GUIDANCE ON THE IMPACT EVALUATION OF AI INTERVENTIONS

AUGUST 2024

WWW.FRONTIER-ECONOMICS.COM

Contents

Summary

This document provides guidance on the impact evaluation of Artificial Intelligence (AI) interventions. This guidance complements HM Treasury's guidance on evaluation in Central Government (the *Magenta Book*).[1] Consistent with the *Magenta Book*, impact evaluation is identified as the systematic assessment of the outcomes of an intervention with the aim of establishing whether, to what extent, how and why an intervention has resulted in its intended impacts.[2]

The key emerging best practice principles for designing proportionate impact evaluations of AI interventions are as follows.

1. Consider the evaluation as early as possible in the process of designing an AI intervention and define the overall objectives of the evaluation.

   i. Begin with the premise that evaluating the impact of an AI intervention can provide valuable insights and learning opportunities.

   ii. Tailor your evaluation to the specific context of the intervention, considering the level of risk involved and the potential for learning.

2. Develop a fully specified and comprehensive Theory of Change.

   i. Identify a comprehensive set of potential outputs, outcomes and impacts from the AI intervention. Be clear about the risks and potential unintended consequences, the mechanisms by which the intervention is expected to generate its results, the assumptions underlying this Theory of Change, and the current evidence for these assumptions.

   ii. Work closely with key stakeholders (such as the team developing the intervention and potential users) and draw on evidence from preliminary assurance exercises where possible.

3. When choosing the evaluation approach, evaluators should:

   i. first, explore options to implement an experimental method, ruling out those that are not feasible or appropriate

[1] https://www.gov.uk/government/publications/the-magenta-book

[2] Please note that this does not include guidance on assessing the capabilities of AI systems against technical benchmarks or assessing the technical components of AI systems, such as training data and model architectures.

   ii. be open to alternative approaches beyond Random Controlled Trials (RCTs), such as quasi-experimental designs

   iii. consider the use of theory-based approaches, especially in cases where the intervention is part of a complex system, the impacts of the intervention are hard to predict, and/or where it is important to understand how and for whom the intervention works

4. For all evaluation approaches, clearly describe the type of comparison group used.

   i. This is essential for being able to interpret the results of the evaluation and understand to what the intervention has been compared.

   ii. For example, suppose the evaluation is assessing the impact of the intervention compared to a 'business-as-usual' comparison group. In that case, the nature of this business-as-usual provision should be well understood.

5. Take into account the iterative process of developing and deploying AI interventions. Evaluation is critical at all stages.

   i. During the initial roll-out and large-scale testing, use rapid evaluation methods to assess immediate effects and make necessary adjustments.

   ii. After the full roll-out of the intervention, conduct regular evaluations to assess medium and long-term outcomes while also continuing to use rapid methods, if useful, to assess the immediate effects of changes to the intervention.

6. Given the iterative and evolving nature of many AI interventions, evaluation plans should be explicitly designed to be flexible and robust to changes in implementation.

   i. Design the evaluation plan to accommodate changes, ensuring it remains relevant and informative even if the intervention adapts over time.

   ii. Be clear about what can be learnt at each stage of evaluation and as a whole over the evaluation lifetime.

7. Give explicit consideration to variation in the impact of the intervention for different groups.

   i. Ensure that experimental and quasi-experimental approaches are designed to allow identifying different impacts for different groups.

   ii. Consider whether theory-based evaluation methods could help identify the drivers of differential impacts between groups.

8. Give explicit consideration to the role of public attitudes and perceptions.

i. Identify how public attitudes and perceptions could influence the impact of the intervention when scoping the evaluation.

ii. Ensure the evaluation methodology is appropriate for assessing the role of public attitudes and perceptions.

9. Think early on about how to establish a clearly defined baseline to support the evaluation, considering what data already exists and what may need collecting.

   i. Consider how best to gather baseline evidence on complex or subjective processes that are being replaced or enhanced by AI.

   ii. If the evaluation approach involves comparing the outcomes of the AI intervention against business-as-usual, document precisely what the business-as-usual consists of, before the AI intervention is implemented.

1 Introduction

Recent growth in the capabilities of Artificial Intelligence (AI) technologies has led to increased interest in the use of AI in Government. Evaluation of AI use in government (including process, impact and value for money questions) is necessary to understand the impact of AI systems compared to the status quo, improve current interventions, inform future policy development and ensure the Government is accountable to the public.

This guidance outlines key principles of best practice in carrying out a robust impact evaluation of programmes and initiatives utilising AI systems in central Government or the delivery of public services (AI interventions).[3]

For the purpose of this guidance, impact evaluation is defined as the systematic assessment of the outcomes of an intervention with the aim of establishing whether, to what extent, how and why an intervention has resulted in its intended impacts.[4] Therefore, this document does not include guidance on assessing the capabilities of AI systems against technical benchmarks or assessing the technical components of AI systems, such as training data and model architectures. These activities are crucial to generate evidence that AI systems are effective and safe. They can also provide useful information to evaluate the overall impact of using AI systems on the central government activities and public services where they are being deployed (the focus of this guidance). However, robust impact evaluation of AI interventions requires a distinct approach and broader sources of evidence in line with the best practice set out in the *Magenta Book*.

The key principles of robust impact evaluation are no different for AI interventions than for any other type of government programme. However, AI interventions can present additional opportunities and challenges for evaluation, which will be addressed in this guidance.

This guidance complements, and should be read in conjunction with, the *Magenta Book*. On its own, it is not designed to equip readers with all the skills required to develop an effective impact evaluation of an AI intervention. As with any evaluation, it is important to consult evaluation specialists and analysts in the relevant department as early as possible in the process of designing an AI intervention and planning its evaluation, and to set out appropriate resources for evaluation at the delivery planning stage.

[3] An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. (Source: OECD AI Principles). The definition of AI used in this guidance includes, but is not limited to, generative AI. Please note that this guidance does not cover testing of the capabilities of an AI system against technical benchmarks.

[4] This definition is consistent with the definition of impact evaluation provided in HM Treasury guidance on evaluation in government (the *Magenta Book*). Please note that this does not include guidance on best practice for using AI in the public sector or assuring the safety of AI tools or systems. For guidance on this, please see A guide to using artificial intelligence in the public sector (PDF, 3.7MB)

In the annexes to this guidance, four hypothetical case studies illustrate some key opportunities and challenges in evaluating AI interventions. These hypothetical case studies are:

1. Using an AI system to help check applications for grant funding.
2. Using a Large Language Model (LLM)-based application to help civil servants summarise and analyse a large number of documents.
3. Deploying a Chatbot interface on a government website.
4. Using an AI system to prioritise support for patients with a chronic disease

2 Opportunities and challenges for the evaluation of AI interventions

Due to the novelty of AI, few impact evaluations of these interventions have been carried out to date. Robust and rigorous impact evaluation can fill key evidence gaps, allowing us to learn whether and how AI interventions could be improved, to improve transparency and trust in the use of AI, and to demonstrate to Government and external stakeholders the impact and value for money of AI interventions.

As set out in the *Magenta Book*, criteria for 'priority' interventions that require more substantial evaluation include high-profile policies, interventions with high levels of uncertainty/risk (including possible negative consequences), and interventions with high learning potential.[5] Many AI interventions are likely to fit this profile due to their untested nature and unique risks and benefits. This means many AI interventions require more substantial evaluation than similarly sized business-as-usual interventions.

The digital nature of AI interventions provides opportunities for robust evaluation using experimental methods, such as Randomised Controlled Trials (RCTs). However, the fast-moving nature of AI interventions poses challenges to implementing these robust approaches in practice. These challenges can be overcome by building the evaluation into the intervention's design, aligning the evaluation's delivery with the delivery of the intervention and building flexibility into the approach.

The specific characteristics of AI interventions also mean that it is essential for evaluations to take into account the breadth and unpredictability of the intended and unintended outcomes of AI interventions, the potential variability in these outcomes for different groups, the role that public attitudes and perceptions play in shaping the impact of AI interventions, and the need to establish a clearly defined baseline or counterfactual against which the intervention is compared.

Figure 1 overleaf provides an overview of the key challenges and opportunities in evaluating the impact of AI interventions and their implications for evaluation.

[5] Uncertainty and learning potential will both be higher where the likely outcomes and mechanisms through which the intervention would generate those outcomes are not yet well understood. The learning potential may also depend on existing evidence from similar interventions (less existing evidence would imply greater learning potential from a new evaluation) and the extent to which evaluating the intervention would fill current evidence gaps; the applicability of the intervention to other policy areas, services and departments (broader applicability would imply greater learning potential); and the extent to which the impact of the intervention depends on the specific context in which it is applied (a more context-dependent impact may imply greater or smaller learning potential depending on existing evidence).

Figure 1 Summary of challenges and opportunities for impact evaluation of AI interventions

Opportunity / challenge		Implications for impact evaluation
Opportunities to control assignment of the intervention	▪ The digital nature of AI interventions means that it will often be feasible to control precisely who receives the intervention, when and how (more often than non-digital interventions)	Where appropriate, use experimental and quasi-experimental evaluation techniques, including RCTs, to evaluate AI interventions
Opportunities to understand impact through theory-based methods	▪ AI interventions can be implemented along with other changes, often in complex systems ▪ There is limited existing evidence on how and for whom outcomes of AI interventions are achieved ▪ There can be broad and unpredictable outcomes of AI interventions	Theory-based methods can be used to evaluate the impact of AI interventions that present these challenges.
Potential for broad and unpredictable outcomes	▪ As described in the box above, AI systems can have broad applications and many potential outcomes, intended and unintended	Prioritise developing a robust Theory of Change.
Fast-moving and evolving nature of AI interventions	▪ AI interventions may adapt and change more than other interventions due to: the use of iteration and learning by doing in their development; the self-learning capabilities of some AI systems, which may lead to their improvement or degradation over time	Think about evaluation as early as possible and align phases of evaluation with the development of the intervention. Conduct several stages of evaluation; and be transparent about what can be learnt from evaluation in any stage.
Variation in impact	▪ AI interventions may have different impacts on different groups for a variety of reasons (e.g. context sensitivity, training data sensitivity, misalignment risk, or inclusivity/accessibility)	Consider the possibility of varying impacts in defining the Theory of Change. Choose evaluation approaches that enabling assessing if and how impacts vary.
Public attitudes and perceptions	▪ In AI interventions, users (within government or external) interact with AI systems. Therefore, their behaviours and perceptions of the tool have a significant role to play in driving the overall impact of AI interventions	Consider the role of public attitudes and perceptions in the Theory of Change. Implement the chosen evaluation approach so that it is possible to assess the role of/impacts on public perceptions.
Establishing a baseline	▪ It may be challenging to identify and measure relevant baselines, particularly where AI enables entirely new activities, and replaces or supports complex or subjective activities.	Think about baselining early in the design and delivery of the intervention. Consider carefully how baseline information will be used in the evaluation.

This guidance elaborates on these opportunities and challenges, providing principles of best practice for:

10. choosing the evaluation approach (section 2.1)
11. developing the Theory of Change (section 2.2)
12. establishing a baseline (section 2.3)
13. the application of an evaluation approach to fast-moving and evolving AI interventions (section 2.4)
14. measuring whether impacts differ for different groups (section 2.5)
15. measuring public attitudes and perceptions (section 2.6)

2.1 Choosing the evaluation approach

Impact evaluation seeks to understand to what extent observed outcomes can be attributed to the intervention being evaluated. There are three main types of impact evaluation methods: experimental, quasi-experimental and theory-based. Evaluators should select the methods that can achieve the most robust impact evaluation possible while ensuring this is proportionate to the characteristics of the intervention being evaluated.[6] AI interventions offer specific opportunities to use experimental and quasi-experimental approaches, but there are cases where theory-based methods might be preferable or add value.

2.1.1 Opportunities to use experimental methods

Experimental methods, such as RCTs, randomly allocate the population into experimental groups, for example, a treatment group (exposed to the intervention) and a control group (that is not). Random assignment means that, on average, the groups are expected to have the same characteristics and differ only in whether or not they were exposed to the intervention. This means that any differences in group outcomes can confidently be attributed to the intervention.

AI interventions are delivered digitally, which gives evaluators more control over who can use the AI system and when. This can offer considerable opportunities to use experimental

[6] As described earlier in this section, AI interventions are likely to require substantial evaluation due to the limited evidence on their impact available to date and their unique risks and benefits. However, the scale, potential for learning and risks and benefits will vary between AI interventions, with some warranting more substantial evaluations than others. If an impact evaluation of the AI system being used has been conducted before, it is important to conduct further evaluation wherever the tool/intervention is applied to a new context or changed significantly. Evaluation colleagues can support in identifying proportionate approaches.

methods relative to non-digital interventions (for example, building new infrastructure).[7] Therefore, evaluators should strongly consider the use of these methods when evaluating AI interventions. When appropriately designed, experimental methods offer an extremely robust way to quantify the impacts of an intervention.

Options to experimentally evaluate the impact of AI interventions could include:

- randomising who has access to the AI system upon roll-out by assigning potential users to a treatment group (that has access to the AI system) or a control group (that does not have access)
- randomising the timing of access to the AI system by assigning potential users to a treatment group that has access to the AI system at an early stage (e.g. as part of a pilot) or a control group that only gains access at a later time
- randomising encouragement to use the AI system by granting all potential users access to the AI system, but randomly selecting potential users into a treatment group that either receives encouragement, information, and/or training about using the AI system or a control group that does not receive it

Evaluators should consider how the chosen randomisation strategy can practically be integrated into the intervention delivery plan. For example, if plans for delivering the intervention include a gradual roll-out, randomising the access timing might fit best within the overall intervention design. This also includes deciding the level of randomisation (e.g. at an individual or site level). As with any research design (especially those that involve withholding the intervention from specific groups), it is vital to consider ethics. Evaluation teams are encouraged to consult with an appropriate ethics review committee during the design stage to ensure that plans receive appropriate ethical scrutiny.[8] Other conditions determine the feasibility of an experimental approach: whether the target population receiving the intervention is well defined, whether the outcomes of interest are well defined, how stable the intervention is over time, and what the likely sample sizes are.[9]

Best practice guidance on delivering RCTs includes guidance provided by the Cabinet Office Behavioural Insights Team in 2012;[10] the *Emergency Medicine Journal* (Kendal, J. M., 2003);[11]

---

[7] Moreover, digital interventions offer a greater opportunity to randomise treatment at the individual level rather than at the level of groups (clusters) of individuals. Thereby reducing the sample sizes required to achieve a certain level of statistical power.

[8] For further guidance, please see Government Social Research Professional Guidance - Ethical Assurance for Social and Behavioural Research in Government (PDF, 720KB)

[9] Table 1 in Annex A.1 provides a practical example of assessing the usefulness of an RCT.

[10] Cabinet Office. (2012). Test, learn, adapt: Developing public policy with randomised controlled trials (PDF, 3.0MB)

[11] Kendall, J. M. (2003). Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal*, *20*(2), 164–168.

and a best practice RCT methodology checklist published by the National Institute for Health and Care Excellence (NICE, 2012).[12]

Annexes A.1 and A.2 of this guidance provide hypothetical case study examples where potential users of an AI system are allocated randomly to treatment and control groups. In Annex A.1, this is done upon the full roll-out of the AI system, and the case study describes the considerations made in assessing the feasibility of an RCT, including consideration of required sample sizes. Annex A.2 describes an example of implementing an RCT at the pilot stage of a fast-moving and iterative intervention.

2.1.2 Using quasi-experimental methods

Quasi-experimental methods use statistical techniques to identify a comparison group similar to the treatment group but unaffected by the intervention. Generally, the two groups will differ in known ways that can be accounted for analytically. These methods are useful where random assignment of the intervention to treatment and control groups is not possible, for example, in cases where a tool has already been rolled out without integrating an experimental evaluation approach into the delivery plan.

Quasi-experimental methods include, among others:

- statistical matching, which compares the outcomes of the treatment group to those of a control group that is similar to the treatment group in terms of one or more 'matching variables'.[13]
- regression discontinuity design (RDD), which estimates the impact of an intervention by using a cut-off threshold to assign the intervention.
- difference-in-differences (DiD), which assesses how the evolution of the outcome of interest differs over time between a group that received the intervention and a group that did not.
- synthetic control methods, which use historical data to construct a 'synthetic clone' of a group receiving a particular intervention.

Further description and examples of the application of these methods are available in Annex A of the *Magenta Book*.

Quasi-experimental methods might be feasible when an experimental approach is not feasible or appropriate. As in the case of experimental methods, the digital nature of AI interventions offers opportunities to control who has access to or is affected by an AI system in ways that enable robust evaluation. For example, the intervention can be designed so that a predefined

[12] National Institute for Health and Care Excellence, The guidelines manual Appendix C: Methodology checklist: randomised controlled trials.

[13] Matching variables may include, for example, measures of the distance between treated and potential controls, such as the Mahalanobis distance, or likelihood of participation in treatment, such as the propensity score.

cut-off is used to determine who uses the AI system or is affected by its use. This could enable the use of a Regression Discontinuity Design approach, where the outcomes for those just above the cut-off are compared to the outcomes for those just below.

Alternatively, the roll-out of the intervention can be staggered in a way that enables the use of a Difference-in-Differences approach. This would enable the outcomes of a group of people who have had access to or been affected by the use of the AI system to be compared to a group that has not, taking measurements before and after the intervention is introduced.

Whichever method or combination of methods is chosen, the type of comparison group used should be clearly described. This is essential for being able to interpret the results of the evaluation and understand what the intervention has been compared to. For example, if the evaluation is assessing the impact of the intervention compared to a 'business-as-usual' comparison group, then the nature of this business-as-usual provision should be well understood.

Moreover, the digital nature of AI interventions offers opportunities to collect rich data that can be used for evaluation purposes. AI systems are frequently integrated into data and digital service workflows, which accumulate or produce substantial amounts of data. This data can be used in evaluation across all methods but may be particularly important for quasi-experimental approaches. This is because quasi-experimental approaches require a sufficient quantity and quality of data to construct control groups and account for differences between the treatment and control groups. An example of this is described in Annex A.4.

2.1.3 Using theory-based methods

Theory-based evaluation methods use a well-defined Theory of Change and triangulate various evidence sources to rigorously assess how, why, for whom and in what context change occurred due to the intervention. A theory-based approach can be combined with experimental or quasi-experimental methods when the evaluator is interested in understanding why an intervention did or did not have an impact and how this may vary across contexts or user groups. This is often a key consideration for complex interventions or simple interventions in complex environments. Unlike experimental and quasi-experimental methods, theory-based methods do not aim to give a precise quantification of the impact of the intervention. Examples of theory-based approaches include:

- contribution analysis, which seeks to understand to what extent the intervention has contributed to the observed outcome, combining a range of evidence to test the Theory of Change.
- realist evaluation, which focuses on testing hypotheses about how the intervention may have led to a given outcome of a specific mechanism under specific circumstances.
- qualitative comparative analysis, which systematically analyses qualitative case study data to evidence the link between an outcome and combinations of factors or characteristics.

- outcome harvesting, which involves collecting evidence of change and then working backwards to assess what has contributed to the change.

Further description and examples of the application of these methods are available in Annex A of the *Magenta Book*.

Although AI interventions provide opportunities to tailor their roll-out and collect data in a way that enables the use of experimental and quasi-experimental methods, there will be cases where theory-based approaches are best suited to evaluate the impact of an AI intervention. These may include instances when:

- AI interventions are part of complex systems and/or complex changes
- there is uncertainty on the likely outcomes of AI interventions
- it is particularly important to understand 'how' and 'for whom' AI interventions achieve (or do not achieve) their outcomes

AI interventions that are part of complex systems and/or complex changes

In some AI interventions, AI is introduced in a complex system with many different components and interactions and/or alongside other concurrent changes. This could include broader digitalisation of a government service or an overhaul of an existing digital service.

In these cases, theory-based methods such as contribution analysis can provide a robust assessment by collecting data from multiple sources to gather evidence against the Theory of Change. This analysis triangulates across the different sources of evidence to understand the specific contribution of the AI intervention, considering other factors contributing to the change. Annex A.3 provides an example of using contribution analysis to evaluate the impact of an AI-enabled Chatbot.

When there is uncertainty on the likely outcomes of AI interventions

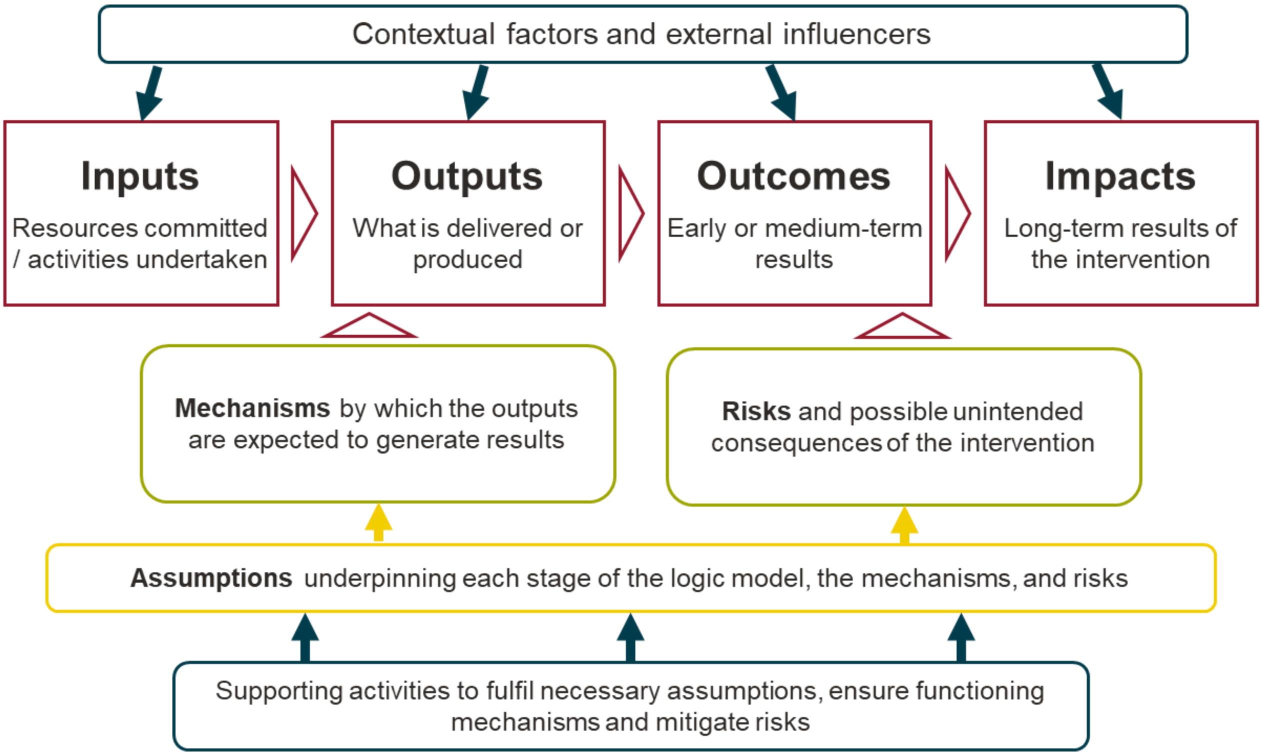
In some cases, defining the specific outcomes of an AI intervention can be challenging due to factors such as the novelty of AI, its broad applicability, and the role of public attitudes and perceptions in determining how people interact with AI (as discussed in section 2.6). When some or all of the potential outcomes of an AI intervention are not known in advance, theory-based approaches, such as outcome harvesting, can offer more flexibility than quantitative impact evaluation methods to detect changes in unexpected outcomes. Outcome harvesting involves collecting evidence of change throughout the intervention delivery period and then tracing back to understand the AI intervention's contribution.

When it is particularly important to understand 'how' and 'for whom' AI interventions achieve (or do not achieve) their outcomes

Theory-based methods can also be appropriate in developing an understanding of how and for whom the outcomes of an AI intervention have or have not been achieved. This can be especially useful when evaluators want to establish how the context has influenced the outcomes of the intervention. This helps assess to what extent the intervention can be successfully applied in a different context. Realist evaluation methods can help answer these questions. A realist approach sets out 'causal hypotheses' in the form of *Context* + *Mechanism* = *Outcome*. These hypotheses are then tested using evidence to determine how the mechanism operates. This approach can also be combined with outcome harvesting, which captures evidence of the impact throughout the intervention delivery period.

2.2 Developing the Theory of Change

With any evaluation, a key first step is developing a Theory of Change for the intervention. The diagram in Figure 2 overleaf gives a stylised example of a Theory of Change, including the key components that evaluators should consider.



A Theory of Change should be developed for an AI intervention in the same way as for any other type of intervention. However, developing a Theory of Change for an AI intervention can pose some additional challenges.

Firstly, with AI systems, there are technical challenges in understanding how and why they make certain decisions (sometimes referred to as the 'black-box' nature of AI systems). In particular, AI systems may learn to pursue objectives in a way associated with undesirable or unintended outcomes (sometimes referred to as the 'AI alignment' problem). For example, AI systems are often trained on data that accurately reflects existing biases in our society. Consequently, they may make predictions or recommendations that are discriminatory towards certain groups. These factors mean it may be more challenging to identify all risks and potential unintended effects when developing the Theory of Change.

Secondly, public attitudes and perceptions play an important role in AI interventions, and it can be difficult to predict how users will interact with AI systems. This creates an additional challenge for evaluators in accurately describing the mechanism by which the intervention is expected to generate results.

Lastly, given the relative novelty of AI interventions, there may be limited existing evidence for many of the key assumptions underlying the Theory of Change. This may include limited evidence on the risks and potential unintended effects of the intervention as well as limited evidence on the mechanisms underlying the Theory of Change.

To conduct robust evaluations in this context, it is crucial to:

- develop a fully specified and comprehensive Theory of Change
- work closely with key stakeholders (such as the team developing the intervention) to draw on evidence from preliminary assurance exercises, where possible

It is also important to consider whether a process and/or theory-based evaluation approach can be integrated into the design to help explore and understand unintended outcomes that may arise from the black-box nature of AI systems.

Develop a fully specified and comprehensive Theory of Change

A simplified Theory of Change focusing only on the intended outputs, outcomes and impacts of the intervention will likely miss potential unintended impacts. It will not adequately consider the evidence for key assumptions underlying the Theory of Change. In mapping out the Theory of Change, evaluators should seek to develop a full understanding of:

- how the intervention is expected to work in practice, including the problem it seeks to address, the intended outcomes and impacts, and the groups expected to be impacted
- the mechanisms by which the impacts of the intervention are expected to be realised, including the main actors and the conditions required for the intervention to succeed
- the assumptions underlying how the intervention is expected to work and the strength of evidence for these assumptions
- the risks associated with the intervention and potential unintended impacts that could result
- the wider context, such as other policy changes or changes in economic, social and environmental factors, as well as supporting activities that may help realise the intended impacts of the intervention

The digital nature of AI interventions offers opportunities to build an automated collection of monitoring data to help identify early signs of whether any unintended consequences are materialising.

Where potential unintended negative consequences are identified, evaluators should consider how these can best be measured and assessed. Where there are key assumptions in the Theory of Change that have limited evidence, evaluators should seek to evidence these assumptions as part of the impact evaluation.

When developing the Theory of Change, evaluators should give particular attention to two factors:

1. Whether the intervention may have different impacts on different groups of people.
2. How public attitudes and perceptions may affect the impact of the intervention.

These issues are discussed further in sections 2.5 and 2.6, respectively.

Work closely with key stakeholders and draw on evidence from preliminary assurance exercises

In developing the Theory of Change, evaluators should also work closely with key stakeholders, such as the team developing the AI tool, the team designing the intervention and the end users of the tool. These stakeholders can provide important insights into how the intervention is expected to work, potential unintended effects and the wider context of the intervention. It may also be helpful to consult government evaluation experts who can provide insights into available relevant evidence and evidence gaps for the assumptions underlying the Theory of Change. Consulting with these stakeholders could occur through workshops, potentially including 'pre-mortem' sessions. In these sessions, participants are asked to assume the intervention has failed or gone wrong and propose plausible reasons for its failure, which can help identify potential risks or unintended impacts.

In some cases, preliminary evidence that could be used in developing the Theory of Change may be available from assurance activities conducted to assess the functioning and safety of the AI tool. For example, the AI tool being used in the intervention may have undergone a 'red teaming' exercise where a select group of expert users test a wide variety of inputs in an attempt to expose critical weaknesses, deficiencies or biases in the tool. Such exercises may provide preliminary evidence on the intervention's likely outputs while identifying potential risks or unintended impacts.

2.3 Establishing a baseline

With any evaluation approach, it is important to establish a baseline: this means collecting information about the situation before the intervention roll-out has started.[14] For example, this includes average outcomes before the intervention and other information about the characteristics of participant groups. Baselines play a critical role in impact evaluation approaches.

In addition to helping evaluators understand the context in which the intervention was introduced, other uses of baselines in impact evaluation include:

- providing a basis for impact estimate (for example, in quasi-experimental approaches that rely on pre-post measures, such as differences-in-differences)
- providing covariate data to help improve the precision of RCT impact estimates (thereby lowering sample size requirements)
- providing the data needed to adjust for pre-existing differences between groups in quasi-experimental methods, such as statistical matching

AI interventions may involve additional challenges around establishing a baseline. Firstly, it may be challenging to identify, define and measure baseline information for several reasons:

- AI systems may enable entirely new activities, which could make it challenging to identify what baseline information, if any, is relevant
- AI systems may replace and support complex or subjective activities, which could make it challenging to define and collect baseline information
- the evaluation of AI interventions typically needs to consider a wide range of factors, such as public attitudes and perceptions, variation in outcomes, and impacts on efficiency, accuracy and quality, all of which would ideally have an established baseline against which to assess the impact of the intervention, which means that a lot of data is required for a comprehensive baseline.

Moreover, AI interventions are often developed iteratively and can change rapidly, creating challenges for evaluators in establishing a relevant baseline quickly or before the roll-out of the intervention begins.

Taking these challenges into account, evaluators should:

- consider carefully how baseline information will be used in the evaluation and decide what information should realistically be collected

[14] The exact timing of baseline data collection will depend on the process used to roll out the intervention. As a general principle, baseline data should be collected at a point where the outcomes of interest are not yet affected by the intervention. It is important to recognise that baseline outcomes may sometimes be affected by the intervention before it is actually rolled out in anticipation of the intervention taking place.

- think about how to establish a suitable baseline early on when scoping and designing evaluations, exploring what data may already exist and what factors would be a priority for collecting additional data

  – Consider how baseline data will be used in the evaluation

AI systems can be used to perform entirely new activities and replace or support complex and/or subjective activities. This can make it challenging to identify, define and measure some of the potentially relevant baseline information. Where this is the case, evaluators should:

- where possible, identify changes to existing monitoring systems for the business-as-usual processes that the AI intervention will update or replace so that relevant baseline data is collected[15]
- assess what baseline data can be collected retrospectively
- if the collection of any baseline outcomes is particularly resource intensive, determine the level of priority for this, taking into account the overall evaluation approach

Think about how to establish a baseline early on when scoping and designing evaluations

Due to the fast-moving nature of many AI interventions, baseline information collection should be considered as early as possible in the evaluation design. Informed by the development of a comprehensive Theory of Change, evaluators should identify the key evaluation questions and, therefore, key factors for which establishing a baseline is a priority.

Evaluators should work with relevant stakeholders to understand what data is already being collected that could be used to establish a baseline. This could include general management information or relevant surveys from when the intervention was being scoped. Where existing evidence is not available to establish a baseline, evaluators should consider whether primary evidence could be gathered within a suitable time frame. In doing so, evaluators may need to be pragmatic and prioritise capturing baseline data on the factors expected to be most important for robust evaluation of that particular intervention.

2.4 Embedding evaluation in evolving and fast-moving interventions

Evaluation can be particularly impactful in the context of fast-moving and evolving interventions due to the greater potential for learning to be actioned and to feed into the iterative development of the intervention. However, the fast-moving and evolving nature of AI interventions presents a key challenge to robust impact evaluation for the following reasons.

[15] Be aware that this may differ across sites if the intervention is being delivered in more than one context.

Firstly, due to their digital nature, AI interventions can generally be deployed quickly (assuming appropriate infrastructure is already in place). This can pose a challenge for evaluation because it reduces the window of opportunity for designing the evaluation before an AI intervention is rolled out.

AI interventions are also likely to evolve during their implementation as a result of iteration and learning-by-doing. This creates further challenges for evaluation as it potentially means that the intervention is changing while being evaluated, making it particularly challenging to attribute changes in long-term outcomes to the intervention. This is typical with many digital interventions but is particularly likely to be the case for AI interventions because:

- AI interventions involve relatively new and untested technologies
- some AI systems themselves are capable of 'learning' and developing over time, which might either increase their effectiveness in achieving the intended outcomes, or lead to a divergence from the intended outcomes
- AI technologies are developing rapidly and, as such, an AI intervention may change over time to deploy superior technical options that were not available when the intervention was originally conceived

To conduct robust evaluations in this context, it is crucial to:

- think about evaluation as early as possible and embed evaluation thinking within the design of the intervention, building flexibility from the outset to recognise that the intervention may evolve
- align the phases of evaluation as closely as possible with the phases of designing and deploying an AI intervention
- conduct regular iterative evaluations
- be transparent about what can be learnt from evaluation at any stage

Think about evaluation early and embed evaluation thinking within the design

Recognising that AI interventions may evolve over time, evaluators and their evaluation approaches need to be flexible. They should focus more on small-scale testing, process evaluations and evaluation methods that can rapidly assess incremental iterations of the intervention.[16] In practice, this can be achieved through early application of evaluation expertise in the intervention design and delivery.

Often, early iterations of an AI system will first go through initial testing and assurance exercises to assess the quality of its outputs and the potential biases and risks (which continue into later stages of roll-out). Subsequently, the impact evaluation team can work with the

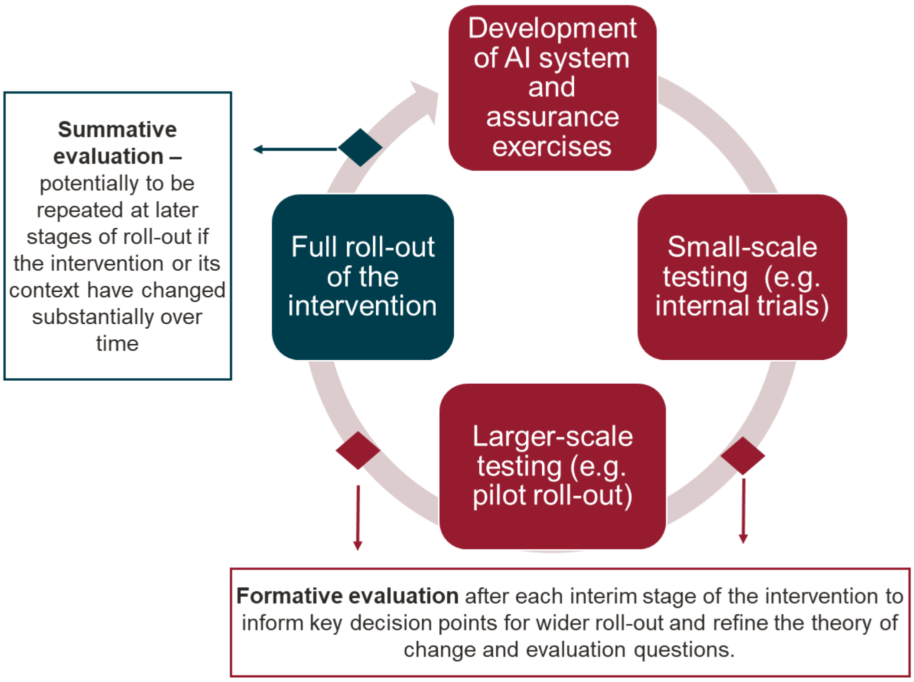
[16] This includes specific methods designed for rapid evaluation, discussed further below, in addition to the rapid implementation of any of the evaluation methods discussed earlier in this guidance. For example, in some cases, RCTs can be completed within a few months.

development team to map out an initial Theory of Change and use this to inform the design of controlled internal testing with a small group of users. The findings of this testing may then be used to improve the AI system or the planned design of the intervention, refine the Theory of Change, and/or make key decisions on whether to progress the roll-out further. This is referred to as ‘formative evaluation’.

Evaluate iteratively, in alignment with the delivery and evolution of the AI intervention

AI interventions typically follow a cyclical process that includes development, deployment and continuous evolution, as illustrated in Figure 3 below. The process begins with either the development of a new AI system or the adoption of an existing one (sometimes referred to as the ‘Alpha phase’). This is often accompanied by the creation of a service prototype in which the AI is embedded (sometimes referred to as the ‘Beta phase’). The approach to development may vary, with some projects employing agile methodologies to allow for rapid iteration and responsiveness to user feedback.

Figure 3 Example of AI intervention delivery lifecycle



Once the AI system/service prototype is ready, it is rolled out incrementally to a small group of users to gather initial insights and then to a broader audience. This phased approach helps identify potential issues and make necessary adjustments before full-scale implementation.

Evaluation plays a critical role at each stage of development and delivery.

- **During small-scale and larger-scale testing**

Especially during initial roll-out and testing at a larger scale, when the intervention is likely to evolve most quickly, evaluators may make use of rapid evaluation methods to collate and synthesise evidence, being realistic about what is feasible and proportionate to evaluate at each stage. Rapid evaluation methods typically include concurrent workstreams whereby data collection and analysis occur in tandem, the use of multiple methods and multidisciplinary teams, and close involvement of stakeholders in design, data collection, analysis, and reporting.[17] Such methods have been used effectively in evaluations of COVID-19 fast response mechanisms.[18] At these stages, rapid evaluations can be used to assess the immediate effects of using the AI system to help determine whether the intervention should move to the next stage (testing at scale and full roll-out), what changes, if any, should be made prior to the next stage, and refine the Theory of Change and the evaluation questions to be answered in future stages. Annex A.2 provides an example of evaluating an AI intervention during an initial roll-out.

- **During and after the full roll-out**

A comprehensive evaluation of the overall impact of the intervention, including medium and long-term outcomes, should typically follow the initial testing phases. However, in many cases, AI interventions will continue evolving after the full roll-out. This could result from explicit cycles of development and testing and/or due to the AI system having self-learning capabilities. Therefore, evaluation after full roll-out should also include:

- continuing to conduct rapid evaluations to assess the immediate impact of changes to the intervention as it continues to develop over time
- repeating comprehensive evaluations at regular intervals as proportionate to the evolution of the intervention and its context and the characteristics of the intervention (as discussed at the beginning of section 2, the *Magenta Book* provides guidance on assessing the proportionality of an evaluation).

Be transparent about what can be learnt from evaluation at any stage

When learning from early and interim/formative evaluations, it is important to ask the following questions:

1. To what extent will the findings hold when the AI intervention is scaled up, for example, when moving from an initial pilot to roll-out at scale?

[17] These methods include, for example, action research, adaptive evaluations and A/B testing. For more on rapid evaluation methods, please see https://www.betterevaluation.org/methods-approaches/approaches/rapid-evaluation.

[18] For example, Gawaya, M., Terrill, D., & Williams, E. (2022). Using rapid evaluation methods to assess service delivery changes: Lessons learned for evaluation practice during the COVID-19 pandemic. *Evaluation Journal of Australasia*, *22*(1), 30–48.

2. To what extent will the findings hold over time, given the potential for AI systems to evolve and for people interacting with the systems to change their behaviour over time?
3. To what extent are the findings externally valid, that is: to what extent will they hold for similar interventions, or when the AI intervention is applied in different contexts?

AI interventions present some specific challenges to the scalability and external validity of evaluation findings. These include:

- the technical challenges involved in understanding why an AI system makes certain decisions means it can be very challenging to assess whether certain impacts are likely to be realised in other settings or at scale
- the rapid change of AI technology means that the estimated effects of using AI may change substantially over time as the technology evolves

Evaluators should, therefore, be transparent about any limitations in understanding how the findings of the current evaluation might translate to new policies or interventions when sharing or reporting the evaluation results. It is also important to specify which AI model was used in the intervention and how it was used.

2.5 Measuring whether impacts differ for different groups

Another challenge for the evaluation of AI interventions is that the impact of an AI system may vary substantially between different tasks, different contexts and for different groups. For example, emerging evidence indicates this is currently the case for LLMs.[19]

While all interventions may have different impacts on different groups, with AI, this variation may be particularly pronounced and difficult to anticipate. With AI interventions, differences in impact can arise due to one or all of the following factors:

- bias or poor representation in the training data
- misalignment (i.e. where the AI tool pursues the intended objective in a way that generates unintended impacts)
- inclusivity and accessibility issues for users of the AI tool (e.g. the tool may be difficult to use for individuals with some disabilities)

This being said, it is also possible that AI systems could act to reduce variation in outcomes between different groups. This is because, in some cases, an appropriately trained AI tool may be less biased (or biased in different ways) than a human performing the same task.[20]

[19] Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper, (24-013).

[20] For example, Kleinberg et al. (2018) show how machine learning could be used to improve criminal sentencing in the United States, including reducing the proportion of black and Hispanic defendants who are incarcerated. Source: Kleinberg, J.,

The possibility of different impacts occurring for different groups should be identified as part of the Theory of Change development, drawing on input from the team designing the intervention and evidence from assurance exercises, as described above. Evaluators should also explicitly consider whether variation in impact can be adequately assessed by the proposed methodology. In particular, evaluators should:

- ensure that experimental and quasi-experimental approaches are designed in a way that can identify different impacts for different groups
- consider supplementing experimental and quasi-experimental approaches with theory-based evaluation methods to help understand how and why impacts vary between groups

Ensure that experimental and quasi-experimental approaches are designed in a way that can identify different impacts for different groups

When designed and implemented appropriately, experimental and quasi-experimental methods can be used to robustly assess the average impact of an intervention for an entire group and specific sub-groups.

To do this, evaluators should ensure that sample sizes are sufficient for key sub-groups so that differential impacts can be reliably estimated. The necessary sample sizes will depend partly on the size of the expected impact and the degree of statistical confidence required. Smaller expected impacts will generally require larger samples to identify differential impacts to the same degree of statistical confidence.

Ensuring adequate samples for key sub-groups may be more challenging in the early stages of an intervention's roll-out, when the total number of people exposed to the intervention may be relatively small. As such, it is important that evaluators consider these issues as early as possible during evaluation design.

Where important sub-groups have a relatively low representation in the target population, evaluators should consider over-sampling these sub-groups to ensure adequate sample sizes.[21] Where this is done, evaluators will subsequently need to account for this over-sampling when re-weighting their results and estimating the overall impact of the intervention.

Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237–293.

[21] In situations where evaluation resources are limited, decisions over whether to over-sample from particular groups will need to be taken with particular care. It is important to consider whether any over-sampling from one group would require reducing the sample size from any other group in order to maintain the same level of cost, and what implications this might have.

Consider whether theory-based methods could help identify the drivers of different impacts for different groups

For interventions where differential impacts are a particular concern, it may be useful to complement experimental and quasi-experimental methods with a theory-based approach. For example, realist evaluation methods could be used to help identify the context and mechanisms that are generating different outcomes for different groups. This would involve gathering primary evidence on whether, why and how users from different sub-groups are engaging with the AI tool and what contextual factors are influencing this. Relevant contextual factors could include skills and training, as well as attitudes and perceptions around AI (as discussed further in Section 2.6).

Alternatively, process tracing could be used to test whether the mechanisms underlying the Theory of Change are working as expected. This would involve identifying the outputs and outcomes that should be observed if the Theory of Change were true and seeking evidence of whether these have occurred. In the case of generative AI tools, this could involve working with the team developing the AI tool to sample the outputs being generated by the AI tool over the course of the intervention, for example, a sample of the chat logs generated by an AI Chatbot. By doing so, it may be possible to identify examples or patterns of bias and representational harms that suggest the Theory of Change is not functioning as expected or intended.

Annex A.4 describes a hypothetical case study of an AI intervention where the primary evaluation questions include investigating how the impact of the AI intervention varies across different groups.

2.6 Measuring public attitudes and perceptions

Public attitudes and perceptions are likely to play a key role in most AI interventions and create additional challenges for evaluators. For example, individuals may have strong feelings about the use of AI in certain applications and certain contexts.[22] It can also be difficult to predict how users will interact with AI systems due to:

- their relative novelty and, therefore, the lack of evidence on how users typically interact with different types of AI systems
- their technical complexity and differences in the level of public understanding of how they work.

[22] See, for example, research by the Department for Transport into public attitudes towards AI for consultation and correspondence in different settings: Using AI in consultations and correspondence: Thinks Insight & Strategy research report (PDF, 979KB)

Given these difficulties and given that public attitudes and behaviours can have significant implications for the impact of AI interventions, it is important for evaluators to:

- identify how public attitudes and perceptions could influence the impact of the intervention when scoping the evaluation
- ensure the evaluation methodology is appropriate for assessing the role of public attitudes and perceptions

– Identify how public attitudes and perceptions could influence the impact of the intervention when scoping the evaluation

In scoping the evaluation and developing the Theory of Change, specific attention should be paid to whether and how public attitudes and perceptions could influence the impact of the intervention. For example, could it be the case that users become overly reliant on the AI tool or overly trusting of its outputs, even where attempts are made to present the limited accuracy of the tool to users? Conversely, could it be that users underestimate the accuracy of the tool and, therefore, do not use it, leading to limited impact?

Evaluators should also consider whether the intervention itself could impact public attitudes and perceptions around AI. Successful AI interventions could have positive impacts on public attitudes that may lead to greater public engagement with future AI interventions.

These issues should be considered as part of developing a fully specified Theory of Change, as discussed above. Consulting stakeholders such as the team developing the AI intervention, end users of the AI and evaluation specialists within the government may be helpful in identifying where and how public attitudes are most likely to play a role.

Ensure the evaluation methodology is appropriate for assessing the role of public attitudes and perceptions

Where public attitudes, perceptions and behaviours are expected to be important for the impact of a particular AI intervention, evaluators should consider what additional evidence they will need to gather to understand these factors.

Where experimental methods are used, evaluators could build user surveys into the RCT design to collect data on attitudes and perceptions towards AI tools. This data could be compared against the results of the RCT to see how differences in attitudes correlate with the outcomes observed. This could even include pre- and post-surveys to see if attitudes changed for those exposed to the intervention.

Evaluators could also gather qualitative data from focus groups and interviews with users. These consultations could explore a structured set of questions around attitudes towards AI in different contexts (including the context of the intervention at hand), perceptions of the accuracy and quality of the AI tools outputs, and why users did or did not engage with the intervention.

Given the complexity of these issues, supplementing experimental or quasi-experimental approaches with theory-based approaches that triangulate a range of evidence sources may be particularly effective. For example, realist evaluation could be used to explore the contextual factors that may influence attitudes and perceptions, as well as the mechanisms by which these influence the outcomes. Alternatively, contribution tracing could be used to triangulate quantitative and qualitative evidence and understand whether specific aspects of the Theory of Change around user behaviour are likely to be true.

Annex A.3 describes an example evaluation where theory-based approaches are used to evaluate an AI intervention.

3 Conclusions

This document has provided an overview of emergent principles of best practice for evaluating the impact of using AI systems in central Government and public services, in line with HM Treasury's *Magenta Book*.

In the early stages of the design of an AI intervention, it is important to consider impact evaluation as early as possible, plan evaluation phases and methods with the iterative and evolving nature of AI interventions, and develop a comprehensive Theory of Change by working closely with key stakeholders and drawing on evidence from preliminary assurance exercises where possible.

When selecting evaluation approaches, consider experimental methods for quantitative impacts, exploring feasible options and alternative designs, like quasi-experimental methods. Theory-based approaches are useful for complex systems and complement experimental or quasi-experimental methods. AI interventions require iterative evaluation throughout the development, deployment and evolution stages, using rapid methods for immediate effects and comprehensive evaluations for long-term outcomes. Evaluation plans should be flexible to accommodate changes and consider variations in impact across different groups. Public attitudes and perceptions should be factored into the evaluation, and establishing a clear baseline early on is crucial for effective assessment.

This is a fast-moving and exciting area with numerous opportunities to learn from the evaluation of AI interventions. This guidance describes some of the key challenges and opportunities for evaluating AI interventions, providing advice on how to best address them. However, this guidance on its own is not intended to equip readers with all the skills required to develop an effective impact evaluation. Therefore, for those involved in designing or delivering an AI intervention, it is crucial to consult with evaluation experts in the relevant departments to ensure that the impact of the intervention is evaluated robustly.

Annex A: Hypothetical evaluation case studies

Practical example #1 – an AI system to help assess applications for grant funding

This case study describes key aspects of the evaluation of a hypothetical AI intervention.

The AI intervention

A government department is about to introduce a grant funding programme to support a range of investments that individuals can make to reduce their greenhouse gas emissions. The grant applications need to be accompanied by photographic evidence.

Applications are assessed by grant officers, who carry out a number of checks to determine whether an application is eligible to receive funding. The department has decided to deploy an AI system that flags potentially fraudulent cases based on patterns of multiple grant submissions and/or detecting photos that were not identical but showed the same evidence. The AI system provides possible cases of fraud to the grant officers. It is up to grant officers to conduct further review of flagged applications and decide whether to accept or reject them.

The AI system has been developed and tested for accuracy, and it is about to be rolled out. An evaluation team is tasked with designing and carrying out an impact evaluation of this AI intervention.

Key challenges and opportunities for evaluation

The evaluation team identifies two key opportunities and challenges for the scoping and delivery of this evaluation. Firstly, the evaluation scoping is taking place prior to the roll-out of the intervention, which means that there is scope to randomise access, timing, or encouragement to use the AI system to evaluate its impact.

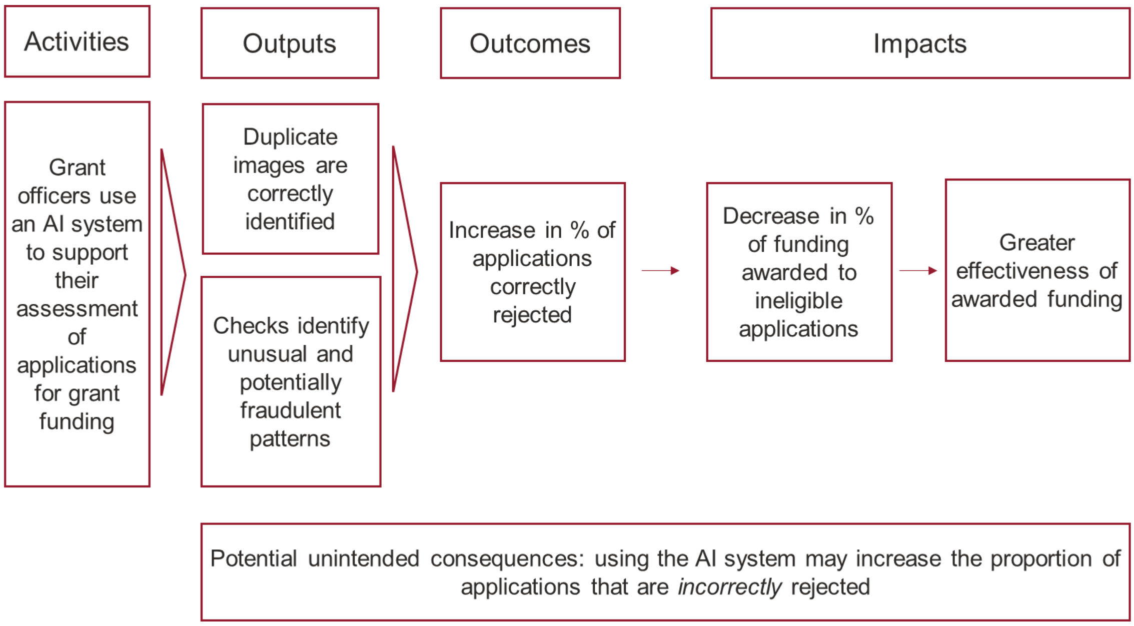
Secondly, there are likely to be challenges in measuring some of the relevant outcomes. Ideally, the evaluation team would assess whether the proportion of correctly rejected applications differs between the treatment and control groups. However, while observing how many applications were rejected is straightforward, assessing whether the rejection was the correct decision is more challenging. This means that the primary outcome of this trial will need to be inferred using data on appeals to rejected applications rather than directly observed.

The challenge with measuring outcomes for this study also creates difficulties in establishing a baseline for the rate of fraudulent applications being accepted before this intervention began. This lack of clear information on the baseline situation makes it harder to understand the scale of the problem this intervention is intended to solve, which has implications for how the valuation results will be contextualised.

Scoping the evaluation

The evaluation team identifies the likely intended outcomes of this AI intervention and the potential unintended outcomes. These are represented in the high-level Theory of Change below (Figure 4).

Figure 4 High-level Theory of Change



Note: This is a simplified Theory of Change that focuses on specific outcomes in order to illustrate the chosen evaluation approach.

Note that along with the intended effects of the AI system, the Theory of Change also identifies potential unintended consequences. Although grant officers are required to perform additional checks on applications flagged by the AI system, there is still a risk that using the system may lead to more applications being incorrectly rejected. This could happen, for example, if officers are over-reliant on the AI-generated flags relative to their own assessment and if the AI system is not sufficiently accurate in flagging suspicious patterns or duplicate images.

Based on conversations with key stakeholders, the evaluation team determines that the primary evaluation question is:

1. Has using the AI system increased the proportion of correctly rejected applications?

The secondary evaluation question is:

2. Has any increase in correctly rejected applications been achieved without also increasing the proportion of incorrectly rejected applications?

The evaluation approach

The evaluation team considers possible options to evaluate the impact of the intervention, including experimental, quasi-experimental and theory-based approaches. The digital nature of this AI intervention means it is possible to control precisely which grant officers can use the AI system, creating an opportunity to use an experimental approach. The evaluation team determines that a Randomised Controlled Trial (RCT) is feasible and appropriate (see table below for further detail).[23]

The main steps to implement an RCT are:

1. plan and implement the creation of a treatment and a control group[24]
2. plan and implement data collection
3. plan and implement analysis of the data collected

Table 1 Assessment of the usefulness of RCT for fraud prevention case study

Conditions for RCT	Assessment
Is the target population receiving the intervention well-defined?	Yes – the immediate target population are the grant officers screening grant funding applications with applicants being the ultimate recipients.
Are the outcomes of interest well-defined?	Yes – the key intended outcome is a decrease in the proportion of ineligible applications that receive funding.
Is it practically feasible to randomly assign the 'treatment'?	Yes – there are a number of options, discussed below.
Is the intervention stable over time?	Yes – the eligibility criteria and the process of evaluating applications are not expected to change substantially over time.

[23] For further guidance on when using an RCT would be appropriate, please see section 3.5 of the *Magenta Book*.

[24] Or, in more complex designs, multiple treatment and control groups. This includes taking into account ethical considerations and seeking informed consent from individuals to participate in the RCT.

Conditions for RCT	Assessment
Are the sample sizes likely to be sufficiently large?	Yes – the evaluation approach described in this case study assumes that the proportion of ineligible applications is likely to be large enough that it is realistic to detect differences between the treatment and the control group (see sections below).
Are there ethical reasons why assigning the intervention randomly may not be appropriate?	None identified.

Implementing the evaluation approach

Creating treatment and control groups

The evaluation team works with colleagues who are designing and deploying the AI intervention to plan the RCT. Grant officers are assigned to a 'treatment' group (who will use the AI system to screen the applications for funding) or a 'control' group (who will not be able to use the AI system).

The number of individuals in the treatment and control groups (the sample size for the RCT) should be large enough to give the evaluator confidence that the sample size has a reasonably high chance of detecting the true effect of the intervention. [25]

For the sake of illustration, the rest of this example assumes that a large number of grant officers are working on the funding applications. If the number of officers was relatively small, it could still be possible to undertake an RCT, as described in the box below.

Other options to randomise the AI intervention

In this example evaluation, there would be two alternatives to the randomisation strategy described above. These alternatives could allow running an RCT even if the number of grant officers assessing the funding applications is relatively small.

[25] The required sample size depends on a number of factors including the expected size of the effect being evaluated. Existing resources on determining required sample sizes and conducting RCTs include guidance developed by the Behavioural Insights Team and guidance from the Office for Health Improvement and Disparities on evaluating digital products.

The first alternative would be to randomise when officers have access to the AI system. For example, imagine it will take 12 grant officers four weeks (20 working days) to assess the applications. Grant officers could be randomly divided into equal groups using or not using the AI system every day. At the end of the period, data for the treatment and the control groups would each include information on decisions made in 60 officer-day pairs (six officers, ten working days).[26]

The second alternative would be to use each application as the unit of randomisation. In this case, grant officers could be shown the results of the AI system's checks only for some of the applications they are assessing (the treatment group). The impact could be evaluated by comparing the rejection rate of applications assessed using the AI system versus the rejection rate in the control group.

A downside of these options is that they require grant officers to go back and forth between two different ways of assessing the grant applications. This could cause some confusion and lead to different outcomes compared to a more realistic setting where a grant officer consistently uses one approach (with or without AI assistance).

Data collection

The evaluation team wants to assess whether using the AI system increases the proportion of grant funding applications that are correctly rejected: the proportion of all applications that (i) were rejected by grant officers and (ii) were truly ineligible to receive funding (because they were fraudulent or involved errors made in good faith by applicants).

The monitoring systems for this intervention record applications that were rejected. However, it is more challenging to determine whether an application was truly ineligible.

Therefore, the evaluation team decides that the key data to be collected for the RCT is:

- the proportion of applications that were rejected in the treatment and control group (primary outcome)
- the rate of successful appeals: the proportion of rejected applications that were reconsidered and overturned upon appeal from the applicant in the treatment and control group (secondary outcome)

Collecting information about appeals helps check whether using the AI system has the unintended effect of increasing the rate of false positives (applications rejected as ineligible

[26] It would also be possible to tweak this strategy, for example, officers could be allocated to treatment and control groups that change each week rather than each day.

which were, in fact, eligible for funding). The rate of successful appeal is an indirect measure of the rate of false positives.[27]

As the rate of successful appeals is an indirect measure of false positives, the evaluation team suggests that, when rolling out the AI system, it is important to make sure that grant officers carry out further checks when an application is flagged as potentially fraudulent before rejecting it. This informs the design of the monitoring system for the intervention, so that it includes the collection of this along with other monitoring data.[28]

Analysis

The evaluation team defines an analysis plan and implements it once the data has been collected. The analysis plan measures the impact of the AI intervention as the difference in outcomes (described above) between the treatment group and the control group.

The analysis finds that:

- both treatment and control groups have assessed around 20,000 applications each
- the treatment group has rejected 10% of applications, compared to 5% in the control group. This indicates that a further 1,000 applications were rejected thanks to the use of the AI system
- there is no difference in the rate of rejected applications that were appealed, and the rate of successful appeals between the two groups

Learnings from the evaluation

Overall, these findings suggest that the AI intervention has increased the proportion of funding applications that have been correctly rejected by grant officers. As described above, the evaluation has encountered some challenges in measuring precisely the primary intended outcome of the intervention (the proportion of correctly rejected applications). However, the analysis of appeal rates suggests that the increase in rejection rates did not come at the cost of rejecting more legitimate applications – especially considering that additional checks were put in place to make sure that grant officers are not over-reliant on the AI system.

---

[27] A downside of these indirect measures is that not all applicants who have had an application for funding incorrectly rejected would appeal the decision. The indirect measures may, therefore, underestimate the rate of false positives. A more direct measure of the rate of rejection of truly ineligible applications and of eligible applications (false positives) could be based on ex-post checks of application outcomes run by a third group of individuals (not part of the treatment or control group) or by the evaluators. Collecting this information could be time-consuming, but it could be particularly important if there are any concerns that using the AI system could lead to a material increase in the number of incorrectly rejected applications.

[28] Section 4 of the *Magenta Book* provides guidance on planning and implementing data collection for evaluation purposes. Other data could include information on the characteristics of applications and grant officers.

Practical example #2 – using an LLM-based application to help civil servants analyse large amounts of information

This case study describes key aspects of the evaluation of a hypothetical AI intervention.

The AI intervention

A government department is preparing to deploy a generative AI application based on a Large Language Model (LLM) that would help process large amounts of information.[29] For example, the application could be used to analyse and summarise a high volume of documents by civil servants who write briefing notes for ministers. The use of LLMs could improve the efficiency and consistency of this work by completing text analysis and drafting summaries faster than the human alternative.

Whilst LLMs have many benefits, there are associated risks. LLMs may not be able to understand specific contexts or particular nuances that a human could. They may also 'hallucinate', where they confidently assert incorrect information. Additionally, LLMs may carry inherent biases, which could result in an unpredictable skew in the summaries of documents. Therefore, the department plans to deploy a phased 'test-and-learn' approach to evaluate this intervention. The LLM application will first be released to a group of early users and then gradually rolled out to others in the department.

There is a lot of interest in the use of the LLM tool and enthusiasm for the team to deploy it. However, they want to ensure that they can assess the impact and any risks before extending roll-out. As such, the evaluation team tasked with scoping and implementing an impact evaluation of this intervention will need to work at pace.

Key challenges and opportunities for evaluation

The evaluation team identifies two key opportunities and challenges for the scoping and delivery of this evaluation.

Firstly, given the planned roll-out, there is an opportunity to use experimental methods to robustly identify the impact of the intervention.

However, the evaluation team will need to align the evaluation with the iterative nature of the intervention while also working at pace. The evaluation team will also need to work closely with the delivery team to develop a Theory of Change for the intervention, clearly identifying the intended outcomes, risks, mechanisms, assumptions and underlying evidence.

[29] An LLM is a type of artificial intelligence capable of general-purpose language generation. It 'learns' from large datasets of text documents to predict and generate responses based on the input received. It can interpret and respond to text inputs in a human-like manner, making It user-friendly and suitable for use without technical expertise.

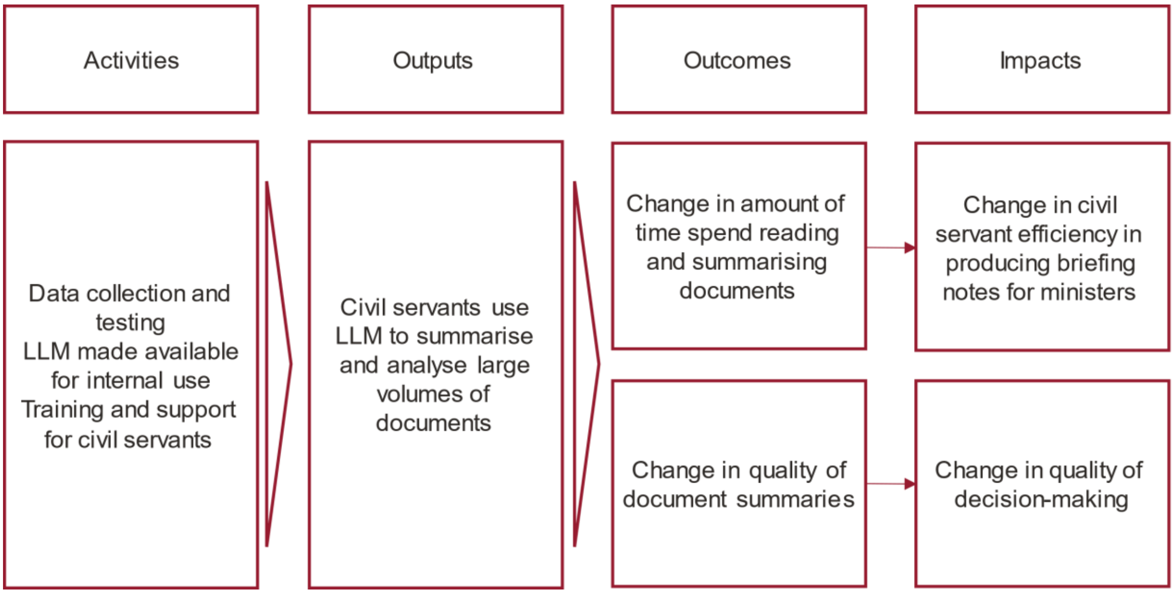
Scoping the evaluation

The evaluation team works closely with the delivery team to ensure evaluation is built into the phased roll-out of the intervention. It is agreed that at each progressive stage of the roll-out, the evaluators will gather and assess evidence on the impact of the intervention, refining the Theory of Change and evaluation questions. This formative evaluation will refine the intervention and inform key decision points on further roll-out phases. In order to iterate more quickly and ensure evaluation findings are effectively employed in refining the intervention, the evaluation team involves the delivery team closely in design, data collection, analysis and reporting.

LLM applications can be used for a wide range of purposes, and the overall objective of the intervention, as stated above (to 'help with the processing of large amounts of information'), is broad. Therefore, to evaluate the intervention robustly, it is important to define its intended outcomes more precisely. To do this, the evaluation team engages in discussion with the sponsors of the intervention and the AI development team.

Based on these discussions, a rapid literature review on LLMs and insights from assurance exercises carried out by the AI development team, the evaluators identify both the likely intended outcomes of this AI intervention and the potential unintended outcomes. Some of these are represented in the high-level Theory of Change below (Figure 5).

Figure 5 High-level Theory of Change



Note: This is a simplified Theory of Change that focuses on specific outcomes in order to illustrate the chosen evaluation approach.

Based on the Theory of Change, the evaluation team determines the primary evaluation questions to be:

1. Has using the AI system improved the efficiency of civil servants in producing briefing notes for ministers?
2. To what extent does the use of LLMs lead to a change in the quality of document summaries?

The evaluation approach

The evaluation team considers possible options to evaluate the impact of the intervention, including experimental, quasi-experimental and theory-based approaches. Since the plan for rolling out the LLM application includes releasing it to progressively larger groups of early users, the evaluation team sees an opportunity to evaluate the intervention through an RCT.

The main steps to implement the RCT are:

1. Plan and implement the creation of a treatment and a control group[30]
2. Plan and implement data collection
3. Plan and implement analysis of the data collected

Implementing the evaluation approach

Creating treatment and control groups

The evaluation team works with the delivery team to plan the RCT. First, they identify a list of civil servants whose responsibilities include drafting briefing notes for ministers. Then, they randomly allocate these civil servants to a 'treatment' group (the early users of the LLM application) and to a 'control' group (who will not have access to the LLM application at this stage).

The evaluation team has chosen this approach to carry out the RCT because it enables them to produce evidence on the impact of using the LLM relatively quickly before full roll-out. This approach also allows feedback from participants to be gathered, which can be used to inform the roll-out and refine the final design of the LLM application.

Data collection

The evaluation team wants to assess whether using the LLM application makes civil servants more efficient at producing briefing notes and whether it impacts the quality of their output.

[30] Or, in more complex designs, multiple treatment and control groups. This includes taking into account ethical considerations and seeking informed consent from individuals to participate in the RCT.

Data for the analysis will be collected from four sources:

1. An initial survey of the treatment and control groups will capture their background characteristics, information on the types of briefing notes they produce and their perceptions of the use of LLMs.
2. The civil servants will be asked to record the number of briefing statements they produce over the test period and how much time they spent doing so.
3. A follow-up survey will gather information on whether and how the treatment group used the LLM tools to produce briefing statements.
4. A random selection of briefing statements produced by the treatment and control groups will be assessed and scored for accuracy and quality by the supervisor or line manager of the civil servant producing the briefing note and by an independent expert in the relevant subject area, using a pre-defined scoring scheme. While it may not be possible to hide from the supervisor whether the note was produced with the aid of the LLM tool, the independent expert will not be informed whether or not an LLM tool was used.[31]

To complement the RCT design, the evaluators also propose to:

- interview a small sample of the treatment group to gather qualitative information on how they approached the tasks, how they used the LLM and how it could lead to efficiency savings in their day-to-day work
- assess a random selection of the outputs produced by the LLM in the treatment group to check whether the LLM produced any 'hallucinations' (i.e. generated false information) and, if so, whether any of this false information was used in the final briefing statements produced by the treatment group

Data analysis

Comparing the results for the treatment and control groups for the first phase of the roll-out, the evaluators find that:

- both the treatment and control groups completed approximately five briefing notes on average during the test period
- the treatment group spent an average of 3 hours and 15 minutes per briefing note, compared to the control group, which spent 4 hours and 55 minutes on average[32]

[31] As measuring quality in this context is challenging, beyond using a pre-defined scoring scheme, the evaluators could also consider asking several reviewers to score the notes, rather than just one, and/or using a different AI tool to provide an assessment of quality. They could then use the average of all graders' scorers.

[32] Although sample sizes were relatively small for this phase of the roll-out, this difference was still found to be statistically significant at the 5% level. That said, statistical significance on its own might not be a good measure by which to validate the sample-based results to infer impacts on the wider population. Further validation performance measures should be considered.

- over 80% of those in the treatment group reported using the LLM tool for the majority of the briefing notes produced during the test period
- of the briefing notes evaluated by an expert panel, the treatment group scored an average of 80% on accuracy and quality measures, while the control group scored an average of 81%

Learnings from the evaluation

The evaluation suggests that the LLM reduced the average time to complete tasks and did not come at the cost of lower-quality responses. The evaluators use these findings to provide initial indications of the potential time-saving had the tool been rolled out across the whole civil service, multiplying time saved by an estimate of the number of briefing notes produced per year. However, they clearly indicate that these preliminary findings may not necessarily hold at scale, and additional testing is required.

Based on these findings and working closely with the delivery team, the department decides to progress the roll-out of the tool to a larger group of users, repeating the same evaluation design. Small adjustments are made to the tool's user interface based on comments received from civil servants and will be evaluated in the next phase.

Practical example #3 – using a Chatbot for providing citizen user support

This case study describes key aspects of the evaluation of a hypothetical AI intervention.

The AI Intervention

A Chatbot powered by an LLM was launched and added as a feature to a website to provide citizens with needed user support. The Chatbot draws from published information on a given website.[33] The Chatbot is intended to help users find the information they need by letting them ask questions about the website in the way they would write or speak in everyday life. The Chatbot responds with summaries of information and signposts users to the best place to find information. Before users start the interaction with the Chatbot, they are informed that they are about to interact with an AI-powered chat service that mimics interaction with a human service provider. The tool replaces an older 'rule-based' Chatbot, which uses a word search algorithm to suggest links to relevant information pages. The introduction of the AI-enabled Chatbot is part of a broader overhaul of the website, including new information about the provided services. The older rule-based Chatbot has not been updated to include the new information.

The aim of using AI technology in the Chatbot is to improve the users' experience of the website and the quality of information they receive. This is expected to reduce the users' need to contact the call centre for further help.

Key challenges and opportunities for evaluation

The evaluation team identifies two key challenges for the evaluation of this AI intervention:

Firstly, because the AI intervention was one of several website updates made in the same time period these concurrent changes (e.g. updates to the information pages) will make it difficult to attribute the outcomes (e.g. greater user satisfaction) to the AI intervention.

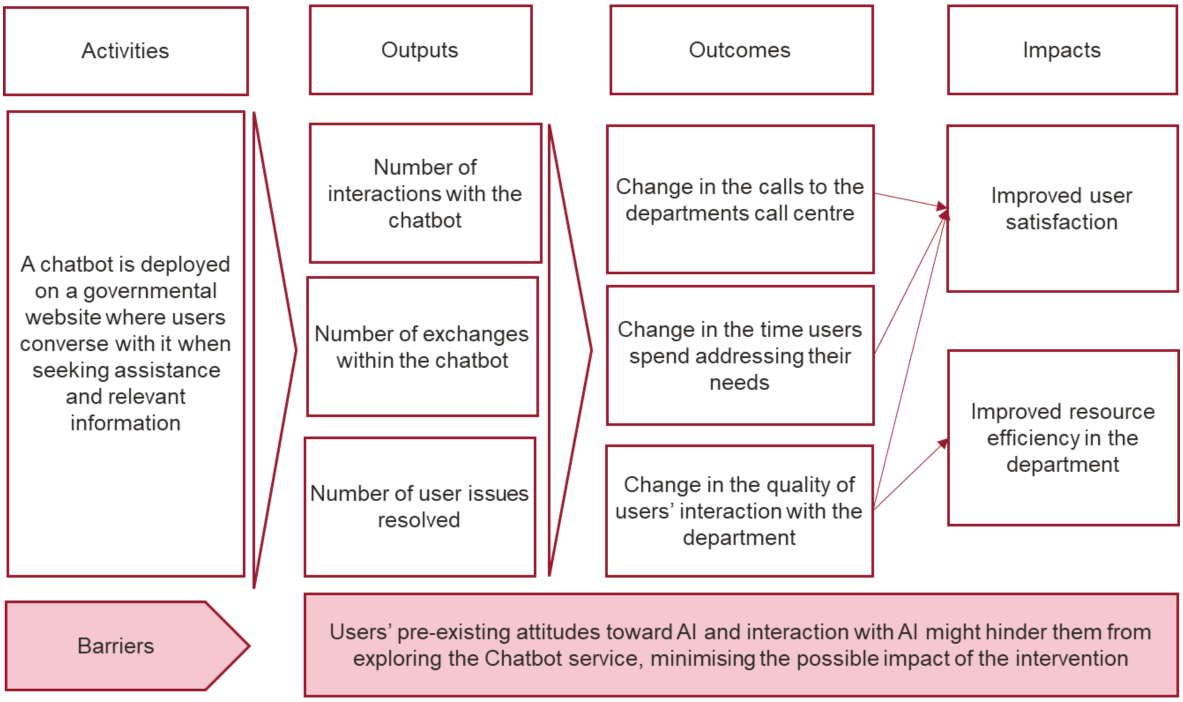
Secondly, since the Chatbot will be used by members of the public, public attitudes towards AI are likely to influence the impact of the intervention. Therefore, evaluators should understand these mechanisms and design an evaluation approach that takes public attitudes into account.

[33] The accuracy of the new technology was tested in the development stages before its publication.

Scoping the evaluation

The evaluation team sets out the identified outcomes of this AI intervention and presents them in a high-level Theory of Change, as shown below (Figure 6).

Figure 6 High-level Theory of Change



Note: This is a simplified Theory of Change that focuses on specific outcomes to illustrate the chosen evaluation approach.

The evaluation team identifies that users' preconceptions of AI and Chatbots in general (whether rule-based or AI-enabled) might create barriers to maximising the impact of this intervention. Some users may have negative opinions of AI and not want to engage with the Chatbot. Upon learning that they are interacting with an AI, users might not want to continue the conversation with the Chatbot as they might believe it will not be able to produce the relevant information. Since this might hinder the intervention's impact, the evaluation team notes it will be important to assess if this barrier exists and, if so, to what extent.[34]

In light of the Theory of Change, the evaluation team determines that the primary evaluation questions are:

1. Did the AI Chatbot increase user satisfaction with the new Chatbot (utilising the AI technology) compared to the older Chatbot version?

[34] The accuracy of the information provided by the Chatbot is being assessed through a separate RCT component of the evaluation during the development stages of the tool and throughout implementation.

2. Did the AI Chatbot reduce the number of calls that are received in the call centre compared to the volume of calls in the period of the older version of the Chatbot?

The secondary evaluation question is:

3. Did users' pre-existing attitudes toward AI limit their utilisation of the Chatbot?

The evaluation approach

This impact evaluation is part of a wider evaluation framework that includes randomised testing of the quality of the AI tool and the accuracy of its answers (pre-, during and post-development). The evaluation team has now been tasked with assessing the wider impact that the tool might have when implemented as part of a policy intervention.

The evaluation team considers different possible impact evaluation methods and determines that a theory-based approach, using a contribution analysis, is most appropriate. This is because key stakeholders for the evaluation are particularly interested in understanding how the context in which the intervention has been rolled out shaped its results and exploring the interactions between the Chatbot and other concurrent changes that have taken place.

After setting out the evaluation questions to be answered and developing a Theory of Change (as above), implementing a contribution analysis involves gathering evidence on the Theory of Change, assembling a contribution narrative that sets out how credible it is that the intervention has contributed to the observed outcomes, identifying gaps in the evidence on the contribution narrative, and iterating on these last steps.[35]

To gather evidence for this contribution analysis, the evaluation team undertook the following steps:

1. Observe changes in the outcomes of interest. In this case, the evaluation team identified user satisfaction and the number of user calls to the help centre as the primary outcomes.

2. Identify and observe factors other than the intervention that might have influenced the outcomes.This includes broader changes to the website, users' perceptions of AI and users' perceptions of the services they are seeking to access through the website.

3. Identify and evidence the mechanisms through which the intervention might have influenced the outcomes: In this case, the team identified three such mechanisms:

    a. AI provides more accurate and relevant information to users compared to the older version of the Chatbot

[35] For a full description of the six steps in contribution analysis, please see Mayne, J. (2008) *Contribution Analysis: An approach to exploring cause and effect*. Brief 16, Institutional Learning and Change (ILAC).

b. AI improves ease of use by presenting only the relevant information in the chat

c. AI mimics human interaction, which might improve user engagement with the tool[36]

Implementing the evaluation approach

Data collection

The evaluation team gathers the following data on primary outcomes from the website's monitoring information:

- responses to several survey questions posed to users at the end of the interaction with the Chatbots (the rule-based Chatbot before the intervention and the AI-enabled Chatbot afterwards), including a 1-5 scale answer to the question 'How satisfied are you with the responses you received overall?'[37]
- number of calls to the help centre over time

The evaluation team also collects information on the mechanisms through which the intervention and other factors may have influenced the evolution of the outcomes above. The information is gathered through monitoring data, website users' surveys and interviews with the website team.

The monitoring data collected to inform the understanding of the potential mechanisms includes:

- the number of interactions with the Chatbot (before and after the intervention)
- the number of text exchanges per conversation (before and after the intervention)

An increase in the number of interactions with the Chatbot for a given user over a given period could indicate an improvement in user attitude towards the tool. In particular, if the average number of interactions increased after the introduction of AI, it might suggest that those interactions are positive, making users more inclined to reach out to the Chatbot in future interactions.[38] An increase in the number of text exchanges per conversation could suggest

[36] The evaluation team expects that evidence should be available to support the first two mechanisms. The first should be available from the AI tool testing phase, showing the accuracy of the AI tool. The relationship between ease of use and higher engagement is well-known, and the mechanism is documented. However, the evaluation team noted that the third mechanism, which is related to the unique nature of AI, might have been less studied so far. As such, the evaluation team noted they might need to collect further evidence about this mechanism.

[37] Data was also collected about additional questions to gather further insights and evidence about the mechanism of the new Chatbot that improves their satisfaction. Questions were also designed to help understand alternative drivers that might lead to improvement of users' satisfaction or any issues that might be hindering a greater impact from the tool.

[38] The data is available as users need to log on to their restricted area before interacting with the chatbot. This means that the number of interactions per period per user is available before and after the intervention.

that users find the answers useful and converse with the chat. That said, too many exchanges would also suggest that the right information is not being given quickly enough or that users are struggling to frame questions.

The survey on users of the website collects data on:

- whether the user is aware of the Chatbot and whether they have interacted with it
- users' attitudes towards the use of AI in public-facing services (for example, to what extent they think the use of AI to improve public services is appropriate, or whether they have interacted with Chatbots before and how helpful they found them)
- how users prefer to receive the information they need about the service (for example, from a conversation versus reading guidance or instructions)

Topics explored in interviews with the website team include:

- the broader changes that were made to the website and their likely effects
- how they expect users to interact with the Chatbot
- the context in terms of user satisfaction with the website, its drivers and any insights gathered from recent user research

Analysis

The evaluation team analyses the data and assesses the changes in the outcomes before and after the introduction of the AI. It finds that:

- user satisfaction has increased since the introduction of the AI Chatbot
- the number of calls to the help centre has decreased after the introduction of the AI Chatbot

The evaluation team seeks to understand to what extent the introduction of the Chatbot has contributed to these changes, alongside other factors. Key findings from the evaluation teams' analysis include the following:

- The average number of interactions with the Chatbot per user increased over time after the introduction of the AI tool.[39]
- The user survey showed that only a very small proportion of users is opposed in principle to AI being used in public services, and only a small proportion of users have had a negative experience interacting with Chatbots in the past.

[39] The evaluation team monitored the observed changes over a longer period of time to ensure the observed changes were not related to any short-term impact associated with changes in the user interface.

- Users' views on their preferred way of getting information on online services are mixed, with many reporting that they find conversations most helpful, but an almost equal proportion reporting that they prefer to read written guidance or instructions.

These findings support the hypothesis that using an AI-enabled Chatbot would increase user engagement, and suggests that users' perceptions of AI and Chatbots are not likely to pose a significant barrier to engagement. However, the mixed results on users' preferences for accessing information on online services suggests that the conversational nature of the Chatbot will be helpful for many but not all users.

Moreover, the website team indicates that the website overhaul was conducted two months before the AI component of the help Chatbot was introduced. The evaluation team has observed a more pronounced change in the user satisfaction scores trend after the introduction of AI and a smaller improvement in this trend after the overhaul two months prior. This suggests that the introduction of the AI Chatbot has contributed more profoundly to improving user satisfaction.[40]

The older Chatbot typically directed users to a web page to access information. The need to click through might have hindered users from seeking the relevant information (additional effort). The team suggested that the AI Chatbot functionality that presents only the relevant information in the chat box will lead to quicker presentation of information, improving users' understanding and satisfaction. The evaluation team has corroborated this finding through existing literature that evidences the link between the reduction in the number of actions needed from the user and the higher engagement that they present online.

Learning from the evaluation

Taking all this information together, the evaluation team prepared a contribution narrative:

The evidence supports the claim that the intervention contributed to improvements in user satisfaction and reduced the number of calls to the help centre. The effort needed to access the relevant information was reduced as the relevant information now appears in the same Chatbot in front of the user. AI's ability to mimic interaction with humans has also contributed to users' engagement with Chatbots, which has led to a higher proportion of users receiving the needed information. Other improvements to the website have also contributed to the accuracy of the information that users can find online, but they are unlikely to fully account for the changes observed in user satisfaction and use of the help centre.

[40] The evaluation team notes that overall improvements might require longer to materialise, as implementation issues at early stages might lead to lower user satisfaction.

Practical example #4 – supporting patients with a chronic disease

This case study describes key aspects of the evaluation of a hypothetical AI intervention.

The AI intervention

The NHS has recently implemented a programme to encourage the use of an AI-enabled digital intervention ('Patient support AI') that can help improve the health outcomes of patients suffering from Chronic Obstructive Pulmonary Disease (COPD).

COPD is a diagnosis that refers to several respiratory conditions.[41.] The main symptoms of this chronic disease include breathlessness, chesty cough, frequent chest infections and persistent wheezing.[42] While specific symptoms may differ among patients, exacerbations of symptoms can negatively affect the overall health of the patients. COPD exacerbations are the leading cause of patient death and hospitalisation.[43] Exacerbations can be prevented by early detection of deterioration and timely treatment, which can effectively lower the severity of exacerbations and prevent hospitalisation and death.

The Patient support AI programme includes a wearable device that records relevant patient indicators (e.g. blood pressure and blood oxygenation levels). The data is shared with an app that patients have on their phones. The AI system analyses the recorded health data and identifies cases where proactive medical attention may prevent an upcoming exacerbation. These cases are flagged to the patient's GP, who can contact the patient to suggest actions that might avoid an exacerbation and subsequent hospitalisation. By flagging earlier points of medical intervention, the tool is able to reduce the number of exacerbations among COPD patients.

Patient support AI is now available to all GPs in England; the roll-out was done in two stages. In the first stage, 300 GP practices registered to participate in a pilot. After a year, Patient support AI was used by all GP practices in England to monitor patients with chronic disease (where the patient has consented to this).

It has now been a year since patient support AI was made available to all patients, and an evaluation team has been tasked to design and implement an evaluation of this AI system.

[41] Including but not limited to emphysema and chronic obstructive airways disease.

[42] Source: [Chronic obstructive pulmonary disease (COPD) - NHS (www.nhs.uk)]

[43] Source: Flattet, Y., Garin, N., Serratrice, J., Perrier, A., Stirnemann, J., & Carballo, S. (2017). Determining prognosis in acute exacerbation of COPD. *International journal of chronic obstructive pulmonary disease*, 467-475.

Key challenges and opportunities for evaluation

The evaluation team identifies two key challenges for the evaluation of this AI intervention:

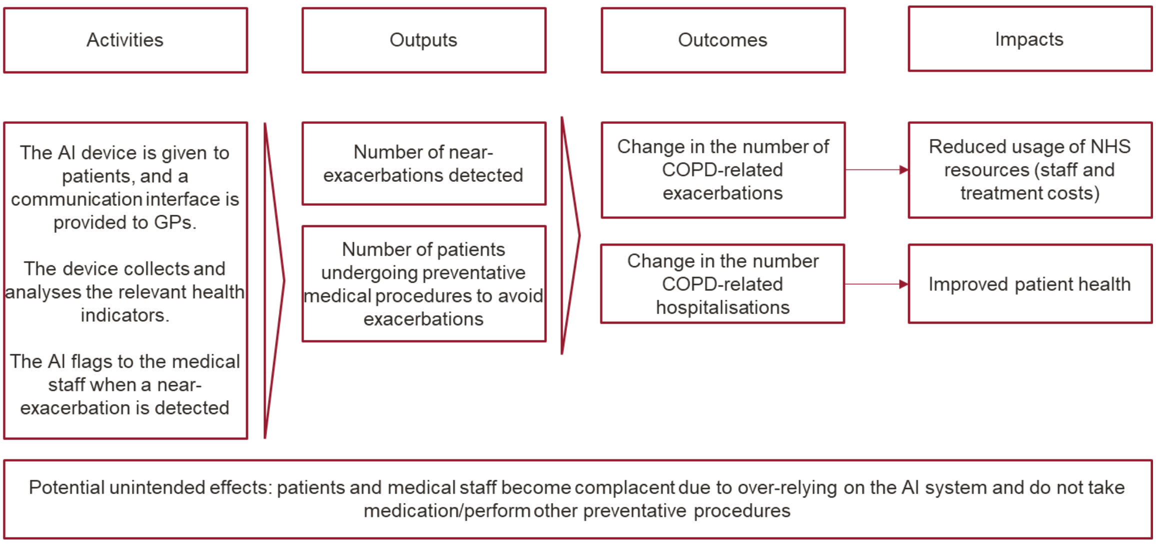
Firstly, the impact of using the AI system may differ across patients since health conditions and the effectiveness of treatment might vary between different patient groups. If the AI tool was predominantly trained on a patient population with a specific set of characteristics, it might not be as efficient in flagging upcoming exacerbations for patients with other characteristics (different age, sex, etc.). Those differences might have already been tested as part of clinical trials. However, the evaluation team would still like to test if differences in outcomes arise once the tool is rolled out into the real world.

Secondly, because roll-out has already happened, it is not possible to conduct an RCT, and the team has not been able to collect baseline data before the intervention roll-out. Therefore, the team will likely need to rely on secondary, routinely collected data for the evaluation.

Scoping the evaluation

The evaluation team identifies the likely outcomes of this AI intervention. These are represented in the high-level Theory of Change below (Figure 7).

Figure 7 High-level Theory of Change



Note: This is a simplified Theory of Change that focuses on specific outcomes to illustrate the chosen evaluation approach.

Based on conversations with key stakeholders, the evaluation team determined that the primary evaluation questions are:[44]

1. Did the Patient support AI reduce the number of COPD-related hospitalisations?
2. Did the Patient support AI reduce the number of COPD exacerbations?

The secondary evaluation question is:

3. Did the impact of the Patient support AI tool differ for some patients?

The evaluation approach

The evaluation team considers possible options to evaluate the impact of the intervention, including experimental, quasi-experimental and theory-based approaches. The nature of this AI intervention would have made an RCT technically feasible had this been planned from the start and subjected to appropriate ethical scrutiny from an ethics review. However, in this case, the AI system has already been rolled out to a group of GPs that were not selected randomly, so an RCT is not possible.

The team determines that a quasi-experimental approach is an appropriate way to evaluate the intervention. The period when the pilot was introduced to only some of the GPs can be leveraged to compare outcomes between patients of treated and untreated groups. Outcomes of the intervention are expected to have already materialised in the ‘treated’ group (those who received the intervention earlier, as part of the pilot) but not yet in the ‘untreated’ group (those who received the ‘regular’ COPD treatment and the AI tool only later, as part of the full roll-out).

Ideally, the evaluation team would have been involved earlier in the delivery of the intervention. The team could have worked with delivery colleagues to design an RCT or suggest tweaks to the roll-out plan that could have helped with the implementation of a quasi-experimental approach.[45]

However, robust evaluation is still feasible. The steps that the evaluation team need to undertake to implement a quasi-experimental design are:

1. Define the treatment group and a feasible approach for identifying the comparison group
2. Plan and implement data collection
3. Plan and implement analysis of the data collected

[44] Note: for simplicity, this case study focuses on evaluating the impact of the intervention on its outcomes (exacerbations and hospitalisations) rather than the ultimate impacts on patient health.

[45] It is worth noting that in considering an RCT, the evaluation team would have considered potential ethical concerns. Indeed, restricting the roll-out of a tool that can improve patients' health would involve some ethical risks.

Implementing the evaluation approach

GPs actively chose to register their interest in the pilot, which means that the allocation was not random. Therefore, simply comparing patients who received the treatment because they were registered with GPs who were part of the pilot with other patients is unlikely to provide a robust estimate of the AI tool's impact. For example, GPs who care for more patients with severe COPD may have been more likely to register their interest in the pilot. Their patients are also more likely to require emergency care. As such, the evaluation design needs to account for GP characteristics that might have affected being part of the treatment group.

In light of this, the evaluation team decides to deploy a matched difference-in-differences approach to evaluate the AI intervention. The approach involves constructing a comparison group of patients who did not receive the treatment (as they were registered with GPs who were not in the pilot) but were similar to those who received the treatment. In particular, the matching technique identifies a group of patients with similar personal characteristics to those in the treatment group (in terms of age, sex, socioeconomic background, the severity of the COPD and the use of other prophylactic measures) and similar GP characteristics (number of COPD patients, age of the medical staff, etc.). Then, changes over time in the outcomes of the treated patients are compared with changes over time in the outcomes of patients in the comparison group (making this a difference-in-differences estimation). [46] [47] [48]

The approach to matching is set out in advance and specified in a pre-analysis plan to enhance the credibility of the analysis.

Data collection

The evaluation team collects data on the following outcomes (before and after the roll-out of the pilot):

- number of reported COPD exacerbations across the patients of each GP
- number of COPD-related hospitalisations across the patients of each GP

[46] Since the treatment was assigned to the GP level and not the patient level. The proposed approach might require clustering of standard error at the GP level to assess the statistical confidence of the results.

[47] In practice, it might be challenging to gain approval for accessing patient-level data for non-clinical research, given the high risk that personal medical records bear. However, if, in practice, the evaluation team had assessed that access to patient-level data was not possible, they would have considered an evaluation design on the GP level. In this scenario, the evaluation team would still like to control the patient characteristics of the treated GPs. In this case, the data can be collected on the GP level (i.e. number of COPD patients, the proportion of severe COPD patients, etc.), and the matching can then be done between the GPs rather than between the patients.

[48] Using a difference-in-differences approach involves an assumption that, in the absence of the intervention, the outcomes of the treatment group and the comparison group would have changed at a similar rate (the 'common trends' assumption). This cannot be tested directly, as it is not possible to observe how the outcomes of the treated group would have evolved without the intervention. However, if there is a similarity in pre-intervention trends in the two groups, this can raise confidence that the assumption of common trends will be met.

The team also gathers data on the characteristics of GPs and patients who have and have not taken part in the pilot and the local areas in which they operate. The team also reviews the available evidence and discusses with stakeholders what patient attributes may affect the impact of the AI tool on their health. The table overleaf presents a list of the possible data points that the evaluation team used for this evaluation.

Table 2 Examples of data collected by the evaluation team

Category of information	Example indicators
Characteristics of patients registered with the GP practice	16. Number of COPD patients registered. 17. Indicators of severity of COPD among patients. 18. COPD patient characteristics (e.g. age, sex, socioeconomic background, ethnicity, comorbidities, history of smoking, and participation in preventative activities). 19. COPD patient-related activities (smoking, participation in other prophylactic activities, etc.).
Characteristics of GP practices	20. Average staff age (which can correlate with GPs' aptitude towards using cutting-edge technology). 21. Staff skills indicators such as level of education and certifications (which can be a driver of staff's desire and ability to engage with new technologies).[49] 22. GP contract type/financial information about the GP (which can influence GPs' ability to dedicate staff to new programmes).
Characteristics of the local area	23. Level of deprivation in the areas serviced by the GP (which can be a driver of worse/better health outcomes for patients). 24. Air quality in the area.

The sample of GP practices included in the pilot is sizeable (300 out of around 6,000 practices in England).[50] With an average of about 2,300 patients per GP[51] and 1.9% COPD patients,[52] the treatment group was found to include around 13,000 patients, which is a sizeable number.

[49] More direct measures of GP's attitudes towards technologies would also be useful and may be available through survey data for a sample of GPs.

[50] Source: BMA (2024), Pressures in general practice data analysis

[51] Source: Royal College of General Practitioners, Key general practice statistics and insight, June 2024.

[52] Source: House of Commons (2021), Support for people with chronic obstructive pulmonary disease, Research Brieifing.

The large treatment group size also allowed the team to investigate variations in outcomes between different patient group characteristics (e.g. gender, ethnicity and age).

Analysis

Based on the data collected, the evaluation team constructs a comparison group of patients who did not participate in the pilot. Having done this, the team:

1. estimates the average impact of the AI intervention by comparing the annual change in COPD-related exacerbations and hospitalisations between the treatment and control groups
2. creates sub-groups of the treatment and comparison group patients (e.g. by sex, ethnicity and comorbidities)
3. whether the impact of the AI intervention in the treatment sub-group (point 2) is different from the impact estimated on the whole treatment group (point 1)

The results of this analysis show that the number of exacerbations and hospitalisations among patients who participated in the pilot group decreased faster compared to the comparison group. The estimated impact is a reduction of 15% in annual exacerbations and 10% in COPD-related hospitalisations compared to the comparison group.[53] The estimated effects of patient support AI in the patient sub-groups that were tested were very close to the figures above. This suggests that the impact of the tool does not vary between the tested patient groups.

Learnings from the evaluation

Based on the findings, the evaluation team concludes that the introduction of the Patient support AI tool has had a positive impact by reducing exacerbations and hospitalisations of COPD patients. It also concluded that no evidence currently suggests that impacts differ between the subpopulations.

The evaluation team notes that it would be valuable to continue monitoring the impact of the AI intervention in the future. Further evaluations can test if the observed impact changes over time.

[53] Technical note: the findings were statistically significant at the 5% level. That said, statistical significance on its own might not be a good measure to validate the sample-based results to infer impacts on the wider population. Further validation performance measures should be considered.

Glossary

AI intervention: a programme or initiative using AI systems in central Government or the delivery of public services.

AI system: consistent with the OECD AI principles, this guidance defines an AI system as a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. The definition of AI used in this guidance includes, but is not limited to, generative AI.

Baseline: the status quo before the roll-out of an (AI) intervention has started.

Contribution analysis: a theory-based evaluation approach that seeks to understand to what extent the intervention has contributed to the observed outcome, combining a range of evidence to test the Theory of Change.

Covariate: an independent variable included in statistical analysis to control for the effects of variables that might influence the outcome, even though they are not the primary focus of the analysis.

Difference-in-Differences: a quasi-experimental approach that assesses how the evolution of the outcome of interest over time differs between a group that received the intervention and a group that did not.

Formative evaluation: an evaluation conducted during the implementation of an intervention, intending to inform decisions on whether and how to make improvements to the intervention.

Hallucination: a situation where an AI tool produces incorrect or misleading information.

Impact evaluation: the systematic assessment of the outcomes of an intervention with the aim of establishing whether, to what extent, how and why an intervention has resulted in its intended impacts.

Outcome harvesting: a theory-based evaluation approach that involves collecting evidence of change and then working backwards to assess what has contributed to that change.

Qualitative Comparative Analysis: a theory-based evaluation approach that analyses systematically qualitative case study data to evidence the link between an outcome and combinations of factors or characteristics.

Randomised Controlled Trial: an evaluation approach that involves randomly allocating an intervention into two or more groups.

Realist evaluation: a theory-based evaluation approach that focuses on testing hypotheses about how the intervention may have led to a given outcome, a specific mechanism and under specific circumstances.

Regression Discontinuity Design: a quasi-experimental evaluation approach that estimates the impact of an intervention by using a cut-off threshold to assign the intervention.

Statistical Matching: a quasi-experimental approach that compares the outcomes of the treatment group to those of a comparison group that is similar to the treatment group in terms of one or more 'matching variables'.

Summative evaluation: an evaluation conducted during or after the implementation of an intervention, with the primary aim to assess the overall impact of the intervention. It often intends to inform decisions about whether to stop or continue an intervention.

Synthetic Control methods: a group of quasi-experimental approaches that use historical data to construct a 'synthetic clone' of a group receiving a particular intervention

WWW.FRONTIER-ECONOMICS.COM