Validation and verification

The validation and verification of artificial intelligence (AI) tools

First published 17 June 2025

8 mins read

These initial tests are to establish that the tool or system:

- is safe and legal, and complies with regulatory and other professional standards that apply to your force
- does what it is supposed to do

The **Data-driven technologies authorised professional practice** (APP) sets out the overarching principles and requirements for validation of data-driven technologies. It includes the requirements that:

- you should complete an algorithmic transparency report
- all machine-learning tools and systems should be tested on your force's data, regardless of how much testing it has undergone before

For advice on your testing plans, you can ask the Centre for Police Productivity at the College of Policing, which has set up the Evaluation Advice and Review Panel, where research and evaluation specialists can provide advice and scrutiny of forces' plans.

Key elements of validating the tool or system

In addition to the information in this section, see the Department for Science, Innovation and Technology's **Introduction to AI assurance**.

Risk assessments and compliance

You should complete a risk assessment, an equality impact assessment and a data protection impact assessment. You should also establish with your legal team that the technology complies with the force's other legal and regulatory duties.

Bias audit

A bias audit is where an algorithm is applied to force data and outcomes are measured to identify whether any potential biases have been introduced. You would divide your test data into sets containing data from groups that you have identified as being at risk of bias and run the model to see if it generates disproportionate outcomes. Remember that the risk of bias is not just about minority groups in the general population, but also about people who are under-represented in the specific data set the model has been trained on. For example, an algorithm trained on domestic abuse data might not work accurately on female perpetrators or male victims. This testing can be done by analysts within your force. You can also seek advice from the College of Policing.

If bias testing reveals unjustified disproportionalities, appendix A of the Department for Science, Technology and Innovation's **Review into bias in algorithmic decision-making** sets out bias mitigation strategies and links to detailed explanations of how these work. However, your recourse to these may be limited if the tool or system is an off-the-shelf third-party model and you do not have access to the solution code for the model.

Interpretability and explicability testing

A further test is to run the tool or system and see if you can predict and explain the decisions it makes. This can be done by the delivery team in conjunction with subject matter experts and/or those who will use the tool in practice.

Penetration testing and red-teaming

Led by your information security team, penetration testing identifies and probes the tool or system's vulnerabilities, and gauges conformity with your force's information security obligations. Red-teaming is an adversarial approach to penetration testing that might, for example, emulate a hacker trying to break into the system and access sensitive information. The aim is not just to test the AI's vulnerabilities (which can be difficult to fully assess because of their 'black box'), but the overall response of your organisation to malign attacks. These types of tests can be carried out by internal information security teams or provided by independent **accredited NCSC CHECK penetration testing teams.**

Verifying the tool or system

Formal verification involves ‘lab’ testing to ascertain that the product does what it is supposed to do. This will involve testing the tool or system on a data set that it has not been trained on.

Verification metrics

You will need to understand what your performance indicators are, as well as the appropriate metric. These will vary according to the type of AI being used.

For tools that involve pattern analysis, prediction, classification and information retrieval, the correct metric will be precision and recall. This is a measure of both:

- the percentage of correct predictions
- the measure of false positives – for every correct prediction, how many false positives were also returned. Go to **Appendix 1** for further explanation of how this works and how to apply it

To assess whether the rate of false positives or negatives invalidates the model, you will need to weigh their costs and harms in the deployment context. For example, if false negatives put someone at risk of death or serious harm, you are likely to have a very low tolerance of them. However, if the cost of a false positive means that you distribute anti-burglary leaflets to one more postcode than you needed too, you may have a higher tolerance of this, provided the intervention is effective overall.

Do not use accuracy (correct predictions out of all predictions made) as a metric for these tools and systems. It is not suitable for contexts when you are trying to establish the probability of rare events, especially where the cost of false negatives is severe.

Testing generative AI

A different approach will be needed for verifying generative AI, such as chatbots and other tools, which performing complex, subjective tasks such as summarising, synthesising and generating content. You will typically need to test for risks and safety, which is especially important if outputs are going to reach or interact with the public, as well as quality and performance.

Risks and safety checks are about ensuring that the tool does not generate content that is harmful – for example, by generating content that is hateful, sexual, violent or related to self-harm – or that contains protected copyright infringing content. For more information go to Microsoft’s

Observability in generative AI.

For assessing the quality of generative AI outputs, you will need to decide what performance metrics you are looking for before you test. These are likely to be factors such as:

- groundedness (how well generated answers align with information from the source data)
- relevance (the extent to which responses directly relate to queries)
- fluency and coherence (whether the output flows smoothly and is grammatically correct)

When assessing quality and performance of the generative AI tool, you will usually be looking for the degree of consistency with a human comparator – or whatever your business-as-usual is – doing the same task. You are not looking for perfection. For example, if you are bringing in a tool that claims to convert interview transcripts into witness statements, you would then give the same transcript to the tool and to a member of your force who does this routinely as part of their job, and then compare outputs. This comparison could be done by randomly presenting the versions to a subject matter expert to see which ones they prefer.

You would need to do this with multiple example transcripts that reflect the range of subject matter and difficulty that the tool will encounter, and then run it several times over the same text as a further gauge of predictability and consistency. For risk and safety testing and for quality and performance testing, there are automated and even some AI-assisted tools that your data scientists can explore to help. However, there should be human review as well. Commonly used automated generative AI solutions include:

- F1 score
- bilingual evaluation understudy (BLEU) score
- recall-oriented understudy for gisting evaluation (ROUGE) score
- metric for evaluation of translation with explicit ordering (METEOR)

These are just a couple of examples of how approaches to evaluating AI tools and systems vary from product to product.

Sources of verification testing

Whatever the type of tool, you should consider whether its complexity or the risks associated with it require independent expertise. For verification testing, this can be provided by the same sources

that have traditionally supplied evaluation support to policing, universities and established research consultancies.

Results of verification

If the testing reveals glitches and unwanted patterns, data scientists can tweak the tool through feature engineering, data splitting and modelling to improve the process, and then test again.

Results of the verification and validation should be submitted to – and scrutinised by – the ethical oversight mechanism that you have put in place before you make your decision to proceed. A decision to deploy the AI-based tool or system must be signed off by the force's head of information security in compliance with legal requirements, and by the chief constable.

Beta testing

The tests described above are what the **Data-driven technologies APP** refers to as 'alpha' testing. The APP also advises beta testing, where the tool is tested in an operational environment with a limited number of users to see if it functions as expected in the real world. If it is a tool that interacts with the public, such as a chatbot, this would involve inviting members of the public to try it and give feedback, ensuring you explain what it is and that it is in testing phase. If it is a tool that officers use in their work, you could select a group of users to try in on live data. However, as the APP stresses, you must ensure that the outputs do not have an impact on members of the public.

Prior to any testing in the real world, ensure that the tool or system is locked, otherwise it could continue to evolve in ways that you cannot predict and are difficult to test. You can continue developing it in the background and make further improvements based on the results on the beta testing, and then test the new iterations. Short cycles of iterative testing can be an effective means of understanding and refining how an AI model works. For each test cycle, make sure that the model is locked as soon as it makes contact with the operational environment.

Tags

Artificial intelligence