the use of these tools.

We are concerned here with predictive tools, which assess *future risk* (typically of reoffending, violence, or serious harm), rather than matching algorithms (such as those used in facial recognition). We have also focussed here on risk assessment tools that estimate individual risk, but would suggest that the ethical and methodological challenges explored are relevant to tools that predict risk spatially (such as Geolitica, formerly PredPol), which are used to allocate police resources across geographic areas to target 'hot spots'. We also do not address systems used to assess police officers themselves, such as early intervention systems designed to predict misconduct or flag officers at risk of complaints, although many issues raised here are likely to be relevant.

Method

This paper draws on an interdisciplinary expert working group convened at the Ethox Centre, University of Oxford. The shared motivation of the group was to move beyond polarized debates about the use of risk assessment tools in policing, and instead to identify and clarify the key ethical 'sticking points' that continue to hinder their responsible use. The aim was not to advocate for or against any particular tool, but to develop a practical framework to support more transparent, defensible, and ethically informed decision making around their design, deployment, and governance.

Development of the working group

Experts were initially identified through academic ethics groups at the University of Oxford, including Ethox and the Oxford Uehiro Centre for Practical Ethics. Additional experts were then purposively recruited from other departments and institutions through recommendations from early participants, to ensure breadth across relevant disciplines and applied expertise. The group included academics spanning philosophy, statistics, law, psychology, psychiatry, and criminology, alongside police stakeholders with operational and governance responsibilities for risk assessment and decision making. Thirteen members attended an in-person workshop in Oxford in November 2024, and a further five experts contributed asynchronously through individual discussions and written feedback on draft material.

Generation and synthesis of themes

The process was iterative. Early individual discussions with contributors informed an initial scoping set of candidate ethical issues and a provisional thematic structure (initially eight themes). These emerging themes were shared with the group at the workshop, where experts then engaged in a facilitated discussion to develop and refine them. Discussions included how to map each theme to its implications for key influential bodies (tool developers, decision-makers, and oversight committees). Particular attention was paid to points of disagreement, especially where these reflected underlying value judgements or trade-offs (e.g., between different conceptions of fairness, and the value of threshold scores to frontline police staff). The workshop was audio recorded, with participant agreement.

Following the workshop, the lead author reviewed the recording and integrated it with notes from earlier discussions and asynchronous feedback. The emergent themes were iteratively revised, and their scope and associated ethical risks were clarified. Revised drafts were circulated to contributors

for comments, with targeted follow-up to resolve ambiguities. The final set of 10 themes represents a consensus-seeking synthesis intended to be practically useful for those making and overseeing decisions about the use and implementation of risk assessment tools, rather than an exhaustive catalogue of all possible ethical concerns.

Case studies—risk assessment tools

The tools presented here are those in which the members of our working group have no vested interest, to enhance objectivity.

A significant challenge in selecting illustrative case studies was the limited availability of publicly accessible, independent evidence: many tools used in policing contexts are proprietary, and developer-produced validation reports (where they exist at all) are rarely publicly available. A further complication is that police forces may have legitimate operational reasons for limiting transparency around the specific criteria and weightings used in risk assessment tools. Public disclosure of the factors that trigger a high-risk classification could, in principle, allow individuals to conceal or manipulate those indicators. This creates a genuine tension between the transparency required for meaningful ethical oversight and the confidentiality that effective law enforcement may require.

We selected four case study tools. The Harm assessment risk tool (HART), Offender Assessment System (OASys), and Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) are actuarial risk assessment tools that estimate the likelihood of an outcome, such as violence or reoffending, by allocating numerical scores to risk factors associated with that outcome (Fazel and Wolf 2018). These factors may be static, such as sex at birth, or dynamic, such as current substance use. HART and COMPAS use machine learning methods, which allow them to handle complex, non-linear relationships in data but introduce opacity into their decision processes. OASys is informed by simpler regression models. The Active Risk Management System (ARMS), by contrast, is a dynamic structured professional judgement framework: it does not generate an algorithmic risk score but instead guides practitioners through the assessment of a defined set of risk and protective factors, producing a priority-for-work rating rather than a categorical prediction. However, ARMS integrates actuarial scores calculated from the Risk Matrix 2000 in policing contexts. Together, these tools illustrate the range of approaches currently in use and proposed for policing contexts, from opaque proprietary tools to practitioner-led structured assessments.

It is worth noting that COMPAS operates in US corrections settings rather than law enforcement, but its inclusion is justified on two grounds: the ethical challenges it illustrates are directly transferable to tools now being considered for use in policing; and the volume of independent research on COMPAS makes it uniquely useful as a case study in a field where independent public-facing evidence is scarce. It is reasonable to anticipate that, as more complex machine learning models become embedded across public services, policing will follow: the ethical considerations raised here therefore need to apply across structured professional judgement frameworks, regression models, and machine learning systems alike. This is particularly important given recent legislative proposals such as the House of Lords' Public Authority Algorithmic and Automated Decision-Making Systems Bill, which emphasises the need for human agency and oversight in automated decisions (UK Parliament 2024).

Finally, we also considered the inclusion of DASH (the Domestic Abuse, Stalking, and Honour-Based Violence Risk Identification Tool), as this is a tool used widely in UK police forces to assess risk in domestic abuse cases. DASH was not developed as a predictive tool: its stated purpose is to structure a conversation between officer and victim around known risk factors, to inform safeguarding and risk management rather than to generate a probability estimate (Richards 2009). Nevertheless, members of our working group raised concerns that in practice DASH are often applied as though it were predictive. Risk gradings derived from DASH directly determine access to safeguarding resources and referral to Multi-Agency Risk Assessment Conferences, which are the primary mechanism through which high-risk victims receive coordinated multi-agency support and intervention, with standard and medium risk classifications typically receiving little or no specialist follow-up (HMIC 2014). In this way, whatever its original design intent, DASH functions operationally as a predictive instrument. Evidence that has examined its predictive ability has found that officer risk predictions based on DASH perform poorly—in one study, not much better than chance (Turner et al. 2019). Where a risk grading determines who receives intervention specifically aimed at preventing future harm, DASH acts, in effect, predictively regardless of whether a numerical probability is attached to it, and the classification carries the same operational weight as a formal risk score. We decided not to include DASH as a primary case study because it was not designed to be used predictively, and predictive tools are the focus of this paper. However, the DASH literature points to a concern that instruments developed for one purpose (e.g. risk management) can come to function as predictive instruments in practice, without the scrutiny expected for a tool explicitly designed for that role. Where a risk classification shapes access to safeguarding and specialist support, this gap between design intent and operational use can be problematic, and the ethical obligations we identify here should apply wherever a tool functions predictively in practice, regardless of what it was originally designed for.

Harm assessment risk tool

HART is a predictive policing instrument developed by Durham Constabulary in collaboration with the University of Cambridge, which was used between 2018 and 2021. HART adopted machine learning algorithms (random forests) to analyse historical criminal data and offender characteristics and predict harm (Oswald et al. 2018; University of Cambridge 2018). Random forests aggregate the outputs of multiple decision trees, and are able to handle complex, non-linear relationships within data (Breiman 2001). HART was a proprietary tool, and to our knowledge the source code is not publicly available. It was deployed within Durham Constabulary's operational framework in the UK as part of a broader move

towards data-driven law enforcement practices. It defined individuals as being 'low', 'moderate', or 'high' risk, based on their past offences.

In 2021, Durham constabulary discontinued the use of HART, apparently 'due to the resources required to constantly refine and refresh the model to comply with appropriate ethical and legal oversight and governance' (Fair Trials 2022).

Active Risk Management System

Active Risk Management System (ARMS) is a dynamic risk assessment framework introduced in 2014 as the national standard for the assessment and community management of registered sex offenders in England and Wales. It is administered by specialist Management of Sexual and Violent Offenders (MOSOVO) officers and is used across all 43 police forces (Kewley and Blandford 2017; College of Policing 2025).

Unlike HART or COMPAS, ARMS does not *solely* produce an actuarial risk level. Instead, practitioners assess 11 factors—six risk indicators and five protective factors—and integrate these with a static actuarial score produced by an accompanying actuarial tool called the 'Risk Matrix 2000'. The risk indicators include sexual pre-occupation, offence-related sexual interest, hostile orientation, opportunity to offend, emotional congruence with children, and poor self-management. The protective factors include social investment, commitment to desist, intimate relationships, employment and positive routine, and social influences (Kewley and Blandford 2017; College of Policing 2025). ARMS guides frontline staff to assign a priority-for-work rating (high, medium, or low priority) that informs the frequency and nature of subsequent home visits, and the risk management plan.

A national evaluation of ARMS published in 2020 found that the tool's intended benefits had not been fully realized in practice, largely because of the resource implications of administering it, siloed working between police and probation, and inconsistencies in training across forces (Mann and Lundrigan 2021). The evaluation also identified difficulties in rating certain factors, and highlighted gaps in the tool's applicability to subgroups including registered sex offenders with mental health conditions, learning disabilities, and those who are transgender.

Offender Assessment System

OASys is a structured risk assessment tool, designed in 2001 and used widely within the National Offender Management Service (NOMS) in England and Wales (NOMS 2015). It is designed to incorporate validated risk factors alongside clinical judgement. The assessments include both static and dynamic risk factors (National Offender Management Service 2015).

OASys is organized into four central components: an analysis of offending-related factors, a risk of serious harm analysis, a summary sheet, and a sentence plan. The offending-related factors component includes 13 sections covering criminal history, analysis of (current) offences, assessment of 10 dynamic risk factors, and suitability to undertake sentence-related activities (e.g. unpaid work, offending behaviour programmes). Each dynamic risk factor is assessed using between 4 and 10 questions, each scored on a 0/2 or 0/1/2 basis (Howard and Dixon 2012). Summary scores are then presented for the offender violence prediction, and the general reoffending predictor (Howard and Dixon 2013). OASys is completed in part by a criminal justice practitioner, and in part as self-report by the individual. The risk assessment was designed and validated using regression models to select static and dynamic risk factors (Howard and Dixon 2012). Practitioners use summary scores alongside their professional judgement to assign risk ratings, which guide decision making.

Individuals are classed as being 'standard', 'medium', or 'high' risk based on their summary scores and practitioner judgement. The definitions are as follows:

Standard: Current evidence does not indicate likelihood of causing serious harm.

Medium: There are identifiable indicators of risk of serious harm. The offender has the potential to cause serious harm but is unlikely to do so unless there is a change in circumstances, for example, failure to take medication, loss of accommodation, relationship breakdown, drug, or alcohol misuse.

High: There are identifiable indicators of risk of serious harm. The potential event could happen at any time and the impact would be serious.

Correctional Offender Management Profiling for Alternative Sanctions

COMPAS is used widely in the USA and was developed by Northpointe (which rebranded to Equivant in 2017). It has been used to assess more than 1 million people since it was first introduced in 1998 (Dressel and Farid 2018). The COMPAS tool is proprietary, and limited information is available regarding its design and use. The tool uses Machine Learning methods (Brennan and Dieterich 2018); however, the source code is protected under the trade secrets law and has not been made publicly available, as upheld in the Wisconsin v Loomis (2016) case. The tool consists of 22 scales, grouped into five main categories. A raw score is first calculated and converted into a decile level, which in turn determines the allocation of a low, medium, or high-risk label. Each individual is then given a risk score out of 10 within that category—low (1–4), medium (5–7), or high (8–10) (Blomberg et al. 2010). Separate tools are available for youth, and for women, due to different risk profiles in these groups (Blomberg et al. 2010).

10 Key ethical and methodological challenges

Through our multi-disciplinary workshop, we identified and refined the following 10 themes. Each theme is introduced by outlining a specific risk, followed by a discussion of the associated ethical and methodological issues, and concluding with a summary of relevant implications for those involved in the development, implementation, or oversight of risk assessment tools.

Accuracy and reported metrics

Risk

If accuracy metrics are not properly reported or understood, decision-makers may use poorly validated tools that are inappropriate for the population in which they are applied. This can lead to serious consequences, such as unjustified detention, denial of bail or parole, or inappropriate allocation of resources. Inaccurate tools may fail to identify individuals who require intervention or may wrongly flag individuals as high risk, contributing to outcomes that are unfair, and/or disproportionate.

Discussion

The reporting of relevant metrics for risk assessment tools is generally poor (Fazel and Wolf 2018). As Fazel and Wolf (2018) discuss, when selecting a tool it is important to look for reports measuring both discrimination and calibration. Measures of discrimination indicate the tools' ability to allocate a higher risk score to those who have the outcome (e.g. crime) to those who do not. This could include the area under the curve statistic, and measures of classification (which require cut-offs) such as sensitivity (the proportion of true positives the tool correctly identifies), and specificity (the proportion of true negatives it correctly identifies). Calibration tests how well predicted risk scores compare with actual outcomes. This can be explored visually by graphing predicted versus actual retrospective occurrence of the outcome, as well as through statistical tests to measure miscalibration. Calibration is an important measure largely overlooked in the literature (Seyedsalehi and Fazel 2024).

Estimates of tool performance are often inflated in studies where one of the authors has a financial conflict of interest in the tool, as examined by Fazel et al. (2022a). In addition, for tools such as OASys, much of the published evaluation has been co-authored by tool developers, and there is no evidence of independent evaluation (see National Offender Management Service 2015, for a summary of existing research on the OASys tool). The value of independent evaluation is illustrated by the national assessment of ARMS: commissioned independently of the tool's developers, it identified significant implementation failures that internal monitoring had not surfaced, including inconsistencies in assessment quality across forces and unresolved difficulties in applying the tool to particular subgroups (Mann and Lundrigan 2021). Decision-makers and oversight committees should therefore consider requiring independent evaluation of any tool before deployment, rather than relying solely on developer-produced validation reports.

Implications

Decision-makers should take care to select tools for which there is strong empirical evidence. Decision-makers should require developers of these tools to comprehensively report a range of measures of discrimination and calibration. Decision-makers and/or oversight committees should consider independent assessment of the accuracy of tools, rather than those conducted by tool developers. They should fund independent studies if they do not exist.

False positive and false negative rates

Risk

It is mathematically impossible to optimize overall predictive accuracy while also ensuring that false positive and false negative rates are equal across different demographic groups—unless those groups have identical base rates of the outcome. These competing goals cannot mathematically be achieved simultaneously (except in highly constrained circumstances).

Discussion

As Kleinberg et al. (2016) point out, there are inherent mathematical trade-offs in the definitions of fairness that can be employed by these models. This trade-off exists between calibration—where the given scores mean the same thing in different groups (e.g. a score of seven for a female defendant means the same level of risk as a seven for a male defendant), and parity of false positive and false negative rates across groups. This is because where risk is not equal across groups (e.g. if males have a higher risk of reoffending than females), rates of false positives and false negatives across these groups will not be equally distributed if calibration is achieved.

It is mathematically impossible, except in highly constrained circumstances, to achieve calibration as well as equal rates of false positives and false negatives between groups (Kleinberg et al. 2016). Calibration is critically important; if 'high risk' for a female means a different level of risk to 'high risk' for a male, practitioner interpretation is then required. This would mean that practitioners would choose how much weight to give each feature of the individual, defying the point of using a validated tool. Where individuals fit into multiple groups—e.g. 'male' and 'black', the level of extrapolation required from the given risk score becomes pure guesswork. False positive and false negative rates are therefore a trade-off, which should be determined situationally. It may be that excessive false positives are more acceptable in a tool which informs the provision of additional safeguarding provision for those assessed to be at a higher risk of domestic abuse, but unacceptable in a tool whose output determines an individual being held on remand pre-trial (rather than granted police bail)—i.e. where the outcome is intervention vs detention. The acceptable balance between false positives and false negatives will depend on the stage of the criminal justice process and the ethical weight attached to each type of error (see Douglas et al. 2017). Resource implications should also be considered: in settings where specialist capacity is already stretched, many false positives can carry a practical cost beyond their ethical weight, diverting limited resources away from cases that genuinely require intervention (Trood et al. 2026).

This problem was highlighted in the analysis by Angwin et al. (2016), as the COMPAS tool has unequal rates of false positives for Black and White defendants—it produces more false positives for Black defendants, meaning that their risk of reoffending is often over-predicted, compared to White defendants, who have more false negatives. This is related to differential measured reoffending rates in the underlying data for COMPAS, as Northpointe pointed out in their response to Angwin (Dieterich et al. 2016).

The HART model was explicitly designed to prioritize public safety by tolerating certain types of error. Specifically, it was built to accept a higher rate of false positives—where the individual's risk was over-estimated (termed 'cautious errors'). False negatives (referred to as 'dangerous errors') where genuinely high-risk individuals are mistakenly assessed as low risk, were minimized in the model (Oswald et al. 2018). This reflects a deliberate design choice: the model treats it as less harmful to mistakenly restrict someone who poses little risk than to mistakenly release someone who poses a serious risk. Such value judgements are embedded in the model's structure. Decision-makers selecting or deploying tools like HART must be aware of these underlying trade-offs, as they directly shape how the tool distributes error.

It is worth noting that while the trade-offs between false positive and false negative rates are relevant to all four tools discussed here, they are considerably more difficult to evaluate for ARMS. For HART, OASys, and COMPAS, algorithmically derived risk scores provide a numerical and reproducible output against which subsequent outcomes can be evaluated. For ARMS, the priority-for-work rating reflects an integrated practitioner judgement across a defined set of factors. As the practitioner combines and weights those factors in a subjective way, two officers assessing the same individual can arrive at different ratings while operating within the recommended framework. This makes it more difficult to isolate whether a false positive or false negative reflects a limitation of the tool itself or variation in how it was applied.

Implications

These challenges are not only technical in nature but also raise fundamental ethical questions. Trade-offs between different fairness outcomes—such as parity in false positives versus calibration—cannot be resolved through optimization, nor can they be settled in advance during the design of a tool. Instead, they rely on ethical judgements as to the acceptability of different errors in different contexts.

As such, awareness of these trade-offs is critical when selecting and implementing a tool. But more than that, tools themselves should be designed to allow decision-makers some flexibility to respond to the ethical considerations relevant to their context—for example, the relative acceptability of false positives in a safeguarding setting versus a setting where decisions about detention are made. This flexibility requires accuracy metrics to be reported across all possible cut-offs (e.g. with a ROC curve).

The assumptions built into a model, such as the prioritization of certain error types, should be made explicit and open to scrutiny. Practitioners should be supported to understand how error rates may vary across groups, and oversight bodies should not only monitor these patterns over time but remain open to revising how fairness is weighted and enacted in practice.

External validation in the population of interest

Risk

If tools have not been properly validated, they may not be appropriate for use in the population of interest. For example, a tool validated in only a male sample will not be appropriate for use with females unless additional validation is undertaken.

Discussion

Actuarial risk assessment tools will inevitably perform better on the data in which they were trained, as discussed by Fazel and Wolf (2018). For example, the COMPAS tool was developed and tested primarily in male populations. Hamilton (2019) found that it systematically over-estimated risk for women. This is because women generally reoffend less than men, but the tool was not adjusted to reflect this. Testing performance of the tool in a separate, or external, sample is therefore important in understanding how it will perform in real-world contexts.

The national evaluation of ARMS similarly identified gaps in the tool's applicability to particular subgroups, including registered sex offenders with mental health conditions, learning disabilities, and those who are transgender, suggesting that the tool had not been adequately validated for use with these groups prior to national rollout (Mann and Lundrigan 2021). For further guidance relating to the technical aspects of the tool that should be considered, see Fazel and Wolf (2018).

Implications

Decision-makers should select tools that can evidence validation in external samples, and developers should give considerable thought to fully representative external validation. Decision-makers and oversight committees should ensure that developers share descriptive information about the population on which their tool was developed and tested.

The use of demographic data

Risk

The use of demographic data may introduce bias against demographic groups. This can result in decisions that are influenced, either directly or indirectly, by characteristics such as age, race, sex at birth, or socioeconomic status.

Discussion

Bias in the use of actuarial risk assessment tools can be thought of in layers (Eckhouse et al. 2019). First, bias can be introduced through the model, which may apply weightings incorrectly to demographic variables, creating bias in the prediction for different groups of people.

It is important to note that weighting all variables equally is also inappropriate: this assumes that all variables have the same association with the outcome, which is rarely the case. Tools that adopt an equal-weighting approach—assigning identical weight to each factor in a checklist regardless of its evidential association with the outcome—perform notably more poorly on average than those that derive weights empirically from data (Hamilton et al. 2015). The key challenge is therefore not whether to weight variables, but how to do so transparently and on the basis of robust evidence.

Secondly, bias may exist in the underlying training and validation data, which is then systematically and immeasurably embedded in the predicted outcomes. For example, if the

training data predominantly includes individuals from over-policed communities, the model may learn to associate those communities with higher risk levels, not because of actual differences in behaviour, but due to their disproportionate representation in the data. A further layer of complexity is that the data on which these models have been trained reflect what is reported and recorded, and who is charged—not what has actually occurred. There are disparities in who is formally charged for offences, and these are not randomly distributed (Lammy 2017). Instead, they are shaped by policing priorities, reporting rates, and systemic disparities in how different communities are treated by the criminal justice system. This means that a tool's accuracy must be understood as accuracy in predicting recorded outcomes, which may be an incomplete proxy for the underlying behaviour the tool is intended to assess, in some cases. Good quality tools can mitigate against this potential disparity, particularly if violent crime outcomes are the predictive focus (Skeem and Lowenkamp 2016).

This can lead to perpetuation of these structural biases. There is some emerging evidence that, even when controlling for case characteristics, being held in detention pre-trial is associated with a greater likelihood of being convicted, harsher sentences, and future cyclical justice system involvement (Digard and Swavola 2019). This underscores the high stakes of pre-trial decisions informed by risk assessment tools: false positives—where an individual is wrongly assessed as high risk—can have compounding, long-term consequences that extend far beyond the immediate decision point.

Finally, some tools—especially those using more complex machine learning methods—may perpetuate existing societal biases by replicating them. As Birhane (2021) discusses, machine learning models frequently automate and reinforce historical, unjust, and discriminatory patterns already embedded in society. This is not the creation of bias in the sense of a technical error, as with an unrepresentative dataset or a poorly calibrated model, but rather the perpetuation of structural inequalities through seemingly neutral systems. A clear example of this is the COMPAS tool, which Angwin et al. (2016) found to disproportionately classify Black defendants as high risk of reoffending, even when they had similar criminal histories to white defendants. The model learned from patterns in historical data that already reflect systemic racial disparities, such as longer sentences or higher arrest rates for Black individuals, and encoded these into its risk predictions. It is critical to note here that the removal of an undesirable variable (such as race) does not necessarily remove the ability of a tool to learn racialized attributes by proxy, particularly where more complex machine learning models are used, such as in the case of the HART tool (Amoore 2020; Davies and Douglas 2022).

Implications

Bias in risk assessment tools cannot be eliminated, particularly when they rely on predictors shaped by historically unequal systems, such as criminal history. The aim is not to exclude all biased variables, but to understand how they function within the model and to consider whether their use is ethically and contextually appropriate. Decision-makers should be aware of the perpetuation of historical discrimination in society, making sure that training data are temporally relevant to reflect policy and practice changes, and providing appropriate training for practitioners in the interpretation of these tools. Testing whether the tool retains accuracy across different subgroups is one way to establish whether bias is being introduced against particular groups—but there may be limitations to the statistical power that can be achieved in some settings.

Threshold scores

Risk

Three of the tools (COMPAS, OASys, and HART) examined as case studies here operate a threshold 'risk score'—for example low, medium, or high risk. The Risk Matrix 2000 (which contributes to the ARMS assessment) also uses a similar threshold structure. The use of thresholds could be inappropriate as risk thresholds are essentially arbitrary, can be inconsistent between tools, and acceptable risk may vary between contexts.

Discussion

Any threshold in this context—where one person scoring 29 (e.g.) is designated as 'low risk', and another scoring 30 might be considered 'medium risk'—is an ultimately arbitrary threshold and treating these two people differently when their risk levels are similar is not reasonable. Acceptable risk thresholds may also vary depending on context—a 30 per cent risk that an individual will commit a violent offence may be less acceptable than a 30 per cent risk of an individual committing a non-violent offence (Wynants et al. 2019). Additionally, risk thresholds may vary across tools, leading to inconsistency—someone may be allocated 'high risk' on one tool, and 'medium risk' on another on the basis of inconsistent cut-offs (Kroner and Derrick 2022). Some authors have made attempts to standardize risk thresholds—for example by developing a five-level system, as described by Kroner et al. (2020) and Kroner and Hanson (2022), however these standardized approaches have not gained wide acceptance.

Probability scores across a particular timespan (e.g. likelihood of reoffending within 1 year) are an alternative to threshold scores, which could offer a more nuanced representation of risk, which can be considered contextually. Probability scores also perhaps better represent the dynamic nature of an individual's risk, which may vary depending on current life circumstances. However, probability scores could be more challenging for frontline practitioners to interpret than categorical labels such as 'high risk'. Standardized thresholds are often easier to implement a standardized response operationally, whereas interpreting a probability score requires an understanding of what that number means in context. Without appropriate training, there is a risk that these scores are treated as de facto thresholds—for example assuming that a score above 0.7 equates to 'high risk'. On the other hand, probability scores can allow for more professional discretion, as they enable practitioners to consider individuals on either side of a cut-off point more flexibly (e.g. treating someone with a score of 0.29 similarly to someone with 0.30).

Implications

Risk thresholds are ultimately arbitrary, and ideally decision-makers should be involved in the determination of the allocated cut-offs and their interpretation in context. Probability scores, rather than risk thresholds, offer a more nuanced account of the risk predicted by a tool, which can be contextually interpreted by practitioners. However, appropriate training is required for practitioners to be able to understand and act on probability scores, and guidelines may need to be drawn up that examine possible thresholds.

Transparency

Risk

Tools that are opaque or 'black box' (Pasquale 2015) lead to decisions that are not understood by either the individual or the practitioner. Any bias being introduced by the tool becomes difficult to detect and correct. As models become more complex—particularly when developed using machine learning techniques or by private companies—transparency and interpretability often decrease, even as the volume of data and potential predictive power increase.

Discussion

The publication of source code (internally, if not publicly) is essential in achieving transparency and should be insisted upon by decision-makers. The source code should also be translated in a way that makes it understandable to frontline users. There are instances globally where source code is protected under trade secrets law—as in the case of the COMPAS tool. This makes it impossible to judge how the tool is weighting particular risk factors, and to have oversight of the decisions made by the tool developer.

However, the working of actuarial tools is not solely determined by their source code; they are also shaped through their relationships to the data they learn from. As the training data reflects our societal treatment of particular groups, we cannot separate ourselves and our societal structures from their outputs (Amoore 2020).

The degree of transparency achievable varies significantly with model type. In traditional regression models—of which OASys is a well-documented example—the contribution of each variable to the outcome can be reported directly, making it possible to scrutinize the weighting given to individual risk factors (Howard and Dixon 2012). ARMS, as a structured professional judgement tool, achieves transparency through a different route: the factors assessed and the rationale for their inclusion are explicit, even if the way a practitioner integrates them is not algorithmic. By contrast, machine learning models such as the random forests used in HART, and the proprietary algorithms underlying COMPAS, introduce opacity that cannot be resolved simply by publishing source code: as the model iteratively updates through its relationship with training data, the weighting given to each variable becomes impossible to isolate (Amoore 2020). This suggests a spectrum of transparency across tool types, with structured professional judgement frameworks at one end and complex proprietary ML models at the other.

Implications

Source code should be made available, if not publicly then at a minimum to decision-makers and oversight committees. As discussed by Babuta et al. (2018), all algorithms used in these contexts should be able to be retroactively deconstructed—meaning that it should be possible to work out how a particular decision was generated, which variables were used, and how they were weighted or combined to produce the output.

Relationship to practitioner judgement

Risk

There is a risk that the use of these tools can be perceived as having objective authority, discouraging those interpreting and applying their output from critically engaging with their limitations. Conversely, there is an equally significant risk that practitioners over-ride a tool's output too readily or without adequate justification, reintroducing the inconsistency and subjective bias that these tools were designed to reduce.

Discussion

Some authors (Slobogin 2021) have argued that overriding or adjusting the results of a well-validated risk assessment tool to account for practitioner judgement is not permissible, defeating the purpose of a well-validated risk assessment tool. Slobogin (2021) argues that incorporation of intuition about risk (often based on factors already accounted for by the tool) makes the tool less accurate, diminishing the benefits of using an actuarial tool in the first place. Guay and Parent (2018) found that the predictive validity of upward over-rides (where clinicians increase the given risk score) was better than that of downward over-rides, but that clinical over-rides appeared to reduce the predictive validity of the tool across all measures. Other authors (Fazel et al. 2016, 2022b) argue that these tools should be an adjunct to clinical decision making, as there are several individual-level factors, which tools do not capture, and professional judgement is therefore required.

The nature of this integration varies considerably across tools. ARMS is explicitly designed to incorporate professional judgement throughout the assessment process: the practitioner's reading of the evidence for each factor, combined with the actuarial output from the Risk Matrix 2000, drives the final priority-for-work rating. Tools such as HART, OASys, and COMPAS operate differently, producing a risk classification that practitioners receive and consider acting on, but with less structured guidance on how their own assessment should interact with the tool's output. This distinction affects how the risks of over-reliance and inappropriate over-ride are understood: for ARMS, the concern is less that a practitioner will defer uncritically to an algorithmic output, and more that the structured framework may be applied inconsistently, or that professional judgement may be exercised without sufficient scrutiny of the actuarial estimation of risk.

The question of when and how practitioners should depart from a tool's output is therefore consequential. There is a real danger that permitting professional discretion without clear boundaries legitimizes the kind of *ad hoc*, subjective decisions that actuarial tools were designed specifically to

counteract. This is particularly the case where over-ride is based on features already captured by the tool. At the same time, rigid adherence to a tool's output in all circumstances can be problematic in the case of specific individual-level and contextual factors in someone's circumstances, which are not captured by a tool. In high-stakes decisions, where the consequences of inaccuracy can be serious, practitioners may also feel unable to deviate from the tool's output even where individual/contextual factors would justify doing so (e.g. if the individual is being actively targeted by another individual or a gang).

Decision-makers in the police can also consider other objectives beyond simply assessing risk when deciding on outcomes for individuals (such as the individual's employment or education opportunities, which may be affected by bail decisions).

Implications

The individual variables assessed in risk assessment tools should be fully transparent to practitioners and decision-makers, so that they are not encouraged to over-ride scores on the basis of risk factors already incorporated by the tool (thus 'double counting' these risk factors). All tools should allow for the possibility of professional over-ride, and some that explicitly structure the integration of practitioner judgement, such as ARMS where the practitioner's assessment of each factor is a defined part of the process, provide examples of this. Unstructured discretion of practitioner over-ride risks reintroducing precisely the inconsistency and subjectivity that actuarial tools were designed to reduce. Where professional over-ride is exercised, it should be reserved for specific individual and contextual circumstances not captured by the tool, and should be formally documented with explicit justification. Oversight committees should consider the available training for practitioners, and monitor practitioner views about the tool, to ensure that this relationship between tool and officer judgement is appropriate. They should also monitor over-ride cases, and particularly the frequency and direction of downward over-rides, as one mechanism for assessing whether the tool-practitioner relationship is functioning as intended and whether patterns of over-ride are problematic.

Model framing

Risk

The framing of these tools as 'predictive' can be misleading, suggesting that they can determine what an individual will do in the future. In reality, risk assessment tools estimate the probability of an outcome based on patterns observed in a population—they do not offer certainty at the individual level. In a sense, they identify *current* factors/markers for potential future harms.

Discussion

Actuarial risk assessment tools can never predict what an individual will do; instead, they tell us what, on average, happened to someone similar to this person in the characteristics assessed by the tool, in the society whose data in which the model was trained. When their outputs are misinterpreted as individual predictions, this can lead to over-reliance on risk scores, potentially sidelining professional judgement or context-specific information.

Implications

Frontline officers responsible for interpreting the output of the tools may need additional training to effectively combine their understanding of the model output with their professional expertise (Babuta et al. 2018). This training should focus on an understanding of the output of these models as group-based probabilities and current risk markers, rather than individual-level prediction.

Acceptability and feasibility

Risk

Some risk assessment tools require an excessive amount of information (upwards of 100 variables). This unnecessarily increases burdens on both the practitioner and individual, as fewer variables would provide equal predictive accuracy.

Discussion

The burden a tool places on practitioners is not a peripheral concern: tools that are excessively time-consuming risk disengagement, inconsistent completion, and ultimately degraded data quality. The national evaluation of ARMS found that resource implications were among the most significant barriers to effective implementation, with MOSOVO officers in some forces finding the assessment framework difficult to complete to the required standard given existing caseload pressures (Mann and Lundrigan 2021). This echoes findings from other criminal justice tools: The OASys used in probation settings in the UK has been described by probation practitioners as 'the worst tax form you've ever seen'—burdensome to administer, with disengagement from the tool reported as a direct consequence of its length (Mair et al. 2006). The COMPAS tool collects data across 22 scales covering both risk and needs, despite research consistently identifying a small number of variables—principally sex at birth and age—as the strongest predictors of reoffending (Fazel et al. 2016, 2022b). Tool developers should give serious consideration to the trade-off between maximising predictive precision and minimising the data burden on the practitioners administering the tool and the individuals being assessed.

Implications

Tools should be selected that are not unnecessarily long or burdensome to complete, and do not require more data than is required to carry out their purpose. Tool developers should devote attention to understanding the trade-off between maximising predictive accuracy, and minimising the amount of information required to achieve this.

Retraining strategy

Risk

Retraining a model over time is essential because these models operate in environments where the underlying data and risk factors evolve over time, e.g. through changes in policing

Table 1 Summary of 10 key ethical challenges for risk assessment tools in policing.

1. Accuracy and reported metrics Poorly reported accuracy metrics (especially not reporting of calibration) can mask how well tools actually perform. Decision-makers should demand independent validation, not rely solely on developer-led studies.	6. Transparency Many tools, especially proprietary or machine learning-based ones, function as 'black boxes', making it difficult to understand, audit, or challenge their outputs. Full access to source code and variable weightings is vital for ethical oversight.
2. False positive and false negative rates It is mathematically impossible to balance predictive accuracy and error rates equally across groups, requiring ethical decisions about which trade-offs are acceptable in different contexts.	7. Relationship to practitioner judgement There is a risk of both uncritical deference to tool outputs even where specific individual or contextual factors warrant consideration, and of unjustified over-rides to risk scores, which reintroduce subjectivity.
3. External validation in the population of interest Tools often perform poorly when applied outside their development sample, so external validation in the intended population is essential for fair and accurate use.	8. Model framing Tools framed as 'predictive' can be misunderstood as deterministic individual-level predictions, despite only providing group-based probability estimates.
4. The use of demographic data Bias cannot be fully eliminated from tools that rely on data shaped by structural inequalities. Ethical use means acknowledging this, monitoring subgroup impacts, and making value-based decisions about data use.	9. Acceptability and feasibility Tools that demand excessive data burden both practitioners and subjects without improving accuracy; tools should be efficient, using only what is necessary to maintain predictive value.
5. Threshold scores Categorical risk labels like 'high' or 'low' are often based on arbitrary cut-offs that may vary across contexts and tools, limiting their reliability and interpretability.	10. Retraining strategy Tools must be regularly retrained to remain valid, yet retraining is complicated by interventions (which change outcomes) and shifting contexts; without careful strategy, models may degrade over time or reinforce the effects of their own outputs.

policy, public health interventions, and laws. Without regular updates, a model may become less accurate over time due to changes in behaviour, policy, or social context. This is sometimes referred to as concept drift—where the patterns in the data used to train the model no longer reflect current realities. Retraining is therefore essential to maintain the tool's predictive validity and ensures that its outputs continue to support informed, balanced decision making (Slobogin 2021).

Discussion

There is a significant likelihood that retraining will be confounded by the very interventions the risk assessment tool triggers. Because individuals classified as 'high risk' receive interventions that may mitigate adverse outcomes, the observed data subsequently used for retraining is inherently impacted by these protective actions. This lack of a clear counterfactual makes it challenging to determine whether changes in model performance are due to shifts in underlying risk or merely the effects of intervention. For individuals assessed as high priority through ARMS, for example, the risk management plan may include more frequent home visits, licence conditions, or referral to multi-agency oversight panels, such as the UK's multi-agency public protection arrangements. These interventions are protective by design: they change the conditions under which the individual operates and may well reduce the likelihood of reoffending. This makes it very difficult to evaluate whether the original risk rating was accurate —certain cluster randomized trial designs can potentially do this (which would require substantial resources and national approvals to conduct). If an individual assessed as high priority does not reoffend, this may reflect the effectiveness of the intervention rather than an over-estimate of risk—but the data generated will record it as a false positive regardless. Over time, this can lead to a model that is miscalibrated or appears systematically biased, undermining its predictive accuracy and potentially eroding practitioner trust.

Implications

A key challenge for developers is determining how to assess the impact of a tool on individuals classified at different risk levels, since each level of risk guides the corresponding level of intervention. It is entirely problematic to 'turn-off' the model and allow harm to occur where it could have been predicted or prevented, for the purpose of assessing model accuracy. However, it is also difficult to assess model performance over time. Decision-makers and oversight committees should have a thorough understanding of the limitations of any retraining strategy proposed by the developer. Where tools incorporate dynamic risk factors, regular reassessment of individuals may help to maintain predictive accuracy, especially in the absence of robust retraining protocols (see Yukhnenko et al. 2026). This may also ensure that changes in risk over time are captured and addressed, rather than being treated as static.

Discussion

The ethical challenges surrounding actuarial risk assessment tools in policing are complex, context-dependent, and deeply consequential (Babuta et al. 2018). This paper has outlined ten key challenges that arise in their development, implementation, and oversight—ranging from issues of transparency and fairness to the unintended consequences of retraining strategies. These are summarized in Table 1.

Our discussion of four case studies—HART, OASys, ARMS, and COMPAS—illustrates both the potential benefits and the critical limitations of risk assessment tools across different deployment contexts, from UK policing to US corrections, and across different methodological approaches, from machine learning to structured professional judgement frameworks. This range is deliberate: the ethical challenges identified are not specific to any one tool type, and the implications for developers, decision-makers, and oversight bodies apply across this spectrum.

While actuarial tools may offer improvements over unassisted human judgement, they do not operate in a vacuum; their outputs reflect, and can reinforce, existing structural inequalities. The notion that risk assessment tools are inherently more objective than human decision making must be tempered by an understanding of how these models encode biases, generate trade-offs, and shape outcomes in ways that are not always visible or easily challenged.

Given the weighty implications of decisions informed by these tools—including pre-trial detention, public protection, and the allocation of safeguarding resources—it is imperative that their use is accompanied by robust ethical oversight. To that end:

1. Developers must ensure transparency in model design and make explicit the value-laden choices embedded in error weighting, feature selection, and retraining strategies.
2. Decision-makers should prioritize tools that are externally validated in the populations they serve and be equipped with training to critically interpret outputs in context.
3. Oversight bodies must move beyond one-time approval processes and take responsibility for ongoing monitoring of tool performance, subgroup effects, and unintended consequences. In practice, this might include periodic audits of accuracy metrics across demographic subgroups; systematic tracking of outcomes for individuals assessed by the tool (that could lead to updating of the tool), tracking of practitioner over-ride rates and their direction, structured collection of practitioner feedback on tool usability and perceived validity, and review of whether the target population remains sufficiently similar to the population on which the tool was validated. Where significant drift is identified, oversight bodies should have the authority to require retraining of a tool pending independent review, or strengthening of alternative measures (such as the introduction of other tools).

In the UK, there is a growing case for a central regulatory body or national committee to provide independent evaluation, establish minimum ethical standards, and support jurisdictions in assessing whether a tool is appropriate for use. Academia has an important role to play in this process, acting as an independent force capable of interrogating the claims made by developers, evaluating evidence, and shaping ethical norms.

Ultimately, the goal must be to cultivate a governance framework that prioritizes justice, minimizes harm, and prevents the automation of structural inequities. The tensions outlined in this paper should not be seen as obstacles to the use of actuarial tools, but as an ethical foundation for their responsible, equitable, and transparent deployment.

Acknowledgements

The authors would like to acknowledge the Caroline Miles visiting scholarship at Ethox, University of Oxford, awarded to HK, which facilitated the workshop from which this paper arose. Employees of Thames Valley Police are co-authors on this paper and attended the workshop; however, the views expressed herein do not necessarily represent the views of Thames Valley Police. For the purpose of open access, the authors have applied a 'Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising'.

Conflicts of interest

Thomas Douglas reports a relationship with Merck KGaA that includes funding work unrelated to this manuscript. Seena Fazel has co-authored publications on the development and validation of OxRec, OxRIS, and OxDOV, which are freely available web-based risk assessment tools for use in criminal justice. Tori Pamela Anne Olphin has developed a tool for prediction of domestic abuse risk (DARAT), and also two tools for prediction of solvability of crime. None of the above tools are discussed in this paper. All other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was facilitated by the Caroline Miles visiting scholarship (Ethox Centre, University of Oxford), awarded to H.K. M.S., M.G., and S.F. are supported by the NIHR Biomedical Research Centre: Oxford Health. T.D. is supported by the European Research Council (grant number 819757), and the Uehiro Foundation on Ethics and Education. W.H.W. is supported by the Barrow Cadbury Trust. J.H. is supported by the AHRC (grant number AH/Z506072/1). The funders had no role in the study design, analysis, and interpretation of the data, the writing of the report, or the decision to submit for publication.

References

Amoore, L. (2020) *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. North Carolina: Duke University Press.

Angwin, J. *et al.* (2016) *Machine bias: Risk assessments in criminal sentencing.* ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Babuta, A., Oswald, M. and Rinik, C. (2018). *Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges*. (Whitehall Report 3-18). Royal United Services Institute for Defence and Security Studies. https://researchportal.northumbria.ac.uk/ws/files/21461394/Babuta_et_al_Machine_Learning_Algorithms_and_Police_Decision_Making_OA.pdf

Berk, R., Berk, D. and Drougas, D. (2019) *Machine Learning Risk Assessments in Criminal Justice Settings*. Springer. Switzerland.

Birhane, A. (2021) 'Algorithmic Injustice: A Relational Ethics Approach', *Patterns*, 2: 100205. https://doi.org/10.1016/j.patter.2021.100205

Blomberg, T. *et al.* (2010) *Validation of the COMPAS Risk Assessment Classification Instrument*. Tallahassee, FL: College of Criminology and Criminal Justice, Florida State University.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45: 5–32. https://doi.org/10.1023/A:1010933404324

Brennan, T. and Dieterich, W. (2018) 'Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)', in Singh, J. P., Kroner, D. G., Wormith, J. S., Desmarais, S. L. and Hamilton, Z. (eds.) *Handbook of Recidivism Risk/Needs Assessment Tools*, pp. 49–75. London: Wiley.

Coalition for Critical Technology (2020) *Abolish the #TechtoPrisonPipeline*. https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16, accessed 22 Feb. 2025.

College of Policing. (2013) *Risk: Ten Principles Related to Taking and Reviewing Risk*. https://www.college.police.uk/app/risk/risk

College of Policing (2025) *Identifying, Assessing, and Managing Risk*. https://www.college.police.uk/app/major-investigation-and-public-protection/managing-sexual-offenders-and-violent-offenders/identifying-assessing-and-managing-risk, accessed 2 Mar. 2025.

Danziger, S., Levav, J. and Avnaim-Pesso, L. (2011) 'Extraneous Factors in Judicial Decisions', *Proceedings of the National Academy of Sciences of the United States of America*, 108: 6889–92. https://doi.org/10.1073/pnas.1018033108

Davies, B. and Douglas, T. (2022) 'Learning to Discriminate: The Perfect Proxy Problem in Artificially Intelligent Criminal Sentencing, *in Sentencing and Artificial Intelligence*, pp. 97–121. Oxford: Oxford University Press.

Dieterich, W., Mendoza, C. and Brennan, T. (2016) 'COMPAS risk Scales: Demonstrating Accuracy Equity and Predictive Parity', *Northpointe Inc*, 7: 1–36. https://go.volarisgroup.com/rs/430-MBX-989/images/%20ProPublica_Commentary_Final_070616.pdf

Digard, L. and Swavola, E. (2019) *Justice Denied: The Harmful and Lasting Effects of Pretrial Detention. Vera Evidence Brief*. New York: Vera Institute of Justice.

Douglas, T. *et al.* (2017) 'Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data', *European Psychiatry: The Journal of the Association of European Psychiatrists*, 42: 134–7. https://doi.org/10.1016/j.eurpsy.2016.12.009

Dressel, J. and Farid, H. (2018) 'The Accuracy, Fairness, and Limits of Predicting Recidivism', *Science Advances*, 4: eaao5580. https://doi.org/10.1126/sciadv.aao5580

Eckhouse, L. *et al.* (2019) 'Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment', *Criminal Justice and Behavior*, 46: 185–209. https://doi.org/10.1177/0093854818811379

Fair Trials. (2022) *FOI Reveals Over 12,000 People Profiled by Flawed Durham Police Predictive AI Tool*. https://www.fairtrials.org/articles/news/foi-reveals-over-12000-people-profiled-by-flawed-durham-police-predictive-ai-tool/, accessed 5 Mar. 2025.

Fazel, S. *et al.* (2016) 'Prediction of Violent Reoffending on Release from Prison: Derivation and External Validation of a Scalable Tool', *The Lancet: Psychiatry*, 3: 535–43. https://doi.org/10.1016/S2215-0366(16)00103-6

Fazel, S. *et al.* (2022a) 'The Predictive Performance of Criminal Risk Assessment Tools Used at Sentencing: Systematic Review of Validation Studies', *Journal of Criminal Justice*, 81: 101902. https://doi.org/10.1016/j.jcrimjus.2022.101902

Fazel, S., Sariaslan, A. and Fanshawe, T. (2022b) 'Towards a More Evidence-Based Risk Assessment for People in the Criminal Justice System: The Case of OxRec in the Netherlands', *European Journal on Criminal Policy and Research*, 28: 397–406. https://doi.org/10.1007/s10610-022-09520-y

Fazel, S. and Wolf, A. (2018) 'Selecting a Risk Assessment Tool to use in Practice: A 10-Point Guide', *BMJ Mental Health*, 21: 41–3. https://doi.org/10.1136/eb-2017-102861

Fiesler, C. (2023) *AI has Social Consequences, but who Pays the Price? Tech Companies' Problem with 'Ethical Debt'*. UK: The Conversation.

Guay, J.-P. and Parent, G. (2018) 'Broken Legs, Clinical Overrides, and Recidivism Risk: An Analysis of Decisions to Adjust Risk Levels with the LS/CMI', *Criminal Justice and Behavior*, 45: 82–100. https://doi.org/10.1177/0093854817719482

Hamilton, M. (2019) 'The Sexist Algorithm', *Behavioral Sciences & the Law*, 37: 145–57. https://doi.org/10.1002/bsl.2406

Hamilton, Z. *et al.* (2015) 'Isolating Modeling Effects in Offender Risk Assessment', *Journal of Experimental Criminology*, 11: 299–318. https://doi.org/10.1007/s11292-014-9221-8

Harris, A. P. and Sen, M. (2019) 'Bias and Judging', *Annual Review of Political Science*, 22: 241–59. https://doi.org/10.1146/annurev-polisci-051617-090650

HMIC. (2014) *Everyone's Business: Improving the Police Response to Domestic Abuse*. UK: Her Majesty's Inspectorate of Constabulary.

Holsinger, A. M. *et al.* (2018) 'A Rejoinder to Dressel and Farid: New Study Finds Computer Algorithm is More Accurate Than Humans at Predicting Arrest and as Good as a Group of 20 lay Experts', *Federal Probation*, 82: 50. https://heinonline.org/HOL/Page?handle=hein.journals/fedpro82&div=21&g_sent=1&casa_token=LPH_rEI4VbQAAAAA:ge3CwU1vfu4JGqjJVEIjDFBQdb4j38k-ma6Mczy76NxQslT80Yxq8ear6GlpKovbC4YAHmA

Howard, P. D. and Dixon, L. (2012) 'The Construction and Validation of the OASys Violence Predictor: Advancing Violence Risk Assessment in the English and Welsh Correctional Services', *Criminal Justice and Behavior*, 39: 287–307. https://doi.org/10.1177/0093854811431239

Howard, P. D. and Dixon, L. (2013) 'Identifying Change in the Likelihood of Violent Recidivism: Causal Dynamic Risk Factors in the OASys Violence Predictor', *Law and Human Behavior*, 37: 163–74. https://doi.org/10.1037/lhb0000012

Kewley, S. and Blandford, A. (2017) 'The Development of the Active Risk Management System', *Journal of Criminal Psychology*, 7: 155–67. https://doi.org/10.1108/JCP-10-2016-0034

Kleinberg, J., Mullainathan, S. and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. arXiv 1609.05807. https://doi.org/10.48550/arXiv.1609.05807, 19 September 2017, preprint: not peer reviewed.

Kroner, D. G. and Derrick, B. (2022) 'The Council of State Governments Justice Center Approach to Increasing Risk-Level Consistency in the Application of Risk Assessment Instruments', *Assessment*, 29: 169–80. https://doi.org/10.1177/1073191120958066

Kroner, D. G. and Hanson, R. K. (2022) 'Measuring What Matters: Standardised Risk Levels for Criminal Recidivism Risk, *Challenging Bias in Forensic Psychological Assessment and Testing*, pp. 95–110. London: Routledge.

Kroner, D. G., Morrison, M. M. and Lowder, M. E. (2020) 'A Principled Approach to the Construction of Risk Assessment Categories: The Council of State Governments Justice Center Five-Level System', *International Journal of Offender Therapy and Comparative Criminology*, 64: 1074–90. https://doi.org/10.1177/0306624X19870374

Lammy, D. (2017) *The Lammy Review: An Independent Review into the Treatment of, and Outcomes for, Black, Asian and Minority Ethnic Individuals in the Criminal Justice System*. London: Her Majesty's Government.

Law Society of England and Wales. (2019) *Algorithms in the Criminal Justice System*. London: The Law Society.

Lin, Z. J. *et al.* (2020) 'The Limits of Human Predictions of Recidivism', *Science Advances*, 6: eaaz0652. https://doi.org/10.1126/sciadv.aaz0652

Mair, G., Burke, L. and Taylor, S. (2006) 'The Worst tax Form You've Ever Seen'? Probation Officers' Views About OASys', *Probation Journal*, 53: 7–23. https://doi.org/10.1177/0264550506060861

Mann, N. and Lundrigan, S. (2021) 'Dynamic Assessment of Registered Sexual Offenders: The National Practitioner Perspective on the use of the 'Active Risk Management System' (ARMS)', *Policing & Society*, 31: 1199–216. https://doi.org/10.1080/10439463.2021.1873324

Myhill, A., Hohl, K. and Johnson, K. (2023) 'The 'Officer Effect'in Risk Assessment for Domestic Abuse: Findings from a Mixed Methods Study in England and Wales', *European Journal of Criminology*, 20: 856–77. https://doi.org/10.1177/14773708231156331

National Offender Management Service. (2015) *A Compendium of Research and Analysis on the Offender Assessment System (OASys), 2009–2013*. R. Moore, ed. Ministry of Justice. https://static.poder360.com.br/2025/04/uk-noms-research-analysis-oasys-julho-2015.pdf

Noti, G. and Chen, Y. Learning when to advise human decision makers. arXiv 2209.13578, https://doi.org/10.48550/arXiv.2209.13578, 27 September 2022, preprint: not peer reviewed.

Ogonah, M. G. *et al.* (2023) 'Violence Risk Assessment Instruments in Forensic Psychiatric Populations: A Systematic Review and Meta-Analysis', *The Lancet: Psychiatry*, 10: 780–9. https://doi.org/10.1016/S2215-0366(23)00256-0

Oswald, M. *et al.* (2018) 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' proportionality', *Information & Communications Technology Law*, 27: 223–50. https://doi.org/10.1080/13600834.2018.1458455

Pasquale, F. (2015) *The Black box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Richards, L. (2009) *Domestic Abuse, Stalking and Harassment and Honour Based Violence (DASH, 2009) Risk Identification and Assessment and Management Model*. London: Association of Police Officers (ACPO).

Seyedsalehi, A. and Fazel, S. (2024) 'Suicide Risk Assessment Tools and Prediction Models: New Evidence, Methodological Innovations, Outdated Criticisms', *BMJ Mental Health*, 27. https://doi.org/10.1136/bmjment-2024-300990

Singh, J. P. *et al.* (2014) 'International Perspectives on the Practical Application of Violence Risk Assessment: A Global Survey of 44 Countries', *International Journal of Forensic Mental Health*, 13: 193–206. https://doi.org/10.1080/14999013.2014.922141

Skeem, J. L. and Lowenkamp, C. (2016) 'Race, Risk, & Recidivism: Predictive Bias and Disparate Impact', *Criminology: An Interdisciplinary Journal*, 54: 680–712. https://doi.org/10.1111/1745-9125.12123

Slobogin, C. (2021) *Just Algorithms: Using Science to Reduce Incarceration and Inform a Jurisprudence of Risk*. Cambridge: Cambridge University Press.

Sorell, T. (2024) 'AI-related Data Ethics Oversight in UK Policing', *Policing: A Journal of Policy and Practice*, 18. https://doi.org/10.1093/police/paae016

Trood, M. D. *et al.* (2026) 'The Limits of Predicting Near Lethal and Lethal Family and Intimate Partner Violence', *Journal of Family Violence*, 1–16, https://doi.org/10.1007/s10896-025-01029-2

Turner, E., Medina, J. and Brown, G. (2019) 'Dashing Hopes? The Predictive Accuracy of Domestic Abuse Risk Assessment by Police', *The British Journal of Criminology*, 59: 1013–34. https://doi.org/10.1093/bjc/azy074

UK Parliament (2024) '*Public Authority Algorithmic and Automated Decision-Making Systems Bill*'. https://bills.parliament.uk/bills/3760.

University of Cambridge (2018) '*Helping Police Make Custody Decisions Using Artificial Intelligence*'. Cambridge: University of Cambridge. https://www.cam.ac.uk/research/features/helping-police-make-custody-decisions-using-artificial-intelligence. [Accessed March 2025]

Viljoen, J. L. *et al.* (2025) 'Are Risk Assessment Tools More Accurate Than Unstructured Judgments in Predicting Violent, any, and Sexual Offending? A Meta-Analysis of Direct Comparison Studies', *Behavioral Sciences & the Law*, 43: 75–113. https://doi.org/10.1002/bsl.2698

Wisconsin v Loomis. (2016) *881 N.W.2d 749 (Wis. 2016),* cert. denied, 137 S. Ct. 2290 (2017).

Wynants, L. *et al.* (2019) 'Three Myths About Risk Thresholds for Prediction Models', *BMC Medicine*, 17: 192–7. https://doi.org/10.1186/s12916-019-1425-3

Yukhnenko, D. *et al.* (2026) 'Dynamic Prediction of Reoffending in Individuals Given Community Sentences: Development and Validation of a Novel Risk Monitoring Assessment Tool (OxMore)', *Law and Human Behavior*, 50: 153–81. https://doi.org/10.1037/lhb0000641