PROBabLE Futures

Probabilistic AI Systems in Law Enforcement Futures

The Use of Probabilistic AI in Law Enforcement and the Criminal Justice System of England and Wales

Temitope Lawal, Kyriakos N. Kotsoglou, Evdoxia Taka, Marion Oswald, Michele Sevegnani, Muffy Calder

June 2026

UK Research and Innovation

"we all need to remember why we are there...not to save ourselves time and to be efficient but to get justice for members of the public"

(Interview 21)

Contents

About PROBabLE Futures 5

Executive Summary 6

1. Introduction 14
- 1.1. Background 14
- 1.2. The present study 16
- 1.3. Structure 17

2. Methodology 18

3. The Landscape of AI Deployment 20
- 3.1. Broad experimentation, limited maturity 21
- 3.2. Where AI is delivering value 23
- 3.3. Hidden enterprise AI and the reclassification of automation 26
- 3.4. The commercial market 27

4. The Human in the Loop 29
- 4.1. Trust, oversight, and the limits of the human-in-the-loop construct 30
- 4.2. Automation bias and de-skilling 33
- 4.3. Accountability gap 35
- 4.4. Redistribution of cognition load 37
- 4.5. Workforce readiness 38
- 4.6. Skills deficit 39
- 4.7. Informal use of AI 40
- 4.8. Required assurance beyond 'human-in-the-loop' 40

5. System Integrity Under Uncertainty 41
- 5.1. Performance considerations 42
- 5.2. Data-availability and quality constraints 44
- 5.3. Testing constraints 45
- 5.4. Legal and evidential challenges 46
- 5.5. Structural and institutional barriers 49
- 5.6. Procurement constraints 50
- 5.7. Commercial dynamics 51
- 5.8. Chaining effects 52
- 5.9. Emerging external threats 55

6. Governing for Legitimacy 58
6.1. Governance frameworks 59
6.2. Evaluation challenges and existing infrastructure 61
6.3. Transparency 62
6.4. Funding as a governance lever 64
6.5. Public legitimacy 66

7. Conclusion and Recommendations 70

Bibliography 79

About the Authors 87

About PROBabLE Futures

PROBabLE Futures (Probabilistic AI Systems in Law Enforcement Futures) is a four-year, £4.2M RAi funded Keystone Project led by Northumbria University with Glasgow, Northampton, Leicester, Newcastle and Cambridge Universities.

Our project, working alongside our law enforcement, third sector and commercial partners, is developing a framework to understand the implications of uncertainty and to build confidence in future Probabilistic AI in law enforcement, with the interests of justice and responsibility at its heart. Activities include mapping the AI ecosystem in law enforcement, including the use of large language models; developing guidance, checklist and frameworks to guide assessment of scientific and legal validity, and mock trials with AI outputs as evidence.

PROBabLE Futures is supported by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible AI UK

Cite this report as: Temitope Lawal, Kyriakos N. Kotsoglou, Evdoxia Taka, Marion Oswald, Michele Sevegnani, Muffy Calder "The Use of Probabilistic AI in Law Enforcement and the Criminal Justice System of England and Wales," PROBabLE Futures (June 2026).

Executive Summary

This report provides a detailed account of how probabilistic AI is being introduced, understood, and operationalised across law enforcement and the wider criminal justice system of England and Wales. Drawing on semi-structured interviews with professionals from policing, prosecution, the Bar, government, oversight bodies, and the technology sector, it examines perceptions, their practical use, and the gap between intended design and user experience. The report examines implementation challenges, including the risks arising when AI outputs feed into further automated processes, while also assessing the operational benefits AI can deliver when deployed responsibly. It further considers the funding and innovation structures shaping adoption, evaluates the adequacy of current institutional, evidential, and governance arrangements, and concludes with recommendations for the bodies responsible for overseeing these technologies.

The central finding is twofold. AI systems are already generating significant gains in efficiency, decision-making, and, in some contexts, public protection, particularly where deployment has been problem-led, iterative, and properly evaluated. At the same time, adoption is advancing more rapidly than the institutional, evidential, and governance frameworks needed to support it. The result is an increasing gap between technological capability and the conditions required for its legitimate, reliable, and responsible use.

Key findings

The Landscape of AI deployment

A wide range of algorithmic tools is being experimented with or piloted, yet relatively few have reached operational maturity. Where AI systems have been carefully designed and matched to clearly defined operational problems, they are delivering tangible value—not only in efficiency but also in decision quality, officer welfare, and accessibility. The most consequential forms of AI may, in fact, be the least visible, as they are often embedded within standard enterprise systems and therefore not consciously recognised as AI by their users. The supplier market is dominated by commercial vendors, and the distinction between substantive innovation and the repackaging of existing foundation models[1] is often opaque.

Key finding 1: AI tools are now being trialled or used in a variety of settings, from contact-centre triage and witness statement drafting to redaction, digital forensics, knowledge management, and (automated) facial recognition. Most of this activity is concentrated at the investigation and pre-charge stages of the criminal justice process and remains in early stages of exploration.

Key finding 2: Where AI systems are carefully designed and calibrated to clearly defined operational objectives—rather than being a technology in search of an application—and robustly evaluated and monitored to ensure reliability and validity, they could deliver real value. Officers can save hours of manual work on transcription, redaction, and case review; contact-centre AI call-analysis tools identify vulnerabilities that human teams would otherwise miss; and AI-assisted crime linkage tools have outperformed the expectations of the units that commissioned them.

Key finding 3: Some of the most significant benefits extend beyond mere efficiency, encompassing gains in welfare and inclusion. For instance, AI tools that identify child sexual abuse material reduce officers' direct exposure to traumatic content. Similarly, statement-drafting tools have made tasks accessible to neurodivergent officers who previously found them challenging.

Key finding 4: The most operationally significant AI dimension may also be the least apparent. For many users, AI is experienced not as a separate technology but as a series of incremental modifications to familiar systems and infrastructure. Consequently, AI capabilities embedded within standard enterprise software may not always be recognised as AI by the professionals using them, potentially falling outside transparency frameworks designed to track AI deployment. This issue is compounded by broader definitional ambiguity: what constitutes AI in one institutional context may be understood elsewhere as routine automation, statistical processing, or established operational practice. Where the boundaries of the concept are unclear (as they often are), governance frameworks that rely on identifying AI systems will inevitably understate the true extent of AI use within the institutions they regulate.

Key finding 5: A substantial proportion of AI tools deployed in law enforcement contexts are commercial products developed by external providers. This commercial market comprises a mix of large enterprise vendors and smaller specialist firms. While some vendors are fostering constructive relationships, including peer-to-peer knowledge sharing across police forces, respondents also raised concerns about vendor opacity, supplier-driven timelines, and the difficulty in distinguishing genuine innovation from rebranded foundation models.

The Human in the Loop

The concept of the 'human in the loop'—which underpins much of the current regulatory and procurement architecture—proves, upon examination, to be largely performative, i.e. without substance and the underpinning socio-legal infrastructure. Over-reliance on AI outputs, the gradual erosion of professional skill, and the diffusion of responsibility across complex socio-technical systems combine to create accountability gaps that cannot be resolved merely by the nominal presence of a human decision-maker. Genuinely meaningful human oversight places

demanding requirements on organisations. Such requirements are often unmet because workforce capability has not kept pace with the speed of deployment.

Key finding 6: Meaningful human oversight is achievable, but it depends on deliberate design choices. These include two-stage review processes, interface features that highlight uncertainty, and clear commitments that AI should support rather than replace human judgement.

Key finding 7: The 'human in the loop' construct, as currently applied across much of the regulatory architecture, varies significantly in substance. Where it is well designed, it functions as a genuine safeguard. Where it is not, it risks becoming little more than the presence of a person somewhere in the workflow. This crucial difference is not always apparent to those who rely on the assurance it is intended to provide.

Key finding 8: Over-reliance on AI outputs is a genuine and recurring concern. There is a risk that those who depend on AI tools from the outset may never develop the foundational competencies those tools are intended to support. This concern cannot be dismissed by analogy to earlier technological transitions. The safeguarding function that human actors are expected to perform within AI-assisted systems depends upon humans gaining or retaining capabilities and expertise that such systems may, over time, diminish.

Key finding 9: The accountability gap operates at two levels: individual displacement of responsibility onto the AI system, and systemic diffusion of responsibility across complex arrangements where no single actor is positioned to account for AI-mediated outcomes. Effective accountability therefore requires a distributed architecture spanning the full lifecycle of an AI tool, from design to deployment and review.

Key finding 10: The redistribution of cognitive load is an under-recognised consequence of automation. The removal of lower-intensity tasks concentrates human effort at the more complex end of the spectrum, with real implications for sustained cognitive strain and decision-making quality.

Key finding 11: Workforce capability is developing but lags behind deployment in important respects. Skills gaps are evident at senior leadership, middle management, and frontline levels, and training has not kept pace with the introduction of new tools. In some cases, forces that have invested in approved internal tools have seen a reduction in the unsanctioned use of consumer AI ("shadow AI"), which suggests that providing approved alternatives tends to be more effective than simply banning unapproved ones.

System Integrity Under Uncertainty

Probabilistic AI systems used in criminal justice deal in probabilities, not certainties. This creates problems for a legal system whose foundation is premised on factual determinations requiring decisions, the ability to challenge evidence through cross-examination, and the central role of human testimony. These tensions are compounded by data-quality limitations, unresolved questions of evidential status and disclosure, the emergent practice of chaining AI systems across multiple stages of processing, and the institutional fragmentation of the forty-three-force model. At the same time, external threats—including fabricated evidence, deepfakes, and AI-enabled fraud—are developing at a pace that is likely to exceed current capabilities to detect or contest them in the absence of targeted countermeasures.

Key finding 12: The relationship between AI accuracy and risk may not be linear. This is because as a tool approaches near-perfect accuracy, the people using it may stop checking its outputs, because errors seem so unlikely. As a result, the rare error is more likely to pass undetected (and carry serious consequences) since no one is looking for it anymore. As one respondent put it, at 99.9 per cent reliability, *"very quickly people would stop verifying and the one in a thousand would kill you" (Interview 15)*.

Key finding 13: Errors in large language models, including fabricated citations, are a recurrent concern. The lack of adequate fine-tuning to adapt commercially developed models to the linguistic, legal, and situational specificities of policing produces systematic contextual mismatches rather than isolated errors.

Key finding 14: Data availability and quality are fundamental constraints on the responsible adoption of AI. Police data often consists of partial records shaped by enforcement priorities, and practical barriers include legal restrictions on repurposing live data and the illegality of possessing certain categories of test material (such as child sexual abuse material).

Key finding 15: The opacity and the level of technical expertise required to scrutinise AI-generated or AI-assisted evidence raise concerns for the adversarial process—where one party relies on tools that the other cannot meaningfully scrutinise, the conditions for disclosure and cross-examination are weakened. These concerns are intensified by the risk that AI-assisted workflows could potentially introduce subtle contextual changes into evidential material. The purpose of the criminal justice system must remain in view: as one respondent put it, *"we all need to remember why we are there…not to save ourselves time and to be efficient but to get justice for members of the public" (Interview 21)*.

Key finding 16: The structural fragmentation of policing leads to duplicated procurement, inconsistent governance, and a tendency towards *solutionising*—adopting an AI tool or capability before the underlying problem has been clearly defined. External pressures, including government directives to use AI and commercial influence, drive adoption at a pace that outstrips institutional readiness.

Key finding 17: Procurement processes are simultaneously too rigid for rapid innovation and insufficiently robust for the governance demands generated by AI adoption. To achieve rapid market penetration, commercial providers are able to shape implementation timelines by compressing the time available for structured evaluation and governance. Public bodies often find themselves unable to resist these pressures, which include offers of free proof-of-concept possibilities and the embedding of new AI capabilities within existing contracts and software upgrades. The cumulative effect is to entrench dependency on incumbent suppliers, narrowing the scope for substitution, withdrawal, or iterative governance once AI capabilities have been introduced—with consequences for both innovation and value for money.

Key finding 18: *Chaining*—the sequential operation of AI tools across processing stages—is currently rare but is likely to emerge as a category of risk that single-tool governance frameworks are not yet designed to detect. For example, evidence uploaded to a digital evidence platform may be automatically transcribed by AI as it is reviewed, and that transcription could then feed into a downstream disclosure process that is itself partly automated, with each stage potentially inheriting and compounding errors introduced earlier.

Key finding 19: Fabricated evidence, deepfakes, and AI-enabled fraud constitute emerging threats for which the safeguarding-architecture—encompassing detection, authentication, and training—is not currently in place. A further, consequential threat is that authentic evidence may be presented as fabricated.

Governing for legitimacy

Governance is developing, but unevenly. Promising mechanisms are emerging, such as the use of funding conditions to require transparency and accountability as a prerequisite for investment, alongside centralised evaluation, and benchmarking. These point towards a more coherent framework in which standardised testing and formal approval would precede, rather than follow, operational deployment.

Key finding 20: Internal governance processes for AI adoption are developing. Initiatives such as data ethics impact assessments, harms workshops, cross-agency joint operating procedures, the Responsible AI Checklist for Policing,[2] and the establishment of PoliceAI represent meaningful progress in building national capability for AI evaluation and oversight.

Key finding 21: The cost of independent or specialist internal evaluation is significant and under-resourced. Police forces are often left to choose between proceeding without adequate evidence of effectiveness or abandoning trials whose potential benefits remain undemonstrated.

Key finding 22: Transparency arrangements are inconsistent. The Algorithmic Transparency Recording Standard (ATRS) provides a structured means for public authorities to disclose the AI tools they use, but compliance is not currently mandatory for police forces. A meaningful transparency regime requires not only mandatory participation but records that are substantive enough to enable external scrutiny and accessible to the public.

Key finding 23: AI projects in policing are financed through a combination of central government allocations, police-led innovation funds, and force-level procurement budgets. Conditions attached to central funding (which include requirements for transparency, benchmarking, and adherence to nationally agreed standards) are likely to operate as a significant governance instrument that incentivises responsible AI adoption.

Key finding 24: A contraction in AI investment (i.e. an 'AI winter') is a possibility that some respondents considered likely within the next two years. While a pause could facilitate consolidation, reflection, and institutional learning, it could also destabilise emerging governance structures if funding is withdrawn before sufficient evidence has accumulated and supporting evaluation, capacity-building, and coordination mechanisms have matured.

Key finding 25: Public attitudes towards AI are shaped by institutional trust, perceived accuracy and fairness—not solely by the technical properties of the tools themselves. Public acceptance is strongest where deployment is targeted and specific rather than generalised, and weakest where it is perceived as indiscriminate surveillance.

Recommendations

Based on these findings, the report advances a set of recommendations directed at specific institutional actors. These recommendations are rooted in the experiences and concerns expressed by respondents throughout the study and are organised according to the institutions best positioned to address them. These are listed below and fully set out in *section 7*:

Home Office

1. Use funding as a governance mechanism, not simply a means of supporting AI adoption.
2. Commission a national evaluation programme beyond facial recognition technologies.
3. Address the gap between structural reform and AI deployment in order to reduce duplicative procurement and governance.
4. Develop capability in relation to evidential risk caused by AI manipulation and deepfakes as a national infrastructure priority.

PoliceAI

5. Exercise gatekeeping authority, not merely coordination.
6. Develop a uniform strategy for data quality assessment and management.
7. Define what meaningful evaluation should involve.
8. Develop a risk classification framework for AI tools that is intelligible to end users inspired by the existing intelligence system.
9. Require that AI tools meet minimum design standards for explainability, transparency, and contestability as a condition of procurement.
10. Develop context-specific oversight protocols that govern human interaction with AI outputs at the point of use.
11. Make the PoliceAI public-facing registry mandatory, substantive, and interoperable with the ATRS.
12. Build a national repository of evaluation findings, including failures and withdrawal of tools.
13. Identify the functions requiring retained human competence and develop a centralised training programme accordingly.
14. Address the structural conditions that undermine workforce capability.
15. Rebalance the relationship between policing and technology vendors.
16. Invest in participatory oversight research that moves beyond consultation.

College of Policing

17. Develop tiered AI literacy aligned to role, responsibility, and risk.
18. Treat prompting as a professional skill.
19. Develop Authorised Professional Practice (APP) for the policing functions most affected by AI.

The Crown Prosecution Service

20. Engage with the Ministry of Justice to propose an amendment to the disclosure framework for AI-generated material.
21. Mandate early engagement prior to deploying AI tools that may have evidential implications.

Judiciary

22. Issue a comprehensive Criminal Practice Direction on the procedural handling of AI-processed and AI-generated evidence in criminal proceedings.

Technology Providers

23. Treat transparency as a condition of operating in the criminal justice system.

Research Community

24. Prioritise research on chaining effects, including their data protection implications.
25. Develop the evidence base on AI and the long-term exercise of professional judgement.
26. Conduct a longitudinal follow-up study on the conditions for responsible deployment of AI.

1. Introduction

The integration of probabilistic artificial intelligence (AI) into law enforcement and the broader criminal justice system in England and Wales[3] is arguably one of the most significant institutional developments of the present decade. By 'probabilistic AI', we refer to AI models whose outputs are produced with a degree of confidence rather than as definite outcomes. These systems, for instance, produce risk estimates, predict the next word in a sentence, identify an unknown person in an image, and predict locations of high criminal activity, all with a degree of confidence based on a learning mechanism derived from available past data.[4] AI systems are already being embedded in a range of functions, spanning police–public contact, investigative analysis, evidential processing, and administrative workflows. However, the probabilistic AI systems are expected to have the most significant impact due to their inherent probabilistic nature and the resultant difficulty in tracing how a specific outcome was produced. Therefore, this report specifically focuses on probabilistic AI.

1.1. Background

In the last two decades, there has been a growing body of literature on the use of AI and data-driven technologies in law enforcement, with different disciplines often approaching AI systems from distinct perspectives. Critical socio-legal scholarship has predominantly focused on questions of surveillance, bias, accountability, and civil liberties,[5] while debates on the governance of operational machine-learning risk-prediction tools have commonly focused on the application of human rights, public law and equity frameworks. [6] Adjacent to this, research on the so-called 'uberization' of policing highlights that technological adoption is not deterministic;[7] rather, it is mediated, and at times resisted, by the cultural practices and discretionary routines of police organisations.

The conceptual language used to describe these developments have evolved over time. Earlier accounts of surveillance societies and the 'surveillant assemblage' described how discrete data flows are abstracted from territorial setting, combined and reassembled into virtual 'data doubles' capable of facilitating new forms of monitoring, social sorting, and categorical suspicion.[8] Subsequent studies examined the operational reality of big data surveillance and the shift from intelligence-led policing towards predictive analytics, automated alert systems, and algorithmically mediated forms of decision-making.[9] Within policing practice, AI systems have been utilised across a wide range of functions, including facial recognition, redaction, intelligence and demand forecasting, risk assessment, digital forensics, automated triage, and, increasingly, generative AI applications intended to support investigative

and administrative tasks.[10] Proponents argue that carefully designed AI systems can improve resource allocation, enhance investigative capacity, process vast volumes of information rapidly, and reduce forms of human error and discretionary bias.[11] In increasingly data-intensive environments, predictive and data-driven policing are heralded as essential tools to maximise efficiency and to shift policing from a reactive posture to a proactive, preventative one.[12]

However, debates concerning these developments remain deeply contested. Because predictive models rely on historical crime data, they are highly susceptible to reproducing the biases embedded in past policing practices. Lum and Isaac's critique of feedback effects in place-based prediction, for instance, illustrates how algorithms trained on biased historical police records create self-reinforcing feedback loops that reflect patterns of enforcement rather than actual crime.[13] Related to this, other scholars have examined how algorithmic predictions about crime are implicated in producing the very patterns they purport to describe.[14] As Moses and Chan argue, the practice of predictive policing fundamentally alters the data being collected.[15] When algorithmic outputs drive police deployment decisions, the increased presence of officers in a specific targeted neighbourhood naturally spurs a higher rate of recorded crime or arrests in that area.[16]

At the same time, these concerns are not universally accepted as inevitable consequences of AI deployment. Some perspectives suggest that feedback loops, discriminatory outcomes, and bias amplification can, in certain contexts, be mitigated through careful system design, fairness-aware machine learning techniques, and robust institutional oversight mechanisms.[17] Others argue that debates surrounding predictive policing are more complex than accounts portraying algorithmic systems as inherently discriminatory. Instead, outcomes depend heavily upon the quality of underlying datasets, institutional implementation practices, reporting behaviour, and whether predictive models rely upon highly discretionary arrest records or citizen-initiated calls for service.[18]

Turning from the systems themselves to those who use them, research on human-AI interaction produces a set of findings on operator behaviour. Studies have shown how automation complacency and bias have led operators to over-rely on algorithmic recommendations and use automated outputs as substitutes for independent information processing, resulting in errors of both omission and commission.[19] Fundamentally, the underlying issue is trust calibration. Appropriate reliance depends on aligning human trust with the actual capabilities of the system.[20] Research across different contexts indicates that cognitive economy and an unwarranted belief in technological infallibility drive a persistent automation bias.[21] Alon-Barkat and Busuioc have extended these findings to public-sector decision-making, revealing a more nuanced pattern of selective adherence where

decision-makers are more likely to follow algorithmic recommendations when those outputs confirm pre-existing stereotypes.[22] In the context of Generative AI, specifically Large Language Models (LLMs), Carnat's recent work highlights how automation bias persists and intensifies because the highly anthropomorphic features of chatbots, combined with their capacity to generate plausible but factually incorrect outputs, unduly elevate user trust.[23] Concurrently, treating reliance as a graded phenomenon reflects how susceptibility to AI guidance is usually shaped by an individual's pre-existing positive attitudes toward AI and specific tool designs.[24]

Beyond questions of operator behaviour, literature on algorithmic accountability proceeds from the concept of a 'black box' society, where proprietary secrecy and technical opacity place algorithmic decision-making beyond ordinary mechanisms of scrutiny.[25] Building on this, Busuioc identifies three non-negotiable conditions for meaningful algorithmic oversight: access to information (transparency), the ability to interrogate the logic behind decisions (explanation/justification), and the capacity for consequences and redress.[26] The implementation of these conditions, however, remains contested. The right to explanation under data protection law has attracted criticism as both legally restrictive and technically unsuited to the kinds of harm that complex machine-learning systems produce.[27] Furthermore, operationalising abstract accountability principles within public-sector organisations requires alignment with institutional cultures, resource constraints, and existing workflows.[28] Perhaps, more fundamentally, transparency alone is insufficient where decision-makers, courts, or members of the public lack the necessary technical expertise, time, or systemic power to evaluate algorithmic systems effectively.[29]

Closely related to these accountability concerns, evidence-law scholarship developed rapidly in response to documented technological failures, most notably the Post Office Horizon scandal.[30] As a result, the long-standing legal presumption that computers operate reliably has come under sustained challenge. In parallel, literature on public attitudes draws on procedural-justice theory and its application to policing,[31] where a consistent finding is that public acceptance of AI systems depends less on the technical properties of the tools themselves and more on perceptions of institutional legitimacy, fairness, transparency, and trustworthiness.[32]

1.2. The present study

The adoption of probabilistic AI is largely incremental and, in some instances, operationally beneficial. At its best, AI offers law enforcement agencies new capabilities, such as optimising patrol routes, synthesising massive volumes of unstructured digital intelligence, and reducing administrative burdens.[33] However, the rapid expansion and deployment of these technologies can outpace the development of the legal, institutional, and evidential frameworks necessary for their responsible

governance. This would not merely result in a temporary regulatory gap, but rather a structural misalignment between the inherently probabilistic character of AI systems and the normative foundations underpinning criminal procedure.

Through the accounts of respondents working within law enforcement and the broader criminal justice system, this study provides a detailed snapshot of how probabilistic AI is being introduced, understood, and operationalised. It explores the conceptual understanding of these systems and the challenges and risks of implementation, while also weighing them against the genuine operational benefits of AI and the conditions under which those benefits can be realised responsibly. Drawing on interviews with professionals across policing, prosecution, government, oversight bodies, and the technology sector, the study addresses the following *research questions:*

1. In what ways are probabilistic AI tools used across law enforcement and the broader criminal justice system of England and Wales?
2. How are probabilistic AI tools developed, procured, tested, and evaluated, and what factors shape their governance and accountability?
3. How do law enforcement professionals interact with and act upon AI-generated outputs, particularly probabilistic information, and how does this influence decision-making and risk management?
4. What are the organisational, legal, and ethical challenges associated with interoperability (including chaining effects between multiple AI systems), risk, liability, and the wider governance of AI in law enforcement?
5. What are the anticipated future applications of probabilistic AI in law enforcement, and what systemic implications do they present for law enforcement and the criminal justice system?

1.3. Structure

This report is structured around four themes. It begins by examining the *landscape of AI deployment* and the conditions under which these tools deliver operational benefits. It then turns to the *role of human oversight* and the limitations of existing conceptual frameworks. The third theme addresses the *interaction between probabilistic AI and the evidential and doctrinal foundations of the criminal process,* including the risks associated with system interdependencies. The final theme considers the *emerging governance architecture,* encompassing developments in evaluation, funding, and public accountability. Outcomes from this thematic analysis provide the basis for a set of recommendations aimed at developing a more coherent framework for the responsible use of AI in law enforcement and the broader criminal justice system.

2. Methodology

For this study, empirical data were generated through semi-structured interviews conducted between May 2025 and January 2026. Respondents were recruited via purposive sampling to reflect a wide range of professional roles across law enforcement and the broader criminal justice system. The sample included police officers of varying ranks, civilian police staff, and representatives from criminal justice and governmental organisations, members of the Bar, academic researchers, technology policy think tanks, and commercial technology providers working with law enforcement. To broaden perspectives, snowball sampling was also used; respondents were invited to refer additional qualified experts. Participants signed a consent form, and the study was approved by the Ethics Committee of the Northumbria University.

Informed by the research questions, interviews were conducted using a semi-structured topic guide, as shown in the box below.

Interview topic guide

1. Awareness and use of AI in the criminal justice system
2. Procurement and implementation of AI systems
3. AI reliance, probabilistic nature, and decision-making
4. Interoperability and chaining effect of AI
5. Future applications of AI in the criminal justice system

This approach allowed later interviews to probe themes emerging from earlier discussions, while maintaining comparability across the dataset. Interviews typically lasted between forty-five and sixty minutes. With participants' consent, they were audio-recorded and transcribed. Follow-up interviews were undertaken where initial discussions identified material developments warranting further enquiry.

Transcripts were analysed using a hybrid coding framework.[34] This combined top-down (deductive) codes, derived from the project's research questions, with bottom-up (inductive) codes generated directly from the data. The framework was developed iteratively across four (4) overarching themes, with sub-themes refined through cross-coder discussion. The analysis is supplemented by primary policy materials, including the NPCC AI Strategy[35] and AI Playbook[36]; the PROBabLE Futures/NPCC Responsible AI Checklist[37]; the Home Office consultation on the regulation of facial recognition and related technologies,[38] together with our

project's response;[39] our submission to the Ministry of Justice's consultation on software-generated evidence;[40] and the Home Office police reform white paper, which, among other things, established a National Centre for AI in Policing (PoliceAI).[41] Academic literature is incorporated where it bears directly on the empirical findings.

This study has methodological limitations. The first is the absence of probation service representation in the interview sample. Despite using both purposive and snowball sampling, no representative from the probation service agreed to participate. This is worth noting, given the centrality of probabilistic risk assessment tools in that context, including the Offender Assessment System (OASys)[42] and its proposed replacement, the Assessment of Risks, Needs and Strengths (ARNS) project. The ARNS project has been piloted since December 2024 and is scheduled for national rollout later in 2026.[43] Publicly available reporting has raised serious concerns about the development and accuracy of these tools,[44] and other work on AI in probation[45] provides further, albeit still limited, information. The second limitation concerns the scope of the sample more generally. Although respondents were drawn from a diverse range of roles and institutions, the sample cannot be taken to represent the full spectrum of professional perspectives across law enforcement and the broader criminal justice system. The findings are accordingly presented as analytically substantive rather than statistically generalisable.

The Landscape of AI Deployment

"[the] definition [of AI] shifts over time quite a lot and we need to be careful exactly what we're talking about" (Interview 15)

"[t]he vast majority of [AI] tools [...] being deployed by the police, is open source or commercial built by someone else" (Interview 3)

"AI is like teenage sex: everyone says they're doing it, but no one's really doing it" (Interview 19)

"those integrated AI features of existing enterprise tools are probably going to be the biggest impact [...] even if they're probably going to be the least noticed" (Interview 13)

"it's very difficult to be a clever, a smart buyer, because every vendor is coming to you and saying that their tools are AI enabled [...] I have these frustrating conversations with sales people where they basically refuse to tell you what's under the hood" (Interview 13)

3. The Landscape of AI Deployment

The *first* theme explores the AI landscape in criminal justice. This landscape is characterised by a proliferation of pilot projects, proof-of-concept trials, and exploratory discussions, with only a small number of tools having reached operational maturity. Much of what is marketed as 'AI' proves, on closer inspection, to be neither especially novel nor meaningfully transformative. There are obscure and institutionally embedded components; indeed, some of the most consequential forms of AI are effectively invisible to their users, subsumed within enterprise systems or no longer recognised as AI due to their routine and long-standing deployment. The specific conditions under which AI does, and does not, generate operational value emerge from the interviews and are presented here.

3.1. Broad experimentation, limited maturity

Key Finding 1

AI tools are now being trialled or used in a variety of settings, from contact-centre triage and witness statement drafting to redaction, digital forensics, knowledge management, and facial recognition. However, most of this activity is concentrated at the investigation and pre-charge stages of the criminal justice process, and remains in early stages of exploration.

The interview data indicate that AI is being explored across a far wider range of applications than is commonly acknowledged, yet the depth of deployment remains markedly shallower than this breadth of activity might suggest. The present landscape is thus characterised less by entrenched operational systems than by a proliferation of pilot projects, proof-of-concept trials, and exploratory initiatives. Only a limited number of tools have progressed to a level of operational maturity sufficient to sustain routine use. As one respondent observed, with deliberate bluntness, AI in policing is *"like teenage sex: everyone says they're doing it, but no one's really doing it" (Interview 19)*.

The applications described by respondents fall into several functional categories. Tools designed to reduce administrative burden constitute the most prevalent category, with Microsoft Copilot identified as the most frequently used enterprise application. Respondents reported its deployment for meeting summarisation, drafting terms of reference, and email management *(Interview 21)*. A further example is a correspondence drafting tool used within a government department, which extracts information from its case-management system to populate letter templates for victim correspondence. This system also records the extent of post-generation amendments made by reviewers, which is used as a performance

metric *(Interview 21)*. Beyond administrative automation, respondents described a broader set of applications spanning contact management, public engagement, witness statement drafting, digital forensics, automated redaction, AI-augmented knowledge work, and object and facial recognition.[46] However, most of the identified projects remain at an exploratory or pilot stage, a finding consistent with our project's preliminary mapping of the probabilistic AI ecosystem.[47] One respondent characterised the majority of tools currently in production as commercial off-the-shelf systems, with the remainder *"in multiple stages of low TRL levels of alpha still testing" (Interview 19)*. Overall, a substantial proportion of current use is concentrated at the investigation and pre-charge stages of the criminal process, while downstream applications in court and post-conviction contexts remain comparatively underdeveloped *(Interview 25)*.

Table 1: *Categories of AI applications in law enforcement across England & Wales emerging from the interviews*

Category	Definition	Application
Administrative automation	AI applied to routine operational tasks to reduce workforce demand	Transcription, translation, automated redaction, correspondence drafting
Investigative support	AI applied to the analytical work of criminal investigations	Digital forensics natural language processing, device data extraction, data integration, pattern matching, crime linkage analysis
Public-facing tools	AI deployed at the interface between policing institutions and the public	Contact triage, post-call analysis, real-time call handler support, chatbots, victim portals, enhanced video response
Evidential processing	AI that generates or transforms material entering the evidential chain	Witness statement generation, case file preparation, disclosure schedule automation
Knowledge augmentation	AI that supports professional judgement through retrieval and analysis	Legal research, policy retrieval, retrieval augmented generation tools
Surveillance and identification	AI applied to the detection, tracking or identification of individuals or objects	Facial recognition, object detection, CCTV filtering, automatic number plate recognition
Enterprise AI	AI capabilities embedded in standard organisational software	AI features introduced through software upgrades and contract renewals, integrated meeting summarisation

Among the applications surveyed, automated redaction tools[48] emerged as the most frequently cited example of systems that have progressed from pilot phase to operational deployment at scale. A pilot conducted by one police force in 2023 reported high accuracy and consistently low false-positive rates *(Interview 8)*, and a national rollout is now being pursued through a procurement framework. One respondent indicated that a leading redaction tool is now in use across thirty-two or thirty-three forces *(Interview 22)*. Witness statement generation, by contrast, represents the most legally significant application identified in this study. Several forces are trialling systems that transcribe officer–witness interactions and generate draft statements structured around the "points to prove" for the relevant offence *(Interview 2)*. However, at least one force has had to discontinue its trial of such a tool due to *"...the disbenefits it created in terms of the unused materials [it generated]...[which] led us to start thinking whether or not we were going to see some true efficiencies in using this software" (Interview 2b)*. Similar functionality is being trialled across more than one force, with different commercial providers offering comparable approaches *(Interview 2b)*. While the potential for faster statement production and reduced administrative burden is promising, respondents also noted significant risks:

> *"What we don't want is a victim to give us their words in terms of what happens and then a really intelligent AI system create this statement that William Shakespeare could have written, and it's not in the victim's voice" (Interview 2).*

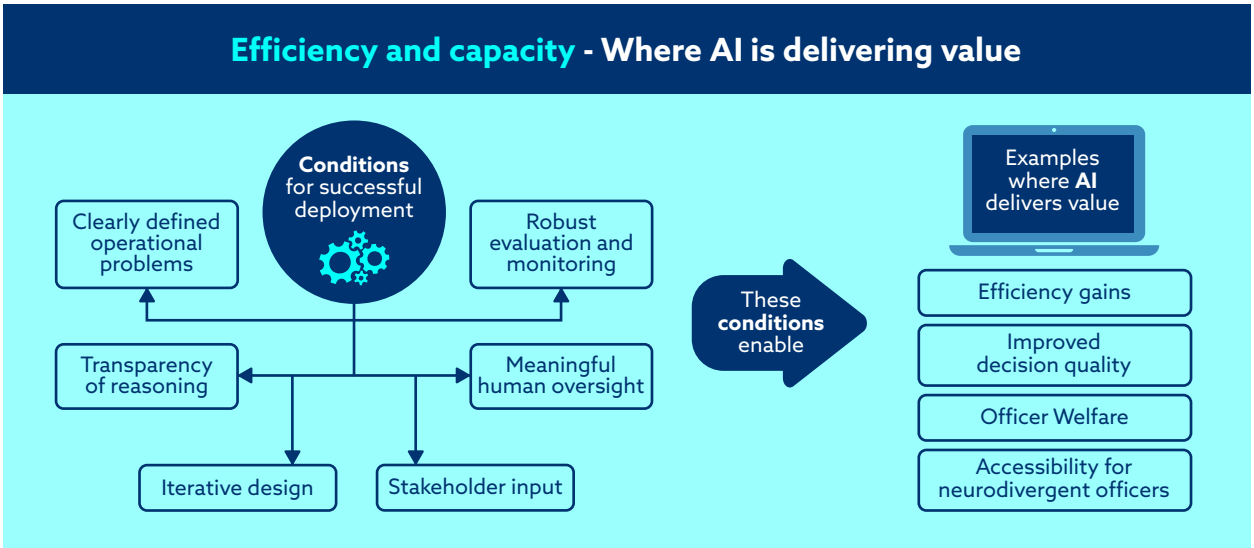
3.2. Where AI is delivering value

Key Finding 2

Where AI systems are carefully designed and calibrated to clearly defined operational objectives—rather than being a technology in search of an application—and robustly evaluated and monitored to ensure reliability and validity, they could deliver real value. Officers save hours of manual work on transcription, redaction, and case review; contact-centre AI call analysis tools identify vulnerabilities that human teams would otherwise miss; and AI-assisted crime linkage tools have outperformed the expectations of the units that commissioned them.

Where systems are carefully designed and aligned with clearly defined operational problems, the benefits are both tangible and, in some cases, substantial. Voice-to-text functionality, for example, allows officers to dictate rather than type, generating meaningful efficiencies in the production of routine documentation *(Interview 2)*. Automated redaction removes hours of *"repeat and very tedious"* manual work *(Interview 7)*. AI-enabled post-call review tools transcribe, categorise, and triage

individual calls, including identifying calls that are *"about something that is not the responsibility of policing and then signposts the caller to the correct organisation" (Interview 20)*. It is then able to aggregate this analysis across the totality of demand to produce intelligence on where capacity is *"being consumed"*, and where *"opportunities to improve the caller experience"* lie *(Interview 20)*. Similarly, AI-assisted crime linkage tools reduce the time analysts spend on case processing, enabling specialist units to respond to a higher volume of serious crime cases *(Interview 23)*. In forensic contexts, respondents pointed to the potential for AI to compress what would otherwise involve forty hours of CCTV review into a targeted fifteen-minute task *(Interview 19)*. These examples underscore a central point: where applications are narrowly scoped, operationally grounded, and deployed against well-specified tasks, AI can deliver measurable gains in efficiency and capacity.



Key Finding 3

Some of the most significant benefits extend beyond mere efficiency, encompassing gains in welfare and inclusion. For instance, AI tools that identify child sexual abuse material reduce officers' direct exposure to traumatic content. Similarly, statement-drafting tools have made tasks accessible to neurodivergent officers who previously found them challenging.

Distinct from these efficiency gains, improvements in decision quality are also evident, and may be of greater significance. A post-call analysis tool, for example, identified indicators of vulnerability in 21 per cent more calls within a reference dataset than had been flagged by human operators. When those calls were subsequently reviewed, the operators accepted that these indicators had been missed *(Interview 20)*. In a domain where undetected vulnerability may carry

life-threatening consequences, this constitutes a benefit of a qualitatively different order from administrative efficiency alone. Similarly, the crime linkage tool developed for a specialist serious crime unit achieved accuracy levels exceeding the expectations of its developers, to the point that its operational value was not subject to internal dispute. These findings suggest that, in appropriately bounded contexts, AI can contribute not only to speed and scale, but also to the substantive quality of decision-making *(Interview 23)*.

Beyond efficiency and decision quality, two further benefits identified in this study warrant separate attention, as they hold direct practical and human significance and do not fit neatly within either of these categories. The first concerns officer welfare in investigations involving child sexual abuse material. Tools capable of identifying such material on seized devices can significantly reduce the volume of distressing imagery officers are required to review *(Interview 16)*. Given the well-documented risk of secondary trauma[49] associated with such exposure, any system that reliably limits this burden—without compromising evidential integrity—represents a substantial occupational benefit. The second relates to accessibility for neurodiverse officers. In one trial of a statement-drafting tool, a dyslexic officer described it as *"really beneficial"* in enabling them to produce statements that had previously been difficult to complete *(Interview 2)*. Where AI systems can reduce functional barriers of this kind, the benefit accrues both to the individuals concerned and to the organisations within which they operate, by allowing existing professional competence to be more fully realised.[50]

It is significant that the conditions under which these benefits arise are, in large part, identifiable in advance. The successful deployments examined in this study did not begin with a technology in search of an application, but with clearly specified operational problems. They were developed iteratively, with sustained input from relevant stakeholders, and incorporated accuracy and transparency of reasoning as design features rather than retrospective assurances. Crucially, they treated human oversight as a substantive safeguard rather than a formal designation—a point discussed further in section 4 of this report. The implication is that where these conditions are absent, the resulting deficiencies are not best understood as inherent limitations of the technology itself, but as consequences of choices made at the stages of procurement, design, and implementation.

3.3. Hidden enterprise AI and the reclassification of automation

Key Finding 4

The most operationally significant AI may also be the least apparent. This is because AI is often encountered not as a discrete tool, but as an incremental enhancement to otherwise familiar infrastructure. Consequently, AI capabilities embedded within standard enterprise software may not always be recognised as AI by the professionals using them, potentially falling outside transparency frameworks designed to track AI deployment. This issue is compounded by broader definitional ambiguity: what constitutes AI in one institutional context may be understood elsewhere as routine automation, statistical processing, or established operational practice. Where the boundaries of the concept are unclear, governance frameworks that rely on identifying AI systems will inevitably understate the true extent of AI use within the institutions they regulate.

Although bespoke tools attract the greater share of public and policy attention, the most significant impact of AI in practice may arise not from these systems, but from features embedded within standard enterprise software. As one respondent observed, *"those integrated AI features of existing enterprise tools are probably going to be the biggest impact [...] even if they're probably going to be the least noticed" (Interview 13)*. In this sense, the most operationally consequential forms of AI may also be the least visible. Some organisations have sought to address this issue through systematic checks at each software upgrade and contract renewal, aiming to preserve visibility over the AI functionalities introduced into existing systems *(Interview 18)*. This reflects an emerging recognition that AI is frequently encountered not as a discrete tool, but as an incremental enhancement to otherwise familiar infrastructure. The implications for governance are significant. If the most consequential AI is precisely that which professionals do not consciously recognise as AI, then governance frameworks that depend upon the identification and registration of discrete AI systems are likely to materially understate the extent of AI use within the institutions they are intended to regulate.

Relatedly, the problem of reclassification further compounds these difficulties. One respondent highlighted automatic number plate recognition as an instructive example: although the system reproduces a form of machine-based inference, it has been in use for so long that *"it's not considered to be artificial intelligence" (Interview 13)*. In this way, what counts as AI in one institutional context may be understood elsewhere as routine automation, statistical processing, or established operational practice. This has direct implications for any governance framework that depends upon the conscious identification of AI systems by users or deployers. This ambiguity is reinforced by the conceptual heterogeneity evident across the interviews. Definitions of AI varied considerably among respondents. Among those offering

more precise distinctions, one characterised AI as systems capable of being *"trained to make connections across data that replicates a human type decision"*, as opposed to models that *"take static lumps of data and throw out patterns"* without such a trained component *(Interview 24)*. Others were more explicit about the limits of definitional clarity, with one respondent noting that *"[a] lot of people get confused around automation and AI and I'm still trying to get my head around that as well" (Interview 18)*.

A common concern was that 'AI' has become a vacuous term, encompassing everything from rule-based automation to generative language models. This semantic expansion creates instability in discussions, particularly concerning risk and governance. As one respondent observed, the meaning *"shifts over time quite a lot"* and *"people now use AI purely to mean generative AI" (Interview 15)*. Others struggled to identify a clear boundary between automation and AI. One suggested that *"[o]nce you get into the thinking and the reasoning you're straying into AI" (Interview 21)*, while another accepted that the concept would remain *"a movable feast"*, cautioning against getting *"hung up on semantics" (Interview 13)*. Within this broader conceptual uncertainty, a particularly salient distinction emerged between narrow- and general-purpose forms of AI: Artificial Narrow Intelligence (ANI) as opposed to Artificial General Intelligence (AGI). One respondent described organisational efforts to develop a language differentiating between *"a more controlled and constrained application of AI where the results are more binary"* and systems *"where you could ask it anything and it would give you any answer" (Interview 14)*. In practice, they noted, their organisation relied primarily on *"models that do very specific things and only that thing" (Interview 14)*. This distinction between narrow and general-purpose AI has direct implications for risk classification:[51] governance requirements cannot be applied uniformly across all AI systems; they must be calibrated to the design, scope, and integration of the specific tool in question.

3.4. The commercial market

Key Finding 5

A substantial proportion of AI tools deployed by in law enforcement contexts are commercial products developed by external providers. This commercial market comprises a mix of large enterprise vendors and smaller specialist firms. While some vendors are fostering constructive relationships, including facilitating peer-to-peer knowledge sharing across police forces, respondents also raised concerns about vendor opacity, supplier-driven timelines, and the difficulty in distinguishing genuine innovation from rebranded foundation models.

Turning from definitional to structural concerns, a substantial proportion of AI tools deployed in law enforcement contexts are commercial products developed by external providers. As one respondent put it, *"[t]he vast majority of [AI] tools [...]*

being deployed by the police, is open source or commercial built by someone else" (Interview 3), a view consistent with our project's preliminary mapping exercise.[52] The market itself comprises a mixture of large enterprise vendors and smaller specialist firms supplying tools for transcription, witness statement generation, redaction, contact management, and investigative analysis. Reliance on commercial providers introduces a recurrent set of dynamics, including vendor opacity, supplier-driven implementation timelines, and the practical difficulty of being a "smart buyer" in a market where almost every product is presented as AI-enabled *(Interview 13)*.

Several respondents expressed concern about the quality and verifiability of vendor claims across a range of supplier relationships. One described the challenge of being "a smart buyer" when sales representatives "basically refuse to tell you what's under the hood" *(Interview 13)*. Another observed that many products marketed to policing amount, in substance, to rebranded applications of existing foundation models, with limited evidence of meaningful fine-tuning or original development *(Interview 7)*. A more pointed critique questioned whether some prominent suppliers should properly be characterised as AI companies at all *(Interview 19)*. It would nevertheless be inaccurate to present this landscape as uniformly problematic. At least one vendor respondent described a structured engagement model in which learning from individual force deployments is systematically shared across forces and through national bodies, supported by a deployment blueprint and mechanisms for peer-to-peer knowledge exchange *(Interview 14)*. This approach illustrates the potential role of suppliers in strengthening, rather than weakening, the wider governance ecosystem, where incentives are appropriately aligned. It also suggests that the difficulties identified above are not inherent to the commercial relationship per se but are more closely associated with the absence of institutional structures capable of channelling that relationship towards cumulative learning and shared standards.

The Human in the Loop

"policing is designed to still be delivered by a human [...] We are so far away from taking away that autonomy" (Interview 5)

"I want us to ban the term 'human in the loop 'cause I think it's entirely meaningless and becoming more meaningless as we move forward" (Interview 19)

"just because AI's advancing at a breathtaking pace doesn't take away our responsibilities to at least understand how does it work and what are the risks of using it" (Interview 5)

"It's [human in the loop] become a bit of a buzz word within marketing as a safeguard" (Interview 17)

"please remember this: AI is designed to offer you advice and guidance, it's not to make a decision for you... you are responsible for any actions or decisions you make" (Interview 12)

"overreliance on this technology will cause atrophy in skills that would otherwise have been used" (Interview 3)

4. The Human in the Loop

The *second* theme concerns how AI is mediated by the humans who work alongside it. The "human in the loop" has become the dominant frame through which law enforcement and the broader criminal justice system seeks to reassure itself, the public, and the courts that AI deployment remains under meaningful control. The interview findings suggest, however, that this framing performs more rhetorical than operational work. The conditions under which human oversight is genuinely substantive, rather than merely nominal, are demanding and often unmet. Automation bias emerges not as an individual failing, but as a structural feature of human–machine interaction. The level of workforce capability required to render oversight effective lags behind the pace of technological deployment. At the same time, the surrounding accountability architecture diffuses responsibility in ways that no single operator is positioned to correct.

4.1. Trust, oversight, and the limits of the human-in-the-loop construct

Key Finding 6

Meaningful human oversight is achievable, but it depends on deliberate design choices. These include two-stage review processes, interface features that highlight uncertainty, and clear commitments that AI should support rather than replace human judgement.

The capacity of a human reviewer to exercise meaningful oversight depends, in the first instance, on whether there is a sound basis for trusting the AI system's outputs. Concerns about trust, as articulated by respondents, centred on questions of accuracy, bias, opacity, and the robustness of existing quality assurance processes. One framed the issue in terms familiar from forensic practice, cautioning against *"over assurance of what you're being told is right"*, and observing that the revalidation and verification procedures routinely applied in forensic science are not consistently replicated in the context of AI *(Interview 8)*. Another raised the question of evidential provenance, noting that even where an AI-generated witness statement is formally endorsed by both officer and witness, *"the AI is ultimately the tool that was writing it" (Interview 2)*, and that the evidential chain ought to reflect that fact. A third emphasised that the pace of technological development does not *"take away our responsibilities to at least understand how does it work and what are the risks" (Interview 5)*. The point that connects these observations is that if trust in an AI system is assumed rather than established—through testing, verification, and an honest account of what the AI system can and cannot do—then the person reviewing its outputs has no real basis on which to challenge them. Oversight without that foundation is oversight in name only.

Respondents also identified a distinct *psychological* dimension to trust. One legal practitioner illustrated the risks of both *"over trust and under trust"* through a hypothetical scenario: an individual who has *"[lost] a run of cases"* or *"been told off by their manager"* is asked to act as a human check *(Interview 17)*. In such circumstances, diminished confidence in one's own judgement is likely to produce deference to the system's output rather than meaningful scrutiny *(Interview 17)*. A related observation concerned the social dynamics through which trust is formed. As one respondent noted, confidence in a tool *"depends who's promoting it"*, and endorsement from a respected figure may lead others to conclude that if the tool is good enough for that individual, it is also good enough for them *(Interview 25)*. Professional trust in AI is thus mediated by trust in professional authority. This has direct implications for governance design. Where assurance frameworks depend on individual professional judgement, they must account for the social and reputational structures within which that judgement is exercised. Absent such recognition, the formal presence of a *"human in the loop"* may do little to ensure that oversight is, in fact, independent or critical.

Key Finding 7

The** 'human in the loop' **construct, as currently applied across much of the regulatory architecture, varies significantly in substance. Where it is well designed, it functions as a genuine safeguard. Where it is not, it risks becoming little more than the presence of a person somewhere in the workflow. This crucial difference is not always apparent to those who rely on the assurance it is intended to provide.

Within this domain, the concept of 'human in the loop' emerged as one of the most contentious ideas. Some respondents described arrangements that approximated meaningful oversight in practice. These included a letter-drafting process incorporating a two-stage review: one officer reviews and amends the AI-generated draft, followed by a manager conducting a second review before the letter is issued *(Interview 18)*; a senior policing respondent who emphasised that the function of policing is *"designed to still be delivered by a human"*, with any removal of autonomy *"not on our roadmap at all" (Interview 5)*; and a prosecutorial respondent who would be *"extremely concerned"* were a computer-based system permitted to initiate a prosecution *(Interview 10)*. In at least one AI product, this principle was operationalised at the interface design level, with outputs accompanied by explicit warnings that the system was intended to provide advice and guidance rather than determine outcomes, and that a defence of *"AI told me to do it"* would not *"cut it"* in court *(Interview 12)*. These examples indicate that meaningful human oversight is achievable. They also demonstrate, however, that it is not secured merely by formally inserting a human into the process. Instead, it depends upon deliberate institutional and technical design choices that structure how human judgement is exercised in relation to system outputs.

Other respondents were more direct in their criticism of the phrase as it is currently employed. A police officer with governance responsibilities suggested *"ban[ning] the term 'human in the loop'"* altogether, on the basis that it had become *"entirely meaningless"* and reflected *"an incredibly lazy approach to technically assuring these products are good enough" (Interview 19)*. The concern, in this account, is not with the underlying principle of human oversight, but with its reduction to a formulaic label that substitutes for substantive assurance. A legal practitioner offered a concrete illustration: in the case of a transcription tool, human oversight consisted of listening to audio while viewing AI-generated subtitles. This, they argued, was *"completely inadequate because of confirmation bias"*, since the reviewer is *"literally being told what the words are that you're hearing" (Interview 17)*. In such circumstances, the structure of the task itself inhibits independent verification. The same respondent recalled a developer's warning that the *"human over the loop quite easily becomes your human scapegoat for poorly designed systems" (Interview 17)*. They further criticised the rhetorical function of the phrase in vendor marketing, where it has become *"a bit of a buzz word"* signifying little more than the absence of full automation *(Interview 17)*. The findings presented here demonstrate that the term, as currently used, provides a false sense of assurance rather than a genuine safeguard. The solution, however, is not to abandon the concept of human oversight—which remains essential to the responsible use of AI in law enforcement and the broader criminal justice system—but to specify what it must entail in practice: the training, tools, time, decisional authority, and system design features without which oversight is nominal rather than real. The recommendations in section 7 set out what those conditions should look like in practice.[53]

The conditions for meaningful oversight are therefore considerably more demanding than the 'human-in-the-loop' label suggests. They include, first, the operator's possession of sufficient foundational competence to recognise when a system is malfunctioning or producing implausible outputs.[54] Second, they require interface design features that highlight uncertainty rather than suppress it. As one respondent noted, the provision of *"confidence scores, measurements of uncertainty in an accessible way"* is critical; without such features, *"the operator is put in a position where they're supposed to be exercising human oversight… [but] they're not empowered or able to do that" (Interview 17)*. Third, meaningful oversight depends on the mitigation of cognitive and organisational constraints—such as fatigue, isolation, and diminished confidence—that incline reviewers to defer to system outputs rather than interrogate them. Taken together, these requirements point to a substantial programme of design and institutional work, further outlined in the recommendations section of this report. One respondent offered a practical illustration of what this programme of design work might look like in practice. For instance, a post-call analysis tool being piloted in two police forces was described as incorporating what the respondent termed "transparency of reasoning"—the system explains in plain English why it has categorised an element of a conversation in a

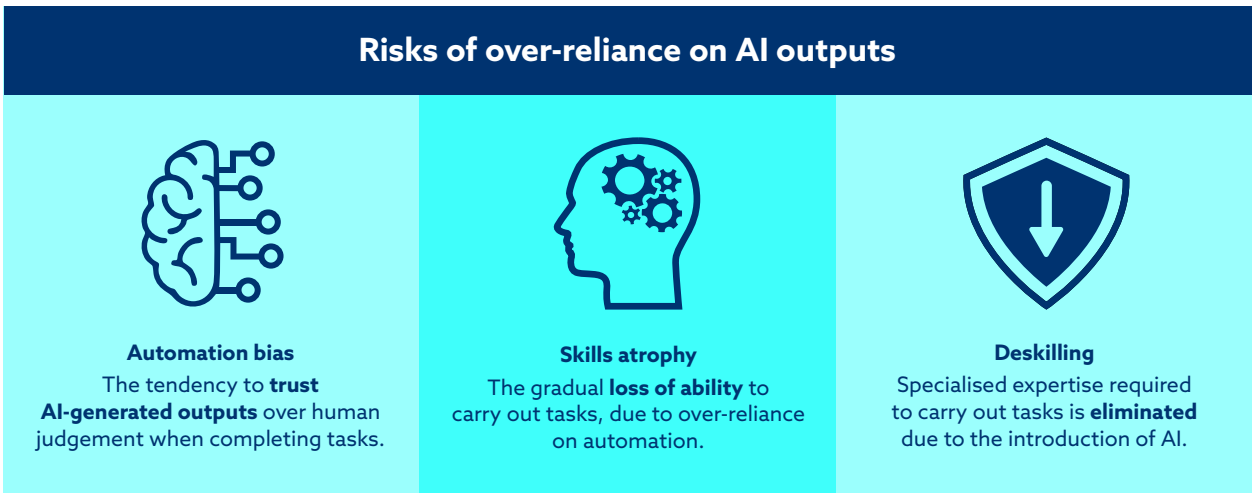
particular way—alongside *"transparency of accuracy"*, which flags outputs in relation to the system's own confidence and indicates where uncertainty is present *(Interview 20)*. The tool also includes a feedback mechanism through which users can record disagreement with an output, with cases then routed back into a review process and the resulting data used to refine subsequent performance (Interview 20). These features are notable not because they resolve the challenges identified above, but they demonstrate that the conditions for meaningful oversight can be deliberately built into a system rather than assumed to follow from the presence of a human reviewer. The distinction between the nominal and substantive oversight is, in this instance, a design choice. The implications of that choice for procurement standards and tool design are addressed as part of our recommendations in section 7.[55]

4.2. Automation bias and de-skilling

Key Finding 8

Over-reliance on AI outputs is a genuine and recurring concern. There is a risk that those who depend on AI tools from the outset may never develop the foundational competencies those tools are intended to support. This concern cannot be dismissed by analogy to earlier technological transitions since the safeguarding function that human actors are expected to perform within AI-assisted systems depends precisely upon the capabilities that such systems may, over time, diminish.

The risk of over-reliance on AI outputs—commonly described as automation bias—was raised by nearly every respondent in this study. Automation bias is a well-established cognitive phenomenon in which human operators tend to privilege recommendations generated by automated systems over their own independent judgement.[56] At the individual level, the concern is that officers will defer to AI outputs without adequate scrutiny, leading to losses in situational awareness and a drift towards cognitive passivity.[57] At the organisational level, the concern is more structural: that the progressive automation of tasks previously requiring professional judgement will erode the underlying capacity to perform those tasks when the technology fails or is withdrawn. As one respondent put it, *"overreliance on this technology will cause atrophy in skills that would otherwise have been used" (Interview 3)*. This concern was linked directly to operationally essential functions, including witness statement generation, and to broader questions of institutional resilience and business continuity. Several respondents expressed particular unease about the implications for early-career officers, identifying de-skilling as a most consequential risk. Those who rely on AI tools from the outset may never acquire the foundational competencies these tools are intended to support.



As one respondent argued:

> *"brand new officers shouldn't use [the AI tool]. Initially they should be going out and taking statements in the traditional way just to build that level of experience up with how a statement should be written" (Interview 2).*

The same respondent made the continuity risk explicit:

> *"ultimately what we don't want is further down the line to completely deskill a workforce because if for whatever reason a computer system goes down [...] you want people to be able to pick up a pen and write a statement if they need to" (Interview 2).*

Similar concerns arose in other operational domains, including the maintenance of call-handling capability in circumstances where the AI systems on which such functions increasingly depend may be unavailable. As one respondent put it:

> *"what are you going to do if this AI tool stops working? Have you got sufficient skills and technical capability as a backup to run a call centre as in the old days? Do people know how to answer these calls if it's all being outsourced AI?" (Interview 3)"*

It is, however, important to note that one respondent advanced a countervailing view, drawing on a broader analogy with technological displacement:

> *"we did think about the potential of deskilling and what we thought was it's like if people don't know how to ride a horse and cart anymore that's because we have motorcars...if we don't need to know then it doesn't matter that we're deskilled there" (Interview 4).*

This position reflects a familiar account of technological progress in which older competencies are rendered obsolete without practical loss. Its applicability in the present context is nonetheless limited. It proceeds on the assumption that AI systems will remain continuously available and sufficiently reliable—an assumption that sits uneasily with the operational realities of policing, where system-failure, evidential challenge, or degraded conditions cannot be excluded. More fundamentally, the analogy obscures a critical distinction. As Busuioc has observed, the retention of a human "in the loop" risks becoming a hollow safeguard where the human intermediary lacks the expertise required to identify when a system is malfunctioning.[58] The role assigned to human operators in AI-mediated processes is not merely residual; it is supervisory and, at times, corrective. It therefore presupposes a level of retained competence that ordinary forms of technological displacement do not demand. On this view, the de-skilling concern cannot be dismissed by analogy to earlier technological transitions. The safeguarding function that human actors are expected to perform within AI-assisted systems depends upon the very capabilities that such systems may, over time, diminish.

4.3. Accountability gap

Key Finding 9

The accountability gap operates at two levels: individual displacement of responsibility onto the AI system, and systemic diffusion of responsibility across complex arrangements where no single actor is positioned to account for AI-mediated outcomes. Effective accountability therefore requires a distributed architecture spanning the full lifecycle of an AI tool, from design to deployment and review.

Beyond the loss of capability, respondents identified the more elusive problem of an 'accountability gap', coupled with the practical difficulty of overriding AI recommendations even where professional judgement points in a different direction. As one respondent observed: *"It takes a pretty bold human to decide in that instance not to follow the advice of an AI tool that he or she has been told performs better than human judgement" (Interview 15)*. This hesitation is exacerbated by the fallacy of technological protection: the misplaced belief that the use of computational systems neutralises human bias and confers an unwarranted presumption of objectivity on AI outputs.[59] In practice, the presentation of those outputs often amplifies this effect. Unlike human decision-makers, AI systems typically communicate without the markers of uncertainty—hesitation, qualification, revision—that signal tentativeness in ordinary interaction. The result is an attribution of trust in the system than its actual accuracy justifies.[60] In these conditions, automated tools can function as "moral buffers": because the decision appears

to have been made, or endorsed by the system, users feel less personally responsible for it and are more willing to defer to it on consequential matters.[61] One respondent captured this dynamic in a judicial context, asking at what point a judge, confronted with a highly accurate AI recommendation, would feel able to trust their own judgement over that of a system capable of processing vastly greater quantities of data. The question goes not only to individual psychology but also to the integrity of decision-making under conditions of asymmetrical informational authority *(Interview 15)*. The design implication is that AI systems supporting consequential decisions should present their underlying reasoning (e.g., chain-of-thought) and explanations about it, prediction confidence, and rigorous estimations of the underlying uncertainty if possible, for the human to reason with. Furthermore, workflows should require the decision-maker to construct their own reasoning rather than simply endorse or reject the system's output.[62]

The concerns arising from the accountability gap highlighted above manifest in two forms. At the *individual level*, respondents described a tendency for responsibility to be displaced onto the system itself. As one put it, *"it would be too easy for officers now to say not my fault, the AI said to make that decision" (Interview 9)*. This mode of attribution stands in tension with established administrative law principles, regarding improper delegation and fettering of discretion.[63] At the *systemic level*, the concern is one of *diffused responsibility*, in which no single actor bears meaningful accountability for AI-mediated outcomes. As one respondent observed,

> *"if you're expecting police officers to be computer scientists then your system will fail [...] there needs to be a level in which there's responsibility distribution throughout a system so that the entire weight of the responsibility of understanding where to spot the issue doesn't lie with the end operator" (Interview 17).*

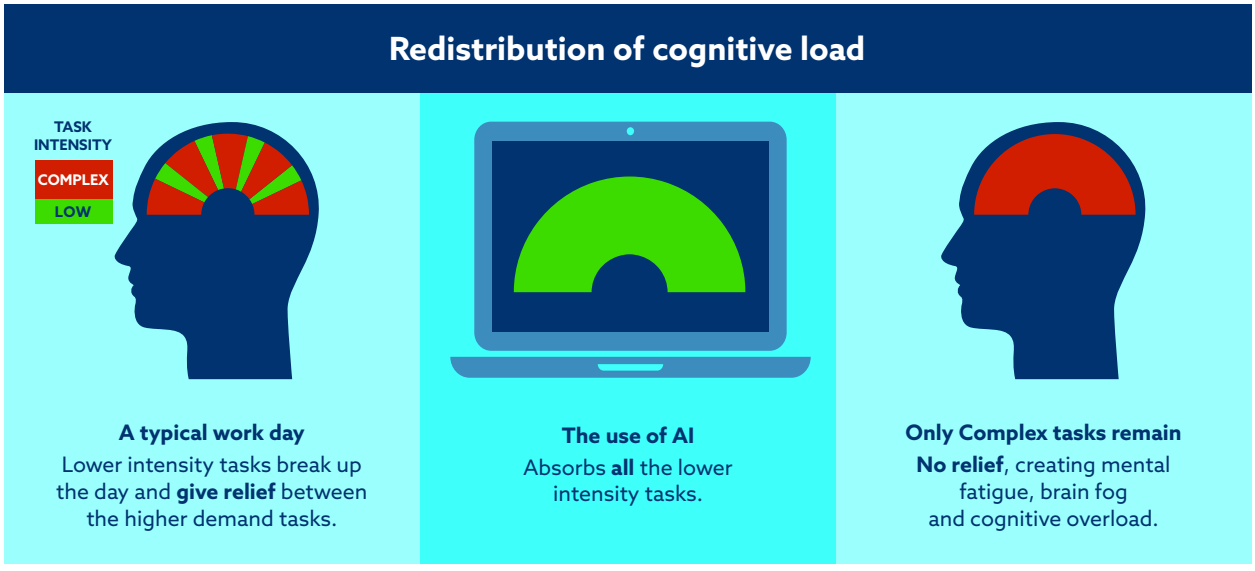
When applied to risk-prediction tools, the same respondent emphasised that an individual officer reviewing outputs *"cannot identify systemic bias in the data nor be expected to be a safeguard alone against discriminatory outputs"*. The point extends further: such an officer may equally be unable to judge whether the AI output is a relevant consideration that should be taken into account in the decision at hand.[64] This is because such a judgement presupposes an understanding of the system's scope, limitations, and known failure conditions—knowledge that no end user can reasonably be expected to possess without institutional support. On this account, effective accountability requires a distributed architecture, with responsibility allocated across mandatory quality assurance processes operating both upstream (in design, training, and validation) and downstream (in deployment, monitoring, and review).[65]

4.4. Redistribution of cognition load

Key Finding 10

The redistribution of cognitive load is an under-recognised consequence of automation. The removal of lower-intensity tasks concentrates human effort at the complex end of the spectrum, with potential implications for sustained cognitive strain and decision-making quality.

A further consequence of routine task automation—distinct from the oversight challenges discussed above but bearing directly on the conditions under which human judgement is exercised —is the redistribution of cognitive load. As AI systems increasingly take over straightforward correspondence and administrative functions, the residual human workload becomes increasingly concentrated at the more complex end of the cognitive spectrum. As one respondent observed, *"the routine things have gone"*, and although this may increase time spent on substantively demanding work, *"those little ones that we sometimes fill the gap [...] give us a bit of a release from the big tasks" (Interview 18)*. The removal of that variety carries its own risk of fatigue. By eliminating these lower-intensity tasks—often overlooked but functionally important—that naturally pace a workday, individuals are subjected to sustained cumulative cognitive strain that can lead to severe mental fatigue, 'brain fog', and cognitive overload, and may ultimately degrade decision-making capacity.[66]



4.5. Workforce readiness

Key Finding 11

Workforce capability is developing but lags behind deployment in important respects. Skills gaps are evident at senior leadership, middle management, and frontline levels, and training has not kept pace with the introduction of new tools. In some cases, forces that have invested in approved internal tools have seen a reduction in the unsanctioned use of consumer AI ("shadow AI"), which suggests that providing approved alternatives tends to be more effective than simply banning unapproved ones.

If the conditions for meaningful oversight are demanding, the workforce capability required to meet them is correspondingly substantial. Respondents consistently identified the skills gap in AI as one of the most significant barriers to responsible adoption. This deficit is not confined to a particular tier but extends across the organisation as a whole. As one respondent put it, *"the kind of skill issues are right across the organisation at all levels [...] at senior level, middle management level and frontline level [...] and also in specialist roles across policing too" (Interview 7)*. Particular concern was expressed in relation to senior leadership. Those tasked with decisions on procurement and governance are often required to exercise judgement in areas for which they lack sufficient technical grounding. As one respondent observed, *"the vast majority of senior police officers have very little technical knowledge when it comes to technology or AI" (Interview 5)*. The result is an institutional vulnerability to supplier-driven narratives and to evaluative processes that are insufficiently rigorous or independently grounded.

Further down the organisation, the challenge is less one of narrow technical expertise than of *interpretive competence*: the capacity to direct AI tools appropriately and to subject their outputs to critical scrutiny. One respondent identified the central governance issue in precisely these terms, pointing to the need for *"interpreting the output of these tools, understanding what is inaccurate [...] and there is no national approach to that at the moment" (Interview 3)*. However, expecting the ultimate user to effectively carry out this interpretation might be unrealistic; instead, tools may need to be designed, tested, and deployed in a way that keeps the human engaged in the drafting and analysis process.[67]

4.6. Skills deficit

Training provision has not kept pace with AI deployment. Even when training exists, some respondents view it as often inadequate, leading to *"a real lack of knowledge and skills around the use of [AI]" (Interview 8)*. In policing, for instance, one respondent noted that available training typically consisted of vendor-provided click-through instructions:

> *"The only training you get to use the systems is the ones provided by the [third party provider] company. And it's very much click here to do this, click here to do that. They don't teach you the trade cart to think about how you could apply it" (Interview 1).*

Another respondent identified *"general digital literacy, AI literacy and training both on general principles and also on the specific tools"* as the significant gap *(Interview 17)*. Embedding new tools in practice has required intensive one-to-one support that organisations have not always anticipated or resourced (Interview 14). The government's One Big Thing initiative—a cross-government programme themed 'AI for All' for 2025[68]—was cited as a positive step, though its impact seemed limited to fostering engagement with Copilot rather than enhancing overall AI literacy *(Interview 18)*. There is also a structural question about the kind of expertise police forces need. One respondent cautioned against the assumption that a single technical lead is sufficient:

> *"you might have one person who's supposed to be the technical expert, but that's never going to be enough because there are several different types of technologists...policing is just starting to get to understand that all of those things are different" (Interview 7).*

The skills deficit is potentially widened when officers are routinely moved across technical domains with minimal preparation: *"you might have been project managing dog training or virtual reality headsets yesterday, tomorrow you're going to be doing large language models...the training is very limited and not helpful" (Interview 9)*. Alongside the skills deficit sits a persistent problem of talent retention that undermines efforts to build internal capacity. Public sector organisations are structurally constrained in their ability to compete with private-sector remuneration for data scientists and AI specialists. As one respondent put it, forces are *"haemorrhaging the really talented people because they're going off to the organisations like your Googles and your Amazons and your Microsofts" (Interview 5)*. The consequence is a pronounced asymmetry: the institutions most exposed to the governance challenges posed by AI are often the least able to retain the expertise required to meet them.[69]

4.7. Informal use of AI

Layered onto these difficulties, the informal use of consumer AI tools—often described as 'shadow AI'[70]—introduces a further dimension of risk. Respondents reported instances of officers using tools such as ChatGPT outside any formal oversight framework, exposing organisations to data leakage and compliance failures. Notably, however, one police force found that deploying Microsoft Copilot internally led to a marked reduction in unsanctioned external use, with staff *"going outside of Copilot"* dropping *"dramatically"* once an approved alternative was available *(Interview 18)*. This suggests that the underlying driver of shadow AI is not deliberate circumvention of policy, but the pursuit of efficiency and innovation in the absence of sanctioned tools.[71] The implication is that prohibition, on its own, is unlikely to be effective. Where legitimate needs are unmet, informal workarounds will predictably emerge. By contrast, the provision of approved, well-governed alternatives appears, arguably, capable of channelling that demand into environments where risks can be managed and organisational standards maintained.

4.8. Required assurance beyond 'human-in-the-loop'

Taken together, the above discussions suggest that process-level disclosure (such as 'a human reviewed this') is *insufficient* as a basis for meaningful assurance. It should be replaced, or at least supplemented, by substantive performance disclosure, including known error rates and the conditions under which errors are most likely to arise. As one respondent put it, officers should be told that a tool *"hallucinates one times out of thirty-three"*, rather than being reassured simply that an output is AI-generated and has been subject to human review *(Interview 19)*. Fundamentally, systems should be designed to challenge end users to exercise 'ownership' of outputs through active engagement with, and construction of, their own interpretation.[72] Furthermore, effective interface design should entail the classification of AI in a similar way to intelligence so that end users know the circumstances in which these systems can be used.[73]

In a similar vein, prompt design could be treated as a structured professional function within policing, supported by validated libraries of prompts rather than ad hoc operator improvisation *(Interview 3)*. None of these requirements can be met through the 'human-in-the-loop' label alone. Each presupposes substantive work at the level of system design, organisational capability, and procurement practice—areas that current deployment cycles have only partially addressed.

System Integrity Under Uncertainty

"There is a massive challenge around [...] the proper use of probabilistic AI for predictive activity because [...] police data by its nature is only that which is connected to the criminal justice processes" (Interview 7)

"we all need to remember why we are there [...] not to save ourselves time and to be efficient but to get justice for members of the public" (Interview 21)

"the tools used by police forces are so fragmented it's bonkers" (Interview 11)

"human dignity needs to be an ever present consideration [...] the way in which data flattens, especially when it's used in a predictive context and when the data of others is used to make decisions about you, not because you've done something but because you're like other people who did something, that's inherently problematic from a human rights and rule of law perspective" (Interview 17)

"[a]nyone that is successful at selling into policing will know sell the solution first and then let them sort out what the problem was after you've sold it" (Interview 20)

5. System Integrity Under Uncertainty

The *third* theme examines what happens when probabilistic AI systems are introduced into a legal system built on the expectation of definitive evidence, the ability to test that evidence through cross-examination, and the central role of human testimony. Whereas the preceding theme focused on the position of the individual user in relation to AI, the present analysis adopts a systemic perspective, considering how these technologies interact with, and potentially disrupt, the institutional logic of the criminal process as a whole. The discussion proceeds as follows: it begins with the performance considerations, data availability and quality, as well as testing constraints that bear upon the reliability of AI outputs. It then turns to the legal and evidential challenges that arise when those outputs are introduced into criminal proceedings. Next, it addresses the structural barriers and procurement dynamics through which AI systems are incorporated into policing. This is followed by an examination of *chaining effects,* whereby sequential AI processing generates compound risks that cannot be captured through the evaluation of individual systems in isolation. To conclude, these developments are situated within a broader threat landscape, including fabricated evidence, deepfakes, and forms of agentic fraud, in relation to which current institutional capacities for detection and rebuttal remain underdeveloped.

5.1. Performance considerations

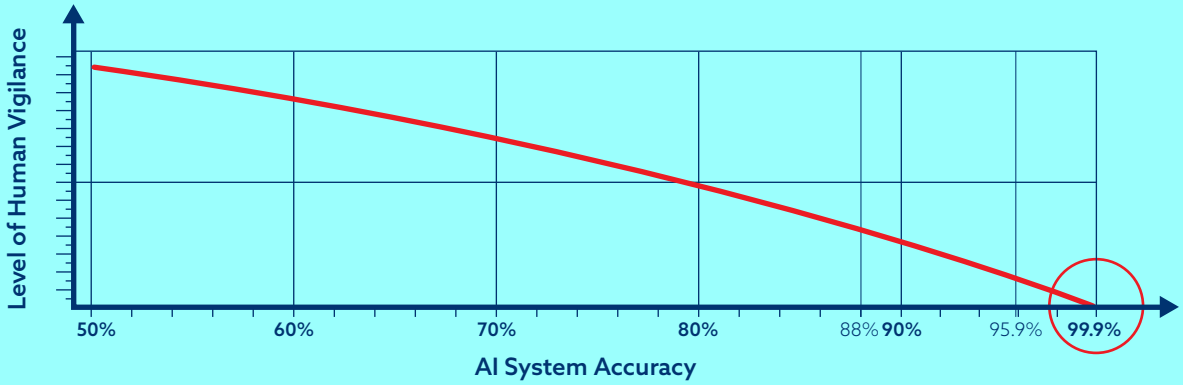
Key Finding 12

The relationship between AI accuracy and risk may not be linear. This is because as a tool approaches near-perfect accuracy, the people using it may stop checking its outputs, because errors seem so unlikely. As a result, the rare error is more likely to pass undetected (and carry serious consequences) since no one is looking for it anymore. As one respondent put it, at 99.9 per cent reliability, "very quickly people would stop verifying and the one in a thousand would kill you" (Interview 15).

The performance of AI systems in policing is uneven with respect to risk, and respondents identified a range of failure modes that resist any simple narrative of linear progress between performance and risk. Errors in large language models emerged as a recurrent concern across multiple organisational contexts.[74] For example, a deployment of Microsoft Copilot was described as exhibiting *"a lot of problems around irresponsible use, because it is"*, as one respondent put it bluntly, *"often hallucinating rubbish" (Interview 3)*. Another recounted an instance in which Copilot generated three full pages describing a concept that did not exist: *"it says it with such authority [...]that you just [think] this thing must exist. No, it doesn't" (Interview 21)*. Of particular concern in legal and investigative settings is the tendency of such

systems to produce fabricated citations. This failure mode is difficult to detect precisely because citation carries an embedded presumption of accuracy within professional practice *(Interview 7)*. In response to these limitations, several respondents described error-management strategies grounded in a presumption of unreliability rather than trust. One noted that they had been trained to *"assume it's wrong"*, treating AI-generated outputs as provisional until independently verified *(Interview 15)*. The same respondent offered a counterintuitive observation: very high levels of accuracy may themselves generate risk.

The paradox of accuracy



At 99.9 per cent reliability, *"very quickly people would stop verifying and the one in a thousand would kill you" (Interview 15)*. This insight reverses the conventional intuition that incremental gains in accuracy straightforwardly reduce risk. Beyond a certain threshold in accuracy, improvements in system performance may erode verification discipline more rapidly than they reduce underlying error rates. Consequently, residual errors, when they occur, are more likely to pass undetected and carry more serious consequences.

Key Finding 13

Errors in large language models, including fabricated citations, are a recurrent concern. The lack of adequate fine-tuning to adapt commercially developed models to the linguistic, legal, and situational specificities of policing produces systematic contextual mismatches rather than isolated errors.

Beyond specific errors, respondents highlighted problems of *contextual mismatch* between commercially developed AI systems—often trained predominantly on US English datasets—and the operational realities of policing in England and Wales. One respondent illustrated this with a concrete example: a US-developed large language model repeatedly classified 'fly-tipping' as animal abuse, *"because that*

phrase doesn't exist in the US and you need to teach it" (Interview 3). The same misclassification was observed in a post-call analysis tool trialled by two separate police forces *(Interview 20)*. That the same error appeared across tools developed by different vendors is significant. It indicates that the underlying issue is not product-specific, but structural, reflecting the US-centric composition of widely used training datasets and the absence of adequate fine-tuning for the deployment environment. Other forms of contextual failure were more subtle but no less consequential. The same post-call analysis tool struggled to distinguish between an immediate threat and one that might materialise in the future *(Interview 20)*. In policing, this distinction is critical: whether a risk is happening now or might happen later determines how urgently officers must respond and what form that response should take. Such failures underscore that effective deployment depends not only on a model's general capability, but on how well it is aligned with the linguistic, legal, and situational realities of the setting in which it is used.

None of this suggests that AI systems should be held to a standard of perfection not applied to human decision-makers. One respondent involved in evaluating commercial AI tools made precisely this point, observing that the difference in performance between the system and the human operators it was designed to assist was, in practice, not especially pronounced *(Interview 16)*. The relevant question is not whether AI is infallible, but whether its performance is acceptable when measured against a realistic and properly specified human baseline rather than an idealised one. That said, this comparative perspective does not diminish the significance of the contextual failures described above. Where errors are systemic rather than occasional—arising from the structural composition of training data rather than individual model limitations—they require a systemic response through fine-tuning, domain-specific evaluation, and procurement requirements that hold vendors accountable for the fitness of their tools within the specific environment in which they will be used.[75]

5.2. Data-availability and quality constraints

Key Finding 14

Data availability and quality are fundamental constraints on the responsible adoption of AI. Police data often consist of partial records shaped by enforcement priorities, and practical barriers include legal restrictions on repurposing live data and the illegality of possessing certain categories of test material (such as child sexual abuse material).

Closely related to the performance considerations discussed above, respondents identified the quality and availability of policing data as a fundamental constraint on system performance. As the wider literature on big data in policing makes clear,

the effectiveness of algorithmic systems is highly dependent on the quality of their raw inputs. Where data are incomplete, outdated, or poorly maintained, the result is the familiar 'garbage in, garbage out' dynamic, with outputs that may be inaccurate, biased, or otherwise unreliable.[76] In practice, forces vary considerably in their data management capabilities, and the datasets available for training are often out of step with current operational realities *(Interview 8)*. One respondent, for example, noted that *"close to four to five years' worth of [historical] data"* was required before an AI system—such as a predictive or pattern-recognition system used for demand forecasting, crime linkage, or risk assessment—could generate outputs of sufficient stability to be operationally useful *(Interview 6)*. A more fundamental concern, however, relates to completeness. As one respondent put it:

> *"There is a massive challenge around [...] the proper use of probabilistic AI for predictive activity because [...] police data by its nature is only that which is connected to the criminal justice processes [...] if you're trying to look at prediction for prevention of harm and you're missing all the data on let's say public health and education and local government [...] you can't do that just from criminal justice data" (Interview 7).*

The consequence is that an AI system's apparent accuracy may be misleading where the data on which it has been trained reflects patterns of enforcement rather than the full picture of the problem it is intended to address. These concerns align with the critiques of predictive policing discussed earlier,[77] and were reinforced by the findings above.

5.3. Testing constraints

A constraint—rarely foregrounded in policy discussions—concerns the generation of test data against which AI systems' performance can be evaluated prior to deployment. Respondents described this not merely as a technical issue, but as a structural constraint on responsible adoption. Policing institutions hold extensive data, but, as one respondent put it, it is *"all live real data"* and therefore cannot be repurposed for development and testing without stringent legal and privacy controls *(Interview 5)*. In some domains, the constraint is more acute. For systems involving the classification of child sexual abuse material (CSAM), standard evaluation methodologies are effectively precluded by the illegality of possessing the relevant data. One such system required development testing in close partnership with the Home Office because, as the respondent explained, *"[i]t's illegal to possess [CSAM], so if we can do it with those, that's the only way anyone can actually work with the stuff" (Interview 16)*. Here, the conditions necessary for rigorous validation are in direct tension with the legal framework governing the underlying material.

5.4. Legal and evidential challenges

Respondents in our interviews consistently identified data governance as a practical and recurring obstacle. In particular, the difficulty of ensuring that data remains subject to UK jurisdiction was frequently raised, especially in relation to cloud-based systems. As one respondent observed, there are *"challenges in terms of the way that artificial intelligence as a technology works and its compatibility with UK law relating to data access and retention"*, with some tools offered by major providers unable to guarantee that data processing would remain confined to the UK *(Interview 11)*. This creates a point of tension between the data management architecture of widely used AI systems and the territorial assumptions embedded in domestic legal frameworks. Where data residency cannot be assured, organisations face tangible difficulty in demonstrating that personal data has been processed lawfully under the UK's data protection framework and that any evidential material derived from it is sufficiently reliable to be adduced in criminal proceedings *(Interview 21)*.

Key Finding 15

The opacity and the level of technical expertise required to scrutinise AI-generated or AI-assisted evidence raise concerns for the adversarial process. Where one party relies on tools that the other cannot meaningfully examine, the conditions for fair contestation are weakened. These concerns are sharpened by the risk that AI-assisted workflows could potentially introduce subtle contextual changes into evidential material. The purpose of the criminal justice system must remain in view: as one respondent put it, *"we all need to remember why we are there…not to save ourselves time and to be efficient but to get justice for members of the public" (Interview 21).*

Beyond these practical obstacles lies a deeper incompatibility between the probabilistic character of generative AI and the structural foundations of modern legal systems. As one respondent observed, the English system *"starts from the point of view of both determinism and cross-examination"*, and the introduction of outputs from *"a slightly non-deterministic tool"* into that framework presents a challenge that has yet to be resolved *(Interview 16)*. In other words, a legal system built on the expectation that evidence can be definitively tested, and its origins traced to an accountable human source (through rationalism) does not readily accommodate outputs whose content may vary with each generation and whose underlying reasoning cannot be examined in the witness box. Another respondent framed this structural tension in similar terms, noting the need to consider *"what the correlative of cross-examining a witness is in relation to computer evidence" (Interview 17)*.

This tension is sharpened by the existing legal presumption that computers operate reliably, under which a defendant who challenges computer evidence is effectively required to prove that the software malfunctioned. That presumption sits uneasily with AI systems whose outputs are inherently statistical, and which can fail in ways that even their own developers cannot fully explain. Where the law assumes that a system produces consistent results that can be traced back and accounted for, AI systems often do not work that way.

The use of AI within the evidential chain also raises a distinct set of procedural concerns. One respondent offered a notably direct assessment of the risks involved:

> *"I would have said don't use AI [in] anything that's in the evidential chain because straight away the issues with the CPS, with the issues around legalities, all of these things have come to the fore" (Interview 9).*

The concern here is not merely caution, but the recognition that AI-mediated processes introduce additional layers of material and potential error into a framework already governed by stringent disclosure obligations. AI-assisted workflows generate intermediate artefacts that do not appear in the final evidential product but may nonetheless fall within the scope of disclosure as unused material under the Criminal Procedure and Investigations Act 1996.[78] In the context of AI-assisted witness statement drafting, this may include the original audio recording, the machine-generated transcription, the AI-generated draft of the MG11 statement, and any amended and final versions of it. The implications are not confined to administrative burden. Where AI-generated transcription fails to capture contextual nuance that a human reviewer would recognise as relevant, the omission itself may become disclosable, introducing a further layer of evidential complexity *(Interview 2; Interview 21)*.

A related risk stems from subtle contextual changes introduced during AI-assisted transcription or drafting. One respondent illustrated this with an example where *"the individual grabbed a bottle of whiskey and shoved it in their coat pocket"* was rendered by the AI system as *"the bottle of whiskey was picked up from the shelf and placed in the basket" (Interview 22)*. This distinction is decisive in establishing the elements of an offence; while the first scenario describes a possible theft, the second describes ordinary shopping. If such a contextual change went unnoticed, it could undermine the prosecution or, in a more serious case, contribute to a wrongful conviction *(Interview 22)*. The example underscores a broader point: in smoothing text into fluent, well-formed prose, AI systems can quietly alter its meaning, and those alterations are not neutral in evidential terms.[79]

Bias presents a further evidential and ethical concern that intersects with the issues outlined above. Although police systems hold a wide range of information (including nuanced intelligence), AI tools are more likely to analyse structured, factual records such as arrests, charges, and convictions. The result, as one respondent observed, is that the data on which AI systems are trained *"might not represent our communities as such" (Interview 8)*. Where those records are drawn disproportionately from communities subject to historically intensive policing, downstream processes risk inheriting and potentially amplifying that distortion. These concerns align with a wider body of scholarship on algorithmic fairness, which has long warned that data-driven systems may reproduce or intensify existing disparities in criminal justice outcomes.[80] The issue is not confined to initial model training, but extends across the evidential chain, where small errors or biases may accumulate and harden into apparently objective outputs.

A related problem concerns the *challengeability* of AI-influenced decisions. Respondents questioned whether individuals subject to investigation or prosecution are, in practice, equipped to interrogate the use of probabilistic systems, given their opacity and the level of technical expertise required to scrutinise them effectively *(Interview 17)*. This raises concerns for the adversarial process: where one party relies on tools that the other cannot meaningfully examine, the conditions for fair contestation are weakened. As one respondent put it, the purpose of the system must remain clear:

> *"we all need to remember why we are there [...] not to save ourselves time and to be efficient but to get justice for members of the public" (Interview 21).*

The risk of dehumanisation in data-driven systems was raised in similar terms. One respondent, drawing on a report on public attitudes to AI in policing, described the concern that police might *"start treating people as data rather than people"*, and articulated the underlying principle as follows:

> *"human dignity needs to be an ever present consideration [...] the way in which data flattens, especially when it's used in a predictive context and when the data of others is used to make decisions about you, not because you've done something but because you're like other people who did something, that's inherently problematic from a human rights and rule of law perspective" (Interview 17).*

5.5. Structural and institutional barriers

Key Finding 16

The structural fragmentation of policing leads to duplicate procurement, inconsistent governance, and a tendency towards 'solutionising'—adopting an AI tool or capability before the underlying problem has been clearly defined. External pressures, including government directives to use AI and commercial influence, drive adoption at a pace that outstrips institutional readiness.

Turning from doctrinal to organisational concerns, the structure of policing in England and Wales presents entrenched difficulties for the coherent adoption of AI. The forty-three-force model was described by respondents as a longstanding source of inefficiency.

As one respondent observed, where two forces are found to be working on the same problem independently, there may well be *"a third force that no one ever knows about doing something that's exactly the same" (Interview 3)*. This duplication was characterised as a structural consequence of a model *"invented before computers and [which] hasn't changed to adapt for the fact that interoperability is now really important" (Interview 3)*. The practical effect is a fragmented procurement environment in which forces develop parallel solutions, duplicate effort, and often end up with incompatible systems *(Interview 9)*. One respondent summarised the position succinctly: *"the tools used by police forces are so fragmented it's bonkers" (Interview 11)*. The recently published Home Office police reform white paper acknowledges this problem and proposes, among other things, a shift towards larger regional units, centralised procurement, and the establishment of a new National Police Service within which PoliceAI will eventually sit.[81] However, the white paper itself envisages an extended implementation timeline, including an independent review of force structures which is not due to report until summer 2026, with the first mergers possible only later in the current Parliament. The benefits of consolidation are therefore still some way off. In the meantime, AI adoption will continue to expand across the existing fragmented institutional architecture, even as that structure is itself under review. PoliceAI is intended to bring consistency and central guidance, but for now, the spread of AI is being shaped by the very structural conditions that reform is ultimately meant to remove.

Cultural tendencies within policing further complicate the prospect of coherent AI governance. AI tools are often introduced through business units rather than formal digital or technology governance structures. Consequently, those responsible for oversight are often engaged only after deployment has already begun *(Interview 11)*. This sequencing problem is reinforced by an institutional tendency towards *solutionising*—adopting a product or capability before the underlying problem

has been clearly defined. As one respondent put it,

> *"[a]nyone that is successful at selling into policing will know sell the solution first and then let them sort out what the problem was after you've sold it" (Interview 20).*

Cutting across these structural and cultural dynamics is a further mismatch between how quickly the commercial and public-sector move. Whereas a private-sector procurement decision may be reached *"in a couple of conversations maybe within a week"*, the equivalent process in policing can take *"three to six months" (Interview 6)*. This gap creates pressure to adopt AI tools faster than existing governance arrangements can manage. As another respondent observed,

> *"there is the government direction to use AI and there's a bit of a panic with we have to use it, we have to use it. Actually, we're not mature enough" (Interview 21).*

A combination of these factors indicates that the pace of adoption is driven less by whether forces are actually ready than by outside pressure to move quickly. The result is a governance environment where decisions about deployment are increasingly shaped by urgency rather than by a measured assessment of whether a force has the capability, understands the risk, and is organisationally prepared to use the tool well.

5.6. Procurement constraints

Key Finding 17

Procurement processes are simultaneously too rigid for rapid innovation and insufficiently robust for the governance demands generated by AI adoption. To achieve rapid market penetration, commercial providers are able to shape implementation timelines by compressing the time available for structured evaluation and governance. Public bodies often find themselves unable to resist these pressures, which include offers of free proof-of-concept possibilities and the embedding of new AI capabilities within existing contracts and software upgrades. The cumulative effect is to entrench dependency on incumbent suppliers, narrowing the scope for substitution, withdrawal, or iterative governance once AI capabilities have been introduced—with consequences for both innovation and value for money.

Procurement processes are, in this respect, simultaneously too rigid for rapid innovation and insufficiently robust for the governance demands generated by AI adoption. Public-sector procurement was described by respondents as *"so rigid and so tight that it does take a lot of flexibility out"*, with the result that

"procurement in public service hinders innovation" (Interview 6). At the same time, the procedural and evidential requirements associated with acquiring even relatively straightforward tools remain extensive. These include data protection impact assessments, detailed security questionnaires addressing encryption in transit and at rest, and compliance thresholds relating to data sovereignty, cyber security accreditation, and network compatibility *(Interview 14; Interview 18)*. While these safeguards are substantively justified given the sensitivity of policing data and systems, they do not address the quality, accuracy, or reliability of the tool's outputs. A system may satisfy every data governance requirement yet still generate results that are inaccurate, biased, or unfit for purpose. The result is a procurement framework that imposes significant transaction costs while largely neglecting the most consequential questions about tool performance.

5.7. Commercial dynamics

The shortcomings in what procurement assesses, highlighted above, are compounded by an imbalance in who controls the process. Policing organisations and technology vendors do not often meet as equals. Respondents described an imbalance in which commercial providers are able to set the pace of implementation in ways that public bodies find difficult to resist. As one noted, it is "in suppliers' interest to push hard and fast" and, in practice, *"sometimes it's the only way forces can get delivery of technology [...] to work to a supplier's timeline" (Interview 9)*. Another offered a broader characterisation of the present situation: *"at the moment it is very much policing being done by technology companies than the other way around" (Interview 19)*. The proliferation of free proof-of-concept arrangements was identified as part of this pattern. While ostensibly lowering barriers to experimentation, such models also reflect vendors' incentives to achieve rapid market penetration, often compressing the time available for structured evaluation and governance *(Interview 9)*.

Additional difficulties arise where AI capabilities are introduced incrementally into existing systems and contracts. In such cases, organisations may have limited practical ability to refuse or disentangle new functionality once it has been embedded within established procurement relationships: *"our challenge is where existing suppliers bring this AI capability into our existing systems and contracts in a way that requires us to try and unpick that and make some choices around it" (Interview 5)*. The most serious consequence is that forces may, as a result, be producing evidential material using AI without having made a deliberate decision to do so—and therefore without the governance, disclosure, or quality assurance arrangements that such a decision would normally trigger. These problems are reinforced by the way procurement is financed. Longer contracts typically reduce unit costs but also reduce flexibility, making it harder for forces to exit or renegotiate arrangements once AI capabilities have been built into it *(Interview 6)*.

5.8. Chaining effects

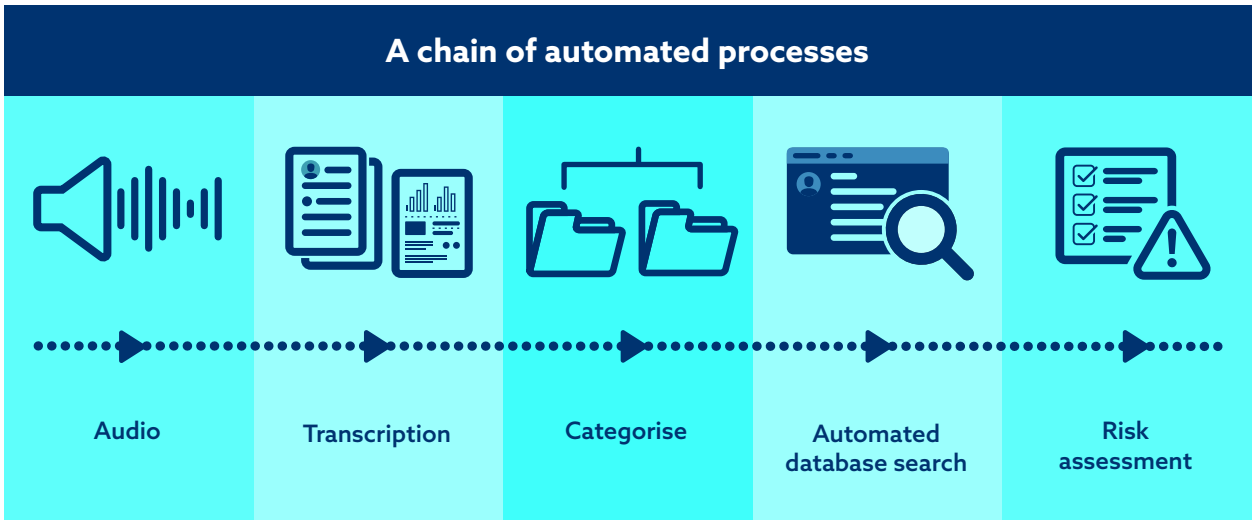
Key Finding 18

***Chaining**—the sequential operation of AI tools across processing stages—is currently rare but is likely to emerge as a category of risk that single-tool governance frameworks are not yet designed to detect. For example, evidence uploaded to a digital evidence platform may be automatically transcribed by AI as it is reviewed, and that transcription could then feed into a downstream disclosure process that is itself partly automated, with each stage potentially inheriting and compounding errors introduced earlier.*

In this study, we investigated the *chaining* of AI systems, specifically where the output of one tool becomes the input for another.[82] We raised this aspect with respondents because, for us, it is both the least well understood concept and, in terms of its potential downstream effects, among the most consequential. Unlike the issues discussed above, which primarily concern individual tools assessed against established legal and governance frameworks, chaining introduces a layer of systemic complexity for which those frameworks were not designed. When multiple AI components operate in sequence, the assumptions, error rates, and interpretive biases embedded in each system interact and accumulate. The resulting effect on accuracy, accountability, and legal compliance may therefore be materially greater than any single-system assessment would suggest. In this sense, risk is not merely transferred across systems but transformed through their interaction.
One respondent acknowledged that chaining in its fullest form is not yet common in policing *(Interview 21)*, although its potential trajectory was clearly recognised. Others were unfamiliar with the concept altogether, with one noting that they had *"not heard of the [term] chaining before" (Interview 9)*. This lack of familiarity is indicative of the limited extent to which systemic, multi-model interaction has been incorporated into current governance thinking.

Despite this prevailing sentiment, our study identified several instances in which sequential AI processing is already occurring in practice. A post-call analysis tool currently being trialled in two police forces, for example, generates two distinct outputs: one designed *"for a human to read"* and another *"optimised for further technological processing" (Interview 20)*. The latter output is then passed into a further system which: *"identifies elements within that call that are capable of being searched for on the police database [...] automatically searches all of the other police databases that we have and brings all the information back for the human to review"*, before proceeding to draft an assessment of the threat, harm, risk and vulnerability *(Interview 20)*.

This sequence—from audio to transcription to categorisation to automated database search, and finally to risk assessment—constitutes an operational chain in which multiple often AI-enabled system components interact in series. Each stage contributes incrementally to a composite output that will inform subsequent decision-making about individuals. The result is not simply the deployment of individual tools within a workflow, but the emergence of a layered system in which responsibility, interpretability, and error propagation become increasingly difficult to isolate at any single point in the process. A further concern is whether the automated nature of this sequence is disclosed in the resulting assessment. If the final output—in this case, a threat, harm, risk, and vulnerability assessment—is passed to a decision-maker without any indication that it was generated through a chain of automated processes, that decision-maker has no basis on which to apply the appropriate degree of scrutiny or to understand the provenance of the information on which they are being asked to act.



A second example involved integrating large language model processing with national policing databases. One respondent described a system combining *"classification, the ability to transcribe data and then using a large language model to pick out specific elements of that data that might indicate higher threat, harm or risk"*, with outputs feeding into data pipelines connected to the Police National Computer and the Police National Database *(Interview 19)*.

Looking ahead, respondents envisaged the emergence of more explicitly *agentic* forensic tools capable of invoking multiple analytical functions in sequence, passing outputs from one stage to the next without human intervention at intermediate points *(Interview 16)*.

Another particularly sensitive area identified was AI-assisted disclosure. As one respondent observed:

> *"[Chaining can occur] if we're thinking about disclosure as a safeguard to police AI but then disclosure itself is being conducted by AI…I think that [AI-assisted disclosure] is going to be a huge area as just the sheer amount of information that police and prosecutors have to go through" (Interview 17).*

The concern here is that disclosure is one of the main safeguards against the misuse of AI in investigations. If disclosure is itself carried out by AI, then the safeguards depends on the reliability of the same kind of technology it is supposed to be checking. In other words, the system meant to catch the errors and the system producing them become one and the same, and disclosure can no longer serve as an independent check.[83]

The governance concern most directly raised by chaining is the propagation of error across sequential stages. Where an inaccuracy arises early in a processing chain, it is not contained within that stage; rather, it is carried forward, potentially amplified, and may shape outputs at each subsequent point in the sequence. As one respondent observed in relation to AI transcription, *"if we don't illuminate [accuracy concerns] and we just assume the transcription is accurate, then potentially everything else could become biased that follows" (Interview 20)*.

A second concern relates to data protection and purpose limitation. Chaining may, in effect, amount to the repurposing of data beyond its original processing scope, with permissions granted for one function implicitly extended to another without explicit authorisation *(Interview 24)*. This raises questions about the adequacy of existing legal safeguards in environments where data is continuously reprocessed across multiple analytical layers. The same respondent further highlighted the risk of compounding error through model interdependence, noting the danger of AI systems training on the outputs of other AI systems without an adequately broad or reliable base of original data. This was described as "doubling your risk", where deficiencies in the foundational dataset of an initial tool are carried forward and potentially magnified through subsequent processing stages *(Interview 24)*.

The 'chaining problem' is not yet widely recognised in AI governance literature. The identification of chaining as a distinct governance concern within the law enforcement context is, to our knowledge, original to this project. It informed our proposal to include chaining within the disclosure obligations relating to AI deployment in policing,[84] and Calder et al have developed the concept further by identifying it as an emerging risk to case progression.[85] Two problems stand out. First, when several systems each handle part of a task, no single system or operator can be said to 'own' the final output, which makes it unclear who is accountable

when something goes wrong. Second, as data passes from one system to the next, it may be used for purposes beyond those for which it was originally authorised, raising data protection concerns. In both cases, responsibility is spread across a chain of technical steps in a way that does not fit the categories on which existing legal or organisational structures rely. As AI develops towards more autonomous, agentic systems—tools able to carry out several functions in sequence without a human approving each step—these chains are likely to become available to policing organisations in the near term. Against that background, the key governance implication is one of timing: the frameworks needed to govern chained systems cannot be developed retrospectively without significant loss of control. They must be put in place before such systems are widely deployed, while there is still scope to shape how they are designed and constrained.

5.9. Emerging external threats

Key Finding 19

Fabricated evidence, deepfakes, and AI-enabled fraud constitute emerging threats for which the safeguarding architecture—encompassing detection, authentication, and training—is not currently in place. A further, consequential threat is that authentic evidence may be presented as fabricated.

Whereas the analysis so far has focused on risks arising from the use of AI within institutional settings, the threats considered in this final section come from a different direction. They concern the use of AI by external actors to create fabricated evidence, disinformation, and fraudulent material that the criminal justice system may then have to receive, assess, and act upon.[86] The integrity of the criminal process is therefore contingent not only on how carefully institutions use AI internally, but also on their capacity to detect and contest AI-generated content introduced from outside. The most immediate concern in this category is fabricated evidence. As one respondent observed:

> *"right now, we could be presented with some evidence that we would have no idea is completely AI generated. It has got so good now [...] and if you acted upon the evidence in front of you, we will have miscarriages of justice" (Interview 6).*

The risk identified here is not speculative in a narrow sense, but structural: it arises from the increasing realism of generative outputs combined with evidential practices that are not yet designed to interrogate their provenance. Public evidence portals are a particularly clear point of vulnerability. In many jurisdictions, members of the public can upload doorbell or dashcam footage for use in investigations and

prosecutions. At present, there is no systematic process for verifying that such material has not been synthetically generated or altered using AI tools *(Interview 1)*.

Deepfakes pose a related and increasingly growing threat;[87] and one respondent anticipated the eventual erosion of the assumption that *"seeing and believing will mean the same in a few years" (Interview 10)*. On one side, fabricated material may be wrongly accepted as authentic and relied upon in investigative or prosecutorial decision-making. On the other, genuine material may be falsely dismissed as fake, introducing doubt where none is warranted. Defendants who have committed an offence may evade liability by casting doubt on authentic recordings, while innocent defendants may be convicted on the basis of fabricated material that the system fails to detect. In both scenarios, the integrity of fact-finding is weakened, and the legitimacy of outcomes is correspondingly undermined. As synthetic audio-visual content become more convincing, public trust in digital evidence as a whole begins to erode, shaping what jurors, witnesses, and the public expect of it *(Interview 18)*. As one respondent observed, *"that whole slightly longer term element of trust in evidence…now potentially legitimately someone might be able to say no, that's not me, that's been generated and the technology to appropriately appraise deep fakes has not really caught up" (Interview 13)*. Overtime, this affects not just individual cases but the wider climate of confidence in which criminal cases are decided.

Agentic AI used for criminal purposes is a further emerging threat. The prospect of increasingly autonomous systems being used by organised crime groups to facilitate fraud and related offending was described by one respondent as *"extremely worrying"* and situated within what they characterised as an ongoing *"race of adaptation"* in cybersecurity *(Interview 10)*. The defensive architecture required to meet this threat is not currently in place. As the same respondent noted, its development *"has to be down to government, and it has to be trusted" (Interview 10)*, reflecting both the scale of the task and the need for any coordinated response to carry public legitimacy. These point to the need to develop capability in relation to evidential risk caused by AI manipulation and deepfakes. This would include the detection of AI-manipulated content, the authentication of evidential material submitted through public reporting channels, the validation of material adduced in court, and the systematic training of police officers, lawyers, and judges in the identification of synthetic content and its evidential risks. Some institutional recognition of this need is emerging. The judiciary published formal AI guidance for judicial office holders in 2025, which acknowledges the risk that AI can be used to produce deepfakes and advises vigilance.[88] Although the guidance addresses the possibility in general terms, it does not establish any procedural safeguards or admissibility standards. Research by the Alan Turing Institute's Centre for Emerging Technology and Security (CETaS) has concluded that UK law enforcement currently

lacks the tools to tackle AI-enabled crime effectively, and has proposed the establishment of an AI Crime Taskforce to develop capability for the proactive disruption of criminal networks.[89] For now, however, these capabilities remain unevenly developed across the criminal justice system, and there is little indication that they are being delivered at a pace commensurate with the threat they are intended to address.

Governing for Legitimacy

"we don't identify the need for an AI tool, we identify the need for a solution to a business problem which may or may not be AI" (Interview 21)

"the key to it all is evaluation and that is agile" (Interview 9)

"we need to have some greater public dialogue around the harms that we're looking to prevent [...] relentlessly talk about it with our communities" (Interview 19)

"policing is a complex beast [...] we want to be innovative, we want to drive forward and we can see all the productivity benefits but at the same time our fundamental objective is to secure confidence from the public in policing" (Interview 5)

"[We] need to include the public at all stages which looks at the importance of public participation from the perspective of legitimacy, democratic accountability and the principle of policing by consent" (Interview 17)

6. Governing for Legitimacy

The *fourth* and final theme of this study examines the institutional architecture upon which the legitimate adoption of probabilistic AI in law enforcement and the broader criminal justice system ultimately depends. The institutional framework is more sophisticated than it was even two years ago, and several of the approaches examined in this section represent genuine and substantive progress. However, compliance remains inconsistent, evaluation capacity is severely constrained, and the structural fragmentation discussed in the previous section means that governance arrangements, which exist in principle, are not always reliably implemented in practice. The subsequent discussion is organised as follows: it begins with the established governance frameworks, evaluation challenges, and existing infrastructure, alongside transparency considerations; it then turns to the funding architecture that is increasingly operating as a central lever of policy direction; and it concludes with the public trust conditions upon which legitimate adoption ultimately depends.

6.1. Governance frameworks

Key Finding 20

Internal governance processes for AI adoption are developing. Initiatives such as data ethics impact assessments, harms workshops, cross-agency joint operating procedures, the Responsible AI Checklist for Policing,[90] and the establishment of PoliceAI represent meaningful progress in building national capability for AI evaluation and oversight.

Several respondents described internal governance processes that reflect a serious and considered approach to AI adoption, and these practices are best understood as evidence of what is achievable within existing institutional constraints rather than as a uniform standard. In more mature governance environments, the process begins not with a technology but with a defined operational problem. As one respondent explained:

> *"we don't identify the need for an AI tool, we identify the need for a solution to a business problem which may or may not be AI" (Interview 21).*

This starting point matters, as it means that the question being asked is not 'what can this tool do?' but 'what problem are we trying to solve and is AI the right way to solve it?'. This orientation aligns with the approach embodied in the Responsible AI Checklist for Policing, which emphasises a prior normative question—whether a tool *should* be used—rather than a primarily capability-driven question of whether it *could* be used.[91]

The internal processes described by respondents involve multiple sequential assessments prior to the approval of any AI tool. One respondent outlined a pathway beginning with a data protection impact assessment, followed by a data ethics impact assessment, and supplemented by a harms workshop (where subject matter experts identify stakeholders, map potential benefits, and systematically rank risks) *(Interview 21)*. The purpose of this sequencing is not merely procedural compliance but structured deliberation over the conditions under which deployment may be justified. Alongside this, a parallel governance track requires project teams to complete documentation addressing moral permissibility, alignment with organisational values, and consistency with wider strategic commitments. Respondents emphasised that these stages are intended to be substantively engaged with, rather than treated as formalities or 'box-ticking' exercises (Interview 21). Cross-agency governance mechanisms have developed alongside these internal arrangements. The Crown Prosecution Service (CPS), for example, has established Joint Operating Procedures through which police forces engage with both the National Police Chiefs' Council (NPCC) and the CPS prior to deploying AI tools that may impact criminal proceedings.[92] These procedures operate on three distinct levels: early-stage consultation when a tool is first being considered; a beta phase during testing; and a final stage at which deployment on live cases and live data is contemplated *(Interview 22)*.

At the national level, institutional development has accelerated considerably. The College of Policing has produced guidance on building AI capability and runs a technical panel through which forces can submit examples of AI deployment for expert scrutiny.[93] The Home Office has also signalled an intention to link future funding to transparency requirements, with one respondent describing the emerging position as follows:

> *"we're not going to give you money if you're not prepared to be transparent and publish on your website all of the different AI tools that you use with a named accountable officer against each of them" (Interview 3).*

However, the most significant institutional development in this space is the announcement in January 2026 of PoliceAI, supported by £115 million in Home Office funding.[94] PoliceAI has been designed to centralise the development, evaluation, and deployment of AI tools, to make approved tools available across forces with associated training and support, to maintain a public-facing registry of AI use; and to function as a national centre of expertise on the criminal misuse of AI.[95] Assuming it delivers on its stated objectives, PoliceAI would directly address what respondents repeatedly identified as a central governance deficit: the absence of a single body with both the authority and technical capacity to evaluate AI systems before they reach operational deployment across the fragmented landscape.

6.2. Evaluation challenges and existing infrastructure

Key Finding 21

The cost of independent or specialist internal evaluation is significant and under-resourced. Police forces are often left to choose between proceeding without adequate evidence of effectiveness or abandoning trials whose potential benefits remain undemonstrated.

Evaluation is, based on the evidence of this study, both the most important and the most under-resourced element of the governance architecture. As one respondent articulated, *"the key to it all is evaluation and that is agile" (Interview 9)*, yet the conditions necessary to support rigorous evaluation remain insufficiently developed. While the College of Policing undertakes both technical and impact evaluations, incorporating cost-effectiveness analysis as a standard component of its work, such methodological rigour is the exception rather than the rule *(Interview 7)*. One respondent noted that a particular AI trial was among only 500 randomised controlled trials in policing, a figure which, in their view, *"tells you there is a distinct lack of testing and certainly not of any new system" (Interview 9)*. The same respondent highlighted a further illustration of the gap between deployment and evaluation: a redaction tool *"is in use, and it is deployed, but it's not evaluated [...] not done academically as far as I know" (Interview 9)*. This pattern—where operational adoption consistently precedes systematic assessment—reinforces the broader finding that evaluation capacity remains uneven and, in many contexts, secondary to implementation.

The structural barriers to evaluation are considerable. Forces often lack both the funding and the organisational capacity required to commission rigorous assessments. One respondent disclosed that independent academic evaluation of an AI trial consumed *"nearly thirty per cent"* of the entire programme budget, and they observed that the findings of that evaluation might prevent the programme from progressing to a subsequent phase *(Interview 6)*. Even when universities offer to conduct evaluations at no financial cost in exchange for data access, the practical demands remain substantial. The work involved in cleaning, preparing, and securely sharing relevant datasets was described by one respondent as *"unfortunately not something that police forces are able to do because they're operationally being effective so that seems a bit of a luxury" (Interview 9)*. Consequently, learning from individual deployments is not systematically captured, consolidated, or disseminated *(Interview 9)* across the wider ecosystem. Evaluation, in other words, is not failing because its importance is disputed; rather, it is constrained by the absence of the financial resources, data infrastructure, and protected operational capacity required to make it routine.

Recognising the limitation of force-level evaluation, the Home Office has indicated an intention to fund national benchmarking exercises and to make compliance with benchmarking standards a condition of central funding. One respondent described a proposed framework under which:

> *"if we got to the stage where, for example, we were planning to give money to policing to use AI enabled translation tooling we would put in provisions such as you can only use tools with this money that have been nationally benchmarked as being sufficiently accurate" (Interview 3).*

A comparable approach has already been trialled in the context of facial recognition, where providers have been required to compete within a standardised testing environment using both real and synthetic datasets *(Interview 19)*. The broader aspiration is to develop *"a sandbox environment where we then test those tools in a standardised way"*, with high-performing systems potentially progressing to national procurement status *(Interview 19)*. Whether this ambition is realised will depend significantly on the capacity, authority, and operational reach of PoliceAI.[96]

6.3. Transparency

Key Finding 22

Transparency arrangements are inconsistent. The Algorithmic Transparency Recording Standard (ATRS) provides a structured means for public authorities to disclose the AI tools they use, but compliance is not currently mandatory for police forces. A meaningful transparency regime requires not only mandatory participation but records that are substantive enough to enable external scrutiny and accessible to the public.

Transparency operates at several levels within AI governance. Regarding process transparency, the Algorithmic Transparency Recording Standard (ATRS) offers a structured mechanism for public authorities to disclose the AI tools they utilise and their deployment purposes.[97] However, respondents noted inconsistent compliance. The CPS has published its algorithmic transparency record for its Correspondence Drafting Tool,[98] whereas many police forces have not published equivalent disclosures, stemming from the fact that policing organisations are not currently mandated to complete the ATRS.[99] Although the ATRS provides a comprehensive system in terms of structure and information requested, the pertinent issue is whether police forces possess the resources to complete and maintain these records, and if it offers interactive functionality that would make it useful for both the police and the public.[100]

At the level of *individual tool design*—distinct from institutional disclosure regimes—one respondent described a deliberate attempt to embed transparency within the system infrastructure, rather than treating it as an external reporting obligation. The practical implications of this distinction were illustrated earlier in section 4.1, where the design of a post-call analysis tool, currently being piloted in two police forces, demonstrates how transparency and contestability can be integrated into a system's interface, rather than being treated as external reporting obligations.

Alongside these developments, the development of formal standards has been one of the areas where governance has made the most progress. The Responsible AI Checklist for Policing has already influenced procurement and deployment practices in some forces, with one respondent describing the checklist and its associated academic engagement as *"a big part of this, being able to ask the right questions" (Interview 19)*. The Home Office has indicated an intention to seek ministerial endorsement of the checklist and to direct forces to adhere to it *(Interview 3)*, while the College of Policing has adopted it as the recommended assessment framework for all forces.[101]

By way of analogy, one respondent compared this emerging approach to the approval regimes governing speed cameras and DNA testing laboratories, suggesting that a comparable system of type-approval for AI tools would provide courts and the public with a more stable basis for confidence in AI-generated outputs *(Interview 25)*. This analogy has its limits, given the greater complexity, adaptability, and context-sensitivity of AI systems, but it nevertheless points towards a governance model in which standardised testing and formal authorisation precede operational deployment rather than following it. In a related development, the Police Reform Bill proposes a new legal framework for law enforcement use of facial recognition and similar technologies, including the creation of a single regulatory body to provide independent advice and oversight.[102] If enacted, this would represent the first instance of a dedicated statutory regime governing a specific category of AI tool in policing, and could serve as a model for the kind of formal authorisation process that the type-approval analogy envisages.

The question of whether a dedicated regulator is needed for AI in policing more broadly —beyond biometrics—was not raised directly by respondents in these terms. However, the governance gaps identified throughout this study point in that direction. Meaningful governance of automated systems requires, at a minimum, three conditions: *access to information*, so that scrutiny is possible; the *capacity to interrogate the logic of decisions*, so that transparency does not collapse into disclosure without comprehension; and the *availability of consequences or redress when systems fail*, so that accountability attaches in practice rather than in the abstract. At present, no single body possesses both the authority and the technical

capacity to enforce all three conditions across the fragmented policing landscape. The establishment of a single expert regulatory body for biometrics under the Police Reform Bill represents the first step towards this kind of institutional architecture. Whether its remit should extend to other categories of AI tool is a question that the findings of this study suggest deserves serious consideration.

6.4. Funding as a governance lever

Key Finding 23

AI projects in policing are financed through a combination of central government allocations, police-led innovation funds, and force-level procurement budgets. Conditions attached to central funding (which include requirements for transparency, benchmarking, and adherence to nationally agreed standards) are likely to operate as a significant governance instrument that incentivises responsible AI adoption.

Based on our findings, the conditions attached to central funding are beginning to function as a significant governance instrument. For instance, business cases for AI funding are subject to assessment by the Home Office who must consider risks, dependencies, and AI-specific risks *(Interview 3)*. Police forces are then required to report monthly on their progress and the management of identified risks, with the continuation of funding contingent on satisfactory performance against these reports *(Interview 3)*. Additional criteria intended to *"incentivise responsible AI adoption"* are actively being developed in the context of the forthcoming spending review *(Interview 3)*. Where formal regulatory powers are absent, limited, or politically contested, funding conditionality provides an alternative governance mechanism. It acts at a crucial point in the decision-making cycle: the moment forces seek resources for deployment, making them most responsive to externally imposed requirements.

Regarding the structure of funding flows, AI-related work in policing is financed through a combination of central government allocations, police-led innovation funds, and force-level procurement budgets. Home Office funding is typically channelled to specific forces or programmes rather than directly to technology vendors *(Interview 3)*, maintaining a formal separation between public financing and commercial provision. Alongside this, police-led innovation funding provides a second major stream of investment, which includes both a STAR fund,[103] supporting research and technology development, and a Test and Learn approach[104] designed to enable shorter, experimental projects to proceed on the explicit basis that failure is an acceptable and anticipated outcome *(Interview 8)*. The Test and Learn stream is particularly significant as it institutionalises a degree of tolerance for unsuccessful pilots, which has traditionally been absent from public-sector technology procurement.

Table 2: Dimensions of funding conditions

Dimension	Strength	Vulnerability
Conditions on funding	Forces are required to demonstrate transparency and manage risk	Requirements do not yet cover evaluation or data quality
Where investment is directed	Focused on lower-risk applications where errors carry fewer consequences	Tools with evidential or legal significance receive less attention
Evaluation	Test and Learn approach accepts that pilots might fail	Evaluation is under-funded and not built into programme budgets
Independence from vendors	Public money goes to forces, not directly to technology companies	Vendor lock-in weakens this separation in practice
Long-term sustainability	New criteria for responsible adoption are being developed	A drop in investment could undermine governance before it takes hold

Within the funded portfolio, AI occupies an increasingly prominent position. Central government has, in the words of one respondent, developed *"a good sense of the areas of intervention where we think AI can make the biggest difference in policing to the efficiency of the process" (Interview 3)*, and four priority domains have been identified. These are: enhancements to police–public contact, including the triage of 101 calls; operator-assist tools supporting call handlers through automated transcription and summarisation; language translation; and the integration of AI into body-worn video systems. Two observations follow from the structuring of these priorities. First, they heavily focus on efficiency gains and reductions in administrative burden, rather than on applications with more direct evidential or legal significance. The AI Opportunities Action Plan, as one respondent noted, *"is a growth strategy"* oriented towards efficiency *(Interview 17)*, and the funded priorities align with this orientation. This indicates a deliberate preference for use cases perceived as operationally beneficial while remaining comparatively insulated from high-stakes legal consequences. Second, this distribution of investment reflects an implicit judgement about risk exposure. Contact centre operations and administrative workflows are areas where error tolerances are relatively higher, and the downstream evidential consequences of system failure are more limited. Therefore, the concentration of funding in these areas can be understood, at least in part, as a rational allocation of governance and operational risk within an emerging and uneven regulatory landscape.

Key Finding 24

A contraction in AI investment (i.e. an 'AI winter') is a possibility that some respondents considered likely within the next two years. While a pause could facilitate consolidation, reflection, and institutional learning, it could also destabilise emerging governance structures if funding is withdrawn before sufficient evidence has accumulated and supporting evaluation, capacity-building, and coordination mechanisms have matured.

The sustainability of the current funding environment remains uncertain. One respondent reported having recently encountered the concept of an AI winter in discussions about investment cycles and suggested that *"in the next two years there's going to be that lull again because investment must stop because nobody's confident in what they're investing in because the outputs haven't been seen yet" (Interview 18)*. The same respondent added that such a pause might *"give [...] us the break that we need to catch up",* noting that *"everything's sprinting along far too fast" (Interview 18)*. While a contraction in investment could create space for consolidation, reflection, and institutional learning, it could also disrupt the continuity needed for meaningful evaluation and governance development. If disillusionment arises before the current generation of tools has had time to demonstrate value in operational settings, funding streams may be withdrawn before sufficient evidence can accumulate. In that scenario, the emerging governance architecture—already reliant on sustained investment in evaluation, capability-building, and institutional coordination—risks being destabilised by the very volatility of the innovation cycle it is intended to regulate.[105]

6.5. Public legitimacy

Key Finding 25

Public attitudes towards AI are shaped by institutional trust, perceived accuracy, and perceived fairness—not solely by the technical properties of the tools themselves. Public acceptance is strongest where deployment is targeted and specific rather than generalised, and weakest where it is perceived as indiscriminate surveillance.

Public confidence occupies a distinctive position in the governance of AI in policing; it is not merely one consideration among others but a *constitutive element* of legitimate policing itself. As one respondent noted, while policing has an interest in productivity and innovation, *"our fundamental objective is to secure confidence from the public in policing" (Interview 5)*. These objectives are not inherently in conflict, but the balance between them has to be managed carefully.

The benefits of AI—particularly in efficiency and operational capacity—must be weighed against public concerns regarding fairness, accountability, and ethical use. When this balance is misjudged, the result may be a direct erosion of public trust and, consequently, the perceived legitimacy of policing institutions.[106] Crucially, such erosion is not easily reversible. Once confidence has been undermined, restoring it requires not only technical corrections but also sustained institutional effort, transparency, and demonstrable accountability over time.

The current state of public understanding is partial and uneven, significantly influenced by underlying levels of institutional trust. One respondent described the public as broadly unagitated but attentive: aware of the potential productivity gains associated with AI yet focused on whether its deployment is *"appropriate, ethical, legal, necessary and also protecting our data" (Interview 5)*. Another offered a more sceptical view, observing that *"the public have a general distrust of public sector and IT" (Interview 20)*. These observations align with empirical research indicating that public attitudes toward AI are shaped less by the technical characteristics of the systems than by the institutional contexts into which they are introduced,[107] and by whether individuals believe AI use will be accurate and fair.[108] Individuals approach public bodies with pre-existing expectations and levels of trust; these baseline attitudes tend to structure their responses to the use of algorithmic tools. In this sense, institutional trust operates as the primary mediating variable, with AI acting less as an independent object of evaluation and more as an extension of the institutions that deploy it.[109]

Public understanding is most developed in the area of facial recognition, where attitudes are more nuanced than a simple privacy-versus-security trade-off implies. Respondents identified a clear distinction between retrospective facial recognition—where there appears to be substantial public support—and live facial recognition *(Interview 5)*. Recent national survey data confirms this hierarchy of acceptance: 97% of the public view RFR as at least sometimes acceptable, compared to 88% for LFR used to locate criminal suspects.[110] The latter raises issues of 'collateral intrusion' and the ability to navigate public spaces without being subject to continuous state identification. Survey data suggest that citizens are capable of holding both positions simultaneously: they recognise the operational utility of facial recognition in catching criminals (75%), locating missing and vulnerable persons (71%), while expressing legitimate, concurrent concerns about the technology being misused or hacked (68%), the risk of false identification (49%), and a lack of control over collected data (46%).[111] As such, public acceptance is highly contextual; it is strongest where AI is deployed for specific, targeted purposes—such as protecting vulnerable individuals or tackling severe offences like terrorism and murder—and in high-risk locations like airports and riots.[112] Conversely, support is weakest where it is perceived to approximate generalised surveillance in everyday, community-level settings like residential streets and schools.[113]

Beyond facial recognition, public concern about AI in policing appears more diffuse. Respondents described it as centring on general anxieties—data security, algorithmic opacity, and the potential amplification of bias—rather than on sustained opposition to specific tools *(Interview 25)*. Predictive policing was a recurring example; one respondent suggested that the public often assumes it is being used at a scale significantly greater than is actually the case *(Interview 19)*. Whether this perception gap is reassuring or troubling is ambiguous, but recent public dialogues highlight a prevailing sense of "powerlessness" among citizens, who often feel the rollout of AI in policing is an inevitable progression regardless of public sentiment.[114]

Several respondents identified communication as a structural weakness in this regard. One described a longstanding tendency in policing to *"police at people rather than with people"*, and emphasised the need to engage communities more actively in the process of technological change *(Interview 6)*. Another expressed the same point more directly: *"we need to have some greater public dialogue around the harms that we're looking to prevent [...] relentlessly talk about it with our communities" (Interview 19)*. Evaluative evidence supports this, showing that the public explicitly desires nuanced government communication that transparently explains the risks and the specific safeguards managing the downsides, rather than merely overstating AI's benefits.[115] This reflects the core tenets of procedural justice theory, which holds that institutional legitimacy and public cooperation depend on fair treatment of individuals, transparency in decision-making, and providing a meaningful "voice" for citizens to participate in processes that affect them.[116]

Emerging empirical work applying this framework to AI suggests that these dynamics carry over into technologically mediated contexts. When members of the public perceive that they have input into how AI systems are governed, deployed, and constrained, their willingness to accept and support the use of such systems by law enforcement increases correspondingly.[117] However, independent evaluations warn that if public consultation is perceived as a mere "PR exercise" or box-ticking endeavour for technologies already slated for deployment, it threatens the foundational principle of policing by consent.[118] In this sense, participation is not merely a normative ideal but an operational variable: it shapes the degree of trust in AI-enabled policing and, by extension, the legitimacy of its outcomes.

Public participation must therefore be understood as exceeding ad-hoc consultation in any meaningful sense. As one respondent argued, governance processes need *"to include the public at all stages"*, and that requirement is grounded in *"the importance of public participation from the perspective of legitimacy, democratic accountability and the principle of policing by consent" (Interview 17)*.

The same respondent noted that *"the public might have an idea as well [...] a view on what proportionate interference might be [...] a view on what good looks like in terms of human rights protections within their communities"*, and argued that these perspectives should inform governance decisions rather than being treated as constraints to be managed after the fact *(Interview 17)*.

On this account, public participation is not an adjunct to governance but a constitutive element of it. It reframes citizens from being objects of technological intervention to co-authors of the normative standards justifying that intervention. Where legitimacy in policing rests on consent, and consent presupposes informed and meaningful participation, governance architectures that marginalise public input are not merely incomplete but normatively deficient.

7. Conclusion and Recommendations

The findings of this study reflect that the legitimacy of AI deployment in law enforcement and the broader criminal justice system is not solely based on technical performance; it also depends on institutions' ability to evaluate systems before deployment, to make their operations comprehensible to those who rely on them, and to uphold forms of accountability that align with the risks introduced. In the absence of these conditions, the use of AI threatens to undermine the very principles—fairness, transparency, and reasoned decision-making—that the criminal justice system aims to uphold.

Based on this understanding, we present the following recommendations. Importantly, these recommendations are rooted in the experiences and concerns expressed by respondents throughout the study and are organised according to the institutions best positioned to address them.

Home Office

1. **Use funding as a governance mechanism, not simply a means of supporting AI adoption.** The findings of this study suggest that the Home Office's most effective mechanism for shaping responsible AI adoption is not guidance alone, but the conditions attached to funding. Funding requirements can compel transparency about the matters that currently remain opaque, including a tool's performance, the training data on which it is built, the methodologies underpinning it, and its evaluation statistics. They can also require coordination across forces in ways that voluntary frameworks cannot. Therefore, the Home Office should make future AI-related funding conditional upon forces demonstrating that tools (whether third-party or developed in-house) have undergone independent evaluation against nationally agreed standards before deployment and comply with the transparency and public-facing registry requirements developed through PoliceAI. As one respondent observed, *"we're not going to give you money if you're not prepared to be transparent" (Interview 3)*. Funding conditionality should also extend to the maintaining minimum specialist staffing levels, ensuring that forces receiving AI-related investment possess the internal technical capacity to understand, maintain and govern the tools they deploy.

2. **Commission a national evaluation programme beyond facial recognition technologies.** One respondent identified the facial recognition procurement framework under which the National Physical Laboratory independently tests systems as the only area where standardised, comparative evaluation currently exists, with all other tool categories procured and tested in isolation *(Interview*

19). The Home Office should, through PoliceAI, extend this national evaluation model to include other tools, such as those used for transcription, translation, risk profiling, object recognition, and vulnerability identification.

3. **Address the gap between structural reform and AI deployment in order to reduce duplicative procurement and governance.** The Home Office police reform White Paper proposes a fundamental restructuring of policing, but the independent review of force structures is not expected to report until summer 2026. The first mergers are expected to be even further away. AI deployment is not pausing. The Home Office should establish interim coordination mechanisms through PoliceAI to prevent the current 43-force model from continuing to generate duplicative procurement and inconsistent governance during the transition period. In particular, interim measures should require forces to consult with PoliceAI to determine whether a comparable tool has already been evaluated elsewhere before initiating a new procurement or proof-of-concept, thereby avoiding persistent duplication.

4. **Develop capability in relation to evidential risk caused by AI manipulation and deepfakes as a national infrastructure priority.** In coordination with the Ministry of Justice, the Home Office should fund and coordinate the development of a national capability, in line with CETaS' proposal for a dedicated AI Crime Taskforce,[119] encompassing: the detection of AI-manipulated content in evidential material; authentication tools for material submitted through public reporting channels; validation mechanisms for digital evidence adduced in court; and a programme of systematic training for police officers, prosecutors, and judges in the identification of AI-generated or manipulated content and its evidential risks. These capabilities should be developed and maintained centrally, possibly under the auspices of PoliceAI, rather than being left to individual police forces. Given the rapid advancement of generative AI, this is not a medium-term aspiration but an immediate operational requirement.

PoliceAI

5. **Exercise gatekeeping authority, not merely coordination.** The structural conditions described by respondents, in which forces procure independently, no standardised testing exists, and national bodies lack the authority to reject tools, suggest a risk that centralised infrastructure could prioritise dissemination over genuine quality control. PoliceAI should be explicitly empowered to reject tools that fail to meet standards of scientific validity, operational reliability, and responsible deployment.

6. **Develop a uniform strategy for data quality assessment and management.** Respondents noted the challenges posed by incomplete, outdated, or poorly

maintained data, variations in forces' data management capabilities, and datasets that are misaligned with current operational realities. A uniform approach for police forces to assess which datasets are appropriate for training new AI tools and to manage and maintain their data for potential future AI model training is required to ensure continuity of innovation.

7. **Define what meaningful evaluation should involve.** Based on findings from this study, any meaningful evaluation should include the following components: testing against realistic human baselines; assessment using both real-world policing data and synthetic data of varying quality; measurement of performance across demographic groups; testing under representative operational conditions (e.g., noisy environments and diverse accents) rather than controlled vendor demonstrations; interdisciplinary review combining technical assessment with qualitative analysis of human interactions with the tool; and ongoing evaluation beyond procurement.[120] PoliceAI should publish the evaluation criteria it applies, so that forces and vendors can anticipate what is required.

8. **Develop a risk classification framework for AI tools that is intelligible to end users inspired by the existing intelligence system.** PoliceAI should develop a framework, analogous to intelligence grading systems,[121] that helps end users understand the nature, reliability, and appropriate conditions of use associated with different AI tools. In developing this framework, PoliceAI should draw on the emerging analytical basis for distinguishing between the risk profiles of narrow and general-purpose AI systems. Calder et al. have demonstrated that the risks of LLM-based tools to case progression are not uniform but vary systematically with the design, scope, and integration of the tool.[122] Constrained systems with pre-coded prompts and fixed tasks[123] present a materially different risk profile from general-purpose systems with open-ended interfaces, and this difference should be reflected in the classification. Tools with probabilistic outputs, unconstrained prompts, unclear training and fine-tuning processes, and opaque positions within broader processing chains should require the most stringent evaluation, transparency, and oversight. The framework should also differentiate between intrinsic risks inherent to the model, and extrinsic risks, dependent on external contexts and inputs,[124] since narrow tools are designed to reduce extrinsic risk, whereas general-purpose tools largely leave it unmanaged.

9. **Require that AI tools meet minimum design standards for explainability, transparency, and contestability as a condition of procurement.** Once a tool's risk classification is established, it should only be procured if it meets design standards commensurate with that classification. These standards should include providing plain-language and intuitive explanations of the reasoning

behind outputs, indications of the system's prediction confidence and, if possible, overall and stage-wise uncertainty, and interactive mechanisms that encourage users to challenge or record disagreements with outputs and route cases for review. PoliceAI should publish these standards and apply them consistently as part of its evaluation and certification process.

10. **Develop context-specific oversight protocols that govern human interaction with AI outputs at the point of use.** Findings from this study indicate that 'human in the loop' has varying meanings across different settings. Therefore, PoliceAI should establish protocols that specify how AI tools present their outputs in a given operational context, with emphasis on designs that require humans to actively construct their judgements, rather than merely verifying AI outputs.[125]

11. **Make the PoliceAI public-facing registry mandatory, substantive, and interoperable with the ATRS.** PoliceAI should work with the Home Office to mandate the use of the registry of AI tools proposed in the police reform white paper by all police forces. This registry must prioritise public accessibility, be interoperable with the ATRS, and go beyond simply recording which tools are in use. It should include descriptions such as name, purpose, user, deployment stage, development type, applicable stage in the criminal justice process, and data used for training. Additionally, a systematic categorisation of each tool based on its capabilities should be reported. Each entry should also include evaluation results, including error rates, known limitations, whether the tool's outputs enter the evidential chain, and whether its outputs feed into other AI systems (to identify chaining risks). This level of disclosure distinguishes a meaningful registry from a mere inventory. The PROBabLE Futures interactive mapping tool[126] can serve as a foundation for this registry, which should be accompanied by clear guidelines on what constitutes AI and ideally an example list of tools to be included.

12. **Build a national repository of evaluation findings, including failures and withdrawal of tools.** Respondents described a system that is *"doing a lot very quickly"* but *"not learning a lot very quickly" (Interview 9)*. Evaluation findings (including negative outcomes, failed pilots, and reasons for the withdrawal or rejection of tools) should be collected centrally and made accessible to all forces. The current practice of isolated evaluations and unshared negative results risks repeating the same mistakes across similar deployments. This repository should be distinct from the register of AI tools, as it focuses on what has been learned from evaluating these tools, regardless of their deployment success.

13. **Identify the functions requiring retained human competence and develop a centralised training programme accordingly.** PoliceAI should establish a centralised skills training programme, distinct from general AI literacy training,

specifically designed to maintain professional competence in the substantive tasks increasingly automated by AI systems. Not all AI-assisted tasks carry the same risk of de-skilling. Certain function—such as statement-taking, vulnerability assessment, and evidential review—demand meaningful human judgement and analytical engagement rather than mere supervisory oversight. PoliceAI should identify which functions fall into this category and ensure that the professionals assigned to them remain suitably skilled through regular practice of the non-automated tasks. This should include a policy prohibiting law enforcement staff from using certain AI tools until they have completed the appropriate training and gained practical experience in performing the tasks without automation.

14. **Address the structural conditions that undermine workforce capability.** PoliceAI should advise forces on the specialist roles they need to sustain, recognising, as findings from this study show, that a single technical lead is insufficient and that forces require several distinct types of expertise. PoliceAI should also develop guidance on rotation policies to prevent the routine redeployment of officers across technical domains before their expertise has been consolidated. When forces cannot recruit or retain specialist staff at competitive salaries, PoliceAI should explore shared-service models, secondment arrangements, or centrally funded specialist posts to prevent the most exposed institutions from bearing the full cost of the disparity between public-sector pay and private-sector demand.

15. **Rebalance the relationship between policing and technology vendors.** One respondent described a commercial dynamic in which vendors drive the pace of adoption, define operational problems in ways that align with their products, and shape the terms of engagement itself: *"at the moment it is very much policing being done…by technology companies [rather] than the other way around" (Interview 19)*. To address this imbalance, PoliceAI should require forces to conduct independent needs assessments before engaging with vendors and ensure that proof-of-concept timelines are governed by evaluation readiness rather than commercial urgency. In addition, vendors should be required, as a condition of engagement, to disclose the underlying foundation models, the provenance of training data, and their testing methodologies and results. At least one vendor respondent described a structured engagement model in which learnings from individual force deployments are systematically shared across forces and national policing bodies through deployment blueprints and peer-to-peer knowledge exchange *(Interview 14)*. This suggests that the difficulties identified are not inherent to the commercial relationship itself; rather, they arise from the absence of institutional structures capable of directing vendor engagement towards cumulative organisational learning and sector-wide accountability.

16. **Invest in participatory oversight research that moves beyond consultation.** Public participation in decisions about AI deployment must extend beyond survey-based consultation. The PROBabLE Futures project's experimental work on participatory oversight should be built upon by PoliceAI to develop frameworks that embed co-production from the earliest stages of decision-making, ensure genuine community influence over governance arrangements, and actively involve communities most likely to be affected—including marginalised and underrepresented groups.

College of Policing

17. **Develop tiered AI literacy aligned to role, responsibility, and risk.** Our findings revealed varying levels of AI understanding across policing roles. Frontline officers requite a practical understanding of the probabilistic nature of AI and its outputs, the ability to direct AI tools effectively and the skills to critically assess those outputs. They also need clear guidance on escalation procedures and the circumstances in which AI systems should not be used. Supervisors must be equipped to critically review AI-assisted work, including the ability to identify outputs that are implausible or inconsistent with the original source material. Senior leaders need sufficient technical literacy to make informed decisions regarding procurement and governance and to resist supplier-driven narratives that cannot be independently verified. Training content across all tiers should cover automation bias, disclosure obligations stemming from AI-assisted workflows, and the known limitations and potential failure conditions of the specific tools in use.

18. **Treat prompting as a professional skill.** The College should implement a national training programme to equip officers with essential skills in prompt engineering. This should be recognised as a vital professional skill rather than a matter of individual improvisation.

19. **Develop Authorised Professional Practice (APP) for the policing functions most affected by AI.** AI is increasingly being applied to core policing functions that carry direct legal and procedural consequences, including disclosure, the preparation of case files, and decisions such as stop and search. At present, there is no national professional guidance setting out how officers should exercise these responsibilities when AI tools are involved. The College of Policing should develop APP to address the use of AI in these areas, specifying how officers should handle AI-generated or AI-assisted material, what must be recorded and disclosed, the points at which independent human judgement is required, and the circumstances in which AI should not be relied upon.

The Crown Prosecution Service

20. **Engage with the Ministry of Justice to propose an amendment to the disclosure framework for AI-generated material.** AI-generated outputs are likely to create additional unused material that falls within the scope of the prosecution's disclosure obligations but are not explicitly addressed in the current framework. The CPS should engage with the Ministry of Justice to propose an amendment to the Criminal Procedure and Investigations Act's *Code of Practice* to include, under section 6.6, a specific category of AI-generated material within the list of materials likely to contain information meeting the test for prosecution disclosure.

21. **Mandate early engagement prior to deploying AI tools that may have evidential implications.** The joint operating procedure through which police forces engage with the NPCC and the CPS is an effective mechanism for coordination. However, some forces reach out to the CPS only after a tool is already in development. One respondent described having *"to pause our trial and engage with CPS"* after the evidential implications became apparent, observing that *"we should have engaged with CPS sooner" (Interview 9)*. The CPS should therefore require its involvement at the earliest conceptual stage of any AI project that may be relevant to the generation of material entering the evidential chain.

Judiciary

22. **Issue a comprehensive Criminal Practice Direction on the procedural handling of AI-processed and AI-generated evidence in criminal proceedings.** The judiciary's formal AI guidance, published in 2025, addresses judges' use of AI tools and identifies potential risks posed by generative AI in courts and tribunals. However, there is currently no procedural framework for handling AI-processed or AI-generated evidence introduced by parties in criminal proceedings. As findings from this study indicate, courts are already encountering AI in adjacent contexts, such as the use of tailored large language models in immigration tribunals. The Criminal Procedure Rules (CrimPR) apply broad, legacy definitions to technological evidence—particularly within the disclosure parameters of Rule 19.4 and the human-centric focus of Rule 19.2. Therefore, a dedicated Criminal Practice Direction should be issued to guide judicial discretion within the existing framework. This practice direction should address: the requirement (broadly outlined in CrimPR Part 3) for parties to notify the court when evidence is generated or processed by AI; the circumstances under which a pre-trial reliability hearing should be convened (potentially treating the underlying system as a core component of the expert's competence under CrimPR Part 19); the information the prosecution must provide regarding the AI system's validation (e.g., performance testing with documented error rates,

fairness assessments with identified biases), and known limitations; and guidance on the weight to be attached to probabilistic outputs. Our recent research proposing a three-tier hierarchical framework for calibrating legal scrutiny based on the complexity, transparency, and regulatory oversight of probabilistic systems[127] could inform the development of such a direction.

Technology Providers

23. **Treat transparency as a condition of operating in the criminal justice system.** Findings in this study indicate that one of the main impediments to governance is vendors' non-disclosure regarding the foundation models their tools are built on, the data they have been trained on, and the processing that occurs within their systems. As one respondent noted: *"They don't let you know what they're using and so actually your ability to then understand and compare apples with apples...you don't really know what is happening, which is a worry" (Interview 19)*. Vendors should therefore anticipate that as centralised evaluation and procurement frameworks mature, tools that do not meet minimum transparency requirements will likely be excluded unless they proactively adhere to transparency obligations.

Research Community

24. **Prioritise research on chaining effects, including their data protection implications.** The sequential use of multiple AI tools across the criminal justice process is under-theorised and under-researched. However, whenever the issue of chaining was brought up with respondents, concerns were often raised about the likelihood of it occurring. Chaining may involve repurposing data beyond its original processing scope, with permissions granted for one function implicitly extended to another without explicit authorisation. The governance frameworks necessary for chained systems cannot be developed retrospectively; they must be articulated in advance while system architectures remain amenable to design and constraints. The research community should prioritise efforts that map existing chains, model error propagation through sequential systems, examine the data protection implications of cross-purpose processing, and develop governance frameworks capable of managing risks that span institutional boundaries.

25. **Develop the evidence base on AI and the long-term exercise of professional judgement.** Findings from this study raise pertinent questions that the current evidence base has yet to address. Empirical research is needed to address gaps regarding how AI tools affect the development, maintenance, and exercise of professional skills over time in policing and the broader law enforcement context. Some of the questions include: Do officers who use AI-assisted

statement generation develop the same investigative instincts as those who take statements manually? Do legal practitioners who rely on AI summaries engage with the underlying material as carefully? Do judges who use AI-drafted bench notes reason differently?

26. **Conduct a longitudinal follow-up study on the conditions for responsible deployment of AI.** The present study captures law enforcement and the criminal justice system at a time of rapid change: strategic frameworks are being established while many tools are still in pilot and exploratory stages, coinciding with the announcement of the most significant structural reforms to policing in nearly 200 years. A follow-up study in due course (two to three years) would allow the research community to track whether the governance, evaluation, and capability gaps identified in this report have been addressed, assess the impact of PoliceAI and the police reform White Paper, and determine whether the conditions for responsible deployment are being met in practice.

Bibliography

[1] Foundation models are large-scale artificial intelligence (AI) models pre-trained on vast amount of general data and can be adapted to a broad range of applications. See Johannes Schneider, Christian Meske, and Pauline Kuss, 'Foundation Models: A New Paradigm for Artificial Intelligence' Business & Information Systems Engineering 66(2) (2024) 221.

[2] PROBabLE Futures and National Police Chiefs' Council, 'Responsible AI Checklist for Policing' (May 2025), available at *https://library.college.police.uk/docs/NPCC/Responsible-AI-checklist-for-policing-2025.pdf*

[3] In this report, we use 'law enforcement' to refer mainly to policing (police forces and the bodies that support and coordinate them), while 'criminal justice system' refers to the wider set of institutions through which a case proceeds, including prosecution, the courts, prisons and probation. Our focus falls predominantly on policing, since this is where probabilistic AI is most actively deployed, but the analysis extends to the downstream stages of the criminal process where policing tools may have evidential or procedural consequences.

[4] For a working definition, see Evdoxia Taka, Temitope Lawal, Muffy Calder, Michele Sevegnani, Kyriakos Kotsoglou, Elizabeth McClory–Tiarks and Marion Oswald, 'Mapping the Probabilistic AI Ecosystem in Criminal Justice in England and Wales – Interim Report' (2025) arXiv, available at *https://arxiv.org/abs/2512.04116* Here, we characterised probabilistic AI systems as 'tools that use learning from data through machine learning (ML) models...[T]hese techniques rather than producing a single-point prediction in their output...predict a probability distribution over all possible outcomes'.

[5] Mary K Feeney and Sian Mughan, 'Policing, Technology, and Public Values: A Public Administration Research Agenda' (2025) 8(1) Perspectives on Public Management and Governance 12; Duncan Purves, 'Fairness in Algorithmic Policing' (2022) 8(4) Journal of the American Philosophical Association 741.

[6] See for instance, Daragh Murray, 'Using Human Rights Law to Inform States' Decisions to Deploy AI' (2020) 114 American Journal of International Law Unbound 158; Jennifer Cobbe, 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making' (2019) 39 Legal Studies 636; Marion Oswald, Jamie Grace, Sheena Urwin and Geoffrey C Barnes, 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental" Proportionality' (2018) 27(3) Information & Communications Technology Law 223.

[7] Ajay Sandhu and Peter Fussey, 'The "Uberization of Policing"? How police negotiate and operationalise predictive policing technology' (2021) 31 Policing and Society 66.

[8] Kevin D Haggerty and Richard V Ericson, 'The Surveillant Assemblage' (2000) 51 British Journal of Sociology 605.

[9] See generally, Sarah Brayne, 'Big Data Surveillance: The Case of Policing' (2017) 82(5) American Sociological Review 977; Sarah Brayne, Predict and Surveil: Data, Discretion, and the Future of Policing (Oxford University Press 2021).

[10] Muir R and O'Connell F, 'Policing and Artificial Intelligence' (Police Foundation, February 2025) 8-13, available at *https://www.police-foundation.org.uk/wp-content/uploads/2010/10/policing-and-ai.pdf.pdf*

[11] Michael Veale and Reuben Binns, 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data' (2017) 4(2) Big Data & Society, 1; Pamela Ugwudike, 'AI Audits for Assessing Design Logics and Building Ethical Systems: The Case of Predictive Policing Algorithms' (2022) 2 AI and Ethics 199; Brayne, 'Big Data Surveillance' (n 9).

[12] Alexander Babuta and Marion Oswald, 'Data Analytics and Algorithms in Policing in England and Wales: Towards a New Policy Framework' (RUSI Occasional Papers, February 2020), available at *https://www.rusi.org/explore-our-research/publications/occasional-papers/data-analytics-and-algorithms-policing-england-and-wales-towards-new-policy-framework*

[13] Kristian Lum and William Isaac, 'To Predict and Serve?' (2016) 13 Significance 14.

[14] Lyria Bennett Moses and Janet Chan, 'Algorithmic Prediction in Policing' (2018) 28(7) Policing and Society 806.

[15] Ibid.

[16] Ibid.

[17] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger and Suresh Venkatasubramanian, 'Runaway Feedback Loops in Predictive Policing' in Sorelle A Friedler and Christo Wilson (eds), Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR 2018) 160; Veale and Binns (n 13); Adriane Chapman, Philip Grylls, Pamela Ugwudike, David Gammack and Jacqui Ayling, 'A Data-Driven Analysis of the Interplay Between Criminological Theory and Predictive Policing Algorithms' in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM 2022) 36.

[18] P Jeffrey Brantingham, Matthew Valasik and George O Mohler, 'Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Controlled Trial' (2018) 5(1) Statistics and Public Policy 1.

[19] Raja Parasuraman and Dietrich H Manzey, 'Complacency and Bias in Human Use of Automation' (2010) 52 Human Factors 381.

[20] John D Lee and Katrina A See, 'Trust in Automation: Designing for Appropriate Reliance' (2004) 46(1) Human Factors: The Journal of the Human Factors and Ergonomics Society 50.

[21] Linda J Skitka, Kathleen L Mosier and Mark Burdick, 'Does Automation Bias Decision-Making?' (1999) 51(5) International Journal of Human-Computer Studies 991.

[22] Saar Alon-Barkat and Madalina Busuioc, 'Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' (2023) 33 Journal of Public Administration Research and Theory 153.

[23] Irina Carnat, 'Human, All Too Human: Accounting for Automation Bias in Generative Large Language Models' (2024) 14(4) International Data Privacy Law 299.

[24] Joe Pearson, Itiel E Dror, Emma Jayes, Grace-Rose Whordley, Georgina Mason and Sophie Nightingale, 'Examining Human Reliance on Artificial Intelligence in Decision Making' (2026) 16 Scientific Reports 5345.

[25] See generally, Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (Harvard University Press 2015).

[26] Madalina Busuioc, 'Accountable Artificial Intelligence: Holding Algorithms to Account' (2021) 81 Public Administration Review 825.

[27] Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review 18.

[28] Michael Veale, Max Van Kleek and Reuben Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making' (2018 CHI Conference on Human Factors in Computing Systems, Montreal QC, April 2018) Paper 440.

[29] Jakko Kemper and Daan Kolkman, 'Transparent to Whom?: No algorithmic accountability without a critical audience' (2019) 22(14) Information, Communication & Society 2081.

[30] Paul Marshall, James Christie, Peter Bernard Ladkin, Bev Littlewood, Stephen Mason, Martin Newby, Jonathan Rogers, Harold Thimbleby and Martyn Thomas, 'Recommendations for the Probity of Computer Evidence' (2021) 18 Digital Evidence and Electronic Signature Law Review 18. Also, see generally, Stephen Mason and Daniel Seng (eds), Electronic Evidence and Electronic Signatures (5th edn, IALS 2021).

[31] Ahmet Guler, Sedat Kula and Kaan Boke, 'Examining Public Support for AI in Policing: The Role of Perceived Procedural Justice' (2025) 26(6) Police Practice and Research 673.

[32] See generally, Tom R Tyler, Why People Obey the Law (Yale University Press 1990); Ben Bradford, Julia A Yesberg, Jonathan Jackson and Paul Dawson, 'Live Facial Recognition: Trust and Legitimacy as Predictors of Public Support for Police Use of New Technology' (2020) 60(6) The British Journal of Criminology 1502.

[33] Albert Meijer and Martijn Wessels, 'Predictive Policing: Review of Benefits and Drawbacks' (2019) 42(12) International Journal of Public Administration 1031.

[34] Jennifer Fereday and Eimear Muir-Cochrane, 'Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development' (2006) 5(1) International Journal of Qualitative Methods 80; Kevin Proudfoot, 'Inductive/Deductive Hybrid Thematic Analysis in Mixed Methods Research' (2023) 17(3) Journal of Mixed Methods Research 30.

[35] National Police Chiefs' Council, 'Artificial Intelligence (AI) Strategy' (NPCC, 15 November 2024), available at *https://www.npcc.police.uk/SysSiteAssets/media/downloads/publications/publications-log/science-and-innovation/2025/npcc-ai-strategy.pdf*

[36] National Police Chiefs' Council, 'Artificial Intelligence (AI) Playbook for Policing' (April 2025), available at *https://library.college.police.uk/docs/NPCC/Artificial-intelligence-playbook-policing-2025.pdf*

[37] PROBabLE Futures and National Police Chiefs' Council (n 2).

[38] Home Office, 'Consultation on a New Legal Framework for Law Enforcement Use of Biometrics, Facial Recognition and Similar Technologies' (GOV.UK, 4 December 2025), available at *https://www.gov.uk/government/consultations/legal-framework-for-using-facial-recognition-in-law-enforcement/consultation-on-a-new-legal-framework-for-law-enforcement-use-of-biometrics-facial-recognition-and-similar-technologies-accessible*

[39] Marion Oswald et al, 'Government Consultation on a New Legal Framework for Law Enforcement Use of Biometrics' (PROBabLE Futures, 10 February 2026), available at *https://research.northumbria.ac.uk/probablefutures/wp-content/uploads/2026/02/PROBabLE-Futures-Government-consultation-on-a-new-legal-framework-for-law-enforcement-use-of-biometrics-Final-10.02.2026.pdf*

[40] Temitope Lawal et al, 'PROBabLE Futures Submission in response to Call for Evidence on the 'Use of Evidence Generated by Software in Criminal Proceedings'' (PROBabLE Futures, 14 March 2025), available at *https://researchportal.northumbria.ac.uk/ws/portalfiles/portal/189739643/PROBabLE_Futures_Submission_CfE_DigEv_FINAL_14.04.25.pdf*

[41] Home Office, 'From Local to National: A New Model for Policing' (Home Office, January 2026) para 21, available at *https://www.gov.uk/government/publications/from-local-to-national-a-new-model-for-policing/from-local-to-national-a-new-model-for-policing-accessible*

[42] Although in its current form, OASys does not meet our definitional requirement of what AI is – "tools that use learning from data through machine learning (ML) models." See Taka et al (n 4) 3.

[43] Ministry of Justice, 'Freedom of Information Act Response: Reference 250107008' (28 February 2025), available at https://www.statewatch.org/media/4865/uk-oasys-foi-response-250107008.pdf

[44] Statewatch, 'UK: Over 1,300 people profiled daily by Ministry of Justice AI system to 'predict' re-offending risk' (Statewatch, 9 April 2025), available at https://www.statewatch.org/news/2025/april/uk-over-1-300-people-profiled-daily-by-ministry-of-justice-ai-system-to-predict-re-offending-risk/

[45] Jake Phillips, 'What Does AI Mean for Probation's Future?' (2026) 73(1) Probation Journal 3.

[46] For a Home Office estimate of the potential time savings from administrative AI applications across disclosure, CCTV analysis, case file production, crime recording, and translation, see Home Office, 'From local to national: a new model for policing' (n 41).

[47] Taka et al (n 4).

[48] These tools automatically detect and blur sensitive details (such as identities and licence plates in media files) for disclosure, data security or operational security reasons. See Accelerated Capability Environment, 'Helping Police Redact Sensitive Information in Media Files' (GOV.UK, 13 March 2024), available at https://www.gov.uk/government/case-studies/helping-police-redact-sensitive-information-in-media-files

[49] For a discussion of secondary trauma in officers reviewing child sexual abuse material, see Marije Keulen-de Vos and Funda Sanci, 'Secondary Trauma in Police Officers Investigating Child Sexual Exploitation Materials' in Rosemary Ricciardelli, Joy C MacDermid and Lorna Ferguson (eds), Occupational Stress Injuries: Operational and Organizational Stressors Among Public Safety Personnel (Routledge 2024).

[50] For a broader discussion on neurodivergence in policing, see Michelle O'Reilly, Samuel J Tromans and Alison Drewett, 'The Benefits and Challenges of Autistic Police Force Employees and Those with ADHD Supporting Neurodivergent Colleagues and Neurodivergent Victims, Witnesses and Suspects' (2026) Police Practice and Research 1.

[51] See generally, Muffy Calder et al, 'Responsible AI in Criminal Justice: LLMs in Policing and Risks to Case Progression' (2026) arXiv 1, 20–21, available at https://arxiv.org/pdf/2603.18116

[52] Taka et al (n 4).

[53] Specifically, recommendations 8 (risk classification framework), 9 (minimum design standards), 10 (context-specific oversight protocols), and 13 (retained human competence).

[54] Busuioc (n 26).

[55] See recommendations 9 (minimum design standards), 10 (context-specific oversight protocols), and 13 (retained human competence)

[56] Pearson et al (n 24); Parasuraman and Manzey (n 19); Alon-Barkat and Busuioc (n 22); Carnat (n 23).

[57] Pearson et al (n 24).

[58] Busuioc (n 26).

[59] Pearson et al (n 24).

[60] Ibid.

[61] Busuioc (n 26).

[62] See recommendation 10 (context-specific oversight protocols).

[63] Marion Oswald, 'Algorithm-assisted Decision-making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power' (2018) 376 Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 20170359.

[64] Ibid.

[65] Ibid.

[66] Peng Wang et al, 'Overloaded Minds and Machines: A Cognitive Load Framework for Human–AI Symbiosis' (2026) 59(109) Artificial Intelligence Review 1.

[67] See recommendations 9 (minimum design standards) and 10 (context-specific oversight protocols).

[68] UK Government, 'One Big Thing' (A Modern Civil Service), available at *https://moderncivilservice.campaign.gov.uk/one-big-thing/*

[69] See recommendation 17 (tiered AI literacy) for the content and structure of the training required to address these gaps.

[70] Defined as "the unsanctioned use of AI systems outside approved governance frameworks": Deepak Puthal et al, 'Shadow AI: Cyber Security Implications, Opportunities and Challenges in the Unseen Frontier' (2025) 6(405) SN Computer Science 1.

[71] Puthal et al (n 70).

[72] Nneoma Ogbonna, Mackenzie Jorgensen, Muffy Calder, and Marion Oswald, 'Don't Let AI Do the Policing: Findings from a Multi-Stakeholder Workshop on Evaluating Large Language Models for Policing' (PROBabLE Futures Policy Brief, May 2026) 6, available at *https://research.northumbria.ac.uk/probablefutures/wp-content/uploads/2026/05/LLM-Evaluation-PROBabLE-Futures-Policy-Brief.pdf*; See also recommendation 10 (context-specific oversight protocols).

[73] Marion Oswald, Luke Chambers and Angela Paul, 'Evaluating (Semi)-autonomous Systems in Policing and National Security: A New Framework Based on the Concept of "Intelligence"' in Proceedings of the First International Symposium on Trustworthy Autonomous Systems (TAS '23) (Association for Computing Machinery 2023). See also recommendation 8 (risk classification framework).

[74] Although some respondents described this as 'hallucinations', in this study, we adopt the use of 'error' instead. While hallucination is a popular buzzword, it gives a misleading impression of how AI systems operate. See Søren Dinesen Østergaard and Kristoffer Laigaard Nielbo, 'False Responses From Artificial Intelligence Models Are Not Hallucinations' (2023) 49(5) Schizophrenia Bulletin 1105.

[75] See recommendations 7 (meaningful evaluation), and 15 (rebalance policing-vendor relationship).

[76] Martyna Kusak, 'Quality of Data Sets That Feed AI and Big Data Applications for Law Enforcement' (2022) 23 ERA Forum (No 2, Springer Berlin Heidelberg) 209.

[77] See section 1.1 of this report.

[78] Criminal Procedure and Investigations Act 1996 (s 23(1)) Code of Practice, s 6.

[79] See for instance, Calder et al (n 51); Ogbonna et al (n 72).

[80] Moses and Chan (n 14).

[81] Home Office, 'From local to national: a new model for policing' (n 41).

[82] Also known as 'tool chaining'. See Calder et al (n 51).

[83] Interview 17. See further the project's response to the Criminal Justice Council consultation: Kyriakos Kotsoglou, Temitope Lawal and Marion Oswald, 'Deploying Artificial Intelligence Responsibly' (PROBabLE Futures, 13 April 2026), available at *https://researchportal.northumbria.ac.uk/ws/portalfiles/portal/227891708/PROBabLE-Futures-Response-to-CJC-Consultation-1.pdf*

[84] PROBabLE Futures and National Police Chiefs' Council (n 2).

[85] Calder et al (n 51).

[86] Joe Burton, Ardi Janjeva, Simon Moseley and Alice, "AI and Serious Online Crime," (CETaS Research Reports, March 2025), available at *https://www.turing.ac.uk/news/uk-law-enforcement-inadequately-equipped-tackle-ai-enabled-crime*

[87] Maria-Paz Sandoval, Maria de Almeida Vau, John Solaas and Luano Rodrigues, 'Threat of Deepfakes to the Criminal Justice System: A Systematic Review' (2024) 13(41) Crime Science 1; Karl Jones and Bethan Jones, 'How Robust is the United Kingdom Justice System Against the Advance of Deepfake Audio and Video?' (2022) 57(9–12) Electrotechnica and Electronica (E+E) 103; Don Fallis, 'The Epistemic Threat of Deepfakes' (2021) 34 Philosophy & Technology 623.

[88] Courts and Tribunals Judiciary, 'Artificial Intelligence (AI) Guidance for Judicial Office Holders' (Courts and Tribunals Judiciary, 31 October 2025) 7, available at *https://www.judiciary.uk/wp-content/uploads/2025/10/Artificial-Intelligence-AI-Guidance-for-Judicial-Office-Holders-2.pdf*

[89] Burton et al (n 86).

[90] PROBabLE Futures and National Police Chiefs' Council (n 2).

[91] PROBabLE Futures and National Police Chiefs' Council (n 2).

[92] National Police Chiefs' Council, 'Session 7 – Joint DDaT and Science and Innovation Committees' (NPCC, 9–10 July 2025) 24, available at *https://www.npcc.police.uk/SysSiteAssets/media/downloads/publications/disclosure-logs/npcc-central-office/2025/2412-2025-session-7-15092025.pdf*

[93] College of Policing, 'Building AI-Enabled Tools and Systems' (College of Policing, 17 June 2025), available at *https://www.college.police.uk/guidance/building-ai-enabled-tools-and-systems*

[94] Home Office, 'From local to national: a new model for policing' (n 41).

[95] National Police Chiefs' Council, 'AI Centre for Policing Will Help Catch More Criminals Quicker' (NPCC, 27 January 2026), available at *https://news.npcc.police.uk/releases/new-gbp-115m-ai-centre-for-policing-will-help-catch-more-criminals-quicker*

[96] Recommendations 1 (funding as governance), 2 (national benchmarking), and 7 (meaningful evaluation) address the conditions required to actualise this ambition.

[97] Miri Zilka et al, '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 (EAAMO '23) (Association for Computing Machinery 2023).

[98] UK Government, 'The Crown Prosecution Service: Correspondence Drafting Tool' (GOV.UK, 28 August 2025), available at *https://www.gov.uk/algorithmic-transparency-records/the-crown-prosecution-service-correspondence-drafting-tool*

[99] Although some forces, including Hampshire and Thames Valley, and West Midlands Police have made two ATRS submissions. See UK Government, 'Hampshire and Thames Valley Police: DARAT' (Gov.UK, 29 February 2024), available at https://www.gov.uk/algorithmic-transparency-records/hampshire-and-thames-valley-police-darat; 'West Midlands Police: exploratory analysis of sexual convictions' (Gov.UK, 29 February, 2024), available at *https://www.gov.uk/algorithmic-transparency-records/west-midlands-police-exploratory-analysis-of-sexual-convictions*

[100] Zilka et al (n 97).

[101] PROBabLE Futures and National Police Chiefs' Council (n 2).

[102] UK Prime Minister's Office, The King's Speech 2026 (Prime Minister's Office, 13 May 2026) 69, available at *https://assets.publishing.service.gov.uk/media/6a046665c0cc74b4523e4d3b/The_King_s_Speech_2026_-_background_briefing_notes.pdf*

[103] Office of the Police Chief Scientific Adviser, 'Police STAR Fund' (Science & Technology in Policing, 2026), available at *https://science.police.uk/opportunities/police-star-fund/*

[104] Department for Science, Innovation and Technology, 'Government to Take a "Test and Learn" Approach with Spending on AI and Digital to Push Innovation' (GOV.UK, 10 March 2025), available at *https://www.gov.uk/government/news/government-to-take-a-test-and-learn-approach-with-spending-on-ai-and-digital-to-push-innovation*

[105] See specifically recommendations 1 (funding as governance), 2 (national evaluation programme), and 7 (meaningful evaluation) that set out the conditions required to strengthen and extend funding.

[106] Guler, Kula and Boke (n 31).

[107] Kaylyn Jackson Schiff, Daniel S Schiff, Ian T Adams, Joshua McCrain and Scott M Mourtgos, 'Institutional Factors Driving Citizen Perceptions of AI in Government: Evidence from a Survey Experiment on Policing' (2025) 85(2) Public Administration Review 451.

[108] Rosamund Powell, Marion Oswald and Ardi Janjeva, 'UK Public Attitudes to National Security Data Processing: Assessing Human and Machine Intrusion' (CETaS Research Reports, April 2025), available at *https://cetas.turing.ac.uk/sites/default/files/2025-04/cetas_research_report_-_public_attitudes.pdf*

[109] Ibid.

[110] Home Office, 'Public Attitudes to Police Use of Facial Recognition Technology' (GOV.UK, 4 December 2025), available at *https://www.gov.uk/government/publications/public-attitudes-to-police-use-of-facial-recognition-technology/public-attitudes-to-police-use-of-facial-recognition-technology*

[111] Ibid.

[112] Ibid.

[113] Ibid.

[114] Sophie Reid, 'Home Office AI in Policing Dialogue Evaluation: Final Evaluation Report' (Sciencewise, January 2026), available at *https://sciencewise.org.uk/wp-content/uploads/2026/02/Home-Office-UKRI-Sciencewise-AI-in-policing-dialogue-evaluation-final-report_Jan-2026-Sophie-Reid.docx.pdf*

[115] Ibid.

[116] Guler, Kula and Boke (n 31).

[117] Ibid.

[118] Reid (n 114).

[119] Burton et al (n 86).

[120] Ogbonna et al (n 72).

[121] Oswald, Chambers, and Paul (n 73).

[122] Calder et al (n 51).

[123] Ibid, at 16-17.

[124] Ibid, at 10.

[125] Ogbonna et al (n 72) 6.

[126] PROBabLE Futures, 'Mapping AI Tools in Criminal Justice across England and Wales' (2026), available at *https://probablefutures.github.io/ai-mapping/*

[127] Lawal et al (n 40).

About the Authors

Dr Temitope Lawal is a Research Fellow in Law at Northumbria University

Dr. Kyriakos N. Kotsoglou is an Associate Professor of Law (Criminal Evidence) at Northumbria University

Dr Evdoxia Taka is a Research Associate in Computing Science at The University of Glasgow

Professor Marion Oswald MBE is a Professor of Law at Northumbria University and PROBabLE Futures Principal Investigator

Dr Michele Sevegnani is a Senior Lecturer in Computing Science at The University of Glasgow

Professor Dame Muffy Calder is a Professor of Formal Methods in Computing Science at The University of Glasgow

PROBabLE Futures

Probabilistic AI Systems in Law Enforcement Futures

https://research.northumbria.ac.uk/probablefutures/

probablefutures@northumbria.ac.uk

https://www.linkedin.com/company/probablefutures/

https://bsky.app/profile/probablefutures.bsky.social

RAi | UKRI UK Research and Innovation

Northumbria University NEWCASTLE | University of Glasgow | University of Northampton | Newcastle University | University of Cambridge | University of Leicester