

NPL Call for Evidence: AI testing, evaluation and assurance in the UK

Response from Dr Mackenzie Jorgensen, Dr Evdoxia Taka, Dr Michele Sevegnani, Prof Kyle Montague, Prof Muffy Calder, Prof Marion Oswald

This is a submission in response to the NPL call for evidence on AI testing, evaluation and assurance in the UK published in September 2026. The submission has been prepared by a team of academic researchers with extensive expertise in the technical, operational, legal, and social implications of AI use in law enforcement, national security and other contexts. Our work is anchored in the UKRI Responsible AI (RAi) UK Keystone Project, “PROBABLE Futures,” a four-year, £4.2M research initiative focused on evaluating probabilistic AI systems across the criminal justice sector.

Part 2: Challenges, unmet needs and opportunities

2A: Ecosystem-wide challenges and capability gaps

1. Drawing on your experience to date, where do you see the most significant unmet needs or capability gaps in AI testing and evaluation for assurance?

Please identify, where you can, the types of AI, sectors, use cases, risk domains or evaluation capabilities affected.

The PROBABLE Futures team recommends that NPL explores automatic speech recognition systems (ASR), large language models (LLM), agentic AI, and facial recognition technology (FRT) as AI technologies to evaluate. These are technologies used by the police in England and Wales and we need to develop better evaluation methods to ensure these adopted AI systems are fit for purpose. ASR systems can be used to transcribe conversations from body worn cameras worn by police officers. LLMs like Copilot are used to support police staff in everyday tasks like notetaking. Online chatbots like Bobbi can provide members of the public with non-emergency support. Live FRT flags individuals on a watchlist walking in front of FRT police vans, and officers may then approach them. What is deemed as fit for purpose or an acceptable capability level for an AI depends on the actual deployment setting, the operational context, the legal requirements and tests that must be satisfied (e.g. reasonable grounds) and the associated levels of risk—all nontrivial elements that need close attention. Thus, the testing regime **cannot** be separated from the intended operational context.

Through research undertaken by PROBABLE Futures team members, we have identified the need for methods that:

- test the effectiveness of human-AI workflows (e.g. measure the quality of human decision-making when assisted by AI in comparison to when not and test whether there is a comparative overall performance benefit) [1,6,7];
- evaluate effective and meaningful human oversight (e.g. measure and mitigate the risk of automation bias, measure the levels of understanding about AI uncertainty and risks of human operators) [1,3,6,7];
- identify and mitigate downstream impacts of ASR, LLMs and other generative AI applications (e.g., an LLM generated witness statement that could make mistakes leading to issues in the case progression or trial, and inclusion and fairness concerns for disabled people) [2,4,7];

- assess the cascading effects from sequencing AI systems (e.g. AI-assisted transcription, followed by AI-assisted witness statement drafting, followed by AI detection of criminal activity) [2,7];
- provide reasonable explanations why an error has happened (e.g. a missed detection by an FRT system happened because “of bad lighting conditions” instead of the “the match score was below the threshold”) [3,6];
- identify the implications and unintended consequences of AI systems; and measure the risk for organisations by the deskilling of end-users [1,7], as well as the risks to wider members of society subject to these AI systems.

Our research team highlights two additional evaluation related capabilities that need improvements in the policing and criminal justice domain. Evaluators (and ideally designers of the AI systems) should include affected groups and communities who may be impacted if the AI is adopted [7,8]. For instance, evaluations for an AI technology that will predominantly interact with vulnerable people, e.g. survivors of gender-based violence, will need to include those very people in testing, so that potential risks and harm can be mitigated before deployment. Additionally, evaluators need data on outcomes and in terms of ASR and LLMs, comparator data on evaluators of human completing the same task [7]. Test data needs to be labelled well and be collected from the same context that the AI might be deployed within to more precisely reflect the actual deployment conditions [3,7]. PoliceAI, a new body within the College of Policing, can potentially support with the data needed for rigorous testing.

2B: Your own organisation’s needs or interests - AI testing and evaluation for assurance

Intermediary

For research, standards, regulatory or other ecosystem organisations.

1. What expertise, evidence, or capability could you or your organisation potentially contribute to addressing the testing and evaluation challenges identified above?

For example, this could include research expertise, datasets, standards expertise or programme delivery.

The PROBABLE Futures research project brings research and interdisciplinary knowledge on the following: law, policing, responsible AI, computer science, statistics, social science, and social justice. Our team’s particular domains of focus are AI within law enforcement and AI in the criminal justice system in England and Wales. We study and evaluate the adoption, evaluation, and governance of AI within those domains.

For example, drawing on semi-structured interviews with professionals from policing, prosecution, the Bar, government, oversight bodies, and the technology sector, we investigated how probabilistic AI is being introduced, understood, and operationalised across law enforcement and the wider criminal justice system of England and Wales [1]. The central finding was 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.

We developed a first-of-its-kind [open register of AI tools](#) identified through our research as being used, trialled, or experimentally deployed within the Criminal Justice System across England and Wales. The register enhances transparency by providing stakeholders with a publicly accessible overview of current AI applications and their adoption across the justice sector. Our findings from the analysis of the data in the register and the interview transcripts suggest broad experimentation with limited maturity and a substantial proportion of use concentrated at the pre-charge stages of the criminal process, a strong reliance on private sector providers, and a growing interest in generative technologies, specifically LLMs, but evaluation is limited [1,5].

Motivated by the growing interest in the use of LLMs in policing, we have developed a practical approach to identifying risks, grounded in the policing and legal system of England and Wales. We identified 15 policing tasks that could be implemented using LLMs and 17 risks from their use, and illustrated with over 40 examples of impact on case progression [4].

We identified key issues that the trial of live facial recognition (LFR) should test and set out practical ways to test them [6]. The issues move from the most fundamental questions — *whether a deployment is necessary and proportionate, and who is on the watchlist* — through the accuracy of the system and the way officers act on its alerts, to fairness across communities, the experience of the public, community engagement, and the handling of images and data. A new strand of research in PROBABLE Futures focuses on explaining the decisions of FRT systems by quantifying the uncertainty, stress-testing systems under controlled perturbations, and testing whether counterfactual explanations accurately reflect system behaviour.

In November 2025, our team led a 1.5-day multi-stakeholder workshop in London focused on discussing what evaluations should look like and include for ASR and LLM technologies in policing [7]. We considered four specific use cases: transcription, translation, video summarisation, and information retrieval and question answering. We brought together academics (from linguistics, law, and computer science), police stakeholders (from forces and policing bodies), government stakeholders, and charity stakeholders (e.g., who aim to mitigate violence towards women and girls) to discuss the evaluation requirements and considerations of these four use cases. Our findings highlight that system-level evaluations are insufficient; we must include the human users who use the system too and evaluate how they interact with the AI. Additionally, effectiveness was flagged as more important than efficiency and productivity, and what is “good enough” or “effective” is extremely context dependent. We also assert that if rigorous evaluations cannot happen, then AI systems should not be deployed in policing contexts.

Part 3: Intervention design

We are exploring a range of ways a publicly funded programme could support the development and piloting of scientifically rigorous, repeatable and comparable AI testing and evaluation techniques underpinning AI assurance. The types of interventions below are illustrative examples to prompt your thinking, not a fixed or final set. We would like to understand which you think would be most useful, which you might take part in, and what conditions would need to be in place. We would also like to hear about approaches we haven't listed.

- a) Collaborative research sprints: a series of short, intensive events bringing assurance providers, deployers and researchers together to tackle a defined assurance challenge.
- b) Matched testing pilot: pairing an assurance provider with a deploying organisation to conduct testing and evaluation of a deployment-ready AI use case, contributing towards codification of best practice.
- c) Innovation challenge fund: a competitive grant where teams apply to solve a defined problem developed by a sponsor (e.g. governmental body or business) in testing and evaluation, with winning proposals funded to develop and trial their solution, contributing towards the codification of best practice.
- d) AI assurance methods accelerator: giving third-party AI assurance providers dedicated access to a network of technical experts from NPL and strategic partner organisations to workshop methodological challenges, refine and validate emerging assurance approaches and build evidence on their effectiveness, repeatability and limitations.

1. Which approach do you think would be most valuable overall and why?

The PROBABLE Futures research team recommends bridging b (a matched testing pilot) and d (AI assurance methods accelerator) together into a single programme. Evaluation techniques need to be based in context and, ideally, in particular use cases. Thus, while running a methods accelerator on its own can be useful and interesting, it would lack the use case and context necessary to make the evaluation techniques generated even more meaningful for the AI assurance community and relevant stakeholders [7]. We would recommend that multiple methods (e.g., assessing calibration and demographic differences) be applied and tested within a particular domain and use case. It is crucial when deploying within law enforcement that the legal and operational context is considered up-front to ensure that the relevant legal tests and constraints are understood, plus the consequences of deploying AI for other actors in the law enforcement chain, for example for the prosecution authorities.

3. Which, if any, would you or your organisation be interested in participating in and in what capacity?

For example, as an assurance provider, as a deploying organisation hosting a test, as a research or delivery partner, or as an advisor.

The PROBABLE Futures team would be interested in being a research partner and advisor, subject to the availability of resources and funding.

5. Are there particular areas you think we should focus on as part of this programme?

For example, types of AI, AI testing and evaluation techniques, or risk domains.

Given the rise and adoption of LLMs and facial recognition technologies in policing and the criminal justice system, we would encourage the NPL to focus on building AI testing and evaluation techniques for those AI. Policing and national security are key and high-risk domains which make the rigorous evaluations of them even more important.

Part 4: Closing questions

NPL does not intend to publish individual responses to this Call for Evidence. We expect

to publish an aggregated summary of the insights gathered, which may take the form of a report or a blog post. We will not attribute findings to individual respondents or organisations.

1. May we include your organisation, or your name if you're responding in a personal capacity, in a published list of contributors to this call for evidence?

Yes, the researchers from the PROBabLE Futures team who have contributed to this call response are happy to be identified and acknowledged as contributors.

2. May we contact you to clarify or discuss any part of your response?

Yes.

3. Would you like to be kept informed about the outcomes of this Call for Evidence and relevant next steps or future opportunities to engage with the Centre for AI Measurement?

Yes.

4. If you answered 'Yes' to either of the previous two questions, please provide your contact details.

Yes, Mackenzie Jorgensen (from PROBabLE Futures):
mackenzie.jorgensen@northumbria.ac.uk.

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.

References

- [1] Temitope Lawal, Kyriakos Kotsoglou, Evdoxia Taka, Marion Oswald, Michele Sevegnani, Muffy Calder (2026). The Use of Probabilistic AI in Law Enforcement and the Criminal Justice System of England and Wales. Northumbria University. Report. <https://doi.org/10.25398/rd.northumbria.32769762.v1>
- [2] Evdoxia Taka, Temitope Lawal, Michele Sevegnani (2026). The Need for Assessment of the Cascade Effects from Sequencing AI Systems in Policing. Accepted for presentation to the Inclusive AI Workshop of the 28th ACM International Conference on Multimodal Interaction (ICMI) Conference, Oct 2026, Napoli, Italy
- [3] Marion Oswald, Carole McCartney, Kyriakos N. Kotsoglou, Alexandra R. Taylor, Muffy Calder, Jennifer Ekene Dunkwu, Mackenzie Jorgensen, Evdoxia Taka, Temitope Lawal, Claire Paterson-Young, Katherine Jones, Adrian Weller, Kyle Montague (2026). PROBabLE Future's response to the Home Office consultation on a new legal framework for law enforcement use of biometrics, facial recognition and similar technologies published in December 2025.



- [4] Muffy Calder, Marion Oswald, Elizabeth McClory-Tiarks, Michele Sevegnani, Evdoxia Taka (2026). Responsible AI in criminal justice: LLMs in policing and risks to case progression. Available: <https://arxiv.org/abs/2603.18116>
- [5] Evdoxia Taka, Temitope Lawal, Muffy Calder, Michele Sevegnani, Kyriakos Kotsoglou, Elizabeth McClory-Tiarks, Marion Oswald (2025). Mapping the Probabilistic AI Ecosystem in Criminal Justice in England and Wales. Available: <https://arxiv.org/abs/2512.04116>
- [6] Jennifer Dunkwu, Claire Paterson-Young, Kyriakos Kotsoglou, Evdoxia Taka, Marion Oswald, Carole McCartney (2026). Guidelines for Police Live Facial Recognition Trials. Available: <https://pure.northampton.ac.uk/en/publications/guidelines-for-police-live-facial-recognition-trials/>
- [7] Nneoma Ogbonna, Mackenzie Jorgensen, Muffy Calder, Marion Oswald (2026). Don't Let AI Do the Policing: Findings from a Multi-Stakeholder Workshop on Evaluating Large Language Models for Policing. Available: <https://rai.ac.uk/wp-content/uploads/2026/05/LLM-Evaluation-PROBABLE-Futures-Policy-Brief.pdf>
- [8] Jennifer Dunkwu, Claire Paterson-Young, Michael Maher (2025). Scoping Review of Frameworks for Effective Participatory Oversight of AI in Law Enforcement. Available: https://pure.northampton.ac.uk/ws/portalfiles/portal/89868104/Dunkwu_et_al_2025_Scoping_review_of_frameworks_for_effective_participatory_oversight_of_AI_in_Law_Enforcement.pdf