AI in the Metropolitan Police Service

A report by the London Policing Ethics Panel

July 2026

This report has been prepared with the assistance of AI for research and editing purposes. All material has been independently verified by the authors, and the London Policing Ethics Panel accepts full responsibility for its content. The images have been generated using AI assistance.

Table of Contents

Table of Contents 3

Executive summary 4

Introduction 9

An applied ethics framework 10

- Human dignity and equality domain 11
- Transparency & accountability domain 12
- Technical integrity domain 13
- Legal & Human Rights compliance domain 13
- Human oversight & control domain 14
- Social responsibility & public value domain 14
- Resources 14

Ethical implications in how humans interact with AI 14

- Implementing AI in everyday routines: the example of co-pilot. 17

The bigger picture: anticipating benefits and risks from AI 21

- The organisation and its people 21
  - Reducing the burden of administration 21
  - How reliable is the human in the loop? 22
  - Who is accountable? 22
  - Building capacity for AI implementation 23
  - Others using AI to communicate with the Met 23
- Serving state and civil society: AI and legitimate authority 24
- Serving victims: minimising vulnerability 25
- Serving persons of interest: maintain respect for humanity, act proportionately and fairly. 26

Ethical tensions and recommendations 26

- Trustworthiness and innovation 26
- Transparency and trust 30
- Accountability and the ‘liability sink’ 31

Executive summary

This report provides an ethical review of artificial intelligence adoption by the Metropolitan Police Service (MPS). It is intended for use by MPS, MOPAC and the wider public as the MPS expands its AI capabilities. The report synthesises existing applied ethics frameworks for AI in policing, examines the behavioural dynamics of human-AI interaction in practice, and identifies three key ethical tensions requiring active resolution.

The report draws on the Panel's experience of and reported research on Microsoft Copilot, which is the most widely deployed AI tool in the MPS to date, as an illustrative case study. It then extrapolates lessons to the broader landscape of AI in policing. It concludes with seven recommendations.

Approach

We adopt the College of Policing and Centre for Data Ethics and Innovation definition of AI: technologies capable of performing tasks that would otherwise require human intelligence, such as visual perception, speech recognition, and language translation. Current MPS AI deployment includes:

- Microsoft Copilot chat, available organisation-wide (web-grounded, enterprise data-protected)
- Microsoft M365 Copilot, accessible to approximately 350 licensed users, able to draw on internal MPS data systems
- Specific localised tools, for example for redaction and video analysis
- Live Facial Recognition, to date the most publicly prominent AI application, covered in separate reports
- Data integration and representation systems currently under exploration

Copilot has attracted little public comment despite widespread deployment, making it an instructive starting point precisely because it is familiar and apparently low risk.

An Applied Ethics Framework

A review of over twenty UK and Australasian frameworks for AI in policing reveals strong convergence around some twenty common ethical principles, organised into six domains:

- Human Dignity and Equality: respect, dignity, impartiality, proportionality, and fairness
- Transparency and Accountability: explainability, transparency, ongoing evaluation, and the right to challenge AI-informed decisions
- Technical Integrity: robustness, safety, accuracy, appropriate granularity, and officer skills and knowledge
- Legal and Human Rights Compliance: privacy, lawfulness, Human Rights Act compliance, and data protection

- Human Oversight and Control: the requirement for a human decision-maker in the loop, and for independent oversight
- Social Responsibility and Public Value: public good, avoidance of harm, sustainability, ownership, and environmental responsibility

These principles are not self-applying. Determining which carry greatest weight in any specific deployment requires careful attention to the purpose of the AI application, the trade-offs involved, and potential long-term and collateral consequences. The NPCC's Responsible AI Checklist incorporates some but not all of these principles.

Ethical Implications of Human-AI Interaction

An important claim in this report is that applied ethical frameworks, while necessary, are not sufficient. AI tools are not neutral or passive: they embed the values and choices of their designers and elicit distinctive human behaviours. *How human users interact with AI outputs creates its own ethical dimension*. This human-AI interaction is increasingly well-evidenced but far less frequently addressed in governance frameworks.

Key documented behavioural risks include:

- Cognitive surrender: the tendency of users, including experienced professionals, to accept AI outputs uncritically.
- Deference to AI outputs and diminution of confidence in own discretion and professional assessment.
- Failure to review AI outputs, particularly when under time pressure, allowing AI confabulation to pass uncorrected.
- Less experienced professionals unable to detect AI omission and confabulation as they lack the professional expertise necessary to do so.
- Deskilling: gradual erosion of written communication skills and capacity for independent judgement through habitual reliance on AI.
- Anchoring: allowing an AI-generated starting point to unduly shape subsequent thinking, even when officers attempt critical review.

These are not theoretical concerns. They are documented consequences of AI adoption in analogous professional settings, including medicine, law, and financial services, and they have direct implications for the quality of MPS decision-making, victim contact, and investigation integrity.

Copilot Pre-Mortem

The report applies a 'pre-mortem' methodology to MPS's use of Copilot — working backwards from hypothetical failure to identify risks and mitigations in advance. MPS's stated goals for Copilot include increasing productivity, facilitating information sharing, improving communications, and supporting training.

The analysis identifies failure points across all intended use cases. These are not merely operational difficulties: they create ethical challenges that accumulate over time and have the capacity to affect MPS employees, victims, persons of interest, and the organisation's trustworthiness. The pre-mortem tables in the full report detail specific risks and mitigation strategies for both current and anticipated uses.

The Organisation and Its People

The productivity case for AI in report-writing and administration is weaker than often claimed. Research points to two structural reasons: officers already use templates for most routine communications, and many communications are sufficiently nuanced that AI requires substantial officer intervention to be fit for purpose.

An ethically serious concern is the accountability gap. The 'human in the loop' model central to most AI governance assumes that users can meaningfully review AI outputs. In practice, this assumption is frequently undermined: by cognitive surrender, by the speed and opacity of AI outputs, and potentially inexperience of officers in roles requiring oversight of AI outputs. The result is what researchers term a 'liability sink': accountability flows to individuals for errors they were not realistically positioned to catch, while the organisational and procurement decisions that created those conditions remain unexamined.

Effective deployment requires sustained investment in prompt engineering training, supervisory cultures that reinforce critical engagement with AI, and clear organisational responsibility for post-deployment monitoring.

AI, Communications, and Public Trust

The report identifies risks to MPS's epistemic authority, that is, the public's perception of MPS communications as trustworthy and authentic. Research suggests that people discount material they believe to be AI-generated, even when it is factually accurate. For a service whose authority depends on being believed, this is not a trivial problem.

The risk is compounded in the context of victim communications. Peer-reviewed research demonstrates directly that victims judge AI-mediated police communications as procedurally less just, and therefore less trustworthy, than human communications, even when the words used are identical. Victims are seeking human connection and recognition, not efficiency alone.

For persons of interest, AI errors carry sharper legal and reputational exposure: risks of directing investigatory resource incorrectly, and in the worst cases contributing to wrongful outcomes.

Our report also notes that MPS will need to manage an increasing volume of AI-assisted communications directed at it from members of the public and officers. This is a demand the pre-LLM infrastructure was not built to handle.

Three Key Ethical Tensions

1. Trustworthiness and Innovation

MPS must balance rigorous governance with the operational imperative to innovate. Failure to provide sanctioned AI risks employees turning to 'shadow AI' (unauthorised consumer tools) with associated data security and governance risks. Public attitude research shows that over 60% of the public are supportive of AI use in policing but this support is contingent on responsible use. Demonstrably robust governance is therefore a precondition for public trust, not an obstacle to innovation.

MPS currently has governance structures to control the introduction of AI (Organisational Design Authority, Technical Design Authority, Data Capability Assurance Framework), but ethical appraisal is not yet embedded as a stand-alone requirement. There is no clear mechanism for public engagement with MPS AI innovation, an omission at odds with the principle of policing by consent.

2. Transparency and Trust

The NPCC covenant requires 'Maximum Transparency by Default' for all AI use. We endorse this principle while acknowledging recent research showing that AI disclosure can, paradoxically, reduce trust in some contexts. However, the same research demonstrates that being discovered to have concealed AI use has an even more damaging effect on trust. The conclusion is clear: MPS should enforce maximum transparency, accepting a measured short-term cost to avoid a potentially catastrophic longer-term one.

3. Accountability and the Liability Sink

The current governance model places accountability for AI outputs on individual officers who may lack the knowledge, experience, or practical capacity to identify errors. This is neither fair nor effective. A partial analogy exists with firearms policing, where personal responsibility is retained but robust training, supervision, and shared risk structures ensure this is meaningful. A credible solution will likely require mandatory disclosure of AI use; monitoring and verification processes; inbuilt prompts to users; training in AI limitations and prompt engineering; a 'just culture' that differentiates innocent AI-induced errors from careless or malicious misuse; and contractual obligations on providers to disclose reliability metrics.

Summary of Recommendations

Recommendation One: Embed the Responsible AI Checklist [page 29]

Ensure that MPS AI governance incorporates all elements of the NPCC's Responsible AI Checklist in whatever way fits most effectively and efficiently into existing MPS approval processes.

Recommendation Two: Align Local Operating Procedures with Strategic Governance [page 29]

Establish processes for assuring that when Local Operating Procedures are developed for specific AI applications, they are consistent with MPS's strategic AI governance goals and processes.

Recommendation Three: Transparency in Administrative AI Decisions [page 29]

Where AI is included in systems for making administrative decisions about people — whether employees or members of the public — this should be done with full transparency, and clear proportionate processes should enable people to challenge decisive AI assessments.

Recommendation Four: Public Engagement Strategy [page 30]

A strategy for public engagement, aiming to build understanding of the steps being taken for responsible AI adoption, should be given appropriate priority by MPS and MOPAC. Public engagement is a prerequisite for trust in AI and consistent with the principle of policing by consent.

Recommendation Five: Enforce Maximum Transparency [page 31]

MPS should take steps to enforce maximum transparency in relation to activities that have been supported with AI, consistent with the NPCC covenant.

Recommendation Six: Resolve the Accountability Gap [page 33]

MPS should resolve and make known to MOPAC and other stakeholders: the ways it systematically limits AI error; how it ensures employees are adequately trained; how it holds AI providers to account for functionality standards; and the principles by which MPS employees will be held accountable for AI-induced errors.

Recommendation Seven: Address Impact in Life Cycle Monitoring [page 33]

MPS should review the unforeseen consequences of AI adoption in its AI life cycle monitoring.

Introduction

This report provides an overview of ethical considerations pertinent to the adoption and implementation of AI by the Metropolitan Police Service. It is intended for use by MPS, MOPAC and the wider public as MPS expands its use of AI.

We build on the very extensive work that has already been done by a range of bodies, including the NPCC and the College of Policing, summarising high level applied ethical principles for AI in policing. (These are common ethical principles applied specifically to the domain of AI in policing.) Our report begins with a synthesis of this plethora of applied ethics frameworks, identifying their common parts.

We then move beyond those high-level principles to consider some ethically relevant behaviours that may arise in the course of implementing AI. By way of contrast with the applied ethics framework, we call this second set of considerations 'interactional ethics' because they arise out of human interaction with AI tools and their outputs as they are actually used. The identified behaviours possess a distinctly ethical flavour because, if things are to go better rather than worse when AI is used in policing, they will demand attentiveness, ethical awareness and self-discipline in individuals and organisations deploying the tools. We discuss MPS adoption of Microsoft's Co-pilot to surface these interactions, and we then move to extrapolate the lessons from copilot to the expanding world of police AI.

We conclude by discussing three key ethical tensions that call for active resolution if MPS implementation of AI is to secure public trust, achieve its policing aims, and secure workforce support. These are trustworthiness and innovation, transparency and trust, and accountability for individuals and organisations.

What we mean by AI

We have adopted the definition provided by the College of Policing and the government's Centre for Data Ethics and Innovation: "technologies with the ability to perform tasks that would otherwise require human intelligence, such as visual perception, speech recognition, and language translation".[1] The CDEI argued "It is essential for the police service to adopt new technology in a way that prevents and detects crime, keeps the public safe and safeguards national security, while maintaining legitimacy and respecting individual rights. At a time of profound technological change, the police must adapt and innovate to continue operating effectively in a rapidly evolving landscape".

To our knowledge, at the date of writing the widest use of AI by the MPS has been Microsoft Copilot. Copilot has attracted little if any public comment. A very high-profile use of AI by the MPS, on the other hand, has been the deployment of Live Facial Recognition and we have ourselves published two reports on its adoption.

We will discuss Copilot later in this report, using it both as an example of how AI may currently be used in practice and also of how future AI might function in police organisations. Basic Microsoft Copilot chat has been made available across MPS, and a limited number of users have access to the more powerful Microsoft M365 Copilot. Both are within a 'tenant' arrangement which protects MPS data and keeps it within organisational boundaries. Copilot chat is grounded in data from the web only (not organisational content like files or emails) and is protected by enterprise data protection (the tenancy arrangement) to make it safe for work use. A further 350 licensed users have access to an 'organisation-grounded' version of Copilot which is able to draw on the Met's own data systems, again within the tenancy arrangement. (M365 Copilot was piloted in two phases, and we are grateful to MPS for sharing the findings.)

Alongside some specific tools for specific purposes such as redaction and video analysis, at the date of writing MPS was also exploring adoption of third-party data integration and representation systems. Such systems collate (integrate) data from fragmented data archives and then offer them (represent them) in a user-friendly way to support extraction of operational insights. AI is applied on top of the integrating element of these systems to help users seek specific insights.

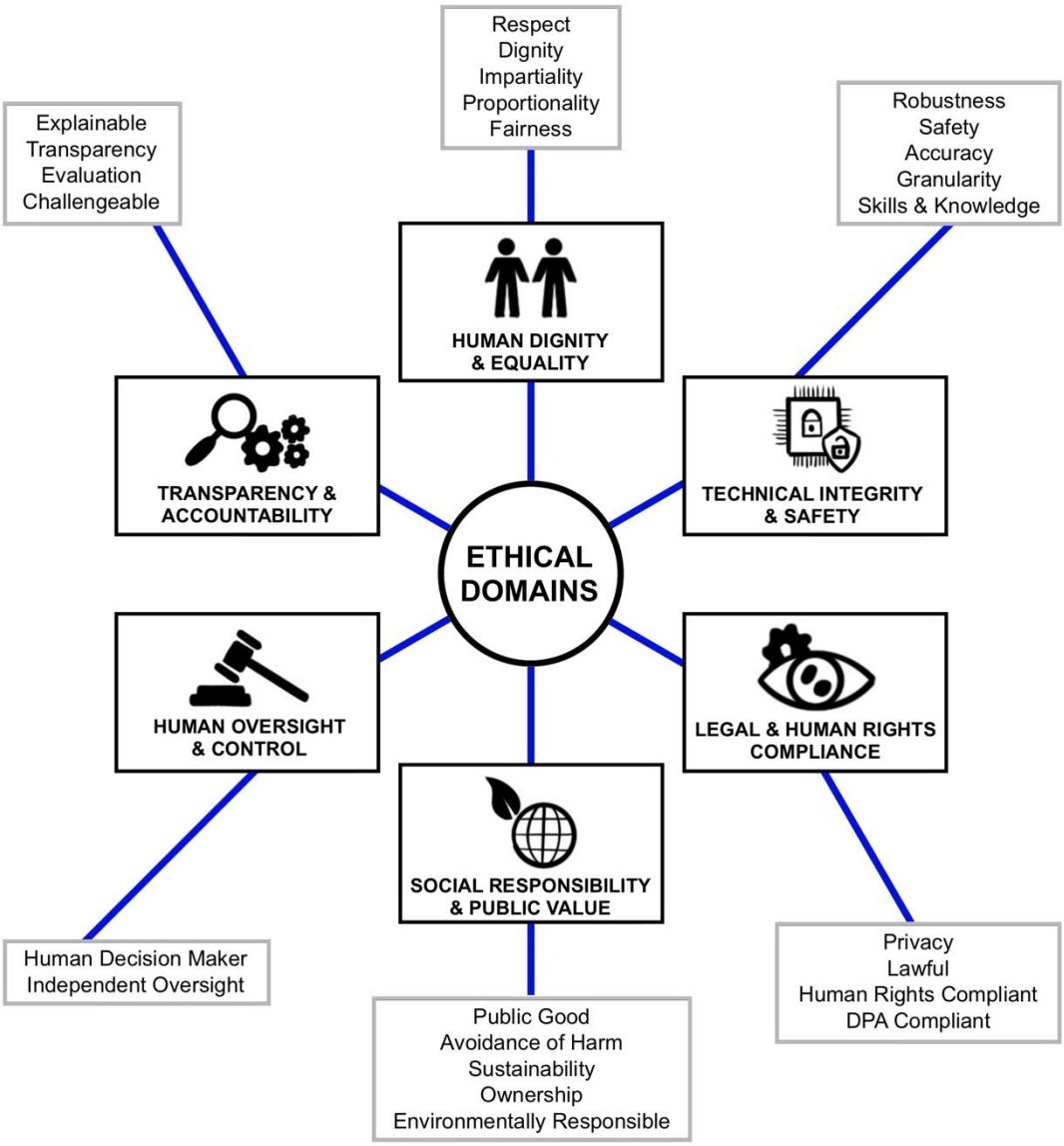
An applied ethics framework

The rapid development of AI applications has been accompanied by a proliferation of ethical and governance guidelines. Writing in 2023, Corrêa et al had located over 200 documents spanning 37 countries. Proposed frameworks have varied in how precisely norms are defined, what core principles are identified, and how authoritative the issuing source is. Notwithstanding notable changes of emphasis over time (from fairness, reliability, and dignity taking priority, to transparency being now the most frequently cited principle) there has been significant convergence.[2]

Corrêa and colleagues found consistency in the current top five priorities from sources emanating from the UK and USA, although institutional concerns impacted ethical priorities. The main preoccupation of governmental institutions was the need for transparency; civil society organizations (i.e. non-governmental, non-commercial groups that act in the public interest) primarily prioritised the principle of fairness; and private corporations were more concerned with reliability. These differences are instructive, and they point to potential sources of disagreement over what should be emphasised in discussions around AI acceptability.

A review of UK and Australasian frameworks assessing the ethical status of AI in policing specifically, yielded some twenty ethical principles they had in common. These have been incorporated into the infographic below (the sources are listed in Appendix One). The principles are explained after the infographic.

Figure 1: infographic: ethical domains



Determining which principles are most relevant and should be given greatest weight requires consideration of:

- the stated purposes of the application;
- tradeoffs, for instance between potential harms against benefits; transparency and honesty; privacy and openness; efficiency and social responsibility;
- long term as well as short term adverse consequences;
- and possible adverse collateral impacts such as reputational damage.

Human dignity and equality domain

- ***Respect*** so that when AI is in use, people are always treated as worthy of respect. This requirement is in many ways an amalgamation of all of the

principles summarised in this infographic. At the level of the individual encounter, treating someone as worthy of respect means they are dealt with as a whole person not as a data point. For example, an officer using AI output should know what the tool can and cannot tell them, and does not treat a risk score as a verdict; the person subject to an AI-informed decision has some meaningful way of knowing that has happened; and the decision retains genuine human judgment, not blindly following an algorithmic output. Additionally, respect for a person subject to a policing decision requires that someone is genuinely accountable, which means the human-AI interface has to be designed so that accountability does not vanish in the gap between officer, organisation, and AI output.

- ***Dignity*** calls for AI to be used in policing only in ways that treat everyone as having intrinsic value as a human being.
- ***Impartiality*** means using AI in ways that ensure people are treated without fear or favour and users taking active measures to mitigate against potential algorithmic bias (e.g. taking into account that an AI may be less reliable when used for a particular group of people in a population)
- ***Proportionality*** means ensuring that AI is used in a manner which is proportionate, balancing for example incursions on privacy with the public safety need; supporting decision-making that is consistent with the law, balanced by police discretion and subject to justification.
- ***Fairness*** refers to using AI in ways that achieve beneficial outcomes and avoidance of harm equally for all, and ensuring that AI does not replicate prejudices based on protected and other distinguishing characteristics

Transparency & accountability domain

- ***Explainable*** refers to the need to be able to explain, accessibly and intelligibly, the input basis of the algorithm; to ensure that practitioners understand how it works and its likely effects; and whether the AI system can produce a trail and/or justifications for its outputs. The NPCC notes that level of explanation expected will be determined by the function the AI application performs (e.g. such as informing a high-impact decision about an individual); and the outputs required of it (i.e. who needs to understand what regarding the output, and how was this reached).
- ***Transparency*** is the need for openness with the public about how and when AI is used, enabling scrutiny of purpose, justification and limitations. The NPCC covenant stipulates that all use of AI will be subject to ‘Maximum Transparency by Default’ and this will typically include publishing an overview of the algorithms used and the known limitations of AI training data.
- ***Evaluation*** refers to the need to monitor outcomes throughout the AI implementation life cycle and take action to mitigate adverse impacts.

- ***Challengeable*** refers to the need for adequate oversight mechanisms, identification of potential bias, a right of notification when AI is used in decision-making, and responsivity to requests to check the accuracy of personal data.

Technical integrity domain

- ***Robustness*** refers to the scientific reliability of input data, the ability of the system to withstand malicious exploitation[#] and ensuring that the application is used in line with its designed policing purpose.
- ***Safety*** refers to soundness of its operation in non-ideal conditions, its ability to function notwithstanding cyber-attack or other system failure, and resilience to undesirable reinforcement learning behaviour when an application is in use.
- ***Accuracy*** calls attention to the need to ensure that input data are valid, and subsequently validation of the tool itself.
- ***Granularity*** is particularly significant in predictive policing applications. It requires that data references are neither too broad (e.g. postcode-level averages standing in for individual-level data) nor inappropriately precise given data quality limitations.
- ***Skills and knowledge*** means recognising that technical integrity is built not just in the design but also the implementation of an application; calling for appropriate training and testing for police users and decision makers when interpreting and apply outputs, and potentially credentialling for higher risk uses.

Legal & Human Rights compliance domain

- ***Privacy*** means ensuring that personal information is only collected and used for legitimate purposes, cannot be inappropriately accessed, and that it is safeguarded against internal and external threats.
- ***Lawful*** means there must be adherence to a legally sanctioned policing purpose and, in corporate uses, relevant legislation such as employment law.
- ***Human Rights compliant*** refers to use consistent with human rights obligations enshrined in the Human Rights Act 1998 and the European Convention.
- ***Data Protection compliant*** means its use must be in line with data protection legislation and guidance.

[#]NIST describes four varieties of threat to AI systems (referred to as adversarial machine learning [AML]), which require robust procedures to resist: evasion (e.g. adding markings to road signs to make an autonomous vehicle misinterpret them) poisoning (introducing corrupted data during training phases) privacy (attempts to learn sensitive information about the AI or the data it was trained on in order to misuse it) and abuse (insertion of incorrect information into a source, such as a webpage or online document, that an AI then absorbs). https://csrc.nist.gov/pubs/ai/100/2/e2023/final

Human oversight & control domain

- ***Human decision maker*** calls for inclusion of a human in the decision-making loop to ensure that judgements are rooted in human discretion and that there is accountability for action. This also requires the human decision maker to retain a degree of independent reasoning and judgement. This raises questions around the effectiveness of the human in the loop as an ethical safeguard, discussed in our section on human-AI interaction.
- ***Independent oversight*** refers to the desirability of effective governance mechanisms; these should include persons independent of deploying organisations to assess the ethical status of applications and deployments.

Social responsibility & public value domain

- ***Public good*** refers to using AI in policing in ways that produce an overall net good for the public, balancing benefits and harms.
- ***Avoidance of harm*** means anticipating and mitigating potentially adverse consequences of AI use, which could include combating use of AI for criminal purposes.
- ***Sustainability*** means considering the potential for and impacts of transformation in working practices, and ensuring the organisation can maintain key functions with and without use of AI.
- ***Ownership*** means enacting due diligence in understanding the proprietary rights of any AI provider and retaining appropriate data control in deploying organisations.
- ***Environmentally responsible*** means taking account of the energy costs and carbon footprint of both AI and its alternatives

Resources

In Appendix One we list the major documents concerned with the ethics and governance of AI in policing on which we have drawn. We have also included a link to the Responsible AI Checklist developed by the PROBable Futures Project.

Ethical implications in how humans interact with AI

Although widespread adoption of AI is still a relatively new phenomenon, it is possible to discern some distinctive patterns in how humans and AI interact, and to anticipate some potential impacts. While applied ethical frameworks for AI have rapidly proliferated, it has taken longer for users and observers to recognise the ethical significance of the particular ways in which humans and AI interact and how each shapes the other.

The Turing Institute[3] set out clearly how the life cycle of algorithmic systems has three phases: *translating in* of human purposes, values, and choices during the design process; digital processing incorporating these purposes, value and choices;

and finally, *translation out* of purposes, values, and choices in the systems' algorithmic outputs. Whatever the nature of the AI under consideration, recognition of the value choices in the three phases reminds us that AI of all kinds is *value laden*. These tools are *not neutral, not unbiased, and not passive objects*. In use, they create a very particular starting point for a human-machine interaction shaped by the intentions of one and the design of another.

This stark fact means that to judge the ethical operation of AI applications we have to ask what happens when humans first seek and then make use of AI outputs. What is sought will reflect human social goals, what AI puts out will reflect what was built into it, and how humans respond to AI outputs will shape the results of the interaction. As is the case with any digital tool, the already value-laden outputs can be used well, or used carelessly, or used maliciously.

The behavioural elements in human AI interaction are now gaining far greater attention[4], and carry important implications for ethical policing. Understanding behavioural interaction with AI is particularly significant in policing because police data systems contain the most sensitive personal data, processing which promises immense potential to bring about good or ill. Using AI to improve data use in intelligence analysis, decision support, internal and external communications, and internal monitoring and accountability processes carries great potential and could concomitantly generate significant human consequences. Looking to front line operations, police officers are required to act as trustworthy professionals, making critical and sometimes high-risk decisions under pressure, with a high degree of autonomy and discretion. Ensuring that AI is used well in by the human actors in these circumstances is vitally important.

In the table below, we have summarised the key human-AI interactions which are now well-evidenced. We have considered the need for corrective action, and potential systemic effects of unmitigated AI impacts.

Many of these human-AI interactions relate closely to the domain of human dignity and equality in our previous diagram and they also carry implications for efficiency and effectiveness. They are therefore of signal importance to maintaining trustworthiness and trust in policing.

Table 1: Human-AI interaction

	EFFECT	CORRECTIVE	SYSTEMIC
Effects of LLM confabulation	Reliance on erroneous information.	Expectation that human expert will review outputs. Takes time to review. Human may not have sufficient knowledge	Less time saved than anticipated. Erroneous outputs blamed on humans. AI suppliers avoid accountability.
Effects of LLM omissions	Reliance on misleading information.	Problem of unknown unknowns: omission only perceptible to human experts.	Over time, loss of expertise capable of contesting LLM outputs.
Effects of inaccurate summaries, transcriptions	Reliance on misleading information.	Requires double checking and repeating.	Less time saved than anticipated. Unchecked outputs blamed on individuals, not AI.
Effects of cognitive biases in LLMs mirroring human cognitive bias	May be unnoticeable to users as they likely share the same bias. e.g framing effects, stereotyping.	If user is aware, they can correct the agent, but this requires user awareness.	Depends on context, e.g. stereotyping could entrench injustice; framing effects mislead investigators.
Effects of embedded bias in AI derived from training data	Humans unaware of AI discrimination. Humans justify discriminatory action as "objective".	Requires monitoring for bias, which may be imperceptible to users.	Entrenched discrimination.
Effects of LLM self confidence	Misplaced trust prompted by persuasive tone. LLM self-confidence leads to user self-confidence.	Requires user scepticism. Requires double check. Could be offset by using appropriate prompts to LLM.	Outputs blamed on individuals.
Effects of expectations of AI accuracy and reliability	Over reliance on AI. Potential cognitive surrender. Automation bias could undermine exercise of discretion.	Proper training required. Calls for constant conscious awareness of AI limitations.	Erosion of discretionary judgment & consequent injustice. Long term deskilling.
Effects of LLM capability	LLM output passed off as authentic human engagement. AI as source concealed. Lack of candour.	Requires scrupulous honesty, candour about AI output, transparency	Risks loss of trust.
Effects of decision support fallibility	User faced with deciding whether to follow AI direction or own judgement.	The exercise of discretion becomes key skill, but user may not acquire sufficient experience.	Increasing tendency to rely on AI direction with loss of expert discretion.
Effects of AI tools being updated by vendor	Users find known bugs replaced by unknown ones.	Requires operators to learn new limitations with each update.	Uncorrected errors.
Effects of potential AI withdrawal	Reliance on AI has created dependency, both in individuals and systems.		Systems and workflows designed around AI fail.
Effects of AI tools and LLMs being unaccountable for error	Individual users become liable for AI faults. Accountability gap where neither user nor machine is properly accountable.	Liability imposed on individuals.	Injustice to individuals. Lack of accountability.
Human in the loop fallacy	Humans are treated as guardrails in relation to outputs they may not understand. Organisational systems may make expectations of human monitoring ineffective (e.g. production pressure).	Requires individuals & organisations to understand and take responsibility for limitations of AI.	Costs of AI fallibility imposed on individual users & employees. Lack of incentive to AI suppliers to improve if they are not held accountable.

Implementing AI in everyday routines: the example of co-pilot.

In this section we have used Microsoft Copilot's capabilities as a worked example to consider how human-AI interaction could work in practice in policing. Copilot is a useful starting point because it is in use in MPS, shares capabilities in common with other Large Language Models (LLMs), and we can extrapolate from how users interact with Copilot to anticipate how they may interact with other AI. There is value in looking at Copilot precisely because it is familiar, flexible, ostensibly benign, and an early step on the MPS AI journey.[Ω]

We carried out a 'pre-mortem' exercise to consider the ethical implications of common-place interactions with copilot. The concept of a pre-mortem was developed as a means of anticipating how a project might fail, in order that mitigating actions can be taken.[5] Analysis proceeds from the premise that the project in hand has not achieved its goals, and the task is to work backwards to account for why this happened and how it could be rectified. It is an effective way of countering the technology hype and optimism bias common to technology projects, and for identifying mitigations at an early stage.

Microsoft states Copilot offers a wide range of AI-assisted capabilities to optimize workflows, enhance productivity, and assist in making faster more informed decisions.[6] It is suggested that copilot can support tasks[*] such as:

Document creation: drafts reports, presentations, and emails by generating content and offering editing suggestions and personalised outputs;

Data analysis: analyses data trends, generates formulas, and creates visualizations to simplify decision-making;

Project management: tracks tasks and schedules, records meetings, takes notes, and creates action items based on conversations in Teams calls;

Communication: summarizes key actions and lengthy email conversations, drafts emails in a desired tone, and structures communications for the clearest message possible.

MPS issued internal documentation for its users, proposing that Copilot could be used for administrative support, summarising knowledge and research, generating ideas and assisting with planning. Users are expected to check to verify accuracy,

---

[Ω] As noted in our introduction MPS has introduced basic copilot chat across the organisation and also tested an 'organisation-grounded' version of M365 copilot with a cohort of licensed users.

[*] Current Microsoft Copilot Terms of Use include, under "IMPORTANT DISCLOSURES & WARNINGS": "Copilot is for entertainment purposes only. It can make mistakes, and it may not work as intended. Don't rely on Copilot for important advice. Use Copilot at your own risk." As of now, the language still appears on the page, though Microsoft has signalled it intends to remove it as it claims it was 'legacy language'. It appears to us to be there in order to limit liability.

especially if decisions have legal implications. They are instructed to desist from using it inappropriately or for unauthorised purposes, advised what level of operational information may be uploaded for Copilot tasks, and what may not be shared.

In discussion with MPS we clarified that the intended goals for Copilot were, in the first instance, to:

- Increase productivity (e.g. saving time on administration and analysis)
- Facilitate information sharing (e.g. document summaries)
- Improve communication (e.g. aiding in drafting, editing, and clarity)
- Support education and training (e.g. creating quizzes and presentations)
- Enhance employee engagement and wellbeing (e.g. access to modern communication tools and ways of working)

Future potential was seen in:

- Improving customer service (e.g. aiding in drafting, editing, and clarity of communication with victims or complainants)
- Decision support (e.g. LLM analysis, policy summaries)
- Helping to minimise environmental impact.

MPS also shared their own analysis of the M365 copilot (licensed version) pilot with us.

We recognise that use of Copilot could potentially bring real benefits to the organisation and to the individuals using it. Moreover, there is a risk that if MPS does not make such tools available to its people, they could be tempted to adopt their own 'shadow AI' (e.g. consumer LLMs) to assist them in their work. 'Shadow AI' is a recognised organisational and data security threat.[Φ]

We asked therefore what could emerge as points of failure: either barriers to realising the expected benefits, or Copilot introducing new and unexpected difficulties. We considered these possible failure points in light of the applied ethical domains outlined above. We have summarised our analysis in the two tables below. The first reflects the primary aims for introducing copilot, and the second reflects some probable future aims. We offer some illustrative, but not exhaustive, mitigation strategies.

[Φ] 2025 research by Microsoft and Censuswide revealed that 71% of UK employees had used unapproved consumer AI tools at work https://ukstories.microsoft.com/features/rise-in-shadow-ai-tools-raising-security-concerns-for-uk/

Table 2: Outcomes of pre-mortem

	POTENTIAL FAILURE POINTS	SHORT-TERM RISKS	LONG-TERM CONSEQUENCES	EXAMPLE MITIGATIONS	APPLIED ETHICS DOMAINS
01 **Increased Productivity & Speedier Analyses**	Factual errors. Reliance on unverified information. Duplication of effort to check & correct context, accuracy, nuance. Uncritical acceptance of outputs. Inappropriately scaled outputs (e.g. too short or long). Individual time savings do not aggregate to increase productivity.	Reliance on inaccurate information. Altered but not reduced workload. Workflow bottlenecks as AI speeds some processes not others. Unequal distribution of benefits & detriments.	Decreases in productivity. Failure to realise benefits. Misinformation cascade (uncorrected inaccuracies embedded)	Effective user training. Monitoring for verification activity. Clarity on which tasks are best suited to AI. Workflow redesign. Audit/evaluation against goals.	Social responsibility & public value Human oversight & control Technical integrity
02 **Access to and Sharing of Information**	Use of Copilot for unauthorised purposes. Intentional abuse of system. Human error in sharing information. Non-compliance with data protection.	Illicit accessing and sharing of data. Non-compliance with legal requirements.	Civil and criminal actions against police. Damage to public trust.	Rigorous internal governance. Monitoring and auditing use. Effective guidance. Rights-based guardrails.	Technical integrity Legal & human rights Social responsibility & public value
03 **Greater Clarity of Documentation & Improved Communication Style**	Misleading content from LLM confabulation, omission, inaccurate summaries. LLM lack of subtlety and nuance. LLM assigns wrong weight/importance to document elements. Careless moderation. AI-generated content passed off as human-authored.	Unclear or erroneous documentation. Additional workload. Misleading inferences.	Damage to,, organisation. Decrease in satisfaction & confidence. Increase in stress. Lack of confidence in system.	Duty of candour. Appropriate contestability mechanisms for those affected by inaccurate documentation.	Transparency & accountability Human oversight & control Technical integrity
04 **Aid for Internal Education and Training**	Inaccurate learning content. Cognitive surrender (trainees over-rely on AI summaries, instructions etc). Algorithmic aversion (trainees distrust AI support).	Generates inappropriate practice. Undermines self-directed learning.	Damage to public trust. Decrease in workforce satisfaction & confidence.	Robust testing of output. Lifecycle management. Train users in prompt engineering to challenge outputs.	Technical integrity Social responsibility & public value Human oversight & control
05 **Employee Wellbeing**	AI not taken into consideration as a possible reasonable adjustment (e.g. for dyslexia).	Employee grievances.	Litigation risk.	HR review.	Human dignity Legal & HR compliance

Current intended uses of Copilot (above)

Anticipated uses of Copilot (below)

01 **Improve customer service - victims, complainants and alleged perpetrators**	**POTENTIAL FAILURE POINTS** Over-reliance on automated systems/responses undermines relational aspects of policing. Responses that are inauthentic and potentially deceitful. Uncorrected errors.	**SHORT-TERM RISKS** Worse experience for victims and others contacting police. Impairment of individual and organisational learning. Compromising judicial outcomes. Secondary victimisation. Miscarriages of justice.	**LONG-TERM CONSEQUENCES** Damage to public trust. Undermining the criminal justice process.	**EXAMPLE MITIGATIONS** Appropriate design with clear boundaries for use. Appropriate use of the tool. Professional attention and responsibility. Transparency - declaration of use.	**APPLIED ETHICS DOMAINS** Human dignity & equality Social responsibility & public value Transparency & accountability Technical integrity
02 **Decision support**	**POTENTIAL FAILURE POINTS** Over-reliance and uncritical acceptance of output/false confidence. Could mislead without expert knowledge. Confirmation bias. Refusal to accept sound advice as don't trust it.	**SHORT-TERM RISKS** Impedes professional curiosity and pride. Damages professional decision-making. Additional workload. Poor decisions. Lack of validation leading to erroneous conclusions.	**LONG-TERM CONSEQUENCES** Decrease in satisfaction and confidence among staff and officers. Decrease in public trust. Degradation of expertise over time. Reputational damage. Damage to stakeholders and professional partners.	**EXAMPLE MITIGATIONS** Clarity on where responsibility lies. Guidelines on use or non-use. Human oversight. Use in areas with prior expertise. Override capabilities. Training. Verification mechanisms. Explainable AI. Appeal mechanisms.	**APPLIED ETHICS DOMAINS** Transparency & accountability Human oversight & control Technical integrity

The tables above suggest how the possible failure points reflect the ethical domains we have set out in our first ethical framework infographic. We view these potential failure points not just as practical difficulties, but difficulties creating ethical challenges. The ethical impact will become particularly noticeable as practices are embedded and effects accumulate. These have the capacity to affect the MPS's aims, its employees, victims of crime, persons of interest, and ultimately the trustworthiness of the organisation. These effects are not inevitable, but they are risks to be mitigated.

We turn next to considering whether there are learning points from the adoption of Copilot that ought to be considered with wider adoption of AIs that could be used to support internal and external communications of various kinds, and to assist decision-making.

The bigger picture: anticipating benefits and risks from AI

The headline claim for Copilot and other AI aimed at police systems is compelling: it should reduce the time spent analysing copious and disparate sources of information, enhance accuracy and effectiveness, assist with drafting communications and, overall, increase capacity for effective crime reduction work. In this section of our report we have used our usual policing ethics framework to organise our discussion: consideration of the organisation and its people; consideration of the interests of state and civil society; consideration of the needs of victims; and consideration of the needs of people of interest.

The organisation and its people

Reducing the burden of administration

Currently, the evidence does not support more than modest returns on AI use for improving policing administration.[7] Research on AI in police report writing points to two structural reasons for this: officers already use templates for most routine communications, limiting how much AI can add; and many communications are sufficiently specific or nuanced that AI requires substantial intervention by the officer to produce something fit for purpose.[8] The productivity argument may therefore be weaker than it appears, depending on the specific AI adopted and the gains sought.

There is a less visible cost to weigh against efficiency gains in communication. Reliance on Copilot risks eroding the written communication skills that officers would otherwise develop, skills that matter, in practice, to the quality of victim contact, investigation records, and court-ready documentation.[9] Beyond technical deskilling, research identifies a further concern: moral deskilling in decision-making, where habitual reliance on AI-assisted outputs may gradually diminish officers' capacity for independent ethical judgement. These are not hypothetical risks. They are documented consequences of AI adoption in analogous professional settings,[10] and they warrant serious attention in workforce development planning.

How reliable is the human in the loop?

Perhaps one of the greatest risks, operationally and ethically, is what researchers have termed "cognitive surrender": the tendency of users to accept AI-generated outputs with minimal critical appraisal, deferring to AI outputs over their own knowledge, experience, and judgement. The evidence on this is unambiguous: AI assistance improves decision-making when the AI is right, but AI produces outcomes *worse* than unassisted human judgement when the AI is wrong.[11] That asymmetry is the core problem. MPS users, even those with expertise (and less so novices in a specific area of policing) will not always know when AI is misleading them. The question begging an answer is whether arrangements for oversight or supervision will be sufficient to catch misjudgements before they have consequences.

Police policy guidelines place the 'human in the loop' at the centre of the governance model for AI-assisted decisions. The human in the loop is a widely touted principle, but there is real concern whether it can work as intended. There are two structural reasons to doubt it. First the tendency towards "cognitive surrender" means that in reality many officers will be tempted to rely on AI outputs with limited critical questioning, regardless of what policy requires. Second, the speed, scale, and opacity of AI-generated material frequently exceed what a human reviewer can meaningfully assess. The result, documented in recent research on AI accountability, is that the 'human in the loop' becomes a liability mechanism rather than a genuine check[12]: officers are answerable for misjudgements they were not realistically in a position to prevent.[13]

Who is accountable?

Accountability for AI outputs is a governance challenge calling for equitable resolution. Identifying errors in AI-generated outputs requires experience and expertise but a significant proportion of MPS officers are young and relatively inexperienced. Placing responsibility on them to override or correct AI without comprehensive training, supervision, and institutional support, is neither operationally sound nor fair to the officers concerned.

For junior officers in particular, requiring them to take personal responsibility for the quality and ethical soundness of AI-generated material they lack the skills to properly assess places them in an invidious position. The current 'human-in-the-loop' provision, as it stands, risks allocating accountability to individuals while leaving the organisational questions – about training, supervision, vendor obligations, and procurement standards – unaddressed.

Research indicates that AI and automation in policing will have an uneven impact across ranks and roles, with the greatest disruption likely to fall on officers in back-office functions, frontline positions, and non-leadership grades.[14] Large-scale redeployment may not be the immediate risk, but the picture is uneven in a telling way: AI may reduce workload for some officers while materially increasing the risk

exposure of those who are less experienced. Is a deployment model that benefits some officers, while placing others in a more vulnerable position, tenable?

But if individual officers are not to be held accountable for AI, who is? We discuss the so-called 'liability sink' in our final section on ethical challenges and their resolution.

Building capacity for AI implementation

Ethical deployment of AI requires officers to be properly trained in using the tool. For LLMs this will include prompt engineering: structuring queries in ways that produce outputs that are accurate, relevant, and appropriate to context. A prompt template can be developed to help standardise outputs against predefined quality thresholds.[15] Whether training provision is achieving the standard required for ethical implementation requires a candid answer, and training needs to be evaluated drawing on the experience of those using the tool, those supervising its use, and the people on the receiving end of communications it has helped to draft.

Beyond initial training, ethical deployment requires ongoing monitoring. Software updates by vendors can introduce disruption, compatibility failures, or data integrity issues that officers are ill-equipped to manage without support. Post-deployment monitoring needs to be structured around four dimensions: *Functionality* (does the system continue to work as intended?); *Operationality* (does it maintain consistent service across its infrastructure?); *Human Factors* (is it transparent and producing high-quality outputs?); and *Security* (is it secure against attack and misuse?).[16] Absent clear organisational responsibility for post-deployment monitoring, and a realistic provision for the in-house capacity required, there is a risk not only that problems will not be mitigated but that they may not even be recognised.

Research identifies three workforce challenges that any AI-assisted decision-making framework must address: equipping officers to understand what AI can and cannot do, and how it complements rather than replaces their own skills; calibrating officers' expectations and reliance on AI so they remain appropriately sceptical; and enabling officers to make sound judgements about when to use AI assistance and what weight to give its outputs.[17] These are not training objectives that can be met by a short induction module. They require sustained investment in officer development and a supervisory culture that actively reinforces critical engagement with AI tools rather than passive acceptance of them.

Others using AI to communicate with the Met

Whatever MPS chooses to do with AI, its people and the public are probably already enthusiastic users of LLMs. MPS will need to ensure that the organisation has the capacity to handle communications arising from AI-assisted outputs. They could for example be complaints from members of the public about service they have received, or from officers and staff about grievances or other complaints. Public sector bodies are already observing that the infrastructure built for the pre-LLM world

will not be sufficient to deal with submissions and challenges the preparation of which was supported by LLMs.

Serving state and civil society: AI and legitimate authority

There is a principled argument for deploying AI that goes beyond operational convenience. As the former UK Biometrics & Surveillance Camera Commissioner Fraser Sampson has observed, the existence of a technology itself generates pressure to use it.[18] The spectre of future accountability (the assertion that 'they could have done more but didn't') will be familiar to those who have navigated ethical scrutiny of operational decisions.

Where a technology can credibly improve police performance, and the benefits outweigh the risks, there is arguably a positive duty to explore it. The question, then, is not whether to use AI, but whether MPS can demonstrate it has weighed the risks seriously and put in place the governance to manage them.

One long term risk that warrants attention is how AI could impact on MPS's epistemic authority: that is, how far the public view MPS communications as trustworthy, credible, and authentic. When AI tools are used to draft police communications, questions of accountability and trustworthiness begin to arise. For instance, in a letter responding to a complaint, is the communication a genuine apology with remorse, or AI flim-flam? Is a public announcement a genuine insight into MPS's knowledge and intentions, or a statement crafted by AI only to manage hostile sentiment? Is a statement of fact authoritative or an AI confabulation? Is an empathetic tone authentic or just a chimera?

While the risk of AI confabulation is increasingly well understood, the deeper concern goes beyond factual accuracy. Seeking the truth, in the context of police communications, is not just a matter of getting facts right. It is also about trust and reliability. Truth in these contexts emerges from a dialogue between two parties, each of whom is seeking solid ground in what the other states. When dialogue between police and the public works, it is because each is assumed to want to connect with the other. AI has no such interest. People sense this, and research indicates that they tend to *discount* material they believe to be AI-generated, even when it is factually accurate.[19] For a service whose authority rests on being believed, that is not a trivial problem.

A second concern is about the competition police communications face. Research on 'algorithmic aversion' shows that when AI-generated and human-generated material are in competition, people tend to prefer the human source. MPS currently operates in a heavily contested information environment, including narratives around 'two-tier' policing. Where AI-authored police communications sit alongside claims made by human actors, the apparently authentic voice of those human actors may well prevail. The 'police voice' carries authority because it is a human voice held accountable to standards. AI cannot replicate that, and incautious use of it risks undermining the standing of everything communicated in the organisation's name.

The overall concern is the one with the longest shadow: public trust in MPS. If the public come to doubt whether what they hear from MPS is credible, and to question whether there is a genuine human actor behind it, trust will erode – and with it the legitimacy on which policing by consent depends. MPS has no need to be reminded how hard it is to rebuild institutional trust once lost, or how quickly a single high-profile failure can accelerate that process.

The governance question, therefore, is not only whether AI works as intended, but whether its use, if it became publicly known, would be defensible to the people MPS serves. This question of how far AI use is declared or known is an ethical dilemma we return to in our final section.

Serving victims: minimising vulnerability

There are potentially considerable operational and ethical benefits for victims to be gained from careful use of AI. For example, timely, well-drafted follow-up contact with victims is something MPS consistently struggles to deliver, and victims consistently say they need. AI could help officers meet that expectation more reliably. The question is whether the benefits in this area can be secured without introducing risks that outweigh them.

Those risks are not insignificant. The epistemic concerns set out above apply with particular force to communications with victims, where the capacity to trust police forms a critical part of recovery from the harm they have suffered. Research now demonstrates directly that victims judge AI-mediated police communications as procedurally less just, and therefore less trustworthy, than communications from human officers, even when the words used are identical.[20] This is a finding that should give pause. It means that even well-intentioned, well-drafted AI-assisted communications may actively set back the relationship between victims and the service.

The research tells us that victims want more from police than efficiency. Many are seeking, through the process of reporting and investigation, to restore their sense of agency and their connection to society after a traumatic experience.[21] Procedural justice (being treated with respect, dignity and genuine recognition) is how police most reliably provide that. The difficulty is that procedural justice, properly understood, implies intentionality: it means showing that a person in the victim's circumstances is *worthy* of care and respect. Communicating this can look like flim-flam coming from AI, as it is not words but human connection that victims are seeking.

A more immediate risk is that victims may be directly exposed to harm from AI errors. The first mistrial in which police are required to withdraw evidence because of an AI-generated mistake will be a major scandal. It will be damaging for MPS, and it will be seriously harmful to the victim involved. Robust human oversight of AI-generated material in anything that could affect the course of an investigation is therefore not optional; it is a basic requirement of ethical deployment.

None of this is intended as a wholesale rejection of AI in the context of victim communications. There is a genuine case that it could *improve* the quality of communications from officers who find empathy and clarity in written form difficult to convey. The task for senior leaders is to design the governance that captures that benefit while containing the risks – which means, at minimum, being clear about what AI should and should not be used for, ensuring officers remain the accountable authors of communications, and being transparent with victims about how their case is being managed.

Serving persons of interest: maintain respect for humanity, act proportionately and fairly.

Everything said above about victims applies equally to persons of interest. Greater efficiency in police processes is in the interests of suspects too: for instance, fewer people spending prolonged periods in the condition of uncertainty that criminal investigation implies. They have, equally, a right to be treated with dignity and respect, and to trust that the processes applied to them are fair and proportionate.

The risks here, however, carry a sharper legal and reputational edge. AI errors in this context are not only operationally problematic; they risk directing investigatory resource down wrong paths and, in the worst cases, contributing to wrongful outcomes for individuals. There is increasing evidence from high-stakes, heavily regulated professions such as medicine that practitioners over-rely on AI outputs, fail to cross-check or verify them, are deterred from challenging them, and become hesitant to exercise independent judgement.

The governance question – whether oversight arrangements are genuinely capable of catching errors before they cause harm, rather than after – is one that carries direct legal and reputational exposure for the organisation, not just operational risk.

Ethical tensions and recommendations

In this final section, we deal with three ethical tensions that in turn lead to three sets of recommendations:

- Trustworthiness and innovation
- Transparency and trust
- Accountability and the ‘liability sink’

Trustworthiness and innovation

An important lesson to take from the introduction of Live Facial Recognition is that the trustworthiness of AI is best demonstrated from the outset. In the case of LFR, trustworthiness has meant developing appropriate governance processes, establishing technological validity, building user capability, and proving its benefits to the public. These requirements apply with vital necessity to the broader introduction of policing AI.

Two useful studies on public attitudes to AI in policing were published in 2025 (CENTRIC[Ψ] and Thinks Insight).[22] Both suggest that while there is a reasonable degree of support for the use of AI in policing, this support is conditional on it being used appropriately. Equality, fairness and accountability were found to be central to public trust.

CENTRIC's national citizen consultation included input from 10,000 UK based participants. CENTRIC found that a little over 60% of the public supported police use of AI in general and were most comfortable with it where it could be established it was clearly beneficial, low risk and aimed at preventing serious harm. The granular findings are of interest. CENTRIC summarises the evidence on deployment contexts and public expectations as follows:

- *"Highest acceptance of AI use by police relates to the identification of (potential) perpetrators of child sexual exploitation and terrorism; acceptance is lowest for the automation of 999 emergency calls.*
- *Trust levels in AI are highest both for situations in which AI deployment carries low risks for the public and in which AI use can reduce serious harms to vulnerable groups (children).*
- *Strongest independent oversight is expected for situations in which AI makes its own decisions in high-risk situations.*
- *Monitoring by an independent oversight body is expected whenever police deploy AI, not just in contexts of low public acceptance or low trust towards AI use by police."*

And in relation to accountability and governance:

- *"53.3% trust the police to use AI accountably, while around 25% do not.*
- *Explainability (of the AI being used) and justification (for its use) alone are not sufficient to build trust. The most effective enablers of trust are close monitoring and strong accountability mechanisms.*
- *75.5% of respondents support a mandatory accountability process prior to deploying AI in policing, but currently fewer than 20% believe such measures are in place."*

The CENTRIC study found support was contingent on the presence of strong safeguards, with significant concern about algorithmic bias, lack of transparency, misuse of personal data and the risk that errors or discriminatory impacts could disproportionately affect certain communities.

[Ψ] Centre of Excellence in Terrorism, Resilience, Intelligence and Organised Crime Research at Sheffield Hallam University.

The Thinks Insight study was conducted on behalf of the Home Office, and their findings were similar to those in the CENTRIC study. Additionally, whilst AI's potential to improve efficiency and consistency were valued, participants expressed concern about loss of human judgement, reduced empathy and insufficient oversight.

Finally, the Ada Lovelace / Turing Institute 2025 public attitude survey found "the majority of the public (72%) indicated that laws and regulations would increase their comfort with AI technologies …65% of people said that procedures for appealing decisions made by AI would make them feel more comfortable with AI, along with 61% who felt getting information on how AI systems made decisions about them would increase their comfort levels".[23]

The implication of all these surveys is that demonstrably robust public sector governance and public information is required if trust is to be secured. The balance that must be sought is rigorous controls that enforce strict parameters for AI adoption and use, while not stifling beneficial innovation.

Innovation is critical in a context where MPS must husband limited resources in the face of ever higher demand, and when sluggish organisational innovation is a known risk throughout the public sector because it tempts employees to experiment with 'shadow AI'. (Shadow AI is the use of unauthorised AI, such as public LLMs, for work purposes.) It is argued that employees are drawn to shadow AI for understandable reasons: these include operational friction (because sanctioned AI is absent or inferior); 'silent mentoring' (employees concealing their use of AI to avoid devaluing their professional expertise); and low AI literacy (people are unaware of the threats present in using unauthorised AI).

Strategic AI governance is the infrastructure that allows innovation to travel safely, at speed, with public trust. We support, however, the emphasis that NPCC places on using the principles from the Artificial Intelligence Playbook for the UK Government[24] to the effect that a strategic and controlled approach is vital for smoothly integrating AI into policing. While not every element needs to be in place before starting, a foundation encompassing a clear strategy, guiding principles, robust governance, transparent communication, and a plan for sourcing, training, and change management is crucial for securing buy-in and ensuring AI is used safely, responsibly, and effectively.

Currently, MPS strategic AI governance includes three domains of organisational control that AI must pass: the Organisational Design Authority, that assesses how far AI satisfies a business need; a Technical Design Authority, that assesses AI against the requirements for digital infrastructure security; and the Data Capability Assurance Framework that ensures compliance with data protection laws. In addition, a Frontline Policing Technology Board provides a forum for gathering and assessing ideas, and for disseminating information about capabilities already available. There is an early version of the MPS AI Framework available on the MPS website,[25] and a strategic AI Policy document is in the final stages of approval following extensive

internal consultation. Local operating procedures for specific AI will be developed by the Senior Responsible Officers accountable for its introduction (as for example was the case with Live Facial Recognition).

Whilst there has been internal MPS discussion about specifically ethical appraisal of novel AI, this is not currently embedded as a stand-alone requirement in the overall approvals framework. It has however been proposed that elements of the NPCC's Responsible AI Checklist should be integrated into the Data Capability Assurance Framework (DCAF) where they are not already covered.

The widely cited Edelman Trust Barometer is an annual global survey conducted by communications firm Edelman[26], and is worth considering alongside the data from the Ada Lovelace & Turing Institutes and the CENTRIC and Thinks Insights public consultations in the UK. Edelman suggests there has been a decisive shift towards trusting community information sources over and above government, NGOs and formal media. Taken together, surveys imply that public engagement will be a critical factor in public trust in AI, that organisations must be able to demonstrate that they are trustworthy stewards of AI and personal data, and that views about trust in AI which travel across networks and neighbourhoods will be influential. There is not at present any clear mechanism in the MPS or MOPAC that serves to engage the public in assessing, discussing, understanding or indeed advocating for AI innovation in MPS. This is hardly consistent with the principle of policing by consent.

We therefore propose:

Recommendation One

Ensure that MPS strategic AI governance incorporates all elements of the NPCC's Responsible AI Checklist[27], in whatever way fits most effectively and efficiently into MPS approval processes.

Recommendation Two

MPS strategic AI governance should include means for ensuring that when Local Operating Procedures are developed, they are consistent with the MPS's AI governance goals overall, and they are informed by best available evidence on human-AI interaction, for example, the evidence set out in Table 1 above – Human-AI interactions. Additionally, as the NPCC Playbook indicates[28], procedures should show the highest standards of respect and dignity towards individuals and groups as set out in the Code of Ethics.

Recommendation Three

Where AI is included in systems for making administrative decisions about people (either as employees or members of the public) this is done with full transparency about its use in the decision; this should include information about

how decision subjects can access proportionate processes to challenge decisive AI assessments.

Recommendation Four

A comprehensive strategy for public engagement, aiming to build understanding of the steps being taken for responsible AI adoption, be given appropriate priority by MPS and MOPAC. As NPCC points out[29], early and continuous public engagement helps mitigate potential risks. The Panel considers that a thorough public engagement strategy would include, as and when appropriate and according to circumstances, each of the following:

- Externality: advice and insight from external experts (e.g. to advise on human-AI interaction, AI risk, AI potentialities). We note that the NPCC recommends introducing an AI ethical review group which incorporates external experts[30].
- Citizen engagement: prior to a decision to adopt specific AI, to provide insight into potential impacts particularly on less trusting, underserved or minoritised groups (e.g. could be in the form of Citizen Juries).
- Socialisation: to build general public awareness and acceptance of AI use in policing.
- Partnerships: MPS works with a wide range of partners many of which will also be adopting AI and may be further ahead of MPS in terms of gaining public trust and acceptance.

Transparency and trust

We noted under the applied ethics principle of transparency, that the NPCC covenant for AI in policing stipulates that all use of AI will be subject to 'Maximum Transparency by Default'. How should this apply in the context of MPS use of Copilot and similar LLMs or other AI?

In principle, any use of AI should be acknowledged by whomever uses it, with sufficient specificity to make clear what contribution AI made. We have done so in the cover to this document.

We start with this principle because it expresses a clear ethical position that candour (and its cousins, openness and transparency) are moral goods. Candour and transparency buttress honesty and trust; without candour and transparency an organisation or person may in effect mislead by omission. Organisational openness is a precondition for effective accountability. And candour is a virtue that when habitually practised by officers and staff sustains a positive ethical culture in policing.

There are sound operational reasons for candour and openness too: without an organisational culture of openness regarding AI use, it is impossible to maintain effective organisational oversight of threats to data integrity.

Unfortunately, recent research suggests that transparency about AI may have some unwelcome downsides. It would appear that openness about AI use can at present compromise, rather than reinforce, trust.

In a robust series of experiments, researchers found that "AI disclosure paradoxically erodes trust…AI disclosure will diminish trust in an actor across a variety of tasks". However, they also found that that "while AI disclosure jeopardizes trust, the exposure of undisclosed AI usage by a third party has an even more detrimental effect".[31]

In other words, *declaring* AI usage will impact on trust; but being found out for *not* declaring will undermine trust even further.

Trust in policing is a very significant interest for MPS, and, in time, undisclosed use of AI is highly likely to be discovered and made much of by critics. We therefore conclude that MPS should adopt and enforce the NPCC's maximum transparency principle for all its use of Copilot and other AI, accepting that there may be a diminution of trust in some areas that would nevertheless be far greater if AI use were not disclosed.

We therefore propose:

> **Recommendation Five**
>
> In line with NPCC, MPS should enforce maximum transparency by all users both when AI is used in decision making, and when AI is used to support other policing activities. As the NPCC Playbook[32] notes, an important aspect of transparency is explainability, that is, how clearly an AI system's decision-making can be understood by those interpreting or overseeing it. NPCC recommend the Information Commissioner's Office detailed guidance on the arrangements organisations need to have in place to explain how decisions may be made with AI[33].

Accountability and the 'liability sink'

The concept of the 'liability sink' refers to how organisations assign accountability for AI-generated errors to human actors, who may not, in practice, be in a position to exercise the oversight that accountability implies.

We have noted that policing's governance framework centres on the 'human in the loop', requiring a human to review, approve, or take responsibility for AI outputs before they are acted upon. In principle, this preserves professional accountability and provides a check on AI error. In practice, however, the human may not be in a position to challenge the AI error that they sign off. When something goes wrong, accountability flows to that individual (the sink into which liability drains) while the organisational and procurement decisions that created the conditions for the error remain largely unexamined. This is not fair on the individual, if the error was not

capable of being corrected by someone with their level of knowledge and expertise. Neither is it good for the organisation, which loses the opportunity to learn what created the conditions for the error in the first place.

As noted earlier, cognitive surrender implies that officers are likely to defer to AI outputs regardless of what policy formally requires of them. The human-in-the-loop function is compromised not just by lack of knowledge or expertise, but also by the psychological dynamics of working alongside an apparently authoritative system. Additionally, officers placed in the human-in-the-loop role may be the least experienced, and therefore the least equipped to identify when an AI output is wrong, contextually inappropriate, or ethically problematic.

In clinical settings, one proposed solution to the ‘liability sink’ is to ensure that patients are engaged more fully in the clinical decision that might be supported by AI. This solution creates its own problem, in that it just assigns some of the accountability burden to patients, who are the least enfranchised partners in the clinical system.

But the proposed clinical solution points neatly to the very great difficulty in dealing fairly with the ‘liability sink’ in policing. Those worst affected by AI errors are likely to be in the weakest possible position (as victims or suspects for example) to challenge the actions or decisions of the police.

Those who are vulnerable to police action must be able to call police to account. At the same time accountability requires fairness to the MPS officers, staff and volunteers using AI tools procured and provided by the organisation. Moreover, there will be limited gains from investing in AI if the Met’s people decline to use it because they do not trust it. The problem of trust will present in a slightly different way if AI use is mandated within MPS, and people have no choice but to use it however sceptical they are about its outputs. AI governance must therefore find a balance between accountability, fairness, and trust in AI capability.

We think there may be a partial analogy here with firearms policing: officers retain personal responsibility for their decisions, but there are systems in place to train them effectively and support them to make the right decision in the circumstances in which they are placed. The risk is, in effect, shared. Another partial analogy may be found in the Post Office Horizon failures: holding technology providers to account for their failures is challenging when they are not incentivised to be open about its shortcomings.

It is not for the Panel to find the solution to this conundrum. We think it may lie in a combination of mandatory disclosure of AI use, monitoring (to ensure verification is being carried out when AI is used), inbuilt prompts to users to verify content, training in prompt engineering to challenge AI outputs, training in AI limitations, a ‘just culture’ which differentiates innocent AI-induced errors from careless or malicious use of AI, and contracting for providers to make mandatory disclosure of ongoing AI reliability

metrics. There are no doubt steps already taken of which we are unaware, but we believe that an appropriate solution will be required.

Recommendation Six

Resolve and make known to stakeholders (e.g. MOPAC, Police Federation, CPS, local authorities) the ways in which MPS addresses AI error, including:

- How it will limit the occurrence of AI error through technological safeguards in each case;
- How it is satisfied that employees are adequately trained to avoid error when using AI to do their jobs; including ensuring employees are aware of the benefits of appropriate use and the need for verification of AI outputs;
- How it holds providers to account for AI functionality standards; and
- The principles by which MPS employees will be held to account for AI-induced error.

Recommendation Seven

We drew attention above (Transparency and accountability domain: evaluation principle) to the need for life cycle monitoring of AI.

- Life cycle monitoring should include any unforeseen consequences of implementing AI, including the costs of rectifying erroneous outputs as well as the evolution of the technology itself.

Appendix One

1. **College of Policing (COP)**

https://www.college.police.uk/article/new-technology-guidance-and-apps-published

2. **The Alan Turing Institute (Turing)**

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

3. **National Police Chiefs Council (NPCC)**

https://science.police.uk/site/assets/files/4682/ai_principles_1_1_1.pdf

4. **Bioethics and Forensic Ethics Group (BFEG)**

https://www.gov.uk/government/publications/ethical-principles-biometrics-and-forensics-ethics-group/biometrics-and-forensics-ethics-group-principles-accessible-version

5. **ALGOCARE protocol**

Developed by Marion Oswald and colleagues this stands for Advisory, Lawful, Granularity, Ownership, Challengeable, Accuracy, Responsible and Explainable

Oswald , M. et al (2018) Algorithmic Risk Assessment Policing Modules; Lessons from the Durham HART model and Experimental Proportionality. *Information & Communications Technology Law* 27(2): 223-250.

6. **CRIMEFACE**

Developed by Bowers and Johnson and standing for Crime Reactions Implementation Misuse Economics Future Adaptability Context Ethical

Bowers, K., & Johnson, S. (2024). Facing the Future of Crime: A Framework for Police Use of Technology. *The Political Quarterly*, *95*(3), 480-488.

7. **Australia/New Zealand Police Responsible and ethical artificial intelligence framework (ANZPAA)**

https://www.anzpaa.org.au/products/products/australia-new-zealand-responsible-and-ethical-artificial-intelligence-framework

[1] Ethical Decision-Making Guidance for Police Use of Data Analytics CDEI 2021 citing UK House of Lords 2018. https://library.college.police.uk/docs/CDEI-ethical-decision-making-guidance-2021.pdf?
[2] Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
Hagendorff, T. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds & Machines 30, 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
Fjeld, Jessica and Achten, Nele and Hilligoss, Hannah and Nagy, Adam and Srikumar, Madhulika, Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (January 15, 2020). Berkman Klein Center Research Publication No. 2020-1, Available at SSRN: https://ssrn.com/abstract=3518482 or http://dx.doi.org/10.2139/ssrn.3518482
Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., ... & de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns, 4(10). https://www.cell.com/patterns/fulltext/S2666-3899(23)00241-6
[3] Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529
[4] https://www.bi.team/publications/ai-and-human-behaviour/
[5] Klein, G. (2007). "Performing a Project Premortem." *Harvard Business Review*, 85(9): 18–19
[6] https://www.microsoft.com/en-us/microsoft-copilot/copilot-101/what-is-copilot?msockid=1185abc03f0c68772443be413e0a69b3
[7] Alda, E., & Tobar, F. S. (2025). Boosting Police Productivity with AI? An Assessment of Productivity and Efficiency Gains in Report Writing. *CrimRxiv*.
[8] Adams, I. T., Barter, M., McLean, K., Boehme, H. M., & Geary, I. A., Jr. (2024). No man's hand: Artificial intelligence does not improve police report writing speed. Journal of Experimental Criminology, 1–18.
[9] Natali, C., Marconi, L., Dias Duran, L. D., & Cabitza, F. (2025). AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. *Artificial Intelligence Review*, *58*(11), 356.
[10] Heimer, C. A. (2023). Where Law and Morality Meet: Moral Agency and Moral Deskilling in Organizations. In *Handbook of the Sociology of Morality, Volume 2* (pp. 43-57). Cham: Springer International Publishing.
[11] Shaw, S.D. & Nave, G. (2026). "Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender." Wharton School Research Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6097646
[12] Lawton, Tom, Morgan, Phillip, Porter, Zoe et al. (2024) Clinicians risk becoming 'liability sinks' for artificial intelligence. Future Healthcare Journal. 100007. ISSN: 2514-6653 https://doi.org/10.1016/j.fhj.2024.10000
[13] Jovanovic, V. (2026). When Human-in-the-Loop Fails: An Answerability Test for Deployed AI Systems.
[14] Halford, E. (2026). Policing in the age of artificial intelligence: a simulation study modelling workforce automation scenarios. *Police Practice and Research*, *27*(1), 164-187.
[15] Sarzaeim, P., Mahmoud, Q. H., & Azim, A. (2024). A framework for LLM-assisted smart policing system. *IEEE Access*, *12*, 74915-74929.
[16] NIST (2026) Challenges to the Monitoring of Deployed AI Systems, accessed at https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-4.pdf
[17] Steyvers, M., & Kumar, A. (2024). Three challenges for AI-assisted decision-making. Perspectives on Psychological Science, 19(5), 722-734.
[18] https://www.biometricupdate.com/202604/just-because-they-can-the-biometric-conundrum-for-law-enforcement
[19] Altay, S., & Gilardi, F. (2024). People are sceptical of headlines labelled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS nexus*, *3*(10), pgae403; Zhu, T., Weissburg, I., Zhang, K., & Wang, W. Y. (2025, July). Human bias in the face of AI: Examining human judgment against text labelled as AI generated. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 25907-25914).
[20] Bradford, B., Kyprianides, A., Andrews, W., Aston, E., Clayton, E., O'Neill, M., & Wells, H. (2025). 'To whom am I speaking?'; Public responses to crime reporting via live chat with human versus AI police operators. *Policing and Society*, *35*(8), 1036-1052.

[21] Pemberton, A., Aarten, P. G. M., & Mulder, E. (2017). Beyond retribution, restoration and procedural justice: The Big Two of communion and agency in victims' perspectives on justice. *Psychology, Crime & Law*, *23*(7), 682-698.
[22] https://stwebsitesprodeu.blob.core.windows.net/centric-website/CENTRIC_AI+Policing_National-Citizen-Consultation(public-version).Oct2025-WEB.pdf https://cdn.prod.website-files.com/6578727b2437b2ef1c478c1b/6880e03f7e7d4a6e2bcb7ed4_Home%20Office%20CoLab_Ai%20in%20Policing_Cohort%202_report_v13_150725_updated%20logo.pdf
[23] https://attitudestoai.uk/assets/documents/How-do-people-feel-about-AI-2025-Ada-Lovelace-Institute.pdf
[24] https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government
[25] https://www.met.police.uk/police-forces/metropolitan-police/areas/stats-and-data/stats-and-data/met/-framework-for-the-use-of-artificial-intelligence-in-the-met/ and also https://www.met.police.uk/SysSiteAssets/media/downloads/force-content/met/stats-and-data/ai-framework/the-ai-framework-for-the-met.pdf
[26] https://www.edelman.com/sites/g/files/aatuss191/files/uk/2026-01/2026%20Edelman%20Trust%20Barometer_UK%20Report%20(1).pdf
[27] https://library.college.police.uk/docs/NPCC/Artificial-intelligence-playbook-policing-2025.pdf
[28] https://library.college.police.uk/docs/NPCC/Artificial-intelligence-playbook-policing-2025.pdf
[29] https://library.college.police.uk/docs/NPCC/Artificial-intelligence-playbook-policing-2025.pdf
[30] https://library.college.police.uk/docs/NPCC/Artificial-intelligence-playbook-policing-2025.pdf
[31] Schilke, O. and Reimann, M., (2026). How Does AI Disclosure Shape Trust? Unpacking the Role of Legitimacy. *Social Psychology Quarterly*, p.01902725261425997.
[32] https://library.college.police.uk/docs/NPCC/Artificial-intelligence-playbook-policing-2025.pdf
[33] https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/