Impact evaluation

Ensuring artificial intelligence (AI) based tools or systems work as intended

First published 17 June 2025

6 mins read

If a decision has been made to roll out the AI-based tool or system further, it will need to be tested post-deployment, so that you know that it works in practice and can measure its impact.

The extent of testing will depend on the risks associated with it, as well as the extent and results of any previous testing. Make sure that you access the results of previous evaluation.

For higher-risk AI tools and systems, especially those with little or only internal testing behind them, you should consider using an independent third-party evaluation partner – for example, by partnering with your local university. If you are hoping that your force's product might be a candidate for national roll-out, it will almost certainly need independent evaluation.

Funding sources for evaluation

Sources of funding for independent evaluation include:

- the Home Office
- the Office of the Police Chief Scientific Adviser (OPCSA) via STAR and Test and Learn Finds
- the College of Policing What Works Board
- the Science and Innovation Co-ordination Committee

You can also pitch to your supplier that it is in their interest to fund an independent evaluation, as this will enhance its appeal to other forces that may want to procure it and perhaps lay the ground for national roll-out. It is worth getting in touch with either the College or the OPCSA for help in mapping available funding sources.

Designing the evaluation

Think of the benefits that you want to achieve through the AI (such as reduced time, reduced human error, improved victim experience, improved officer and staff confidence) and structure the

evaluation to capture them. You will also need to continue monitoring the risks identified in your scope and initial testing.

The most robust method to evaluate impact is a randomised control trial (RCT). An RCT randomly assigns participants into an experimental group or a control group, so that the only expected difference between the groups is the intervention of the tool or system. The **policing evaluation toolkit** sets out step-by-step how to prepare and implement your trial (including how to randomise) and how to analyse the results. The government's Evaluation Task Force has also produced **Guidance on the impact evaluation of AI interventions**, which contains case studies of evaluating different uses of AI in a public sector context.

Sometimes an RCT will not be feasible because of time or cost, or because the tool or system is part of a raft of other changes being made, rendering its specific impact difficult to isolate. If this is the case, you could try other methods for getting an indication of impact, such as comparing two unrandomised cohorts pre- and post-interventions ('difference in difference' testing). Advice on how to design and carry out trials can be found by contacting the College of Policing, who have set up a dedicated Evaluation Advice and Review Panel.

Additional challenges posed by AI

For the most part, the approach to carrying out an RCT should be the same for AI as for any other intervention, but there are some additional challenges and opportunities. What follows is an overview of some of these differences and how to navigate them.

Bias and differential impact

Even if bias was not present during the earlier lab-testing of the tool or system, it can be introduced at any stage, including during the roll-out. Make sure that you understand how different groups could be affected. Map the groups that could be differentially affected and ensure that the evaluation is designed to capture these impacts. This means that you will be able to analyse, for example, whether the tool appears to reduce, replicate or amplify race disproportionalities in how communities are being policed.

Remember that for public-facing AI, such as chatbots, differential impact can arise from accessibility issues as well as bias. Some groups, such as people with disabilities or elderly people,

can find it harder to interact with AI than others. User feedback surveys and interviews would be a means of capturing their experiences.

Public attitudes

Wherever the tool or system has a direct impact on the public, it would be a good idea to capture public attitudes – for example, through user surveys and focus groups. Public perception of AI can have a major impact on force reputation and the traction of the intervention.

Complex tools and deployment environments

If a tool is performing a particularly difficult or sensitive task and/or doing so in a fluid environment, you might want to supplement RCT with further investigation into drivers of specific outcomes you are seeing. For example, this could include where performance varies from one task to another, or from one location to another. You might need to piece together other evidence, such as user surveys and interviews, to figure out what might be going on. More information on how to evaluate in fast-changing deployment environments can be found in the Evaluation Task Force's **Guidance on the impact evaluation of AI interventions** (paragraph 2.4).

Process evaluation

As set out in **policing evaluation toolkit**, evaluate the process as well as the impact. How was it used? Is it easy to use properly or excessively complicated? This can be done through user interviews, focus groups and observations (in effect, following the user around). For an AI evaluation, you will be particularly interested in how meaningfully the human-in-the-loop exercises their role. Do they just accept the tool's predictions or outputs, or do they supplement these with their own judgement and expertise? If your intervention did not work as you had hoped, a process evaluation can shed light on whether this is the fault of the tool or system, or due to it not being used in the intended or optimal way.

Destination of time saved

If your tool or system is productivity-driven, try to track where hours saved through AI were deployed. You could ask users to document their hours and activities, or you could interview a selection of users about changes in activities and working patterns observed as a result of the intervention. Being able to demonstrate this is key to securing future investment in innovation.

Share your evaluation

Whatever the outcome, sharing evaluations helps policing to build on success and avoid repeating the same mistakes. Send the evaluation to the College of Policing, so it is centrally accessible. This applies both to evaluations conducted by independent experts and those conducted by your force. Sharing an internal evaluation for peer review will go a long way to strengthening the credibility of the AI-based tool or system.

Once you have completed validation, verification and evaluation, the chief constable then has to decide whether the risks flagged by the validation and verification processes are justified by the benefits evidenced in your evaluation.

Tags

Artificial intelligence