

Negotiating Risk Boundaries in AI for Policing Through Mixed-Stakeholder Deliberation

Mackenzie Jorgensen^{1*}, Jo Reilly^{2*}, Alex Sutherland³, Miri Zilka^{2†}

¹Northumbria University, UK

²University of Cambridge, UK

³University of Oxford, UK

[†] mz477@cam.ac.uk

Abstract

AI tools are being increasingly adopted in policing in the UK and worldwide. Racial bias is a known and well-documented risk, yet representatives of affected communities are rarely included in decisions about AI adoption. We present results from a mixed-stakeholder deliberation workshop bringing together 30 community representatives, police officers, and academics to assess the risks of 13 AI use cases in policing, with an explicit focus on racial bias. We found that participants were broadly open to AI adoption, rejecting only three use cases outright – most notably recidivism risk assessment, where objections targeted the premise rather than the implementation. Our analysis reveals that foregrounding racial equity did not narrow the deliberation. Instead, discussions gravitated toward a fundamental set of questions: does this tool actually work, will it deliver genuine benefit, and will that benefit extend to everyone? This integrated reasoning—reminiscent of the curb-cut effect in inclusive design—highlights the benefit of incorporating the racial bias lens into the risk-benefit analysis of AI use cases from the outset.

Appendix available here: <https://arxiv.org/abs/2608.05418>.

Introduction

As law enforcement agencies become increasingly stretched and resource-constrained, artificial intelligence (AI) is viewed as a key strategy for increasing efficiency in operations (National Police Chiefs’ Council 2023; National Police Chiefs’ Council 2025). Agencies have been experimenting with data-driven applications for over a decade — from predictive policing to facial recognition — and recent years have added significant political pressure towards large-scale adoption. Many concerns have arisen alongside this expansion. Among them, the persistence and further amplification of racial bias. The historical and ongoing disproportionality in the criminal justice system — where Black individuals are highly overrepresented in arrests, incarceration, and victimisation — makes racial bias a particularly urgent and hard-to-resolve challenge for AI in policing.

While the number of individual AI applications, even in the UK alone, is numerous (Taka et al. 2025), AI use cases in

policing fall into similar categories. Amongst the most popular and well-known are *predictive policing* and *recidivism risk assessment*, for both of which there is substantial academic literature documenting how they can perpetuate racial bias (Ensign et al. 2018; Zilka et al. 2023b). Responsible AI guidelines in policing have emerged in response, such as the AI Playbook for Policing (National Police Chiefs’ Council 2025), which advises forces to map which groups could be differentially affected and ensure evaluations capture race disproportionalities. However, they offer little guidance on *how* evaluations should be conducted, and in practice, these are rarely performed or shared publicly.

The people best placed to identify these risks *early* are rarely consulted at the design stage. The teams tasked with addressing racial bias within policing are often siloed from those driving AI adoption, and the communities most affected rarely have a seat at the table. These decisions are typically driven by management and technical teams, and do not include those with direct knowledge of how racial disproportionality operates in practice. Earlier intervention, at the ideation stage, offers a more effective opportunity to flag approaches that could embed or exacerbate racial disproportionality before sunk costs make course-correction difficult.

We propose and demonstrate a method for negotiating acceptable risk limits within mixed-stakeholder groups, suitable for the conception stage of AI adoption. During a one-day in-person workshop, we convened 30 stakeholders spanning community representatives, police, and academia to consider 13 AI use cases in policing. Participants deliberated in mixed groups using a red, amber, and green risk framework (Table 1). Community representatives were not required to have technical knowledge; their contributions drew on expertise and lived experience of how racial disproportionality becomes embedded in the criminal justice system. Our aims were twofold: to establish which use cases participants found acceptable for deployment and under what conditions, and to understand the process by which they arrived at these judgments. Participants were broadly open to AI adoption, rejecting only three use cases outright, but were sceptical of AI’s ability to drive meaningful improvements in policing. We found that foregrounding racial equity did not narrow the deliberation: instead, discussions focused on fundamental questions about whether a tool works, for whom, and under what conditions. We ar-

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

gue this is reminiscent of the curb-cut effect in inclusive design (Blackwell 2017; Shneiderman 2020) — that designing with marginalised communities in mind surfaces better questions and can produce better outcomes for everyone.

Background and Related Works

AI Use Cases in Policing

There is substantial interest in AI use in law enforcement, in the UK and worldwide. What ‘counts’ as AI varies, but below we give a brief background to a number of high-profile use cases that were considered in the workshop.

Forecasting Crime Hotspots (“Predictive Policing”). Efforts to identify *where* crime is likely to occur or stable hotspots to guide resource deployments have been undertaken from as early as the 1930s (Shaw and McKay 1942; Braga et al. 2019). This is usually based on the assumption that where crime has occurred repeatedly is where it is likely to happen in the future. However, realistically, accurate prediction of where *and when* crimes will occur is quite challenging (Government of the United Kingdom 2025). The utility of “real-time” hotspot methods is questionable due to the finer the resolution, the variability, and the low number of recorded crimes. Indeed, evaluations of these prediction tools, such as PredPolTM have not shown evidence of crime prevention benefits (Hunt, Saunders, and Hollywood 2014; Saunders, Hunt, and Hollywood 2016). In addition, these tools raise strong concerns about racial inequity (Lum and Isaac 2016; Sankin et al. 2021), even if they improve forecast accuracy (Mohler et al. 2015; Lee, Bradford, and Posch 2024). However, identifying *stable* hotspots of violence can be done with triangulation of police, hospital, and ambulance data (Taylor et al. 2016).

Risk of Reoffending Prediction. Criminal justice systems have relied on the prediction of an individual’s risk of future crime for more than 200 years. Tools evolved from phrenology’s pseudo-science in the 1800s, to actuarial models starting in the 1920–30s, to statistical (e.g., logistic) models dominating for many decades (Berk et al. 2014; Greene et al. 2022). Early examples of ‘AI’ in policing include the HART tool in Durham Constabulary to aid decisions about police bail (Oswald et al. 2018), assessments of ‘dangerousness’ predicting homicide (Berk et al. 2009) and domestic violence (Berk, Sorenson, and Barnes 2016). These tools have consistently raised concerns about racial inequity (Angwin et al. 2016; Harcourt 2010; Zilka et al. 2023b).

Facial Recognition. Facial recognition is now used in the UK for both live operations and retrospective investigations (Home Office 2024). Live facial recognition (LFR) is arguably the most well-known AI application in policing and has been in development for at least 15 years (Davis, Popov, and Surla 2011). LFR has been scrutinised worldwide since its first pilot, with concerns raised about wrongful arrests and ethnic disproportionality (Buolamwini and Gebru 2018). These concerns have been addressed in some instances – the UK’s National Physical Laboratory assessed LFR equity questions (Mansfield 2023), and the Metropolitan Police have examined false positives, finding that only

those on wanted lists were arrested (BBC News 2025). However, recent events put these in question (see Box 2).

Racial Bias in Policing AI Systems

Over the last decade, critical technical and socio-technical research has highlighted bias and related issues in algorithmic tools for policing. Several studies demonstrated that reliance on arrest records as a proxy for underlying crime can cause an exacerbation of racial bias in different types of predictive tools. This is due to the likelihood of a crime becoming known to law enforcement varying significantly with race, sex, and age (Baumer and Lauritsen 2010; Bosick et al. 2012; Butcher, Zilka, and Weller 2022). Zilka et al. (2023b) demonstrated this for tools predicting the risk of re-offending. Within ‘hotspot mapping’ tools, feedback loops can increase patterns of over-policing in already over-policed communities (Ensign et al. 2018; Lum and Isaac 2016; Sankin et al. 2021). For facial recognition, early work documented significantly higher error rates for darker-skinned individuals (Buolamwini and Gebru 2018), and while technical improvements have narrowed some gaps, equity in practice remains contested and deployment-dependent (Radiya-Dixit and Neff 2023).

Participatory Research on AI and Policing

Alongside the critical work, a growing body of work examines how practitioners and communities relate to data-driven policing. Kearney et al. (2024) interviewed 40 Police Scotland practitioners and found that officers were averse to tools like predictive risk assessment and facial recognition, and were concerned with the erosion of community policing through datafication. Zilka et al. (2023a) explored police perspectives on algorithmic transparency and found mixed views, highlighting the importance of transparency for public trust alongside operational concerns. Few works have directly engaged community representatives on AI in policing; exceptions include the Citizens’ Biometrics Council which deliberated on facial recognition (Ada Lovelace Institute 2019); and Haque et al. (2024) who brought together community members, technical experts, and law enforcement around a crime-mapping application, finding that community members were more likely than domain experts to question the core motivation of the tool rather than its technical execution. Ziosi and Pruss (2024) interviewed community organisations, researchers, and public sector actors about the Chicago crime prediction algorithm, finding that community-impacted groups used evidence of algorithmic bias to centre liberation and healing, while public sector actors used the same evidence to reaffirm existing power structures. Scholars have also raised concerns about algorithmic tools deployed without community input, particularly regarding their impact on marginalised groups (Richardson, Schultz, and Crawford 2019; Brayne 2017).

Responsible AI in Policing Frameworks

Acknowledging that we have moved beyond interest in AI and into live use in policing, several practitioner frameworks

USE CASE ASSESSMENT		
Context: Use case: _____	Theme: _____	Group: _____
Risk Classification: ☐ Low risk, can be deployed with minimal checks. ☐ Medium risk, deploy only in specific context and with rigorous checks. ☐ Unacceptable risk, do not deploy.		
Justification:		
Allowed use cases and pre-deployment checks:		
Evaluation and monitoring:		

Table 1: Workshop use case feedback form structure.

for the responsible use of AI in policing have been developed. In the UK, this includes the AI Playbook for Policing (National Police Chiefs’ Council 2025), the Responsible AI Checklist for Policing (National Police Chiefs’ Council and PROBABLE Futures 2025), and the Police Foundation’s review of AI in policing (Muir and O’Connell 2025), alongside broader reviews of AI deployment in policing (Zilka, Sargeant, and Weller 2022; Berk 2021; Lee, Bradford, and Posch 2024). At the regulatory level, the landscape is shifting rapidly yet unevenly across jurisdictions. The EU AI Act represents the most comprehensive risk-based AI legislative response to date. It classifies a range of policing AI applications as either prohibited or high-risk. With high-risk deployments subject to mandatory impact assessments, transparency requirements, and human oversight obligations. Critics have noted, however, that the Act’s exceptions for law enforcement are broad enough to permit many of the practices it ostensibly restricts (European Parliament and Council of the European Union 2024). In the UK, which now, post-Brexit, sits outside the EU regulatory framework, the approach remains sector-led and primarily voluntary. The Algorithmic Transparency Recording Standard (ATRS), made mandatory for central government departments in December 2024, requires public sector organisations to publish information about how and why they use algorithmic tools. However, policing remains operationally independent: only two police force ATRS records had been published as of 2025, and compliance is not currently mandatory for forces (Government Digital Service 2024).

The Police Race Action Plan

UK’s Police Race Action Plan (PRAP) (College of Policing and National Police Chiefs’ Council 2022) was developed jointly by the National Police Chiefs’ Council (NPCC) and the College of Policing to address the race disparities affecting Black people that policing cannot currently fully explain. Launched in 2022, the plan sets out key areas for improve-

ment across policing structured around four themes:

- **Culture and Workforce:** a police service that represents and supports its Black officers, staff and volunteers.
- **Powers and Procedures:** a police service that is fair, respectful, and equitable in its actions towards Black people.
- **Trust and Reconciliation:** a police service that routinely involves Black people in its governance.
- **Safety and Victimisation:** a police service that protects Black people from crime and seeks justice for victims.

The plan’s commitments include introducing mandatory training on racism, anti-racism, and Black history; adopting a new “explain or reform” approach to race disparities in the use of police powers such as stop and search, use of Taser, and other types of force; reviewing misconduct and disciplinary processes to reduce racial disparities; and better enabling Black people to have their voices heard in local communities and within policing itself. The current plan, however, does not include the police’s rapid inclusion of new technologies and the associated risk of increased racial bias. This work bridges this gap directly and supplements PRAP with a specific AI-focused mixed-stakeholder deliberation.

Methods

Participants

Participants were invited to register for the workshop by the organisers based on the recommendations of the central Police Race Action Plan (PRAP) team. Four main groups of participants were invited: (a) community representatives involved in the reduction of racial bias, discrimination, and disproportionality; (b) members of the Police working on reducing disproportionality; (c) members of the Police working on innovation and AI adoption; (d) academics who work on the responsible deployment of AI in policing.

Individuals were recruited from the PRAP team’s networks and rolling recommendations, and—in one case—by

Theme	Use Case	Risk Classification
Culture & Workforce	AI in Recruitment	3, 3, 3, 3, 3
	AI Review of Body-Worn Footage to Identify Misconduct	2, 3, 3, 3, 3
	Bias-Auditing Dashboards	1, 1, 3, 3
Powers & Procedures	Predictive Policing (Hotspot Mapping)	3, 3, 3
	AI-Assisted Classification / Analysis / Summary of Reports	1, 3, 3
	Risk Assessment / Predictive Profiling	5, 5, 5, 5
	Live Facial Recognition	3, 3, 3
Trust & Reconciliation	Community-Sentiment Analysis from Public Data	1, 1, 1, 1, 3
	Sentiment Analysis / Summary of Community Feedback	1, 1, 1, 3, 3
	Public-Facing AI Virtual Assistant	1, 1, 1, 3, 3
Safety & Victimisation	Risk Prediction for Victimisation	3, 3, 3, 5, 5
	AI Transcription / Translation of Calls / Statements / Interviews	1, 3, 3, 3, 5
	AI Advice to Decide if to Carry the Investigation Forward	3, 3, 5, 5, 5

Table 2: A summary of themes and AI use cases in policing discussed amongst participants. For each use case, the risk classifications are reported from the groups that filled in the use case sheets. Risk levels are enumerated as 1 (low risk or green), 3 (medium risk or amber), and 5 (unacceptable risk or red). A number 2 indicates a group marked between low and medium risk.

a nomination from another invitee. Community representatives who have been or are currently engaged with PRAP were the main source drawn upon for community representation. Several academics were invited by the organisers directly. Invitations were sent to ≈ 45 potential participants, of which 35 registered for the workshop, and 30 attended for the day, excluding 2 organisers. Of those, 15 were police, 11 were community representatives, and 4 were academics. In terms of demographics, approximately half of attendees belonged to UK minority ethnic communities; gender balance was roughly even.

Agenda and Use Cases

After a short introduction, the participants were divided into 6 groups for the morning, and a different set of 6 groups for the afternoon. In each session, the groups reviewed a set of AI use cases relevant to policing, divided into four core PRAP themes: (a) *Culture & Workforce* (referring to the recruitment, retention and progression of Black officers and staff), (b) *Powers & Procedures* (relating to fair, respectful and equitable use of police powers), (c) *Trust & Reconciliation* (relating to community involvement and anti-racist culture), and (d) *Safety & Victimisation* (relating to the protection and experiences of Black victims of crime). Before the workshop, we asked the PRAP team to highlight any use cases they wanted included in the workshop. Additional use cases were added by the organisers based on a literature review. The full list of use cases can be found in Table 2.

For each use case, one author created an information sheet that included a deployed example, the intended goals, and known risks associated with the use case. The author at-

tempted to present the information objectively and use non-technical language. The provided information was not comprehensive; it was intended as a conversation starter and leveller, since some participants were less familiar with AI and its associated risks than others. The information sheets for all use cases can be found in the Appendix¹.

Risk Framing and Use Case Feedback

To structure the discussion, we asked participants to complete an assessment worksheet as shown in Table 1. As a group, participants were asked to classify the use cases in terms of risk: (a) *Green* – low risk, can be deployed with minimal checks; (b) *Amber* – medium risk, deploy only in a specific context and with rigorous checks; (c) *Red* – unacceptable risk, do not deploy. Participants were then asked to justify their selection, indicate allowed use cases and pre-deployment checks, and highlight what needs to be evaluated and monitored. In addition to the workshop, participants had an opportunity to anonymously raise further comments and concerns in a post-workshop feedback survey.

Analysis

After the workshop, the 76 completed worksheets were organised by use case, scanned, and typed by one of the authors. The analysis proceeded in two stages. In the first stage, one author conducted a thematic analysis of the worksheet responses for each use case, identifying recurring concerns, points of consensus and disagreement, and the reasoning participants used to justify their risk classifications. Three

¹Appendix available here: <https://arxiv.org/abs/2608.05418>

additional authors, who were also present in the discussions on the day, reviewed and annotated the emerging themes, ensuring they represented not only the worksheets, but the conversations during the workshop (which were not recorded). We note that the analysis relied on the worksheet responses, and recollections were only used to ensure that there was no misrepresentation of the discussions. In the second stage, two authors examined the questions participants appeared to be implicitly asking when arriving at their risk classifications — what we refer to as back-engineered questions. This was cross-checked by two additional authors to ensure the findings were faithful to the discussions on the day. Where direct quotes are used in the results section, they are drawn verbatim from the worksheets. Risk classification was translated to a 5-point scale, with low being 1, medium 3, and unacceptable risk 5, and summarised for each use case. When a group marked two risk categories (e.g., low and medium), the number in-between the categories (e.g., 2) was assigned.

Limitations

Several study limitations should be acknowledged. First, all groups reviewed use cases in the same fixed order, which may have introduced order effects: fatigue, anchoring, or carry-over reasoning from earlier use cases may have shaped how later ones were assessed. Future workshops like this should counterbalance the order of use cases across groups. Second, recruitment was conducted primarily through the networks of the PRAP team, which means participants were largely pre-selected for their existing engagement with issues of racial bias and disproportionality in policing. This was a conscious choice as racial bias in policing was central to the workshop; however, we acknowledge this sample is unlikely to be representative of the broader police workforce. One participant in the feedback survey shared that the youth voice was missing. This workshop did not aim to present the public's views, but of stakeholders representing diverse interests. Third, workshop discussions were not recorded so that participants could feel that their conversations were private and that they could communicate freely. Additionally, authors present on the day observed discussions across groups and were able to cross-check that the thematic findings reported were reflective of the conversations as they occurred, and not only of what participants chose to commit to paper.

Fourth, we acknowledge the possibility of in-group power dynamics. When assigning groups for the workshop, we aimed to balance the police and community representatives and assign one academic to each group. To avoid only dominant voices emerging, we remixed the groups for the afternoon session, such that groups only contained individuals who had not previously been grouped. Organisers moved between groups throughout the sessions, encouraging groups to try to allocate even amounts of time to each use case and supporting quieter individuals to join in. Some community representatives appeared to initially feel aware of their lack of AI technical expertise and needed encouragement to share their views (Harrington, Erete, and Piper 2019). Interestingly, many police present had similar levels of AI technical expertise, but did not seem to need as much encouragement

to share. Police participants' confidence in their understanding of how use cases are deployed in practice could translate to confidence talking about responsible AI use. However, the expertise of community representatives regarding ways in which disproportionality is baked into the criminal justice system did not translate as quickly into confidence talking about the responsible AI use. By the end of the day, all participants appeared to feel confident fully engaging and sharing views. Although seven participants completed the post-workshop survey, none of the answers shared additional use-case-related concerns or indications that participants were not able to fully express their views.

Results of Use Case Discussions

We now report on the analysis of 76 completed worksheets for 13 AI in policing use cases. Not all 6 groups completed all use cases. Use cases were distributed across morning and afternoon sessions, and groups varied in how many they were able to assess within the time available. In addition, not all completed worksheets assigned a risk categorisation.

Theme A: Culture & Workforce

For this theme, there were 3 use cases, all marked between low and medium risk by all groups.

AI in Recruitment. The use of automation in recruitment is on the rise in UK policing. Oleeo, a recruiting platform, claims on their website that their platform is used by 70% of UK police forces (Oleeo 2025). The first use case focused on AI in recruitment of new officers, particularly in risk screening. The main perceived benefits were efficiency and cost savings, and a better experience for the candidates. Another stated advantage was reduction in bias by hiding the candidate's name and gender, also known as "*blind recruitment*".

All groups completed this use case, but only 4 ranked it, assigning medium risk. Participants agreed that diversifying the workforce was mission-critical, but were sceptical of the ability of AI to make headway in this challenge. The existing, non-tech recruitment process was described as "*a dumpster fire, even before AI*". Participants found the concept of "*blind recruitment*" naive and unlikely to lead to meaningful change, a sentiment backed up by the literature (Simons, Adams Bhatti, and Weller 2021). Resistance to blind fairness was a strong common theme across groups, as was the concern that human-in-the-loop mitigations are proposed without clarity on how they would work in practice.

Participants highlighted that training data and assessment criteria reflect existing—non-diverse—norms, which means models can only be trained to identify the types of officers already hired in the past. Participants acknowledged the potential benefits of automation, as there may be real gains in time, money, and candidate experience, but highlighted that this might not translate to genuine progress, i.e., to getting the best potential officers or shaping the force to be what it needs to be to future-proof its relationship with a diverse population. Considering the need for change, there was also a worry that automating the hiring process can make it more rigid and harder to reform. This use case was an example

Individual risk assessment and predictive profiling tools, such as COMPAS (US), and OGRS and OASys (UK), aim to offer an efficient, consistent, and objective risk categorisation with respect to reoffending. They estimate the risk of re-arrest within a fixed time period (e.g., 12 months). The best models achieve nearly 70% accuracy. In the UK tools, race is explicitly excluded as an input but validation takes place for different ethnic groups to ensure comparability and minimise inequity (Moore 2015). Nonetheless, these tools rely on arrest and/or prior convictions as proxy for previous offending (which tends to be higher than official records). But if the probability of being arrested is itself racially biased, or subsequent justice decisions are, tools will reflect and compound that bias (Zilka et al. 2023b).

This use case gathered the most objections. While two groups did not rank it, the other four groups ranked it as an unacceptable risk. Unlike other use cases where objections are rooted in implementation rather than the idea itself, **here the objection is clearly to the idea itself** – “*bias issues are still baked into this*”. Participants also failed to see the practical benefit: “*would it be [able] to prevent reoffending?*” This strong objection is worth noting, particularly as policing rebrands itself as more *predictive* or *precise* as part of its *AI Revolution*. Most tools of this kind focus on re-arrest rather than harm, with no emphasis on rehabilitation, context, or enabling non-reoffending. However, participants acknowledged that the question can be reformulated: rather than predicting who will reoffend, could the question become *what will help this person not to reoffend?* — whether that is housing support, employment, or rehabilitation services. This re-framing is more challenging to implement in practice, but it represents a more constructive and ethically defensible way forward for individualised predictive tools within criminal justice. We note that in the UK, probation risk assessment tools OASys and Asset, include questions on housing, welfare, and benefits with the explicit aim to form part of intervention plans and risk management strategies, which then go on to inform supervision activities (Moore 2015).

Box 1: Strong objection to predicting the risk of re-arrest.

of an overarching theme in the discussions, where participants felt AI is being used as a *shiny fix* for a process with more fundamental problems, without proper evaluations to see whether it brings any real improvements.

AI Review of Body-Worn Camera. This use case was relatively well-received, as it adds accountability and does not replace an existing human process. However, participants noted that the underlying assumption—that all behaviour is recorded—is undermined in practice: cameras are sometimes off, a minimum 30-second buffer exists, and officers, particularly bad actors, may use them selectively. This led to scepticism about the system’s potential added value. However, most groups agreed there is real potential for improving “*trust and confidence*”, and “*standards and professional behaviour*”, especially if the system should be used not only as a *stick* but also as a *carrot*, with “*positive behaviour*” highlighted as well, motivating officers’ acceptance and cooperation. Fundamental questions were raised about how the act of recording might change the behaviour of everyone involved, and whether it could interfere with helpful, open, and empathetic policing. The surveillance implications for the public depend entirely on how the system is used, what its goals are, and what commitments are upheld. For evaluation, monitoring, “*community input*”, and “*dip sampling*” were suggested, though a concern was raised that supervisors may lack the resources to review footage effectively, increasing the risk of over-reliance on AI.

Bias-Auditing Dashboards. Discussions on this use case were briefer than the previous ones, likely due to time constraints. Generally, the prevailing tone was scepticism rooted in a lack of specificity—without understanding how the auditing is done, it is very hard to know whether it can be help-

ful. Some of that scepticism stemmed from the act of quantification itself and the absence of cultural context: “*understanding context is more useful than surface-level disproportionality*”. Several participants suggested limiting the audit to HR “*to keep [it] low risk*” or keeping it project-specific to “*keep [a] tight scope*”. Groups who rated this use case as low risk did so on the basis that it is not a decision-making tool but one that offers an opportunity for reflection and transparency, potentially more useful for monitoring AI decision-making rather than human decision-making. Notably, none of the discussions mentioned the potential for ethical white-washing or other cynical uses of such a tool.

Theme B: Powers & Procedures

Predictive Policing (Hotspot Mapping). Perhaps surprisingly, participants did not object to predictive policing or hotspot mapping per se—the ability to plan and strategise resource deployment was viewed positively. Instead, the objections concerned implementation and usage. Participants highlighted two core issues: 1) the reliance on arrest data as the primary input, and 2) the goal of optimising for enforcement. Arrest data does not faithfully represent harm, victimisation, or community sentiment on feeling unsafe, and crime non-reporting patterns are often systematically biased. Participants echoed concerns from literature (Ensign et al. 2018) about feedback loops, i.e., predictions informing interventions, which generate new data used for further predictions, compounding existing patterns of over-policing of racial minority communities. For almost all groups, discussion of data led naturally to discussion of goals, and a clear consensus that increasing arrests is *not in itself a worthy goal*. Participants feared that this technology supports what might be called a *whack-a-mole model* of crime pre-

vention, noting crime displacement and a lack of connection to wider policing strategy or the social causes of crime. One comment captured this well: there is a need to “*balance between public health and justice approaches*”. On a positive note, participants were optimistic about the potential to use this technology as a resource not just for enforcement but for multi-agency collaboration, strategy, and policymaking, employing wider data sources and a broader range of outputs.

AI-Assisted Processing of Incident Reports. There is a clear appetite for using AI in this context, provided it is deployed safely and responsibly. The motivations articulated by participants—extracting “*patterns and trends*”, informing new officers about cases, and managing large volumes of information—all centre on AI as an assistant supporting human decisions, rather than as an automated decision-maker. However, this is often not the guiding principle in the design of current tools. For AI-generated summaries and Copilot tools, resistance came from several directions: (a) a false sense of efficiency (“*is it quicker or do you have to spend time fixing the products*”, or worse, it may not be verified at all); (b) accountability questions around whether officers can be held responsible for AI-generated reports without undermining the efficiency gains; (c) lack of contextual, cultural, and compassionate understanding; and (d) potential downstream consequences for prosecution and court procedures. Evaluation is a major, non-trivial effort – “*red-team[ing]*”, “*stress-test[ing]*”, “*hallucination*” rates, “*traceability*” of data, and “*information retrieval*” accuracy are all important, but collectively represent a significant undertaking that may quickly become outdated. At its core, the discussion about risk-benefit boiled down to one central question: where, when, and how is AI adoption worth it?

Risk Assessment (Predictive Profiling) & Live Facial Recognition. Individualised Risk Assessment gathered the most objections of any of the discussed use cases. In contrast, the use of Live Facial Recognition was generally well accepted under certain conditions. Both use cases are discussed in detail in Box 1 and Box 2, respectively.

Theme C: Trust & Reconciliation

Sentiment Analysis & Summary of Community Feedback. These use cases are quite similar, with the former focusing on analysing public data such as social media, and the latter on feedback submitted to the police. Both use cases were generally perceived as low to medium risk. The key justification was that the scale of monitoring involved is “*not possible to do manually*”. Therefore, AI analysis can function as an informative layer on top of existing processes rather than replacing them. Described as a useful “*temperature check*” and tool for emerging issues, sentiment analysis was considered acceptable when used to inform rather than directly drive action. The risks identified were consistent across both use cases: the sentiment captured is not a balanced picture and can be skewed by “*keyboard warriors and loud voices*”; the tools could exclude “*some community members*”, e.g., offline people and “*marginalised voices*”; and the tools may struggle with nuance, sarcasm, tone, and

local community context. Evaluation is also a challenge—outputs require comparison to outputs from “*other methods of gathering public opinion of the police*”, e.g., human assessment; also, evaluations should not just be focused on efficiency. The recommendations across groups were to “*explore other data analysis techniques before deployment*”, and to remain clearly aware of the data’s limitations.

Public-Facing AI Virtual Assistant. Classified as low to medium risk, the virtual assistant was considered acceptable when limited to “*non-emergency*”, low-risk, and informative purposes. One group noted that the “*public expect[s] these type of tools*”, and already uses external LLMs to ask questions about information on police websites. A key distinction was drawn between a chatbot that simply directs users to the right part of a website (low risk but also low benefit), and one that engages more substantively, which raises greater concerns around misinformation, safeguarding, and public trust. Essential requirements included: monitoring for effectiveness, “*dwell times and response outcomes*”, and user feedback; a redirection-to-an-operator mechanism; and clear signalling about appropriate use (e.g., do not use in an emergency; do not report a crime through this channel). The concern that “*misinformation can reduce public confidence*” was raised, alongside a cost-benefit question: “*how many people end up calling 101 anyway?*” Accessibility and the ability to detect vulnerability in user queries, something a human operator might more naturally recognise, were also flagged as important design considerations.

Theme D: Safety & Victimisation

Risk Prediction Tools for Victimisation See Box 3.

AI Transcription / Translation of Victim Calls / Statements / Interviews. This use case did not achieve consensus, being rated low risk by one group but unacceptable by another, with several medium risk assessments in between. The group that rated it low focused on compliance with existing regulations (e.g., CPIA²), data security, governance, and officer training – a procedural approach. The group that rated it unacceptable was chiefly concerned with the risk of victims and witnesses being misunderstood, and its potential disparate impact: “*negatively affect the Black community due to cultural misunderstandings*”. This concern about disproportionate impact was shared broadly among the participants, who observed that “*minority groups are already a step behind, so AI increases the discrimination*”. Concerns were raised about where and how transcription would be used, with non-evidential recordings considered more acceptable, and higher-stakes uses, such as evidence presented at court or 999 calls, considered not acceptable. Most groups were sceptical that automated transcription would result in genuine time savings, if done properly: “*how much time is being saved if everything is being checked over?*” Requirements included human verification of output, transparency, consent, and appropriate control for the victim or witness, including the ability to “*stop, start, and [go] back*”. Questions

²The Criminal Procedure and Investigations Act, 1996.

Live Facial Recognition (LFR)—as deployed by the Metropolitan Police (the Met; serving the greater London area)—functions as a real-time aid to help officers locate wanted individuals from a watchlist, with safeguarding as an additional declared benefit. Historically, commercial facial recognition systems were shown to have higher rates of misidentification for Black individuals, and Black women in particular (Buolamwini and Gebru 2018). However, LFR systems have since improved, and the Met claims that their system can be operated at settings where there is no statistically significant difference in demographic performance across groups (Mansfield 2023).

Within our workshop, LFR was considered acceptable by the majority of groups, but only under clearly defined conditions: 1) use only against “*high-harm crime*”; 2) do not use for minor crimes (e.g., “*cannabis use*”); and 3) mitigate racial and “*intersectional*” bias, and check for “*watchlist bias*”. Additional conditions included transparency about who is on the list, where and how the LFR is deployed, and transparency around outcomes and actual bias levels, alongside human oversight, governance, and community consultation. The discussion was more refined than for other use cases, likely due to considerable effort by UK policing to communicate the conditions under which LFR is deployed and to demonstrate positive outcomes. The bias discussion extended beyond technical accuracy to include watchlist composition and deployment locations as potential sources of bias. Evaluation criteria included false and true positives, value of use and outcomes, and bias across multiple dimensions.

After the workshop, in March 2026, Essex Police paused LFR deployments after a commissioned study found significantly different performance for Black individuals (Bland and Verrey 2026). Specifically, it found that using a threshold that prevented false identification of Black individuals resulted in under-policing, i.e., the system misses non-Black wanted individuals more often. This meant fewer false arrests, but a disparate impact on Black individuals on the watchlist. In April 2026, the High Court dismissed a judicial review challenge to the Met’s revised LFR policy, finding it imposed sufficient constraints to meet the ECHR quality of law test; the court acknowledged that a policy authorising discriminatory use could undermine its legality, but was not persuaded this policy does so. Our findings and these events underscore how acceptance of LFR hinges on specific conditions and assurances about its use and impact.

Box 2: Live Facial Recognition deemed acceptable for high-harm offenders.

of accountability were also raised: “*who takes ownership of ensuring that [the transcript] is correct?*”

AI Advice to Decide Whether to Carry an Investigation Forward. Three of five groups classified this use case as unacceptable risk, with the remainder classifying it as medium risk. The most accepting groups saw potential value if used as “*part of a suite of products to support decision-making*”, with the tool providing assistance rather than acting as a *decision-maker* in itself. The groups who found it unacceptable focused on the risk of disparate impact and bias, as well as concerns about gamification and broader ethical concerns: “*seems completely unethical*” and a defiance of “*bare minimum of what policing should be doing*”. Even the more accepting groups expressed concerns about bias, over-reliance, the effect on victims, and whether the tool would deliver true efficiency compared to a human baseline.

Analysis of Risk-Bounding Process

In addition to considerations on the specific use cases, the workshop helped us gain insights into the *process* of establishing risk categories. Participants were asked to classify the use cases; they received no guidance on what to consider, except that it should be in the context of racial bias. Below, we examine the reasoning process that emerged.

Our main finding is that although the explicit framing was racial bias, the reasoning process that emerged was considerably broader. Rather than focusing exclusively on risks

to racial minorities, participants consistently reasoned about whether a tool delivers genuine benefit for everyone, including, but not limited to, marginalised communities. The questions groups asked were therefore not only about who might be harmed, but about whether the tool works, for whom, under what conditions, and with what safeguards. We examine this reasoning process in detail below, and return to its broader implications in the discussion.

Does It Work? The starting point for most groups was whether the tool is actually capable of doing what it claims. In some cases, this was a matter of the technological capability or the availability of good-quality, non-biased data. However, for several use cases—including recidivism and victimisation risk prediction—the objections were fundamental and concerned whether the task itself is tractable: predicting victimisation risk, for instance, is a hard prediction problem regardless of the sophistication of the tool. The question of whether the tool works was therefore not only technical but conceptual – does the use case rest on a sound premise? A sub-question was “*will it effectively help a human decision maker?*” As participants had strong objections to autonomous decision-making, groups consistently questioned whether the tool is enabling the officer to make better judgments and decisions, or effectively deciding for them how to act. This distinction was central to how groups assigned risk: tools perceived as adding an additional layer of information were more accepted, while those perceived at risk of dis-

Risk prediction tools for victimisation—such as Spain’s VioGén (Álvarez et al. 2018; Satariano and Toll Pifarré 2024)—aim to reduce harm by predicting risk in domestic violence cases. However, their accuracy is fundamentally limited, not by technology, but the difficulty of the prediction problem. A particular blind spot is the lack of recording of interventions, which makes cases where violence was successfully prevented incorrectly classified as low risk. In the UK, the checklist-based risk-assessment tool, DASH, was designed to help frontline practitioners identify individuals at highest risk of serious harm, but a series of tragic failures raised concerns that it was missing high-risk victim-survivors, failing to capture dynamic risk, the lived realities of victim-survivors, the cultural nuance affecting minority communities, and has little use in terms of predictive validity (Turner, Medina, and Brown 2019). In 2022, the College of Policing recommended switching to DARA, a tool which has better predictive validity (Messing and Thaller 2013); yet as of August 2025, 20 of 39 UK police forces still use DASH.

Our groups’ views were divided: two rated this use case as unacceptable risk, three as medium risk, and one left it unmarked. Where deemed acceptable, justifications included the tool already being in use in some form, the potential for AI to enable greater scale and consistency, and its value as a “*second voice*” that may reduce human bias. It was emphasised that the tool should function as “*part of a wider toolset / factors for consideration*” to support officer decision-making, as it was “*high risk if used as a sole tool without giving consideration of other factors*”. The subtlety of this assessment was stressed throughout, with cases “*often requir[ing a] nuanced understanding of context to make [a] holistic decision*”. A key requirement was that officers retain full accountability “*to remain accountable for decisions made*”. This was tied directly to bias (“*fear of overreliance, lack of accountability on bias from police*”), and an overall worry that the tool would exacerbate cultural stereotypes: “*Black women are seen as less vulnerable so large bias in training data*”. The importance of participatory development was highlighted: “*bring in victims/survivors - validating questions being asked, gather their perception on this*”, alongside the need for multi-agency response. For evaluation, the need to understand the human baseline was emphasised. However, this is inherently difficult, as successful interventions prevent harm and may therefore resemble misjudgments of risk, making it hard to demonstrate good judgment precisely when it succeeds.

Box 3: Concerns around Risk prediction tools for victimisation exacerbating racial bias and cultural stereotypes.

placing human judgment received higher risk ratings. Participants deemed it crucial that accountability for outcomes and for bias can be meaningfully upheld.

Will It Deliver Genuine Benefit? Even for uses deemed potentially acceptable, groups pressed hard on whether deployment would translate into meaningful benefit in practice. One concern was the false sense of efficiency: groups asked whether time savings are real if AI is used responsibly, i.e., when appropriate verification and correction are accounted for, and whether the tool will actually save money if it opens the organisation up to legal challenges. More importantly, participants questioned whether the tools can deliver benefits beyond efficiency: “*Will this use of AI help alleviate the current structural problems or worsen them?*” Some groups had a more pragmatic approach to considering the risk-benefit analysis, using the flawed status quo as a reference point. Groups considered, given the current process being problematic or already under-resourced, whether AI is likely to help. This was often emphasised in the context of bias: if the officers are biased, can “*giving them AI reduce overall bias, or will it make it worse?*”

Will It Deliver Benefit for Everyone? The third and most consistently applied layer of scrutiny asked whether the benefits of a tool would be distributed equitably, and specifically, whether marginalised communities would share in them, or bear a disproportionate share of the risks. Some of these discussions were rooted in well-known failure

points such as biased data, but much of the discussion focused on how loss of context and nuance around language and culture can lead to disparate impact. Crucially, these questions were framed not only in terms of harm avoidance but also of benefit: would the tool help the force become more equitable? Would it improve outcomes for victims from minority communities? Would it build or erode trust? This inclusive framing meant that racial bias was not treated as a separate checklist item but was woven into the broader assessment of whether the tool delivers on its promises—for everyone, not just those already well-served by existing processes. Questions of community involvement followed naturally: groups asked who defines what *good* means, whether community input has been sought, and whether the change is likely to improve trust and relationships with those most affected by policing.

Measuring Success and Public Benefit. Alongside these questions was a practical one: can we actually measure or estimate whether a tool works, and whether it delivers benefits equitably? This requires, as a starting point, a clear and measurable definition of “*what does success look like?*”, and an understanding of “*what evidence is needed to say that it’s effective?*” A strong condition for acceptance was that use cases be designed to have demonstrable benefits, beyond a simplistic notion of efficiency, and that these are actively monitored, measured, and communicated transparently.

Discussion

This is a critical time for policing, which faces strong competing pressures: the desire and public expectation to deliver a better, more protective service; the urgent imperative to address deep and longstanding racial inequity; and the reality of shrinking resources. AI is frequently positioned as the solution to this tension – a way to do more with less, and a clear win when done responsibly (National Police Chiefs’ Council 2025; Moore et al. 2025). In this framing, efficiency is the primary promise, and concerns around ethics, including racial bias, are treated as blockers to unlocking enormous benefits. The reality, however, is more complex: there are very few, if any, clear examples of AI in policing delivering genuine, measurable public benefit, yet numerous examples of algorithmic tools exacerbating already unacceptable racial disparities (Angwin et al. 2016; Ensign et al. 2018; Sankin et al. 2021; Zilka et al. 2023b). This paper presents evidence that challenges this framing, showing that considering racial bias is not an obstacle but a critical lens to be applied as early as possible when considering AI adoption.

The Importance of Community Consultation

Although many concerns raised by participants align with findings from responsible AI literature, the risk categorisations did not always follow expected lines. For example, hotspot mapping—a use case with well-documented equity concerns (Ensign et al. 2018; Lum and Isaac 2016)—was relatively well received; participants viewed the underlying goal of strategic resource deployment as legitimate and the problems as rooted in implementation rather than the idea itself. Similarly, LFR was generally deemed acceptable, despite considerable attention to racial bias concerns (Buolamwini and Gebru 2018; Radiya-Dixit and Neff 2023). In contrast, recidivism risk assessment, deployed for decades in the UK and worldwide, gathered the strongest objections of any use case discussed, critiquing the idea itself, not only its technical execution. Participants questioned if such tools could deliver benefit to the public, or help with offender rehabilitation. Indeed, these tools were introduced to *reduce* racial bias and redirect low-risk offenders from prison, but evidence shows they have failed on both counts (Kleinberg et al. 2018; Zilka et al. 2023b; Stevenson and Doleac 2024). This distinction—between objections to implementation vs to the premise—is hard to surface by any means other than community engagement, and it reiterates its importance. The workshop also surfaced many cultural and contextual concerns rooted in lived experience, such as adultification and the perception of Black women as less vulnerable.

Racial Equality as a Key to Better Design

Current approaches to responsible AI in policing tend to rely on structured, qualitative evaluation frameworks – responsible AI checklists, impact assessments, and long-form documentation that works through considerations of bias, transparency, accountability, and other principles one by one (e.g., (National Police Chiefs’ Council and PROBABLE Futures 2025; Government Digital Service 2025)). These frameworks serve an important function, but they implicitly treat ethical considerations as a separate line of enquiry

rather than a fundamental lens for deciding whether a use case is worth pursuing at all. The deliberative benefit-risk process we observed in this workshop was strikingly different from this checklist model. Participants did not work through a list of considerations; instead, they reasoned in an integrated way, with many of these considerations flowing naturally towards a set of fundamental questions: does it work? Will it deliver genuine benefit and will that benefit extend to everyone? We found this question-led deliberation process much more reflective of participatory design (Jackson, Crowston, and Østerlund 2023; Labedzka et al. 2026), with the strong underlying vision that a tool that can deliver benefits to marginalised communities will also benefit the police and the public as a whole. This vision mirrors what is known as the curb-cut effect (Blackwell 2017): designing with marginalised users in mind surfaces better questions, and is likely to lead to better outcomes for everyone. We argue that for AI in policing, best practice should be not just designing with marginalised users in mind, but, as much as possible, with them in the room (Hamraie 2017).

People Before Progress

Our findings demonstrate that discussion of AI use cases need not be all-or-nothing (Lawal et al. 2026). Participants were clearly not against technological progress or AI: they ranked only 3 out of 13 use cases with an average risk above medium. In fact, participants in this mixed-stakeholder workshop were more accepting of AI use cases than police professionals alone in a similar study (Kearney et al. 2024). Despite much discussion of human bias and flawed current processes, participants maintained a clear stance in favour of human judgment, consistently distinguishing between tools that inform human judgment and tools that risk displacing it (Elish 2019; Gao et al. 2021). Indeed, designing AI to support rather than supplant human judgment may be the most effective way to realise its benefits while managing the associated risks. Recent work on AI for missing persons investigations illustrates what this may look like in practice: a hybrid system combining LLM-based summarisation with rule-based visualisations and source linking, designed to augment officers’ sensemaking while maintaining autonomy and professional accountability (Labedzka et al. 2026). This shows that AI can be built to genuinely help officers—and, through them, the communities they police—rather than creating a veneer of objectivity while quietly eroding the space for professional judgment.

Conclusion

This work demonstrates that mixed-stakeholder deliberation, centred on racial equity, is both feasible and valuable at the earliest stages of AI adoption in policing. Community representatives brought knowledge that technical review alone cannot replicate – surfacing concerns rooted in lived experience and adding invaluable context. Foregrounding racial bias did not narrow or politicise the deliberation; it deepened it. This reinforces the notion that ethical considerations and participatory approaches should be at the centre of AI adoption — particularly when better public service, not just efficiency, is the true goal.

Acknowledgments

MJ was supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK. AS's time was funded as part of his independent contractor role for the Met Police. MZ acknowledges support from the Leverhulme Trust grant ECF-2021-429, which also covered the costs of the workshop itself. We are very grateful to Abena Akuffo-Kelly, Ashwin Varghese, Dave Hudson, Detective Chief Superintendent Shaun White, Dr Colin Burton, Dr Doirean Wilson, Israr Hussain, Jonathan Vale, Judith Germain, Khi Rafe, Laura Toop, Lizzie Rapinet, Megan Julien-Harris, Michael Maher, Paulette Johnson-Clarke, Sam Darbyshire, Professor Myrtle Emmanuel, and Professor Robert Binns for dedicating their time and sharing their views for this work.

Ethical Considerations Statement

This work involved human participants via an in-person workshop. Attendees took part in a professional or representative capacity rather than as private individuals; data were recorded via group worksheets and conversations were not recorded. All quotations are reported without attribution, with demographic information given only in aggregate. During the workshop, groups were balanced between police and community representatives, with groups remixed between sessions. All participants were provided an anonymous post-workshop feedback channel.

Representation. This workshop does not attempt to represent or reflect the general public opinion, and it should not be read as doing so. Community representatives were recruited through networks connected to the Police Race Action Plan, and most had existing engagement with policing institutions. This was deliberate as it produced participants with deep, specific knowledge of racial disproportionality within and related to policing. However, it also means our participants were people willing to work with the police towards reform. Communities most harmed by policing include those who regard engagement itself as the wrong strategy, and may reject the premise that these tools should be assessed rather than opposed. Those views are absent here, and although this was not deliberate, it is likely a consequence of our recruitment strategy. A workshop constituted differently may have produced different classifications. However, our findings describe many relevant and useful concerns and considerations and the reasoning behind the choice of risk categorisation. We emphasise that our results do not necessarily reflect the range of community opinion, and our participants' acceptances of AI use cases that we report should be read with that in mind.

National context. This work was conducted in the UK, where policing operates under the notion of *policing by consent* and where the vast majority of officers do not routinely carry firearms. Both conditions shape what participants treated as proportionate. Judgements about the acceptable risk, for example of a misidentification, inevitably rest on assumptions about how an encounter is likely to esca-

late. In jurisdictions with routinely armed policing, different foundations for police legitimacy, or different levels of state violence, the same use case will carry different stakes. Our deliberative method is portable; however, the specific risk classifications (Table 2) are not, and should not be transferred across jurisdictions.

Adverse Impact Statement

This paper does not endorse any policing technology. We report how a group of stakeholders reasoned about a set of use cases. We do not claim that the use cases they found acceptable are acceptable, that they are safe or effective, or that they should be adopted. No classification in this paper constitutes a recommendation. Our finding that participants were broadly open to most use cases, if quoted in isolation, could be presented as community endorsement. However, from the findings it is clear that even when use cases were classified as low risk, they were not endorsed and many conditions and caveats were applied to every case. Detaching a classification from those conditions will misrepresent the findings in this paper. We would regard it as a misuse of this work for any police force, vendor, or policymaker to cite it as support for adopting a technology, or as evidence that communities have been consulted about one.

Our contribution is to the question of how these decisions should be made and who should be present when they are, as well as important conditions and concerns.

Positionality Statement

The authorship team approaches this work primarily from an academic perspective, with expertise in responsible AI and algorithmic fairness. One author brings a more practitioner-oriented experience, and another has a background in public policy. All authors believe that racial inequality is a critical issue in policing and the criminal justice system, and this may have shaped the analysis. One author's time was funded through a contractor role with a police force, which we note as a potential conflict of interest. Although members of the PRAP team helped with recruitment and identifying use cases, no police force or funder had editorial control over this paper, reviewed it prior to submission, or had any right of approval over its findings.

References

- Ada Lovelace Institute. 2019. Beyond face value: Public attitudes to facial recognition technology.
- Álvarez, J. L. G.; Ossorio, J. J. L.; Urruela, C.; and Díaz, M. R. 2018. Integral monitoring system in cases of gender violence VioGén System. *Behavior & Law Journal*, 4(1).
- Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. ProPublica.
- Baumer, E. P.; and Lauritsen, J. L. 2010. Reporting crime to the police, 1973–2005: a multivariate analysis of long-term trends in the National Crime Survey (NCS) and National Crime Victimization Survey (NCVS). *Criminology*, 48(1): 131–185.

- BBC News. 2025. Metropolitan Police facial recognition finds only wanted people on lists. <https://www.bbc.co.uk/news/articles/c4gp7j55zxvo>. Accessed: 2025.
- Berk, R.; Bleich, J.; Kapelner, A.; Henderson, J.; Barnes, G.; and Kurtz, E. 2014. Using regression kernels to forecast a failure to appear in court. *arXiv preprint arXiv:1409.1798*.
- Berk, R.; Sherman, L.; Barnes, G.; Kurtz, E.; and Ahlman, L. 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): 191–211.
- Berk, R. A. 2021. Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement. *Annual Review of Criminology*, 4: 209–237.
- 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): 94–115.
- Blackwell, A. G. 2017. The curb-cut effect. *Stanford Social Innovation Review*, 15(1): 28–33.
- Bland, M.; and Verrey, J. 2026. Live Facial Recognition: Accuracy, Watchlists, and Deterrence. Technical report, University of Cambridge. Commissioned by Essex Police. Version 6, 12 March 2026.
- Bosick, S. J.; Rennison, C. M.; Gover, A. R.; and Dodge, M. 2012. Reporting violence to the police: Predictors through the life course. *Journal of Criminal Justice*, 40(6): 441–451.
- Braga, A. A.; Turchan, B. S.; Papachristos, A. V.; and Hureau, D. M. 2019. Hot spots policing of small geographic areas effects on crime. *Campbell Systematic Reviews*, 15(3): e1046.
- Brayne, S. 2017. Big data surveillance: The case of policing. *American sociological review*, 82(5): 977–1008.
- Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.
- Butcher, B.; Zilka, M.; and Weller, A. 2022. Racial Disparities in Arrests for Drug Violations in the US: What Can We Learn from Publicly Available Data? In *ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*.
- College of Policing; and National Police Chiefs’ Council. 2022. Police Race Action Plan: Improving Policing for Black People. Technical report, College of Policing.
- Davis, M.; Popov, S.; and Surlea, C. 2011. Real-time face recognition from surveillance video. In *Intelligent Video Event Analysis and Understanding*, 155–194. Springer.
- Elish, M. C. 2019. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society (pre-print)*.
- Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway Feedback Loops in Predictive Policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 160–171. PMLR.
- European Parliament; and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Entered into force 1 August 2024; phased enforcement from February 2025.
- Gao, R.; Saar-Tsechansky, M.; De-Arteaga, M.; Han, L.; Lee, M. K.; and Lease, M. 2021. Human-AI collaboration with bandit feedback. *arXiv preprint arXiv:2105.10614*.
- Government Digital Service. 2024. Algorithmic Transparency Recording Standard (ATRS): Mandatory Scope and Exemptions Policy. <https://www.gov.uk/government/publications/algorithmic-transparency-recording-standard-mandatory-scope-and-exemptions-policy/algorithmic-transparency-recording-standard-atrs-mandatory-scope-and-exemptions-policy>. Published 17 December 2024.
- Government Digital Service. 2025. Data and AI Ethics Framework. Technical report, Government Digital Service, Department for Science, Innovation and Technology. First published 2018; last updated 18 December 2025.
- Government of the United Kingdom. 2025. AI to help police catch criminals before they strike. <https://www.gov.uk/government/news/ai-to-help-police-catch-criminals-before-they-strike>. Accessed: 2025.
- Greene, T.; Shmueli, G.; Fell, J.; Lin, C.-F.; and Liu, H.-W. 2022. Forks over knives: Predictive inconsistency in criminal justice algorithmic risk assessment tools. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(Supplement_2): S692–S723.
- Hamraie, A. 2017. *Building access*. University of Minnesota Press Minnesota.
- Haque, M. R.; Saxena, D.; Weathington, K.; Chudzik, J.; and Guha, S. 2024. Are we asking the right questions?: Designing for community stakeholders’ interactions with ai in policing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20.
- Harcourt, B. E. 2010. Risk as a Proxy for Race. *Criminology and Public Policy*. Also available as University of Chicago Law & Economics Olin Working Paper No. 535.
- Harrington, C.; Erete, S.; and Piper, A. M. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–25.
- Home Office. 2024. Summary: Accuracy and Equitability Evaluation of IDEMIA Facial Recognition Algorithm for Home Office Strategic Facial Matching. <https://www.gov.uk/government/publications/facial-recognition-technology-tests-national-physical-laboratory/summary-accuracy-and-equitability-evaluation-of-idemia-facial-recognition-algorithm-for-home-office-strategic-facial-matching>. Accessed: 2024.

- Hunt, P.; Saunders, J.; and Hollywood, J. S. 2014. Evaluation of the Shreveport Predictive Policing Experiment. Technical Report RR-531-NIJ, RAND Corporation, Santa Monica, CA. Sponsored by the National Institute of Justice.
- Jackson, C. B.; Crowston, K.; and Østerlund, C. 2023. A participatory modeling framework for algorithmic accountability and transparency. *Government Information Quarterly*, 40(2): 101773.
- Kearney, C.; Hron, J.; Kosci, H.; and Zilka, M. 2024. Beyond Use-Cases: A Participatory Approach to Envisioning Data Science in Law Enforcement. In *ACM Conference on Fairness, Accountability and Transparency (FAccT)*.
- Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2018. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293.
- Labeledzka, P. Z.; Peters, D.; Dudley, J. J.; and Zilka, M. 2026. Human-AI Interaction for Time-Critical Sensemaking in Missing Persons Investigations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*.
- Lawal, T.; Paul, A.; Oswald, M.; and Taylor, A. 2026. Remote AI weapons detection: Reversing ‘stop and search’ to ‘search and stop’? *Legal Studies*.
- Lee, Y.; Bradford, B.; and Posch, K. 2024. The effectiveness of big data-driven predictive policing: Systematic review. *Justice Evaluation Journal*, 7(2): 127–160.
- Lum, K.; and Isaac, W. 2016. To Predict and Serve? *Significance*, 13(5): 14–19.
- Mansfield, T. 2023. Facial Recognition Technology Equitability Study. Technical report, National Physical Laboratory.
- Messing, J. T.; and Thaller, J. 2013. The average predictive validity of intimate partner violence risk assessment instruments. *Journal of interpersonal violence*, 28(7): 1537–1558.
- Mohler, G. O.; Short, M. B.; Malinowski, S.; Johnson, M.; Tita, G. E.; Bertozzi, A. L.; and Brantingham, P. J. 2015. Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 110(512): 1399–1411.
- Moore, C.; Gill, C.; Bliss, N.; Butler, K.; Forrest, S.; Lopresti, D.; Maher, M. L.; Mentis, H.; Shekhar, S.; Stent, A.; et al. 2025. Concerning the Responsible Use of AI in the US Criminal Justice System. *Communications of the ACM*, 68(8): 41–44.
- Moore, R. 2015. A Compendium of Research and Analysis on the Offender Assessment System (OASys) 2009–2013. Technical report, Ministry of Justice, London.
- Muir, R.; and O’Connell, F. 2025. Policing AND Artificial Intelligence. Technical report, The Police Foundation.
- National Police Chiefs’ Council. 2023. Covenant for Using Artificial Intelligence (AI) in Policing.
- National Police Chiefs’ Council; and PROBABLE Futures. 2025. Responsible AI Checklist for Policing. Technical report, National Police Chiefs’ Council.
- National Police Chiefs’ Council. 2025. Artificial Intelligence (AI) Playbook for Policing. Technical report, National Police Chiefs’ Council.
- Oleeeo. 2025. Police Recruitment: Reduce the Time It Takes to Hire Qualified Police Officers. <https://www.oleeeo.com/industries/police/>.
- Oswald, M.; Grace, J.; Urwin, S.; and Barnes, G. C. 2018. Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and ‘Experimental’ Proportionality. *Information & Communications Technology Law*, 27(2): 223–250.
- Radiya-Dixit, E.; and Neff, G. 2023. A sociotechnical audit: Assessing police use of facial recognition. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1334–1346.
- 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. *New York University Law Review Online*, 94: 15.
- Sankin, A.; Mehrota, D.; Mattu, S.; and Gilbertson, A. 2021. Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them. *The Markup*, 2.
- Satariano, A.; and Toll Pifarré, R. 2024. An Algorithm Told Police She Was Safe. Then Her Husband Killed Her. *The New York Times*. Accessed: October 2024.
- Saunders, J.; Hunt, P.; and Hollywood, J. S. 2016. Predictions put into practice: a quasi-experimental evaluation of Chicago’s predictive policing pilot. *Journal of experimental criminology*, 12(3): 347–371.
- Shaw, C. R.; and McKay, H. D. 1942. *Juvenile Delinquency and Urban Areas*. Chicago: University of Chicago Press.
- Shneiderman, B. 2020. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4): 1–31.
- Simons, J.; Adams Bhatti, S.; and Weller, A. 2021. Machine learning and the meaning of equal treatment. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 956–966.
- Stevenson, M. T.; and Doleac, J. L. 2024. Algorithmic risk assessment in the hands of humans. *American Economic Journal: Economic Policy*, 16(4): 382–414.
- Taka, E.; Lawal, T.; Calder, M.; Sevegnani, M.; Kotsoglou, K.; McClory-Tiarks, E.; and Oswald, M. 2025. Mapping the Probabilistic AI Ecosystem in Criminal Justice in England and Wales. *arXiv preprint arXiv:2512.04116*.
- Taylor, A.; Boyle, A.; Sutherland, A.; and Giacomantonio, C. 2016. Using Ambulance Data to Reduce Community Violence: Critical Literature Review. *European Journal of Emergency Medicine*, 23(4): 248–252.
- 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(5): 1013–1034.

Zilka, M.; Ashurst, C.; Chambers, L.; Goodman, E. P.; Ugwudike, P.; and Oswald, M. 2023a. Exploring Police Perspectives on Algorithmic Transparency: A Qualitative Analysis of Police Interviews in the UK. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–19.

Zilka, M.; Fogliato, R.; Hron, J.; Butcher, B.; Ashurst, C.; and Weller, A. 2023b. The Progression of Disparities within the Criminal Justice System: Differential Enforcement and Risk Assessment Instruments. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1553–1569. New York, NY, USA: ACM.

Zilka, M.; Sargeant, H.; and Weller, A. 2022. Transparency, Governance and Regulation of Algorithmic Tools Deployed in the Criminal Justice System: a UK Case Study. In *ACM Conference on AI, Ethics and Society (AIES)*.

Ziosi, M.; and Pruss, D. 2024. Evidence of What, for Whom? The Socially Contested Role of Algorithmic Bias in a Predictive Policing Tool. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM.