Responsible AI Checklist for Policing – WORKING DRAFT v8
(A 'living process' to be developed alongside the outputs of 'PROBabLE Futures')

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], the NPCC Guidance on Advanced Data Analytics in Policing [2] (draft), the Recommendations concerning model/system cards [3] (to follow) and College of Policing Guidance to police forces on building AI tools and systems [4] (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. All the ethical considerations of [2] are covered: Lawful, Transparent, Explainable, Responsible, Accountable, and Robust.

How to use the checklist: There are no correct or incorrect answers; 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 contribute to the development of governance and accountability processes in relation to policing AI.
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. 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.

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. Overall, does it increase opportunities and uphold fairness, justice and the police's impartial service to law?

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

Version 8, 16 October 2024

CHECKLIST

1. Is the tool technically valid, reliable and explainable for the context in which it will be used? Factors will include:

- Has it been trained on lawfully obtained data, of suitable quality, integrity and timeliness, and representative of the inputs expected in this instance? What is the risk that training data could be tampered with or 'poisoned'?
- Is there transparency of training and testing data provenance, design decisions, and (underlying) model optimisations and trade-offs to ensure transparency and repeatability?
- How often will the tool be retrained and is there a process for reassessing new training data and reevaluating the tool's outputs on retraining?
- Has synthetic data been used in training of the tool, and if so, how has been justified, recorded and monitored? Is the synthetic data representative of real world use cases?
- What are the measures of performance and uncertainty (accuracy, sensitivity, precision, specificity, confidence intervals) and how do these vary according to protected characteristics? What adverse consequences might occur as a result of the uncertainties in the tool, and how can these be mitigated?
- What are the tool configuration settings and model parameters, who is authorised to set them and why they are appropriate for the policing function involved? How are settings recorded and explained to tool users?
- How often has it been used in similar circumstances, and to what effect?
- Who will be the witness to explain how the tool works in any legal proceedings or external scrutiny process? Will they have access to sufficient information to be able to give an independent assessment of the tool's performance?
- If a third-party tool, does the contract cover its use as evidence or for disclosure in legal proceedings?
- Is the AI tool, its training data and the data to be analysed handled in a secure environment? Is there any risk that data could be leaked out as a result of training external commercial AI models?

2. Will the use of the tool enable accurate and relevant decisions to be made, and positively support the investigative, preventative and evidential process including disclosure obligations? Factors will include:

- 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?
- How are the outputs presented to users? How are uncertainties, errors and parameter settings identified and communicated, and are these clearly understood? Are users properly training to understand these uncertainties and performance metrics (accuracy, sensitivity, precision, specificity, confidence intervals) and the implications of the probabilistic results for the policing decision or intervention?

- How will the tool's results, methods and operation be recorded 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?
- If the tool will result in a near real-time response, what additional checks will be carried out to verify the tool's outputs before intervention?
- How and which tools are chained and how will meta data be recorded and passed between tools and systems? How will information about uncertainties be maintained and communicated?

3. Is the use of the tool in this instance, and the subsequent use of its outputs, legal and proportionate? (Specific legal advice must be taken) Factors will include:

- What policing function and power justifies the acquisition and analysis of the relevant data, and the deployment of the output in the context of police action?
- Is there legal power to use AI if it changes police decision-making to an automated or semi-automated method?
- Is it understood what additional data/information officers may need to make a decision legally and responsibly that is not contained within the tool?
- What are the implications for equalities responsibilities (including the Public Sector Equality Duty) for any disproportionalities and biases in the tool? Has public engagement and consultation with groups likely to be impacted taken place?
- For how long will the outputs be retained, and for what purpose? How will the outputs be stored so that they can be subject to future scrutiny and disclosure to a prosecutor and defence representative?
- Has performance of the tool been evaluated for context of intended use e.g. investigative discovery vs countering a defence at court?
- If the tool will 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?
- What vulnerabilities, risks and dependencies have been created by the deployment of the tool into the policing system? Is a resilience plan in place in the event of failures occurring?
- Considering i) any use of biometric/sensitive data, ii) the likelihood of inaccurate conclusions, iii) the implications of individual interventions based on the results of the tool and iv) the availability of alternative methods for achieving the policing function that may be less intrusive, is the use of the tool proportionate?
- Have training data been properly and legally acquired or licensed from the intellectual property or data owners/providers or data subjects?
- Are all of 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?

The following case-studies aim to illustrate the use of the checklist, although should not be taken to cover all relevant factors:

Case study 1: modern slavery [based on one of the examples in the NPCC Advanced Data Analytics Guidance]

Loamshire Constabulary's data lab is developing a model to identify networks of people involved in modern slavery. To find data and text that may identify potential victims and perpetrators, it uses a third-party large language model, and to identify the highest risk cases it uses an internally developed neural network to predict which individuals are likely to become high harm modern slavery offenders. Names identified by both tools are referred to specialist officers for further investigation and action.

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

- As the large language model has been acquired from a third-party, Loamshire has no knowledge of the training and testing data (and how this has been acquired or licensed), the model optimisation or its quality when applied to the relevant policing datasets. **This is a red flag**.
- As the predictive tool has been internally developed, **Loamshire will have knowledge of the training, testing and design process**. However, predictive AI tools will be designed to avoid certain types of inaccuracies e.g. false negatives. **This is an orange flag if decisions about interventions and further investigation will be made based on the tool's results and without consideration of which inaccuracies the tool attempts to avoid**.
- The data on which the predictive tool is trained, although containing intelligence, is patchy and limited to lines of enquiry pursued over the last two years. **This is an orange flag.**

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

- The third-party large language model does not provide any explanation of why it has generated a certain result (e.g. *J Bloggs* is a victim of modern slavery) nor any indication of the accuracy, sensitivity, precision and specificity of the results. **This is a red flag.**
- The internally developed predictive tool has been built with a function that displays each output requested from the tool and provides links to the data (which includes intelligence) on which the output is based. It also provides the user with an % estimate of the accuracy of the output. **This information helps the user to understand the output**. However, the tool only provides an overall accuracy rate, and does not give an indication of the other measures of accuracy. The tool also does not create a permanent record of each result, and there is no audit capability for access to the intelligence data. Nor are there any handling conditions around the product of the tool which should itself be treated as intelligence. **This is an orange flag.**

Is the use of the tool in this instance, and the subsequent use of its outputs, legal and proportionate?

- 78% of the potential perpetrators flagged by the large language model are previously unknown to the specialist officers. 62% of those flagged are noted to have a surname of West African origin. **This is a red flag.** No further operational use of the tool should be made until the police can determine the reason(s) for the tool's output. Otherwise, the force risks serious errors occurring in the investigation process and being in breach of its public sector equality duty.

Case study 2: live facial recognition

X Constabulary is proposing to use a Live Facial Recognition system at a number of locations to identify 'priority' offenders from a watchlist that is updated daily. An on-site officer will assess any images that are flagged as a match by the system, and will decide whether or not to stop and question a passer-by. Images are deleted after a very short period of time and flags are only "live" for *n* minutes. Images that are flagged, but not processed within that time period, are then deleted.

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

- The system has been acquired from a third-party, **but it was tested for two other forces by an academic body (although the testing was of equitability i.e. whether performance differs across key demographics, and the test acknowledges that different results may be obtained in different contexts and environmental conditions). X has access to that report**. A key question will be which parameter settings are recommended in the deployment by X and how they relate to these tests or any other testing results and subsequent recommendations. Also, how will the recommended value(s) be set and monitored? **This is an orange flag**.
- As the system has been acquired from a third party, there is no information available as to whether the tool was trained on data similar to that expected to be captured in this deployment**. This is an orange flag**.
- There is no explanation of how the watchlist is generated and communicated to the site daily, who assesses the integrity of the images, and if the volume is viable for this deployment. **This is a red flag**
- There is no clear system and allocated responsibility for recording and evaluating false positives (i.e. false identifications). **This is an orange flag**.
- There is no clear system and allocated responsibility for retrospective evaluation of false negatives (i.e. identifications that were missed by the system). **This is an orange flag**.
- There is no indication of how the system is upgraded, how often, and the scope of upgrades. Will upgrades include installing a re-trained model? **This is a red flag.**

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

- **The decision of whether or not to stop and question is taken by an officer**, but there is no explanation of how officers have been trained on use of the system and what information is presented to them in case of a positive identification. There is no indication of monitoring the volume of positive identification flags and whether a real-time response is possible. **This is an orange flag.**

Is the use of the tool in this instance, and the subsequent use of its outputs, legal and proportionate?

- **The use of the tool greatly extends the force's theoretical capacity**, but it is not clear what are "good" results and whether resources may require to be diverted from other activities. If significant biases become apparent from analysis of the false positives, there is no documented process for suspension of the operational use of the system. **This is an orange flag.**
- The watchlist contains 16,000 offenders and wanted individuals, not limited to those suspected of serious crimes who are likely to be found in the LFR deployment area. The deployment location is chosen by local officers based on local needs. **This is an orange flag.**
- There is no indication of community engagement concerning data deletion and monitoring for biases. **This is a red flag.**

Case study 3: use of existing data to identify/predict violent hot spots

X Constabulary is developing a deep-learning, temporal-spatial predictive model to identify violent "hot spots". The data sources for training draw on existing sets including public transport and road traffic, ANPR, cash machines, CCTV, mobile phone data, weather forecasts, sporting and entertainment fixtures, and database of violent incidents and related intelligence over last *n* years. The tool delivers a prediction over a time period of up to 7 days, over a fixed spatial grid. Predictions are 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.

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

- As the predictive tool has been internally developed, **X will have knowledge of the training, testing and design process**. However, there is no indication of sensitivities, such as how accuracy of a prediction varies with time, e.g. a prediction issued 2 hours ago versus one 48 hours ago, and how this aligns with the expected use. Further, AI tools will be designed to avoid certain types of inaccuracies e.g. false negatives or false positives. **This is an orange flag if decisions about interventions and further**

investigation will be made based on the tool's results and without consideration of time sensitivities and which inaccuracies the tool attempts to avoid.

- There is no indication of monitoring performance and who will evaluate the need for re-training. **This is an orange flag**.
- There is no indication of the additional data/information officers may need when deciding to deploy resources that is not contained within the tool (e.g. known offenders bailed to certain addresses, local villain recently incarcerated, recently arrived new family 'known to police'). **This is a red flag (also relevant to legal and proportionate).**

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

- **The decisions of whether or not to deploy resources are made by users**, but there is no explanation of how users have been trained, especially on how to interpret probabilities and confidence intervals. **This is an orange flag**.
- No evaluation process is in place and therefore there is no method of comparing aided deployment decisions versus non-aided and the subsequent impact/number of solved or deterred crimes so that value for money can be deduced. **This is a red flag.**

Is the use of the tool in this instance, and the subsequent use of its outputs, legal and proportionate?

- The use of the tool is a new capability and it is not clear what are "good" results and how to justify resources being diverted from other activities. If significant biases become apparent from analysis of the false positives, there is no documented process for suspension of operational use of the system. **This is an orange flag.**
- There is no indication of access controls and audits concerning the data, especially intelligence and CCTV, on which predictions are based, and whether or not it contains personal data. **This is a red flag.**

Case study 4: triaging text message crime reports

X Constabulary has agreed to a free trial (which itself introduces risk) of a software tool that analyses text message-based crime reports. The vendor does not disclose how it works but shows testimonies from previous policing clients in Singapore and California. The vendor asks to receive two 8-hour shifts' worth of messages manually categorised by police call handlers in one of the force's two control rooms, in order to 'tune' the tool. The system categorises the messages as 'urgent' (routed to 999 call handlers), 'priority' (top of the queue, and domestic violence victims are 'guaranteed' to be prioritised) or 'routine' (lower down/next working day).

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

- The system has been acquired from a third-party and there is no indication of whether performance differs across different cultural contexts and conditions (e.g. use of slang,

geographical features). It is not clear whether some of this detail is included in the customer testimonies or can be obtained from the vendor. **This is an orange flag.**

- As the system has been acquired from a third party, there is no information available as to whether the tool was trained on data similar to that expected to be captured in this deployment. There is an assumption this was developed as an English language tool, but that is not explicit. Further, it is not clear which two 8-hour shifts have been selected for 'tuning' and who has assessed the quality and appropriateness of this data. **This is an orange flag.**
- There is no indication of monitoring performance and who will evaluate the need for re-training/re-'tuning'. **This is an orange flag.**

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

- There is no explanation of how domestic violence is defined and identified within text messages, and whether subject matter experts have been involved. Does the identification process rely on a probabilistic judgement involving AI, and if so, what does the model optimise? **This is a red flag.**
- There is no clear system and allocated responsibility for retrospective evaluation of false negatives (i.e. text messages that were missed or incorrectly categorised by the system). **This is an orange flag.**
- Text based crime reporting is a new capability and there is no indication of the expected performance for urgent categorisations and whether this will be included in the "under ten seconds" 999 call target. **This is an orange flag.**

Is the use of the tool in this instance, and the subsequent use of its outputs, legal and proportionate?

- **The new capability offers a new way for the public to report crime.**
- What happens at the end of the 'free trial', specifically what has happened to the training data handed over that includes personal data? **This is an orange flag.**
- The use of the tool is a new capability and it is not clear what are "good" results, and how will the successful triaging of priority and domestic violence calls be monitored. What is the risk of a call from someone at risk of serious imminent harm being missed? **This is an orange flag.**