PROBabLE Futures
Probabilistic AI Systems in Law Enforcement Futures

NPCC National Police Chiefs' Council

Responsible AI Checklist for Policing
Version May 2025

(A living process developed by the PROBabLE Futures project)

Responsible AI: is deciding whether an AI tool should be used (as opposed to 'could' or 'can' it be used). In deciding this, technical and statistical aspects of policing AI should not be separated from legal, contextual, operational, and ethical considerations.

Related documents: the NPCC AI Covenant [1] and relevant College of Policing APP and Guidance to police forces on building AI tools and systems [2] (to follow).

Scope of this checklist: any **AI or Advanced/emerging Data Analytics** tool, as defined in [1]. AI is used as a shorthand for both.

Before using the checklist: The use of an AI tool is not an end in itself; before employing this checklist, assure yourself the use of AI contributes positively and proportionally to a specific policing function (such as preventing and detecting crime), and there is no other capability that would achieve the same outcome. There may be no current capability in this area, and while performance of the proposed AI tool is not ideal, this checklist will help determine whether it is good enough. The overarching question is, does the tool increase opportunities and uphold fairness, justice and the police's impartial service to law?

How to use the checklist: this is a practical guide for those writing and evaluating responsible AI assessments and can be incorporated into training for those responsible for decisions around the deployment of AI. It is intended to provide structured content to inform police forces' and Police and Crime Commissioners' governance, accountability and staff training processes in relation to policing AI.

How to complete the checklist: Factors are listed for three questions concerning i) technical validity, ii) operational deployment, and iii) legality and proportionality. Each factor is a prompt for an explanation or justification; answers should be **detailed, robust, and addressed objectively, with risks and uncertainties** acknowledged. There are no correct or incorrect answers, but there will be answers that are good and those that are inadequate. For technical validity, questions 13a-17a and 13b-16b are alternatives, depending on whether the tool was developed by a 3rd party or in-house; only one set is answered. The example scenarios set out sample **good, adequate and unsatisfactory fictional answers** indicating the level of detail and thought that is

required to answer the questions well. If the circumstances change, the questions will need revisiting. Sometimes, the answer to whether an AI tool *should* be used will be 'no', even if it *could* be. Note that some factors may not be relevant in some contexts.

Does the checklist only need to be completed once? No, the use of the checklist should be part of a **rolling process of evaluation** in line with force governance. The checklist should be reviewed at the main stages of the AI tool's development (in-principle, proof of concept, pilot, trial, deployment, operational, ongoing evaluation) as answers and risks may change as the project progresses.

Who should complete the checklist? it is recommended that a force AI lead with the necessary technical, legal or operational expertise is appointed to take overall responsibility for completion of the answers. This should be someone who has the time, resources and authority to obtain necessary information from relevant police teams, functions and units, including from those responsible for model development and review. Some of the required information may have to be obtained through alternative means; for example, through supplier records or impact assessments. It should also be identified who should clear/approve the checklist answers in line with the force's governance and accountability procedures. This should normally be a senior officer or police staff acting as SRO (senior responsible officer), escalating to Chief Officers where necessary.

How do we identify areas of concern in the answers? Next to each factor, the checklist gives brief guidance about why the factor is relevant and what issues of concern might arise. The example scenarios also include sample **good, adequate and unsatisfactory fictional answers**, which will help SROs identify the standard of detail that they should be looking for. It is essential that the answers are both completed and reviewed by staff with appropriate skills and knowledge to identify issues of concern.

Are there AI tools that should NOT be used?

We recommend the police do not deploy emotional AI or any techniques with unproven or significantly contested scientific validity or theoretical basis, such as the use of biometric analysis to determine veracity of an individual. Furthermore, we would recommend caution is exercised in respect of the use of generative AI for report writing, statement creation and other disclosure and evidential purposes, due to the high stakes nature of such use. Specific legal advice should be taken.

Should the checklist answers be independently reviewed? You should obtain your own legal and data protection advice regarding the answers to the checklist and the implications of any AI deployment. It is also strongly recommended that you obtain advice on any new or innovative AI proposals from a suitably expert and independent AI and data ethics advisory panel (such as the West Midlands PCC and West Midlands Police data ethics committee).

Key considerations in setting up a panel are:

- Purpose and Role: Advisory panels review and critique the responsible, ethical and legal implications of AI in policing, ensuring they do not harm individuals or society by offering advice and suggestions based on potential risks.
- Structure and Composition: Advisory panels should consist of a diverse range of experts, including technology professionals, ethicists, legal professionals, community representatives, and police officers, to provide a well-rounded perspective on technological impacts.
- Ongoing Evaluation: Advisory panels should review projects at various stages in development, regularly assessing projects and/or receiving updates on projects to ensure they meet the required objectives while adhering to the ethical and legal standards required.

Full details on setting up an advisory panel is contained in **Appendix A**.

**** Please note*: as evidential and disclosure considerations involving AI tools are of increasing importance, future versions of the checklist will include additional sections and guidance in relation to this area.**

Checklist lead authors: Professor Marion Oswald MBE, Northumbria University and Lead, PROBabLE Futures Project and Professor Dame Muffy Calder, Glasgow University and Co-Lead PROBabLE Futures Project.

May 2025

CHECKLIST

A. Is the AI tool technically valid, reliable and explainable for the context in which it will be used?

Factors will include	Guidance for narrative assessment
1. Describe the scientific methods/ techniques underpinning the tool and why the scientific basis is considered sufficiently reliable and valid?	It is important that there is an understanding of the data science methods used in the internal workings of the model, and assurance that this is based on adequate and accepted methodology. This assurance might be based on papers and evaluations produced by internal data science staff, from technical documentation provided by the commercial provider and/or from independent evaluations. Considerable caution should be taken in relation to technologies that have unproven, contested or pseudoscience theoretical basis such as emotional AI.
2. For predictive tools, is ground truth data available for this domain, or are there credible disputes about objectivity, e.g. the data labelling is subjective rather than objective?	Ground truth is the accurate, verified data that is used to train an AI tool and which can be used to test the accuracy of the model's performance, for example whether an image depicts a particular individual. There will be some situations where there is no agreed ground truth e.g. an assessment of whether someone is lying or telling the truth, and therefore it will not be possible to evaluate the tool's performance. Deploying AI tools in situations where there is no ground truth is not recommended.
3. Has the tool been used in similar circumstances and are there any consequences (positive or negative) that might be relevant to this application?	Previous deployments of the tool should be reviewed and any lessons learned taken into consideration. This should include technical evaluations and how the tool was used in practice by operational staff.
4. Has the tool been produced by fine-tuning another model or is this tool part of a family of models with the same architecture? If so, what limitations, caveats and operating rules have been identified in respect of the parent/family model(s)?	It is crucial that all relevant limitations/caveats are carried over into the interface and operating procedures for this tool.

5. Are design decisions and model optimisations and trade-offs transparent and suitable for this context?	It is important that details of design decisions are available and understood. For example, it is important to know whether the model has been designed to minimise false positives or false negatives, and to consider whether such a design is appropriate for the operational decisions that will be made with the assistance of the tool.
6. What are the tool configuration settings (e.g. thresholds) and who is authorised to set them? How are settings revealed and explained to tool users?	Thresholds and settings will make a difference to how the results of the tool are generated. For example, in a facial recognition tool, a 'match' or a 'hit' will be generated when a certain number or threshold of matching measurements is reached. If this threshold is reduced, then more 'hits' will be generated, but more false positives will be generated too. This can have significant consequences, for instance if an intrusive power is exercised as a consequence of a 'hit'.
7. For predictive tools, what are the available measures of performance and uncertainty (e.g. accuracy, sensitivity, precision, specificity, confidence intervals)? Do these vary according to protected characteristics under the Equality Act 2010? Are any key measures missing or do they raise concerns?	It is important not to rely on an overall figure of accuracy, but all available measures of performance and uncertainty should be reviewed. For example, a model might have a seemingly high accuracy figure, but also a high false positive figure. This could have significant implications for an operational purpose related to risk assessment. Pursuant to the public sector equality duty, performance evaluations should include any relevant protected characteristics and all reasonable steps should be taken to ensure the tool does not have unjustified bias based on race, sex or other protected characteristic.
8. For generative tools, what are the available measures of performance (e.g. qualitative and quantitative evaluation, benchmarking, task efficiency), are any key measures missing or do they raise concerns?	Although generative AI produces a probabilistic result, this is not always obvious from the output e.g. text or an image. An evaluation protocol should be developed by focusing on the operational context in which the tool will be deployed, and including tests for bias, testing outputs across a range of

	scenarios against correct outputs, tests for consistency and robustness in handling complex or ambiguous cases, and identifying vulnerabilities and security issues.
9. What adverse consequences might occur because of the uncertainties in the tool, and how can these be mitigated?	This question requires you to think about the operational decision that will be informed by the output of the tool, and what might happen to an individual in these circumstances. For example, would it be legitimate to rely on an individual predictive tool to deploy offender management interventions if the tool had a high false negative rate? If a generative AI structures and automates statements, but regularly misses one or more 'points to prove', should this be relied upon?
10. List all information assurance standards that you and/or your 3rd party provider have followed and confirm these comply with nationally mandated standards for policing; record any risks that remain. This question applies to both the development of the model and to the deployment of the model (either or both of these may be done in-house or by a 3rd party).	Implementations of IT tools, which include AI tools, should adhere to nationally mandated standards. This question confirms the appropriate standards have been identified and considered.
11. Is it reasonably possible that the training data could be tampered with or that it could be recovered from the model or leaked by or to a 3rd party?	This question refers to the specific information assurance risk of exposure of sensitive or confidential information used to train an AI model e.g. by manipulation of the model to reveal sensitive data, or by inadequate security around training datasets. This should be considered by data science and data security specialists.
12. If use of the tool involves an element, input, process, competence or organisation that requires accreditation or authorisation, has this been obtained?	For example, accreditation for forensic science within policing set by the Forensic Science Regulator. ISO standards may also be relevant.

3rd party AI tool that has been pre-trained

Factors will include	Guidance for narrative assessment
13a. Has it been trained (possibly with additional in-house fine-tuning) on data of suitable quality, integrity and timeliness that is representative of (for predictive) or relevant to (for generative) the expected inputs?	This question asks you to consider the suitability of the way that the model has been trained by the 3rd party and with the use of non-police datasets. Is the training data representative of the data that the tool is going to be asked to assess in a live environment? If not, then issues of accuracy and bias are likely to arise, and wider questions of whether relying on the outputs would be relevant and legal.
14a. Is there transparency of synthetic data (in training or testing) and if it has been used, how has this been justified, recorded and monitored?	Synthetic data is intended to simulate real data. As police datasets are complex, and police AI tools are often designed to help with decisions relating to individuals, considerable caution should be exercised if use of synthetic data is proposed by the 3rd party. However, synthetic data can help address lack of inclusivity in training data.
15a. What is the tool's lifecycle plan, including retraining or continuous learning, and how will you track and implement updates from the 3rd party, including in-house fine-tuning, and evaluate tool suitability after those updates?	What is the supplier's plan for support to and development of the model and its interface? For how long does the contract require the supplier to support the model? Is the support dependent on a particular platform? How will you (re)evaluate the model when updates are released? How will you (re)evaluate the model if subjected to continuous or periodic in-house fine tuning? For self-learning models, how will you maintain assurance that the overall system remains fit for purpose, proportionate and that any data used for self-learning that is subsequently shown to be inaccurate or unlawfully obtained can be extricated from the model?
16a. Is the data for analysis handled in a secure environment? What is the risk that data for analysis could be leaked to a 3rd party provider?	Is the inputted data for analysis handled securely and locally (i.e. not uploaded to a 3rd party server)? Are system backups similarly secure? Could the data for analysis be used to retrain the model? Is this for the benefit of your

	implementation, or used by the supplier to retrain other implementations for other customers, even if claims of anonymisation are made? If so, have you considered the implications of this, and entered into suitable confidentiality contractual provisions.
17a. Does the contract cover its use as evidence? How does the contract address the model's possible disclosure in legal proceedings or other inquiry? Who could be the 3rd party supplier's witness who is able to explain how the tool works in any legal proceedings or external scrutiny process?	Any attempt by the supplier to contract out of disclosure (or other legal information disclosure obligations such as FOI) should be referred to commercial lawyers because if any decisions are to be made by or influenced by the model, or it identifies exculpatory material, then exemption from disclosure cannot be guaranteed. The 3rd party supplier should be made aware that the model's functionality and overall performance/ reliability may need to be explained during the duration of the contract, including during legal proceedings. It is recommended that you require the supplier to provide appropriate assistance in this regard.

In-house AI tool

Factors will include	**Guidance for narrative assessment**
13b. Has it been trained on data of suitable quality, integrity and timeliness that is representative of (for predictive) or relevant to (for generative) the expected inputs?	This question asks you to consider the suitability of the way that the model has been trained. Is the training data representative of the data that the tool is going to be asked to assess in a live environment? If not, then issues of accuracy and bias are likely to arise, and wider questions of whether relying on the outputs would be relevant and legal.
14b. Has synthetic data been used (in training or testing), if so, how has this been justified, recorded and monitored?	Synthetic data is intended to simulate real data. As police datasets are complex, and police AI tools are often designed to help with decisions relating to individuals, considerable caution should be exercised if use of synthetic data is proposed. However, synthetic data can help avoid the use of sensitive

	data or address lack of inclusivity in your training data.
15b. What is the tool's lifecycle plan, including retraining or continuous learning processes, and performance reviews for drift and suitability?	What is the plan for support to and development of the model and its interface? Is the support dependent on a particular platform? How will you (re)evaluate the model when updates are released or it is fine-tuned? For self-learning models, how will you maintain assurance that the overall system remains fit for purpose, proportionate and that any data used for self-learning that is subsequently shown to be inaccurate or unlawfully obtained can be extricated from the model?
16b. Who could be the witness to explain how the tool works in any legal proceedings or external scrutiny process, and will they have access to sufficient information to be able to give an independent assessment of the tool's performance?	We would recommend that this is the model SRO within the force.

B. Will the use of the AI tool enable accurate and relevant decisions to be made, and positively support the investigative, preventative and evidential process including disclosure obligations?

Factors will include	**Guidance for narrative assessment**
1. Is use and operation of the tool aligned with the intended policing function and legal constraints, and if not, what is the justification for this use?	Explain how the use of the tool will effectively support and improve the policing aim, and how compliance with relevant legal frameworks e.g. PACE is complied with in relation to its use. Has performance of the tool been evaluated for context of intended use e.g. investigative discovery vs countering a defence at court?
2. Explain your process for evaluation of the tool, its operational deployment and any adverse outcomes.	How do you intend to re-evaluate the tool's performance once operational? How often will this re-evaluation take place? What dip-sampling or other analysis of results will take place to ensure that the tool's response is reasonable and explainable? How will you assess whether the tool seems to

	have any biases in its outputs or adverse outcomes? Are there any thresholds above/below which the overall necessity and proportionality of the tool would come into question?
3. How are the outputs presented to users and how are handling caveats and confidence information, such as performance metrics, uncertainties, configuration settings, decision weights, automated decision-making considerations etc presented to users? Are users trained to understand these and the implications of probabilistic results for the policing decision or intervention?	It is essential that users are given the information, training and user interfaces to enable them to understand the realities and uncertainties of the results presented by the AI tool, so that they can assess the relevance of the result and how far they should rely on it in their decision-making. Model developers may want to consider using NIM 'Five by five' ratings to help users recognise the relative reliability and completeness of a model's outputs (see Oswald, Chambers and Paul, 2023).
4. If outputs are fed into other tools, or shared with third parties, or if inputs are themselves in part the product of other tools, what handling caveats are in place and how will uncertainties and confidence ratings be communicated?	'Chaining' of models can introduce compound uncertainty to which the SROs of both models should be made aware. In a similar way to intelligence, it will be crucial that uncertainties and confidence levels are communicated to recipients of the AI outputs so that they are clear about the meaning of the information, and so that consequences of use of the outputs by other organisations are considered.
5. How will the tool's results, methods, workings and operation (including configuration) be recorded and explained for evidential and disclosure purposes? Can it be demonstrated that the tool's outputs have been generated with sufficient quality and reliability for evidential purposes? How will you comply with obligations under CPIA in relation to use of the tool? *** *Please note*: as evidential and disclosure considerations involving AI tools are of increasing importance, future versions of the checklist will**	The main obligations for investigators are set out in the Criminal Procedure and Investigations Act 1996 and related codes of practice, including obligations regarding the recording, retaining and revealing of material (including exculpatory material). It is essential that CPIA obligations are considered during the early phases of an AI tool, such as (i) storage and explanation of a tool's operation and workings within the design (ii) retention of source data that has been analysed by the tool (e.g. original audio recordings) (iii) retention and disclosure of evaluation and audit results, false positive/false negative results, or results

include additional sections and guidance in relation to this area.	dismissed or not followed up due to configuration thresholds.
6. What impact will use of the tool have on the activities and responsibilities of other organisations or law enforcement bodies?	For example, might the automation of certain policing activities shift work elsewhere, such as the checking of statements for accuracy and completeness?
7. Is there a process for ensuring that versions of the tool are retained, recorded and stored securely for the purposes of future audit and disclosure?	This should be undertaken as part of standard IT processes, as well as for the purposes of compliant disclosure processes.
8. If the tool will result in a near real-time response, what additional checks will be carried out to verify the outputs before intervention?	Such responses can often involve coercive or intrusive action subject to legal tests (such as reasonable grounds) and therefore verification will be required.

C. Is the use of the AI tool in this instance, and the subsequent use of its outputs, legal and proportionate? (<u>Specific legal advice must be taken</u>)

Factors will include	**Guidance for narrative assessment**
1. What are the legal implications of the development of the tool, and of the introduction of the AI tool into the relevant police decision-making process?	Assessments may include (i) the human rights necessity and proportionality test and issues of intrusion (ii) data protection considerations in relation to the use of training data, the analysis of input data, including any biometric/sensitive data, the use of the outputs, and risks identified in the Data Protection Impact Assessment (iii) implications of the introduction of a semi-automated method by the use of an AI tool.
2. What additional data/information will users need to make a decision legally and responsibly?	Not all relevant data to policing decisions will be processed by an AI tool, and users should be made aware that additional, relevant sources of information (e.g. confidential information about an extraordinary event taking place) should always be considered.
3. What are the implications for equalities responsibilities (including the Public Sector Equality Duty) for any	The *<u>Bridges judgment</u>* (para 201) made it clear that, because LFR is a novel and controversial technology, police forces

disproportionalities and biases in the tool or the use of sensitive personal data? Has public engagement and consultation with groups likely to be impacted taken place?	would wish to take all reasonable steps to satisfy themselves the software does not have a racial or gender bias. This principle will apply to uses of other AI tools, and it will be important to consider what appropriate evaluation and external engagement (e.g. through an data ethics panel) should take place to provide assurance around bias. What risks have been identified by the Equality Impact Assessment?
4. If the tool could result in a real-time response, do you have sufficient resources to respond to every likely 'match', prioritisation, or relevant output? If not, what will be the consequences if a serious threat or risk is missed or is not acted upon? What Article 2 and Article 3 ECHR issues might arise if threats or risks that are identified and not responded to?	AI tools are generally able to process more data than human investigators, therefore producing an increased number of leads or predictions. This should be anticipated and a policy implemented to facilitate how additional leads will be acted upon and risks prioritised.
5. What vulnerabilities, risks and dependencies have been created by the deployment of the tool into the policing system?	Issues to consider include whether there is any fall-back manual process, the risks of over-reliance and therefore of missing serious cases that could have been predicted or detected by the police through other human-based resources.
6. Are there any ethical or legal issues arising in relation to the method of development of the AI tool or the training data?	An example could include a tool that has been trained on data of an oppressed minority population or prisoners. Considerable reputational issues could arise.
7. Have all training data been appropriately licensed from intellectual property owners and has consent been obtained from any data subjects?	This question is particularly relevant to generative AI and the use of training data that is owned by other copyright holders and personal data of individuals. Legal liability could arise if training data has not been appropriately licensed.
8. Are there any contractual or regulatory terms and conditions, restrictions or limitations attached to the tool or to the training data that are relevant to their use?	For example, that the tool or training data cannot be used for certain purposes.
9. Is the model subject to Investigatory Powers Act safeguards?	For example, does it process or is it based upon IPA warranted data?
10. Has the tool and/or its outputs been made subject to protective marking?	If so, set out how compliance will be ensured.

11. Are all the senior officers responsible for deployments of this model, and the Chief Officer with final accountability, specifically aware of the decisions made to approve its use and any risks associated?	Set out the force governance process regarding use of new AI, as per the advice at the start of the checklist.

Examples

Example 1: Violence Hot Spots

- **A police force's own IT department is developing a deep-learning, temporal-spatial predictive model to identify violence "hot spots" for the purpose of officer deployment decisions. This is a new capability and the force has not asked neighbouring forces if they have similar tools, or relevant NPCC leads for advice. It will be hosted on the force's own IT infrastructure.**
- **The data sources for training draw on existing paid-for datasets including public transport and road traffic, ANPR, cash machine usage data, CCTV locations, anonymised mobile phone-based crowd data, weather forecasts, sporting and entertainment fixtures, PNC nominals (suspects, offenders, victims and witnesses), and the force's database of violent incidents and related intelligence over last three years.**
- **The tool delivers a prediction over a time period of up to 7 days into the future, over a fixed spatial grid. Predictions are shown to users as probabilities and confidence intervals; the GUI displays the former with a colour coding and the confidence interval displayed after one click. There is also a function that provides links to the data (which includes intelligence) on which the prediction is based.**

In completing the checklist, there follows below a **'good'**, an **'adequate'** and an **'unsatisfactory'** response to three sample questions.

A6 What are the tool configuration settings (e.g. thresholds) and who is authorised to set them? How are settings revealed and explained to tool users?

Good: The key configuration setting is the thresholds for each colour in which the predictions are displayed (i.e. indicating high/med/low probability). These can only be set by the project manager/(lead) software engineer in coordination with the SRO for tool deployment. The choice of the thresholds is in line with the design decisions (see A5), which is to minimise false positives. This is explained in the on-line manual. There

are two other configuration settings, which can be modified by a user. These are the time period of the prediction (default is 7 days) and size of spatial grids (default is 1 square mile). These settings can be modified through the interface (one click to reach the settings) by any user. This is also explained in the on-line manual.

Adequate: The thresholds for probability colours are set by the project manager/software engineer, in collaboration with the SRO. The time period for predictions and size of spatial grids can be configured by the user. This is explained in the on-line manual.

Unsatisfactory: This is explained in the on-line manual.

A9 What adverse consequences might occur because of the uncertainties in the tool, and how can these be mitigated?

Good: This is a predictive tool and we recognise it will never be 100% accurate. This has been explained to the force's senior officer forum already, and we have taken their feedback on board in designing the training for users and mid-ranking officers. We have identified three particularly uncertain types of input data and have weighted their impact lower. We have added a feature to the tool that specifically highlights any regions/districts of the force area that would be left without a suitable unit available within the mandated response time. We will not use the tool solely to manage complex deployment days (e.g. Premiership football fixtures) but instead use it in the background to build confidence and make its outputs available to Gold commanders. The tool now reminds/prompts users to view confidence information.

Adequate: Uncertainties might mean officers are deployed too far away from an actual unpredicted incident. This is scenario will be included in the in-person training that all users must complete before being allowed to use the tool.

Unsatisfactory: We intend to mention uncertainty in the user training documentation.

A15b What is the tool's lifecycle plan, including retraining or continuous learning processes, and performance reviews for drift and suitability?

Good: For the first three years, performance will be tracked and reviewed twice per year. Unless a significant concern arises at a performance review, we plan to retrain the model after 18 months. We expect the data sources to be unchanged but the data used for retraining will be more up to date. At the point of retraining, we will also consider whether to change any underlying internal weightings/design or modify the interface.

After the initial three year deployment, we will conduct a major review that will include suitability of training data sources and platform lifetime. At that point we will develop a new lifecycle plan.

Adequate: For the first three years, performance will be tracked and reviewed twice per year. Retraining will take place after 18 months.

Unsatisfactory: We will retrain the model after 18 months.

C3 What are the implications for equalities responsibilities (including the Public Sector Equality Duty) for any disproportionalities and biases in the tool or the use of sensitive personal data? Has public engagement and consultation with groups likely to be impacted taken place?

Good: One particular ethnic group is seemingly over-represented in violence offences data and so we have engaged with the PCC about this project. The Equality Impact Assessment only flagged that one group as requiring particular attention. But the force believes some violence offences are under-reported in another ethnic group within the force area – that group has a population concentration in one town and so we have modelled and weighted offence data in that location to match the PCC- and Chief Officer-agreed 'most likely' actual incidence. The PCC has used this project as an example in one of his/her quarterly public engagement talks to illustrate what the force is doing to try to tackle violence. The bought-in data is somewhat reliant on device usage (phones, cars, cashpoints, etc) which we know is lower in the lowest socio-economic groups which, in this force area, has a strong correlation to certain racial groups; but research best practice suggests that there is no reasonable way to mitigate this. The users are required to record when they used the tool to assist in deployment decisions and the MSRO will review these logs alongside actual offence data on a quarterly basis.

Adequate: We have conducted an Equalities Impact Assessment which concluded there were no major issues to address. The bought-in contextual data comes with some assurances about bias mitigation. Because this is about operational police decision making we have not consulted with the public.

Unsatisfactory: This is merely reusing data we had already so we do not need to consider equalities. This tool does use sensitive personal data but does not expose it to a larger user base than before.

Example 2: Transcription of body-camera footage

- **A police force has proposed trialling the automated transcription of body-camera footage based on a free trial of a software tool which is hosted externally. The purpose of using the tool would be to reduce officer time needed to produce reports and statements for record-keeping, and for evidential and disclosure purposes, by producing these automatically.**
- **The tool produces transcripts and chronological statements highlighting 'points to prove' for a specified offence, with speaker labels and exact timestamps. Transcription includes a translation step from over 90 languages and there is a summarisation facility.**
- **The vendor does not disclose how it works but shows testimonies from previous policing clients in Singapore and California. The vendor requires the force to give them 50 hours of typical footage, in order to 'tune' the tool.**

In completing the checklist, there follows below a **'good'**, an **'adequate'** and an **'unsatisfactory'** response to three sample questions.

A1 Describe the scientific methods/ techniques underpinning the tool and why the scientific basis is considered sufficiently reliable and valid?

Good: There are four distinct techniques employed within this tool:
i) audio -> text transcription
ii) audio-> speaker diarisation
iii) text -> text translation
iv) text -> text summarisation
All but the last are well established techniques (both in the UK and internationally) with acceptably objective ground truth data and well-established performance benchmarks. The last is a recent technique and there is little guidance about how to evaluate summarisation in a policing context. We have contacted the clients who provided testimonies to ask for information about their evaluations of the summation function, but we have not yet received a reply. Evaluation of summarisation (generally) is a new area of active research.

Adequate: The tool involves transcription, translation, speaker diarisation, and summarisation. The last technique is relatively new and staff have not been able to evaluate it within this tool.

Unsatisfactory: The tool involves transcription, translation, speaker diarisation, and summarisation.

A13a Has it been trained on data of suitable quality, integrity and timeliness that is representative of (for predictive) or relevant to (for generative) the expected inputs?

Good: No information is available on the training data used by the supplier, for any of the four techniques employed. However, the additional (fine-tuning) training provided by us will be highly representative of the video/audio that will be analysed in the live environment. Further, we requested performance measures from the supplier for the predictive techniques (i.e. carried out before fine-tuning) and these were acceptable, though they did not indicate demographics. There may be issues concerning bias and accuracy, particularly regarding processing of UK dialects and accents. Testimonies from the policing clients in Singapore and California did not mention any issues with local dialects/accents, but this is a potential risk and we will carry out internal evaluations from day one.

Adequate: No information is available on the training data used by the supplier, but the performance measures (for the predictive techniques) are acceptable. The additional (fine-tuning) training data provided by us will be highly representative of the video/audio that will be analysed in the live environment.

Unsatisfactory: This is a trial and so we couldn't obtain any information about training data. If we purchase the product, after the trial, we will request this information.

A16a Is the data for analysis handled in a secure environment? What is the risk that data for analysis could be leaked to a 3rd party?

Good: We have a contract in place with the provider even though this is a free trial. We have conducted usual due diligence checks on the supplier and its directors, and with the necessary GDPR permission have checked the team who will be working with us to the same levels as if they were employed contractors. We have visited their data centre, which only handles UK data (not international clients'), asked questions about how they secure data for similar clients and have read their most recent ISO 27001 inspection report. Our information assurance team and the force's Chief Information Security Officer (CISO) have been consulted, and their advice is fully included in the trial. The MSRO has briefed our PCC about the project and specifically about our data being processed by the 3rd party. We have also taken the unusual step of consulting the ICO advisory team because of the novelty of this proposal, although they considered that we had already put in place appropriate safeguards for the trial phase.

Adequate: We have seen the provider's ISO 27001 certification and they have assured us that our data is only accessible to our users and a very small number of their staff. We have exchanged letters to this regard.

Unsatisfactory: The 3rd party provider employs a large number of ex-police officers and shows a well-secured data centre in its glossy brochure.

B5 How will the tool's results, methods, workings and operation (including configuration) be recorded and explained for evidential and disclosure purposes? Can it be demonstrated that the tool's outputs have been generated with sufficient quality and reliability for evidential purposes? How will you comply with obligations under CPIA in relation to use of the tool?

Good: We have a contract in place with the supplier even though this is a free trial. The contract requires the supplier to provide a high level statement of the tool's workings and if necessary to field a witness, albeit at cost to the force. The tool has a comprehensive audit log which on top of the supplier's ISO27001 certification and other information we have received to date lead us and the force's CISO to conclude that there is no more evidential/disclosure risk with this tool compared to several in-house tools we regularly deploy today. We have reviewed the various features of the tool and will not make use of the automated translation feature in production of statements/evidence without the involvement of an approved translator. We have trialled the tool with our most likely foreign languages and consider its accuracy to be acceptable, although this is one feature we will specifically review within the trial. During the trial period, the project team will review every transcript used for statement purposes prior to the statement being submitted – this is a significant amount of effort and will restrict the size of the trial. For disclosure, the reviewed transcript and raw video/audio will be retained. Disclosure officers and SIOs have been briefed on the project.

Adequate: We have exchanged letters with the supplier, these address the need to prove continuity and provenance of evidence. The supplier's ISO 27001 certification demonstrates data integrity. We will retain raw video and audio so that if needs be we can refer to that as evidence or for disclosure. Users and disclosure officers can make use of approved translators in addition to the tool.

Unsatisfactory: The tool is proprietary and so until a contract is in place we can't have its inner workings explained to us. The officers concerned are responsible for their own statements, and SIOs for disclosure, so we do not need to consider any risks in that regard.

Appendix A

Advisory panels in policing

Dr Claire Paterson-Young, PROBabLE Futures Co-I and Dr Jennifer Dunkwu, PROBabLE Futures Research Fellow, Northampton University

Purpose of an advisory panel in policing

Advisory panels[1] have a role in reviewing the ethical implications of projects that involve data and technology, advising on the ethical and legal risks associated with projects (Oswald et al., 2024). As they perform this function, part of their focus is to consider the rights of those who will be impacted by projects that involve data and technology. This reflects the core purpose of advisory panels in policing which is to be a "critical friend" (Oswald et al., 2024). In this capacity, the advisory panel is not there to simply validate projects but to constructively question and critique the impact of projects to ensure that they do not cause any harm to organisations, individuals, or society. For instance, advisory panels can review projects and offer a range of outcomes that provide insights into improving the project, for example, fully approving the project, approving with minor or major changes, or ultimately rejecting the project if there is a risk of serious harm (Oswald et al., 2024). To set up an advisory panel of this nature, there are key recommendations from Oswald et al. (2024) and the PROBabLE Futures scoping review of frameworks. The advisory panel should not exist at a single point in time but engage in all stages of project development including (Note – questions are for illustrative purposes only and questions for each would be designed in accordance with the checklist):

1) Problem identification – clearly define the problem and the reasons technological solutions have been sought to resolve the problem (i.e., what is the need for a technological solution? What type of solutions are proposed? What are the risks of a technological solution? What is the value of a technological solution? Can the problem be better resolved by other non-technological means?).

[1] See example - The West Midlands Office of the Police and Crime Commissioner (WMOPCC) and West Midlands Police (WMP) data ethics committee.

2) Design – potential solutions to the problem including the different approaches and solutions currently available (i.e., What type of technology and/or data is used? What are the potential risks of using the technology and/or data? How does this technology and/or data resolve the problem? What expertise is available and/or required?)
3) Development – development of the prototype including further details on the development based on the design (i.e., How does the prototype resolve the problem? Is the prototype solely for use in tackling the original project? Why was this prototype selected [benefits/risks]? What is the accuracy rate? Is the protype explainable?)
4) Testing – testing the technology/protype in a controlled and safe environment that represents the deployment environment (i.e., What are the initial findings? Does the testing stage provide proof of concept? Does the technology match the intended purpose? Should it return to an earlier stage? Can outputs be explained? Are there any issues with bias? Are there modifications that should be made?)
5) Deployment – deployment in the real world (i.e., Does the technology perform as expected? How do officers engage with the technology? Is the technology explainable? Are there any issues with bias? What impact is the technology having on policing?)
6) Evaluation and Review – evaluation and review of the tool through independent research/evaluation processes (i.e., Does the technology performance align with the original purpose? Does the technology pose any new risks as it evolves? How is the technology deployed and used in practice? Is there any reluctance in using the technology from officers? What mechanisms are in place for auditing, transparency and oversight?)

Three key areas to consider in setting up advisory panels: clearly defined objectives, identifying composition of the panel, and evaluations and adjustment:

Clearly defined objectives for panels

Provide guidance on the ethical implications of using emerging technologies in law enforcement, ensuring that these technologies are deployed in a way that upholds human rights, fairness, and transparency. Setting clear objectives, expectations, and

responsibilities for the panel from the outset, and communicating these to the public, is the first step to consider in putting together an advisory panel for policing (Oswald et al., 2024). Examples of specific objectives:

- **Ethical Oversight:** Review and guide the ethical use of technology in law enforcement to ensure human rights, privacy, and fairness are protected. This includes safeguarding against biases in technology (i.e., algorithm bias), ensuring these technologies do not disproportionately affect marginalised or vulnerable groups.
- **Policy Review and Development:** Provide input on the development or modification of policies regarding the use of technology, such as facial recognition, predictive policing, and surveillance systems.
- **Transparency and Accountability:** Foster transparency in the use of technology, ensuring the public is informed and the systems are accountable.

A clear term of reference sets the advisory panel up for success and should be developed to outline the roles and responsibilities of members, their values and scope (Oswald et al., 2024). Where possible, members should get familiar with the TOR before the first meeting (College of Policing, Data Ethics Committee, 2024).

Identify the composition of the panel

A key question in designing an advisory panel is who should get involved? Oswald et al. (2024) and PROBabLE Futures scoping review strongly recommend designing a committee with diverse representation to ensure the panel includes a broad range of perspectives for well-rounded and inclusive feedback. Potential members could include:

- Ethicists with knowledge on research, safeguarding, technology and law enforcement.
- Experts with a deep understanding of algorithms, data ethics, and technological capabilities.
- Legal experts with knowledge about rights and responsibilities.

- Community Representatives to ensure that diverse public perspectives are considered, particularly from vulnerable groups (The Leadership Conference on Civil and Human Rights, 2019).
- Academics focused on social justice, technology, and policing practices.
- Police representatives who can share their knowledge and understanding of the policing arena (Oswald et al., 2024)

There should be clear criteria on membership, with members expected to have diverse expertise (i.e., not all members will have expertise in all areas, but the cumulative expertise will provide an interdisciplinary perspective on the impact of technology on society and policing). Balanced-diverse representation can bring about validity and acceptance of the decisions taken on projects (Haddaway et al., 2017).

Evaluate and adjust advisory panel

Regular assessment of the advisory panel's work is required to ensure it is achieving its objectives and making a tangible impact on the ethical use of technology in law enforcement. Part of Tomlinson and Parker (2021) six-step stakeholder engagement framework is to analyse the effectiveness of the relationships between stakeholders and consider any necessary modifications or actions that can improve engagement and consequently enhance project outcomes. Additionally, Oswald et al., (2024) also calls for routine evaluation of whether and how advisory panel's recommendations have been followed.

References

College of Policing, Data Ethics Committee (2024). Organisation including workforce. Available from: https://www.college.police.uk/support-forces/practices/data-ethics-committee [Accessed 13th November 2024].

Haddaway, N. R., Kohl, C., Da Silva, N. R., Schiemann, J., Spök, A., Stewart, R., Sweet, J. B. and Wilhelm, R. (2017) A framework for stakeholder engagement during systematic reviews and maps in environmental management. Environmental

Evidence. 6(1). Available from https://doi.org/10.1186/s13750-017-0089-8 [Accessed 10th October 2024].

Oswald, M., Paterson-Young, C., McBride, P.,Maher, M., Calder, M., Gitanjali, G., Tiarks, E. and Noble, W. (2024) Ethical review to support Responsible Artificial Intelligence (AI) in policing: A preliminary study of West Midlands Police's specialist data ethics review committee. Available from: https://researchportal.northumbria.ac.uk/en/publications/ethical-review-to-support-responsible-artificial-intelligence-ai- [Accessed 20th November 2024].

The Leadership Conference on Civil and Human Rights (2019) New Era of Public Safety: A Guide to Fair, Safe, and Effective Community Policing. Available from: https://civilrights.org/wp-content/uploads/Policing_Full_Report.pdf [Accessed 12th November 2024].

Tomlinson, E and Parker, R. (2021) Six-Step stakeholder engagement framework. Available from:https://training.cochrane.org/sites/training.cochrane.org/files/public/uploads/Six%20Step%20Stakeholder%20Engagement%20Framework.pdf [Accessed 9th November 2024].