Cabinet Office

The Mitigating ‘Hidden’ AI Risks Toolkit

A practical guide to identifying, tracking and mitigating unintended consequences of AI adoption

Delivering the UK Government's first generative Artificial Intelligence tool to be approved for cross-government use

Government Communications has pioneered Assist, the first general purpose AI tool approved for use across the UK Government[1]. Developed in-house by the multidisciplinary Applied Data and Insight Team, this secure and bespoke generative artificial intelligence tool (GenAI) is transforming how government communicators work.

Since the launch of the Assist pilot in November 2023, Assist has unlocked greater productivity, saving thousands of hours of communicators' time, whilst enabling better integration of communications best practice by embedding core Government Communications frameworks, policies and documents into the tool's responses. As a result, Assist has already supported the rapid delivery of efficient, consistent and high-quality public sector communications across more than 200 government organisations.

Through developing and scaling Assist across government, we've learnt many lessons and want to share our insight with those facing similar challenges implementing GenAI tools across other organisations. As a result, we have chosen to publish this toolkit and wider resources, including our sister publication, ***The People Factor: a human-centred approach to scaling AI tools***. By publishing these resources, we aim to support other teams to successfully and safely scale GenAI in their organisations. By sharing our work, we hope to contribute to the ethical and impactful use of AI for public good.

Feedback and collaboration

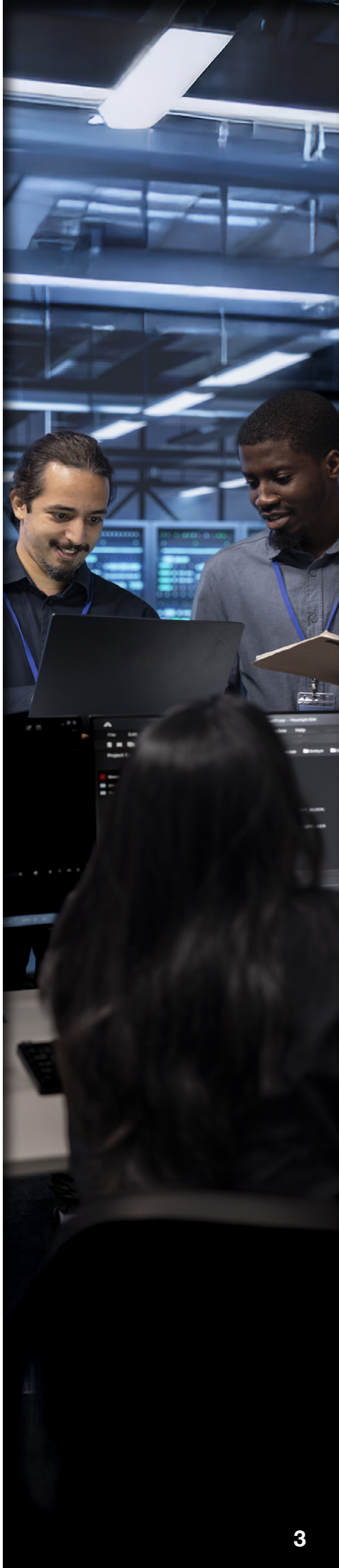
We welcome feedback on the guide as well as opportunities to collaborate with other teams, particularly if you have experience of rolling out AI tools and services. This input is invaluable to the project's continuous development.

Get in touch with us by email: gcs@cabinetoffice.gov.uk.

Learn more about how Government Communications is responsibly harnessing innovations including AI to transform government communications:

- Generative AI policy
- Innovating with Impact Strategy
- Framework for Ethical Innovation

Contents



About this toolkit

The path to AI success isn't just technical – it's cultural, organisational and human. As a result, mitigating the full range of risks associated with rolling out AI tools requires more than just technical safeguards alone.

While our sister guide, ***The People Factor: A human-centred approach to scaling AI tools***, provides a framework for how to scale AI across organisations, this toolkit shows you how to identify, monitor and mitigate the 'hidden' behavioural and organisational risks associated with AI roll-outs. These are the unintended consequences that can arise from how well-intentioned people, teams and organisations interact with AI solutions.

Who is this toolkit for?

This toolkit is designed for individuals and teams responsible for implementing AI tools and services within organisations and those involved in AI governance.

It is intended to be used once you have identified a clear business need for an AI tool and want to ensure that your tool is set up for success. If an AI solution has already been implemented within your organisation, you can use this toolkit to assess risks posed and design a holistic risk management approach.

You can use the Mitigating Hidden AI Risks Toolkit to:

- Assess the barriers your target users and organisation may experience to using your tool safely and responsibly
- Pre-empt the behavioural and organisational risks that could emerge from scaling your AI tools
- Develop robust risk management approaches and mitigation strategies to support users, teams and organisations to use your tool safely and responsibly
- Design effective AI safety training programmes for your users
- Monitor and evaluate the effectiveness of your risk mitigations to ensure you not only minimise risk, but maximise the positive impact of your tool for your organisation

Introduction

When we think about AI safety, we often focus on dramatic scenarios: deepfakes undermining democracy[2], biased algorithms making unfair recruitment decisions[3], or generative AI models hallucinating false information[4]. These are the risks that grab headlines[5].

But what if some of the most significant AI risks come from far more mundane sources?

One well-known lesson from aviation safety is that most accidents aren't caused by dramatic events like engine failure, storms and hijackings. In fact, 60 – 80% of aviation accidents are caused by smaller, less visible problems like poor maintenance, miscommunication between pilots and ingrained organisational or cultural practices[6 7 8].

A pop-science book tells the story of a plane crash, whereby a junior pilot was too polite and deferential to his captain due to their organisation's hierarchical culture[9]. The junior pilot repeatedly hinted that the plane didn't have enough fuel to complete their journey, but never used the word "emergency" directly. The warning was too subtle for the captain to understand before it was too late, and the plane ran out of fuel and crashed.

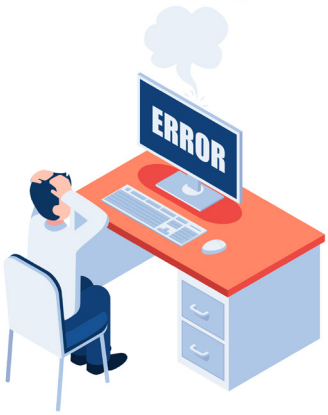
Risks from the use of AI are likely to follow a similar pattern to aviation accidents.

Media-grabbing risks, like deepfakes, are a key focus of AI safety discussions and initiatives because they are visible and predictable – it is easy to anticipate that someone may want to use AI technology for harm. On the other hand, someone privately using a generative AI tool (GenAI) to write a report or to help them with administrative tasks at work – but forgetting or not having the time to check their output for accuracy – is far more likely to fly under the radar unless or until the problem escalates into a crisis.

Risks resulting from the **limitations** of, **attacks** on or **intentional misuse** of AI tools

Highly salient, easy to conceptualise and anticipate

Risks resulting from the **well-intentioned** use of AI tools by people, teams and organisations



Low salience, hard to anticipate, "mundane", but high potential impact ("unknown, unknowns")

For example, imagine an employee, Marco, in 2026.

His organisation has fully automated all of his low-value, 'easy' tasks using AI tools. The expectation is that this frees his efforts to focus 100% of his time on undertaking more complex, 'high value' tasks that are more rewarding. Marco, like others in his team, enjoys having less admin to do and being able to spend more time on meaningful and interesting work.

However, lately he has realised this substitution has come at a cost. In fact, unknown to him, those low effort, 'menial tasks' like writing an email or data entry, were actually quite a relief from his high effort, complex tasks. Now that he's lost these tasks that used to break up his day, he has found it increasingly difficult spending all of his time on cognitively demanding tasks. Unused to this work allocation, he and his other colleagues end up feeling mentally fatigued and stressed, reducing, rather than increasing, their productivity.

We call these 'hidden' risks[10] because they are often less obvious and it will be difficult to identify their origin.

There may be lots of reasons why Marco and his team might be less productive that may have nothing to do with his organisation's AI roll out. As a result, if the 'hidden' link between cognitive fatigue and AI-enabled task automation goes unrecognised, organisations like Marco's might continue to push for full automation of routine tasks not knowing this could be hindering rather than helping them, and without implementing appropriate and effective mitigation strategies to proactively handle this risk.

Importantly, none of the current – predominantly technical – approaches to AI safety are equipped to handle these 'hidden' AI risks.

Technical guardrails or higher quality AI training data won't prevent employee burnout in Marco's scenario. Approaches such as red teaming might help, but without a framework to ensure you're capturing the full spectrum of risks, it is unlikely to provide comprehensive safeguards. These risks from the implementation and use of AI could appear mundane when compared with more salient risks of AI, but can have high potential impacts as AI continues to be embedded into our ways of working.

Addressing gaps in AI safety research – this toolkit

When we started developing Government Communications' bespoke GenAI tool Assist, it was clear to us that there were many potential unintended consequences of integrating AI tools into Government Communications' work that we would need to anticipate and proactively mitigate.

Despite the existence and value of comprehensive catalogues documenting general AI risks[11], there was no practical framework for helping organisations and their teams developing, implementing or governing AI tools to identify and mitigate the specific risks of their specific tools in their specific contexts before, not after, risks have already materialised.

We therefore created our own toolkit, drawing on a wide literature from behavioural and social science, human factors research and emerging AI safety research as well as consulting experts across academic and real-world organisational settings.

We are publishing our approach so that others can use it to identify, track and proactively mitigate the 'hidden' risks of their AI roll-outs too. The result is a **balanced approach which rejects the false binary of 'all good' versus 'all bad' thinking about AI**, instead creating a middle path where risks are systematically identified, evaluated and managed, enabling us, and others, to **make the most of the benefits AI brings whilst mitigating potential downsides**.

What are hidden risks?

Hidden risks in GenAI are unintended consequences that are less salient or which remain invisible until they escalate into obvious problems or crises.

These consequences typically emerge after a very obvious harm has occurred or when a thorough investigation reveals the chain of events leading to the problem. The most well-known examples come from aviation, where black box data helps deconstruct the causes of plane crashes, or healthcare, where inquests and inquiries examine serious medical errors to prevent future incidents.

A significant challenge in anticipating AI risks is the **limited historical examples we can learn from** so another approach is needed. We started with thought experiments[12]—an analytical method used in philosophy to test concepts and assumptions. By integrating behavioural science principles, we developed a series of systematically constructed scenarios that anticipate potential hidden AI risks in specific contexts.

Examples of these are described below.

Imagine a team working in an organisation facing significant pressures to reduce their budgets.

The head of their team, Robert, decides to reduce headcount on the assumption that fewer people are needed to carry out essential tasks once they have access to AI-powered assistants.

Whilst this assumption of resource savings may be accurate for some simple tasks, it may not be a reasonable assumption for others, such as tasks that require deep institutional knowledge, specific expertise and stakeholder relationships that AI tooling cannot replicate.

This assumption may come about due to Robert's limited understanding of the work delivered by his team, and/or his limited understanding of the strengths and limitations of AI-powered assistants currently available.

As a result of the head of the team's decision, the remaining team members become stretched managing both the AI systems and the core work that requires human input. The team lacks psychological safety, so feels unable to safely raise concerns about Robert's decision-making out of concerns for personal impacts. This leads to burnout, inefficiencies and a reduction in service quality.

If this type of poorly informed AI implementation became widespread it could create a systemic productivity paradox—where technology intended to enhance efficiency actually undermines it.

Imagine an analyst, Nithya. She is tasked by her manager to undertake some urgent desk research for a briefing on a new policy her organisation is introducing, which she has limited knowledge on.

Under time pressure, she rushes to use a generative AI tool to speed up her desk research, believing it will help her meet the tight deadline efficiently.

Impressed by accurately-sounding outputs and in a rush, Nithya copies the content directly into her briefing document, checking rapidly for obvious errors but not cross-referencing its contents with official sources due to the limited time she has.

The document is shared with stakeholders who begin citing its contents, only to later discover that the AI-generated information was inaccurate and has already spread throughout the organisation. If significant decisions are taken on the basis of this information, organisational performance could suffer.

Imagine a researcher, Jordan, who needs to compile research on a complex topic and has pre-existing opinions about the likely conclusions.

When using the organisation's GenAI tool, Jordan's prompts reflect his pre-existing views: "Summarise findings that demonstrate the positive impact of X on Y." Since AI models are often fine-tuned based on human feedback throughout their development[13], the AI produces content that emphasises Jordan's preferred evidence while downplaying alternative perspectives.

While intended to improve user satisfaction with outputs and deliver “AI alignment” with human values[14], the alignment process can lead to models that generate sycophantic responses which are “optimised” to respond to users’ beliefs and preferences over a more ‘correct’ truth – in short, telling people what they want to read[15].

Jordan is content with the output as it matches his pre-existing opinions (this is known as confirmation bias) and shares the research summary with decision-makers, who make choices based on this incomplete picture. Months later, the decisions require revision when overlooked factors become evident – these are the factors that were screened out by the reinforcement loop between Jordan's confirmation bias and the AI's people-pleasing design.

Think about Sarah, who has access to an AI tool in her organisation to help her sift applications for a new role in her team.

Her organisation has given all hiring managers optional training in how to use the tool for sifting. A precondition of using the tool in hiring is that all hiring managers quality assure the AI's scores before shortlisting candidates for interview.

The AI tool provides both a numerical score for each application and a brief rationale for the score given. Sarah reviews each of the scores and the rationale, judging that both sound plausible, so she doesn't see any need to change scores.

This seems reasonable to her – why would her organisation encourage using it for this purpose if it wasn't trustworthy or reliable? It also allows the team to save time on the recruitment exercise.

This might not lead to a detrimental outcome if one person does it, but if second assessment becomes more of a "tick box" exercise for all recruitment, this could lead to discrimination embedded within recruitment processes and unequal labour market outcomes if the underlying algorithms are biased and people simply don't know how to detect bias.

Limitations of existing approaches to AI safety

There are broadly three common approaches to AI safety:

1	2	3
De-risking the AI tool itself using technical measures and guardrails	**Ensuring human oversight of AI – a "human in the loop"**	**Assigning risk ownership to the users (or overseers) as part of an overarching AI governance strategy**
Examples include meta-prompts with and without public input[16], reinforcement meta learning from human feedback[17] and AI Red Teaming[18].	This approach – known as "human in the loop"[19] – puts the onus on the user or the person providing oversight to ensure AI is used appropriately and responsibly.	Examples include disclaimers, fair use policies and Terms and Conditions (T&Cs) which require users to agree to conditions such as "I will check my outputs before using them" or "I understand that I am responsible for the appropriate onward use of the outputs".
Whilst these approaches are valuable, they each have limitations:		
		
De-risking tools themselves, including the underlying AI models, is important, but can only take you so far. There will be many 'hidden' risks these approaches don't address or which may backfire.	People can be ineffective at judging the quality of algorithmic outputs and determining whether and how to override outputs[20]. It is difficult for humans to be "in the loop" without the right conditions, such as enough time, relevant expertise and the psychological safety or authority to critically appraise or challenge AI's outputs.	**"May contain nuts"** approaches - focused on providing information of often solely technical risks – are unlikely to be effective alone due to issues such as low readership[21].

Whilst these approaches are valuable, they pose some clear limitations[22].

Technical solutions like higher quality training data may reduce algorithm bias, while meta-prompts[23] could prevent users from eliciting information about building weapons. However, these approaches cannot address the hidden risks illustrated in our examples in the section above —for example, Marco's employee burnout or Robert's unrealistic expectations about AI capabilities. You cannot reduce a person's concern about their job security or job quality when their organisation rolls out an AI tool through the technical development and iteration of an AI tool or service.

Recognising that technical fixes often cannot fix non-technical risks, AI safety research often proposes keeping a "human-in-the-loop" – a form of human oversight[24 25]. However, human oversight of AI outputs without sufficient wraparound interventions won't be sufficient either as users may lack the necessary expertise, time or authority to critically assess or challenge AI outputs, as in Sarah's example.

Figure 1 below provides a handful of other examples of barriers that users may experience in providing effective oversight.

Research shows that even experienced professionals can struggle to effectively oversee AI systems. For example, studies of judges reviewing algorithmic bail recommendations found that 90% of human overrides actually introduced more bias rather than reduced it[26]. Just as many people with nut allergies often miss disclaimers on food containing nuts (with fatal consequences), people will miss disclaimers and Terms and Conditions (T&Cs) on AI tools[27].

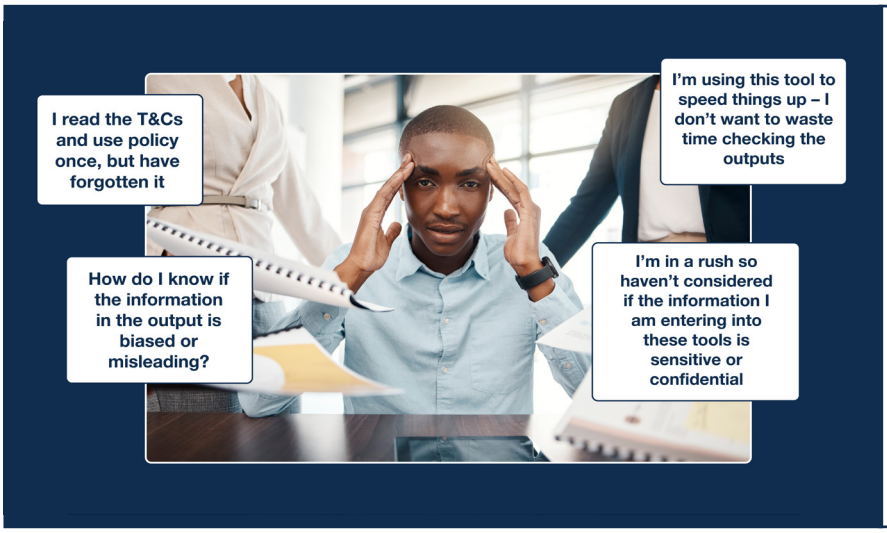


Figure 1: Examples of challenges which show why 'human-in-the-loop' is not a fix-all solution

As these limitations of each of the three existing approaches to AI safety show, **there is a limit to what technical fixes and human oversight can do to mitigate risks in the wide range of ways in which people may use AI tools.**

While some studies have explored what could be classified as 'hidden risks' in GenAI[28 29] , the field currently lacks a comprehensive and practical framework to systematically identify what risks to look for and when for any given AI-powered tool. This gap exposes the entire field of AI risk assessment to bias. Without structured criteria, risk assessment becomes influenced by organisational politics, media attention, individual preferences, or tailored towards specific domain expertise, rather than potential impact.

A new approach: surfacing ‘hidden’ risks

Just as the aviation industry dramatically improved safety, and as a result, consumer confidence, by building a better understanding of human factors involved in air safety, organisations that anticipate and plan for the “hidden risks” of AI roll-outs will be better positioned to realise AI's full potential and mitigate potential negative unintended consequences.

The crucial insight from these disciplines is that any intervention—particularly involving new technology—inevitably contains "unknown unknowns". Rather than reacting to problems after they materialise, we should proactively anticipate potential risks and implement effective preventative measures in advance.

This requires a systematic understanding of the likely **underlying causal mechanisms that lead to unintended consequences** coming about as a result of AI use: these **causal mechanisms are the day-to-day decisions and actions taken by individuals, teams, and organisations (see Figure 2)**. Even when well-intentioned, these decisions can accumulate into unforeseen consequences as people respond to situations in ways that seem reasonable at the time.

A new approach: surfacing ‘hidden’ risks

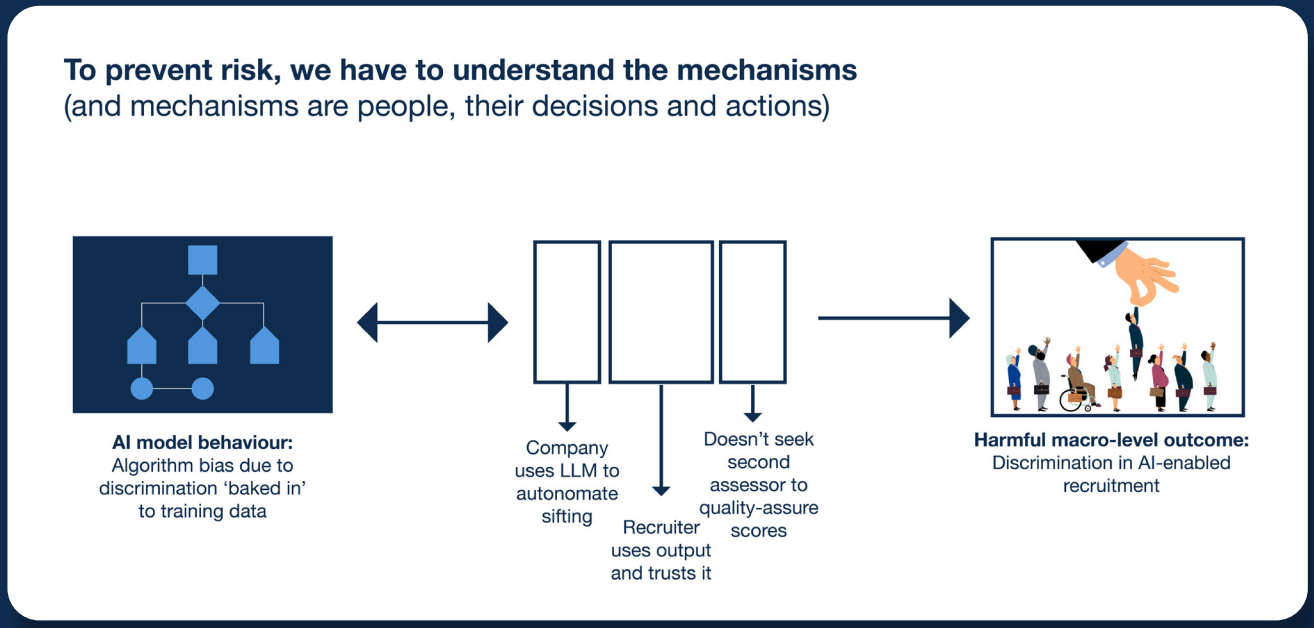


Figure 2. To prevent unintended consequences, we have to understand the mechanisms which create ‘hidden’ risks and which can lead to negative consequences and outcomes.

For example, Figure 2 illustrates the causal mechanisms in the context of Sarah’s example previously covered, whereby Sarah has access to an AI tool to help her to sift through job applications for a new role in her team.

It is true that algorithmic bias makes it technically plausible for discrimination to be ‘baked in’ to the tool. However, it’s not the tool that hires people based on discriminatory outputs. Rather, discriminatory recruitment arises from the way Sarah and her organisation use AI tools in recruitment. For example:

- Sarah’s company decision that procuring an AI tool to help with application **sifting would be an appropriate use case** (for example, compared to other ways to make the recruitment process more efficient);
- Sarah’s organisation **making the training they’ve provided for this tool optional** for Sarah to attend rather than a pre-requisite to tool access;
- Sarah having **undue trust in the outputs of the tool so using it without critically appraising scores meaningfully**, in part due to her organisation’s provision of it - why would her organisation provide it for this reason if it wasn’t trustworthy or reliable?
- Sarah **decided not to seek a second assessor to quality-assure** the scores for the recruitment as a result of her trust in the outputs.

The Mitigating 'Hidden' AI Risks Framework

Surfacing these risks using the Mitigating 'Hidden' AI Risks framework

To help to surface these risks, **the Mitigating 'Hidden' AI Risks framework identifies six categories of 'hidden' behavioural and organisational risks that organisations may face during their AI roll-outs** (see Figure 3).

We developed this framework during our research for Government Communications' AI tool Assist: it builds on existing AI risk typologies (e.g. by MIT[30]) and was refined through extensive stakeholder engagement to ensure it would be useful for wider applications (see the section titled 'Scope and background to this guide' for further information on our method for developing the framework).

As AI technologies evolve and their adoption grows, we will continuously review the framework and iterate on it based on evolving risk assessments.

Figure 3. Six categories of 'hidden' risks arising from organisational AI roll outs

Quality Assurance	Risks arising due to people using inaccurate or average quality outputs in their work.
Task-tool Mismatch	Risks arising due to the use of tools for purposes for which they weren't designed or which it doesn't perform well at.
Perceptions, Emotions and Signalling	Risks arising due to emotional responses induced by AI roll out, people's perceptions and attitudes about AI or the signals sent by an organisation's adoption/use of AI.
Workflow and Organisational Challenges	Risks arising from the work required to embed AI in an organisation or changes to people's ways of working.
Ethics	Risks arising from violations or threats to ethical standards and norms or legal rights (e.g. Equality Act 2010), or that are not in line with organisational guidelines and codes of conduct.
Human Connection and Technological Overreliance	Risks arising from reductions in, or removal of, humans from roles or functions or the overreliance on technical solutions for complex problems.

Six categories of ‘hidden’ risks arising from organisational AI roll outs

These categories have been developed to help you to reflect on some of the causal mechanisms – the types of actions and decisions made by people – by which higher order risks like inequality, discrimination and environmental harms could arise from your tool.

In the section below, we have developed sets of prompting questions for each category to help teams anticipate whether, when and how these risks might occur for your specific tools in your specific context.

You're essentially conducting a "pre-mortem" for GenAI implementation rather than waiting for the "post-mortem" investigation after something goes wrong. To go back to our aviation analogy, instead of retrieving the black box after a crash to understand what happened, you're trying to anticipate what that black box might record if things were to go wrong with your GenAI implementation.

This proactive approach allows you to:

- Identify potential failure points before they occur
- Understand the possible causal chains that could lead to problems
- Implement safeguards and monitoring systems for early detection – systematically, not ad-hoc
- Develop mitigation strategies for various risk scenarios

This preventative mindset shifts the focus from reactive problem-solving to proactive risk management. This is particularly valuable when dealing with powerful, rapidly evolving technologies like GenAI where the consequences can be far-reaching and difficult to reverse once they occur. It also ensures that risk assessment is systematic, which is vital for ensuring that our detection efforts don’t become subjective and potentially biased.

1. Quality Assurance

Risks arising due to inaccurate or average quality outputs

Prompt Questions

- Do they have sufficient subject matter expertise to know if the AI output is accurate or high quality?
- Are there likely to be time pressures which hinder people's ability for them to quality assure outputs?
- Could the quality of work delivered by teams be negatively impacted by people using the tool?
- Could there be pressure to reduce team size or change team expertise based on assumptions about the capability of the tool (the quality and/or efficiency of outputs)?

How could these risks threaten your AI solution's success in delivering positive outcomes?

Examples:

- If people perceive the outputs as low quality they may be less inclined to use tools
- If people dislike quality assuring outputs (e.g. it doesn't interest or excite them), this may deter them from quality assuring or even from using AI
- If poor quality outputs go into the public domain, this could result in reputational damage for your organisation and loss of trust in use of AI

What steps could you take to mitigate risks and optimise the impact of AI?

Examples:

- Test the tool to assess how effective it is at performing specific tasks compared to humans
- Let teams know what types of tasks are likely to require the most quality assurance
- Implement comprehensive AI literacy training programmes for all staff that allow staff to learn by doing and learn from others
- Provide comprehensive briefings and demonstrations for leaders, ensuring that these cover both strengths and limitations
- Encourage hands-on experience with the tools for decision-makers
- Establish systems for monitoring tool performance and impact on work quality
- Identify areas where human expertise remains crucial
- Foster a culture of continuous learning and adaptation to emerging AI technologies

2. Task-tool mismatch

Risks arising due to the use of tools for purposes for which they weren't designed or at which it doesn't perform well

Prompt Questions

- Do people know, or do you anticipate people will know, what the purposes or goals are of your AI tool?
- Have you appropriately informed your users about the tasks your tool is appropriate for, or do you have a communications plan in place to appropriately inform them?
- Could people use the tool for purposes/goals which it wasn't designed for and/or isn't good at? While experimentation is a core part of being innovative, if the tool is used in ways it is not optimised for, this could create risks.
- What positive or negative consequences could arise from these uses?

How could these risks threaten your AI solution's success in delivering positive outcomes?

Examples:

- If people perceive the outputs as low quality (because the tool isn't optimised for that use case) they may be less inclined to use tools
- If poor quality outputs go into the public domain, this could result in a loss of trust in use of AI, making it harder to roll out AI for purposes for which it is genuinely useful
- Poor use cases could limit the positive impact of AI and reduce the potential to build early successful case studies to demonstrate AI potential

What steps could you take to mitigate risks and optimise the impact of AI?

Examples:

- Define problems clearly before developing solutions, ensuring that any implemented tools directly address identified needs rather than being adopted for their own sake
- Make the tool better at the tasks that people are using it for
- If your tool is not good for a task people are using it for and you cannot make it better for that task (at least immediately), ensure you direct them to other tools which are available and suitable for them to use for that use case (i.e. as a substitution, so they use your tool less for that need)
- Take steps to systematically test what the tool is and isn't good at (learning from users as much as possible)
- Let people know what tasks the tool is and isn't good at
- Ban and monitor malicious uses
- Foster a culture of continuous learning and adaptation to emerging AI technologies

3. Perception, Emotions and Signalling

Risks arising due to emotional responses induced by AI roll out, people's perceptions and attitudes about AI or the signals sent by an organisation or institution's adoption/use of AI

Prompt Questions

For internal, back-office tools:

- What emotional response might we observe from staff from rolling out the tool and what behavioural consequences could this cause?
- Could staff perceive the tool as presenting a threat to their job, the type of work they do or skills/ expertise?
- Does the tool replace/augment tasks that people enjoy doing rather than dislike doing? Could this reduce uptake of the tool and therefore limit the positive impact AI could have?
- How might staff with a particularly positive or negative attitude towards AI respond?
- What might an AI roll out signal about the value or importance your organisation places on its staff?
- What might the roll out of the tool imply about your organisation's priorities? (for instance, could it signal that efficiency and speed is more important than quality?)

For public-facing tools:

- What expectations, attitudes or perceptions might the public form as an outcome of your organisation using the public-facing tool?
    - How might it impact the public's expectations or attitudes towards your organisation's services?
    - How could it impact trust in your organisation, or perceptions of your organisation's competency?
- What emotional response might we observe from members of the public using the tool and what behavioural consequences could this cause?

For all tools:

- How might wider narratives about the use of AI in other sectors (e.g. articles written by companies saying that they have built tools to replace specific jobs) influence people's views? How might changes in political or social context shape these views?

How could these risks threaten your AI solution's success in delivering positive outcomes?

Examples:

- If people are concerned tools could replace their jobs, they may be reluctant to adopt and use them
- If the public do not have trust or confidence in the use of AI tools then this could make it harder for organisations to realise the potential benefits

What steps could you take to mitigate risks and optimise the impact of AI?

Examples for internal, back-office tools:

- Take a human-centred approach to AI implementation. For example, conduct thorough user testing to identify and address potential negative reactions, create a bottom-up as opposed to top-down approach to tool development whereby tools are designed according to the preferences and priorities of staff (research from Wharton also suggests that staff will be the best group of people to identify the most valuable use cases), design tools to perform the tasks that staff least want to do themselves (or tasks which staff actively say they would like support with)
- Ensure that communications provides a balanced perspective of AI, drawing on AI experts to build trust in the objectivity of your communications
- Provide clear pathways for users to report issues or concerns, creating a safe environment for people to feel comfortable raising these concerns
- Provide clear information about how AI will impact roles and responsibilities
- Offer upskilling opportunities to help staff work alongside AI tools
- Clearly articulate how AI adoption aligns with broader organisational objectives
- Demonstrate how AI tools can improve both efficiency and quality of work
- Regularly report on the outcomes and benefits of AI use, beyond just efficiency metrics
- Anticipate, monitor and mitigate risks to reduce the likelihood that tools will cause harms that could undermine employee and/or public trust

For public or customer facing tools:

- Develop user-friendly interfaces and clear explanations of AI tool capabilities
- Provide options for human interaction alongside AI-driven services
- Regularly gather and act on public feedback about AI-powered services
- Offer alternative service options for those uncomfortable with AI-driven solutions (not just those with accessibility needs)
- Ensure transparency about when and how AI is being used in your services
- Implement robust safeguards and communicate these to the public
- Showcase successful AI implementations and their benefits to the public
- Show how your organisation is working with industry experts to ensure AI is being used where it can have the most positive effects
- See also section on “Ethics”

4. Workflow and Organisational Challenges

Risks arising from the work required to embed AI in your organisation or changes to people's ways of working, including patchy adoption.

Prompt Questions

- What are the barriers to AI adoption in your organisation? Do you know what these are? Do you have ways of identifying them?
- Do you have the resources and plans in place to support staff to adopt and sustain use of AI tools?
    - Do you have plans in place to motivate people to use tools, and reassure them of any concerns they may have?
    - Do you have plans in place to build organisational and staff capability to use AI tools?
    - What practical barriers might hinder people from using tools and how will you mitigate / remove these?
- Are there parts of your organisation that may struggle to adopt AI or specific groups of people who may struggle to adopt?
- To what extent could low or patchy adoption of AI negatively impact your organisation?
- For example, could low adoption hinder your ability to deliver efficiently relative to others that do adopt AI?
- Will infrastructure, systems and teams be able to cope with the uptake of the tool? (for example, if adoption happens at pace)
- Are there any ways in which objectives (e.g. to drive efficiency) could be undermined, for instance, due to additional behaviours that people may have to undertake in order to embed the tool in their workflow or teams?
    - What additional tasks might teams need to undertake in order to make best use of the tools?

- For example, could the tool add to the time required for people to complete tasks? (for example, if training and QA is required)

- Are there any ways in which the tool could reduce job satisfaction and motivation?

  - For example, could the tool replace easy tasks and leave people with a high volume of cognitively demanding tasks that exceeds people's cognitive loads? Or could the tool replace tasks that people most enjoy doing?

- Do organisational leaders understand the strengths, limitations and appropriate applications of the tool?

- Could use of the tool create dependence and/or erosion of skills that might need to be retained by humans? For example, if AI tools become expensive to access or unavailable due to malicious attacks?

- How might introducing the tool reduce incentives or introduce barriers to collaboration across the organisation?

  - For example, might it reduce engagement with subject matter experts (e.g. research teams) internally or externally in ways that reduce quality of outputs (e.g. if the tools are imperfect)?

How could these risks threaten your AI solution's success in delivering positive outcomes?

Examples:

- If staff are unable to build the skills needed to maximise the impact of AI use (e.g. prompt engineering skills) then this could limit the positive impact of AI in your organisation
- If staff feel that AI is adding to their workloads rather than reducing it, they may be reluctant to adopt and use tools
- If user feedback shows that staff report AI replaces the tasks they most enjoy doing, this could reduce job satisfaction and result in a decline in performance and productivity/ efficiency
- If leaders do not understand the strengths/limitations of the tool, this could result in AI being deployed for poor use cases that could limit the positive impact of AI and reduce the potential to build early successful case studies to demonstrate AI potential

What steps could you take to mitigate risks and optimise the impact of AI?

Examples:

- Give teams secure access to tools
- Provide training and support in prompt engineering
- Identify and address training needs
- Identify and address concerns pro-actively
- Measure and track so adoption challenges can be identified and tackled as appropriate
- Put in place a plan for testing ways to boost adoption and sustained use of your tools
- Develop a phased rollout plan to manage adoption pace
- Gather user feedback to identify and address inefficiencies
- Map out what steps will be required for the organisation, teams and individuals to make the most of the new tools and account for these in roll-out plans e.g. by ensuring staff are given time to undertake training
- Design tools to perform the tasks that staff least want to do themselves (or tasks which staff actively say they would like support with)
- Involve staff in tool development to ensure it enhances rather than replaces satisfying work
- Encourage hands-on experience with the tools for decision-makers
- Establish a panel of trusted experts to advise leadership on AI capabilities and limitations
- Develop contingency plans for scenarios without AI tool access
- Ensure critical skills are documented and regularly updated

5. Ethics

Risks arising from violations or threats to ethical standards and norms or legal rights (e.g. Equality Act 2010), or that are not in line with organisational guidelines and codes of conduct.

Prompt Questions

- What perverse incentives might be created through use of your tool? Specifically, an incentive which produces unintended and undesirable results, often contradicting the goals it was designed to achieve.

- How might use of the tool impact public trust in your organisation?

- How could the use of the tool reinforce or exacerbate discriminatory beliefs or outcomes or existing inequalities?

    - For instance, could the tool affect the nature or quality of work of some groups of people in society more than others?
    - In cases where biased decision making already affects humans, could the tool increase the frequency, speed or extent of these biased decisions or outcomes?
    - How easy or quickly could you identify any harms arising from the tool(s)? Could it be spotted instantly or would it only come to light after it had been in operation/use for an extended period of time?

How could these risks threaten your AI solution's success in delivering positive outcomes?

Examples:

- Discriminatory outputs could cause harm to potentially affected individuals or groups.
- Discriminatory outputs or unethically sourced inputs could cause public uproar that could make it harder to realise the benefits of AI.

What steps could you take to mitigate risks and optimise the impact of AI?

Examples:

- To help build public trust, demonstrate accountability and transparency by regularly publishing accessible reports on AI tool effectiveness, issues encountered and how they have/will be rectified. This includes ensuring the public is informed that these reports have been published.
- To help mitigate unequal and/or discriminatory outcomes, establish systems to monitor for these outcomes and implement strategies to mitigate them. For example, ensuring there is diverse representation in AI development and decision-making teams. Another mitigation is identifying stakeholder groups who may be affected by, or may affect, the design, development and deployment of an AI tool, as a means to ensure meaningful inclusion of those who may be disproportionally at risk from the use of the tool (or its outputs) so that they can be engaged through the process of adoption and use.
- Measure equality of uptake (e.g. are people with certain protected and vulnerable characteristics adopting in lower numbers) so you can take steps to remedy discrepancies
- Identify what barriers some groups may face so that you can incorporate this into your awareness raising, strategic engagement and onboarding processes
- Regularly compare AI-assisted decisions with human-only decisions to identify discrepancies
- Create a safe environment so people feel comfortable raising concerns or risks
- Ensure critical skills are documented and regularly updated

6. Human Connection and Technological Overreliance

Risks arising from reductions in, or removal of, humans from roles or functions or the over reliance on technical solutions for complex problems.

Prompt Questions

- Could a technological solution to the problem the tool aims to address, undermine support for non-technological solutions that may be more effective or better accepted by end users (for example, the public or your customers)?

- Could use of the tool result in a loss of skills/specialist expertise that could be considered particularly important or meaningful to the the public or your customers' specific communities? (For example, the industrial revolution led to a decline in heritage skills such as stonemasonry and thatching which are needed to preserve the UK's history and heritage)

- Could removing or reducing access to humans result in negative unintended consequences?

    - For example, are there tasks, insight, expertise or interpersonal engagement that only a human can provide or fulfil?
    - If public-facing, what impact could removal of humans have on vulnerable people or those who have accessibility needs? Have you done research with these populations to explore and de-risk this?
    - How might thethe public or your customers feel and respond to the reliance on AI for previously human-facing tasks? See Perceptions, Emotions and Signalling risk category.

- Could use of the tool reduce incentives or create barriers to teams or people working with one another in ways that could reduce the quality of people's experiences and/or joy experienced from human connection?

How could these risks threaten your AI solution's success in delivering positive outcomes?

Examples:

- Poor AI use cases could undermine support for AI use in your organisation, reducing the positive impacts AI could have
- Declines in staff motivation could result in losses in productivity and performance

- Poor staff or customer feedback about utility of tools could also undermine projects to roll out AI, making it harder to realise the potential benefits
- Overreliance on AI may reduce collaboration in your organisation

What steps could you take to mitigate risks and optimise the impact of AI?

Examples:

- Identify areas where human expertise remains crucial (in the broad sense - crucial for doing the task effectively or to ensure there is trust and buy-in)
- Keep a human in the loop where it is assessed as being crucial
- Define problems clearly before developing solutions, ensuring that any implemented tools directly address identified needs rather than being adopted for their own sake
- Engage end users or those impacted by the potential use of AI for a use case in decisions that could involve the removal/replacement of a social solution for a technological one to ensure their views and preferences are at the heart of all decisions (e.g. replacing a call centre with a chat bot or providing a therapy app instead of an in-person counselling).

Step-by-step: How to use the Mitigating Hidden AI Risks Framework

This section shows you how you can use our framework in your team or organisation to anticipate and mitigate risks for your own GenAI project, with examples of how we used the tool for Assist.

There are five key steps:

1. Set up a multidisciplinary and diverse working group
2. Surface potential hidden risks for your tool
3. Review and prioritise risks
4. Monitor and develop mitigation strategies for your risks
5. Implement ongoing monitoring and review mechanisms

There are multiple resources in this toolkit to help you complete these steps, including:

- The step-by-step guide detailed in the pages below;
- The Hidden Risks framework with prompt questions and example mitigations above;
- The risk register spreadsheet published alongside this guide, which you can use to document and monitor the risks you identify for your tool and your mitigations for these risks.

Step 1. Set up a multidisciplinary and diverse working group

Form a working group from across your team and/or organisation. Bringing together a diverse group will support you to anticipate behavioural and organisational risks you may not identify alone from your own area, skillset, perspective or viewpoint.

How you can do this:

- Ensure your group is diverse in skillsets, backgrounds and technical skills, including both technical and non-technical stakeholders. The UK Government's Service Standard highlights multidisciplinary teams as a key principle for delivering a service effectively[31].
- Consult and update senior stakeholders on progress. This will ensure senior leaders have governance over and develop a good understanding of the potential risks and planned mitigations for your roll-out. If you are a senior leader, refer to our Tips for Leaders in the accompanying guide, The People Factor: A human-centred approach to scaling AI tools, for more information about how you can support and safeguard your organisation's AI roll-outs.
- Get the right expert input by following best practice guides like this one and/or seeking out specialists in your organisation, especially if you don't have direct access to specialists in your own teams.

How we applied this to Assist:

We assembled a multidisciplinary team with diverse expertise and perspectives to ensure we approached potential risks for Assist's roll-out holistically, going beyond guidance outlined in the UK Government Service Manual[32].

Our team for Assist included:

- Product Manager
- Digital experts
- Data Scientists
- Behavioural Scientists
- AI Engineers
- Developer
- Evaluation specialists
- Service Designer
- Human Resources (HR) specialists
- Learning and Development specialists
- Communications experts
- Service Owner
- Delivery Manager

You might not have access to some of these skillsets in your own team. Notably, native access to these skills and capabilities may not necessarily be needed all the time, and can be drawn in from your wider organisation(s) or through networks.

For example, although we had in-house behavioural and data scientists, we brought in someone who was experienced at digital product delivery from the UK Government's Digital, Data and Technology Profession. Moreover, when we developed our training offer, we consulted Learning and Development specialists within the wider Government Communications organisation, but they did not work continuously on Assist's delivery.

Step 2. Surface potential hidden risks for your tool

Use the prompt questions in each risk category to spot potential "hidden" problems with your AI project.

How you can do this:

- Get a good understanding of how your users are using your AI solution, or how your potential target users may want to use it.
- Interrogate the prompt questions in the toolkit yourself or within your team. For example, a 'Task-Tool Mismatch' risk could arise due to ineffective or defective organisational implementation, such as if a senior leader requests that a tool be used for a task it is not appropriate for, and staff do not or are unable to challenge this.
- You do not have to answer all of the questions – they are intended as a guide or prompt to help you think about how your target users might or could, in future, engage with your tool in ways you do not anticipate or intend them to.
- Ensure you consider the ethical risks. While our toolkit explores this as one of the six risk categories, the Alan Turing Institute's Workbook on AI Safety in Practice provides a very good guide to consider ethics in a safety context (e.g. reliability, performance, robustness, and security)[33]
- Document any and all potential risks in a long-list, including hypothetical scenarios and 'what-if' situations, regardless of whether you perceive they're likely to happen or be true for your AI roll out. Ensure you consider both immediate and longer-term impacts.
- We've developed a risk register template, published alongside this guide, that you can use to catalogue your long-list of identified risks.
- Engage with wider resources and tools to support you to identify risks across specific risk categories related to your AI solution. For example, for quality assurance risks, the UK Government's AI Assurance Toolkit[34] provides a practical framework for surfacing specific AI assurance risks and identifying suitable assurance techniques in combination.

How we applied this to Assist:

We surfaced potential risks for Assist by:

- Surveys and 1:1 interviews with users to understand their perceptions of the tool and what they were using it for
- Analysing random samples of early user prompts (anonymised)
- Using the prompt questions that we created to consider how these desired and actual use cases could backfire – for example, how could things go wrong at different levels of an organisation if more people used the tool in this way?
- For example, we found that many of our early users were using the tool to summarise documents and develop first-drafts of communications plans. As a result, we were able to ask, is the tool good for that task? If not, can we make it better at doing that task, thereby reducing risk?
- Workshops and discussions with colleagues across our team to explore the kinds of risks they felt could arise.

Step 3. Review and prioritise risks

You then need to review and prioritise your risks in order to help you identify a viable strategy for mitigating them. You can do in any way that works for your team or organisation – we've created a risk register template that you can use to help you prioritise your long-list of identified risks.

How you can do this:

- Assess each risk based on its likelihood of occurring and the potential severity of impact if it did occur. You may also want to consider other prioritisation criteria relevant to your organisation.
- Create a prioritised list of risks to progress further. You can group these by the risk category they sit under, as a means for making this manageable to govern and review.

How we applied this to Assist:

As a team, we identified over 100 hypothetical risks that could arise from rolling out Government Communications' AI tool, Assist. This meant we needed to find a way to make this list more manageable to review, prioritise and, latterly, monitor. To do this, we developed and used our own risk register template, which we have published alongside this guide, which helps to catalogue the risks and triage them by a range of filters.

We held workshops as a team to triage the hypothetical risks, prioritising them by our expectations of likelihood of occurrence and severity of impacts. Understandably, it is hard to anticipate whether a risk is likely or how impactful it will be, so we continually reviewed and updated our assessments as time went on.

This helped us to streamline our 100 risks into 33 priority risks across the six risk categories. As time went on, we refreshed and updated our priority risks, demoting some we felt did not materialise or to which we had sufficient mitigations, while adding new risks as and when they emerged which we felt merited monitoring.

Step 4. Monitor and develop mitigation strategies for your risks

Having surfaced your risks, you can now decide how to monitor and mitigate them. This is covered in the next section in more depth, for each of the risk categories. This involves considering what activities need to be put in place to reduce the likelihood of the risk emerging, and can include things such as changes to your AI solution or service, training and support delivery and putting in place policies and processes.

Guiding principles for making 'human-in-the-loop' more effective

As previously explored, there are limitations to relying on human oversight alone (known as 'human in the loop'). This means that there are other approaches required to ensure people have the right conditions to be effective "in the loop" and ensure we can maximise the impact of tools while minimising risks[35].

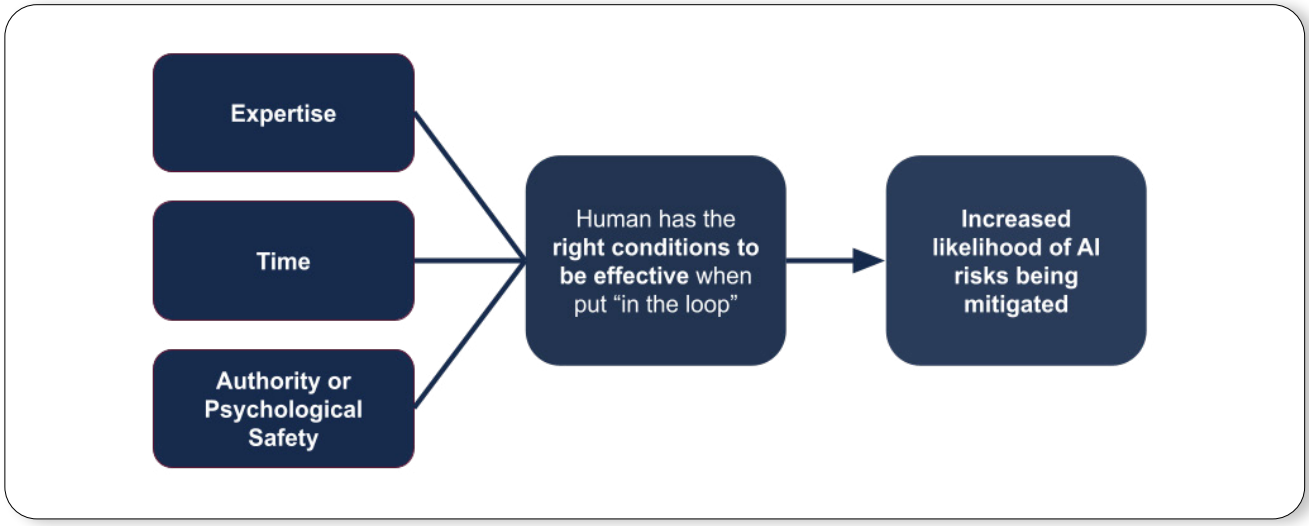
Based on our work on Government Communications' AI tool, Assist, we've identified three guiding principles we follow for our mitigations to ensure that our users have the right conditions to be "in the loop":

1. **Have relevant expertise:** The individual overseeing the AI outputs should have the necessary knowledge and skills to critically evaluate the generated output of a tool. This includes a basic understanding of the underlying algorithm, data sources, and potential biases inherent in the AI system.

2. **Be given adequate time for review:** Sufficient time must be allocated for the human reviewer to thoroughly assess the quality of the AI outputs. Rushed assessments can lead to oversight of critical issues, increasing the risk of erroneous or ineffective decisions based on flawed information.

3. **Have the authority to challenge outputs or how an AI tool is being used:** The human in the loop should have the psychological safety and authority to question and challenge AI-generated outputs or AI use cases, especially when risks or ethical concerns arise. This includes having the seniority or organisational support to escalate issues as needed.

Figure 4. Three principles to ensure that humans have the right conditions to be "in the loop"



How you can do this:

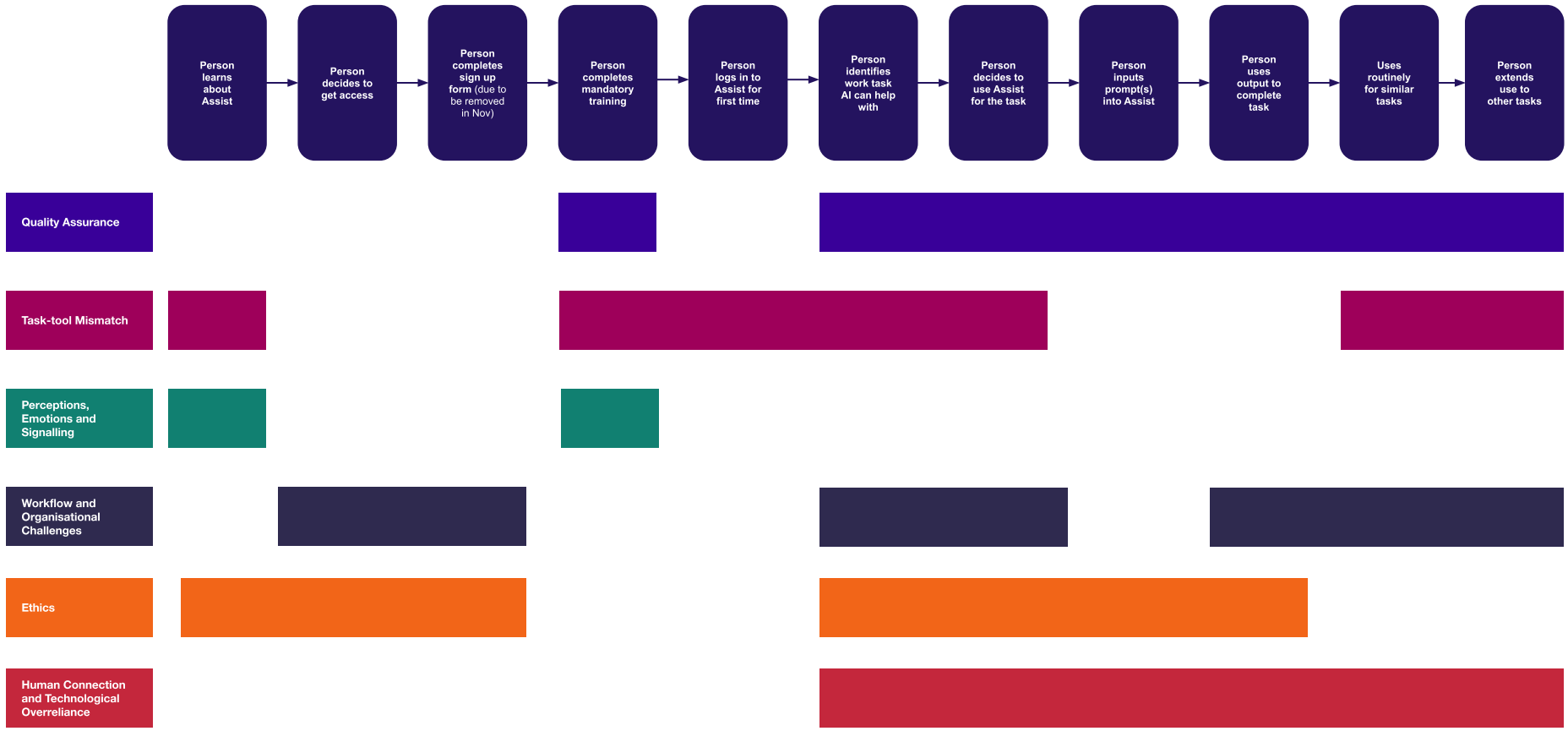
- Appoint a team member to oversee each risk category and coordinate responses. This approach distributes the workload, integrates risk management into daily tasks, and allows team members to develop expertise in their assigned risk area. Additionally, it allocates dedicated time and value to risk monitoring, while empowering team members to challenge assumptions and ask critical questions.
- Identify ways in which you can measure and monitor each prioritised risk. For example, could you use data collected within your organisation already, or use user research you're conducting[36] (e.g. surveys, interviews, testing sessions) as an opportunity to learn more about your identified risks and their prevalence in your roll-out. By monitoring risks, you can see whether mitigations you put in place are improving outcomes or reducing the likelihood of the risks emerging – if not, you can iterate your approach and test out other strategies.
- Map your risks against your users' behavioural journey to adopting and using your AI solution, so you can understand at what points you need to introduce 'protections' or mitigations to reduce their risk. See our example of how we've done this for Government Communications' AI tool, Assist.
- Develop mitigation strategies for each prioritised risk and putting them in place. To help you to develop strategies, the toolkit below outlines some examples of mitigations you could put in place for each risk category, depending on the risks you've identified for your AI roll out.

How we did this for Assist:

Having identified and prioritised risks, we worked together to identify an approach to monitoring and mitigating them.

- We mapped our key risks against how users interact with Assist, right from when they first find out about the tool, right through to possible adoption and regular use. This helped us to spot where to place safeguards.

Mapping risks across the Assist user journey



For example, early research indicated that users might doubt Assist's usefulness early on, so our launch communications highlighted diverse benefits and examples relevant to different teams. You can find more information about mapping your users' behavioural user journey to using your AI solution in the companion guide, The People Factor: A human-centred approach to scaling AI tools, published alongside this publication.

- The team identified how to track key risks using existing data (like login rates) and where we needed new monitoring methods (such as staff survey questions about Assist usage and attitudes).
- We assigned one risk category to each team member to ensure clear accountability.
- We then worked together in 'design thinking' workshops to develop ideas for how we could address these risks.
- We then implemented and tested mitigations for the risks, learning what works using our evaluation methods and iterating our approach. Mitigations included:
    - Bespoke training, support and engagement with senior leaders so they understand the technology, including both its benefits and the risks they should mitigate
    - Webinars and bitesize videos on how to get the most out of the tool, including what people should and shouldn't use it for to support responsible experimentation with the tool
    - Pre-built prompts tailored to Government Communications use cases to help users generate higher quality outputs, lowering the bar to entry for less experienced users of AI tools and supporting high-quality use
    - Collecting and monitoring data on equalities, conducting research with those with low/no use and delivering targeted mitigations to support equality of access and benefits of Government Communications' AI tool, Assist (you can find the questions we used in our primary research and user onboarding in our supplementary resources published alongside this guide)
    - Develop and continually iterate our mandatory training for the tool, to support users to have the right skills and knowledge to use the tool effectively

Step 5. Implement ongoing monitoring and review mechanisms

Hold regular risk review meetings where the team can openly discuss concerns, assess safeguards, and suggest improvements to the AI roll-out. Ensure it is discussed and agreed as a team how regularly to hold these sessions. Ringfencing time to specifically discuss risks and mitigations ensures they won't be neglected and will support your AI solution's successful implementation.

How you can do this:

In your risk review sessions, you can:

- Regularly assess the success and effectiveness of risk mitigations
- Establish a clear, open escalation pathway for emerging risks – over time, new risks may emerge which you may not anticipate and which will require mitigations or reactive responses
- Document lessons learned and successful mitigation strategies which demonstrate positive impacts (i.e. successful and safe scaling)

How we applied this to Assist:

The Government Communications Assist team meets every fortnight to review, discuss and, where appropriate, escalate risks we've identified and have been monitoring. This sits under our 'Optimise' phase of our 'Adopt, Sustain, Optimise' strategy, covered in our complementary publication: The People Factor: A human-centered approach to scaling AI tools. This is a session held to specifically safeguard time to discuss risk management, separate from our wider team progress updates on Assist.

In these sessions, we:

- Ask if anyone has anything to report from their own risk area, or anything they've noticed relating to another risk area.
- We raise it, discuss it and decide together how we're going to act – for example, we usually explore whether we need to collect more information about the concern or risk, or consider whether there is a mitigation we could put in place to address it.
- Share direct user quotes and feedback during team meetings to keep user voices front and centre.

Tips for Teams

1 Risk management and mitigation

Ensure risk beyond data security and GDPR compliance are anticipated, monitored, and mitigated through a comprehensive risk management strategy for AI implementation. Using the process and framework outlined in this guide can help you, alongside other AI safety resources, such as the Alan Turing Institute's online learning module on AI Safety in Practice, developed by their AI Ethics and Governance in Practice Programme[37].

2 Actionable training beyond a "may contain nuts" approach

Go beyond disclaimers and T&Cs and ensure that user training is delivered continually, using a range of formats and touchpoints, and provides specific guidance and advice on how to mitigate the risks for specific tasks or job roles. For example, ensure your training does not just flag the risks of AI without giving concrete instruction of whether and how a user can avoid them creating harms (e.g. algorithm bias).

3 Use case definition and risk assessment

While experimental use of AI solutions can drive innovation, ensure that the intended uses of AI tools are clearly defined and potential pitfalls or unintended consequences are thoroughly analysed. If providing general tools, ensure you know how and what it is being used for so that you can steer people away from poor use cases towards better ones.

4 Leadership understanding

Ensure that your senior leaders understand the capabilities and limitations of your AI solution, so that they can be responsible advocates. Tools need to be deployed for the right tasks, and this can only be achieved if individuals, teams and their seniors understand what are "good" and "bad" (or "less good") uses for your AI tools.

5 Robust impact measurement

Ensure that robust systems are in place to measure efficiency and quality impacts of AI adoption, providing confidence in the accuracy and reliability of these estimates. This goes beyond measuring success metrics alone.

6 User and “non-user” centred design

Ensure that inclusive feedback mechanisms are established to gather input from both AI tool users and non-users, avoiding selection bias and designing tools that cater to the needs of users beyond your ‘early adopters’.

7 Safeguarding time and resource to risk management

Ensure that your team is adequately resourced and safeguards time to balance ongoing delivery tasks with essential activities such as risk assessment, impact measurement, and user feedback analysis. This is a critical condition necessary to enable you to monitor emerging risks, so you can respond effectively and proactively, rather than reactively and ineffectively.

8 Multidisciplinary and diverse teams

Ensure that your team has diverse and multidisciplinary skills and perspectives. For example, including data scientists, machine learning experts, social and behavioural researchers, change management professionals and user researchers, to guarantee a well-rounded approach to AI implementation.

9 Ensure “humans in the loop” have adequate time, expertise and authority

Relying on human oversight alone to mitigate risks is a risk. When developing your risk mitigation strategies, consider whether your users have the relevant expertise to oversee and critically assess the AI outputs, adequate time to review them effectively, and appropriate authority or psychological safety to challenge outputs or the use of your tool for tasks it may be not appropriate for. This includes having the seniority or organisational support to escalate issues as needed. Do not rely on human oversight to mitigate risks - to be effective human oversight must be well trained and empowered.

Scope and limits of this guide

What kinds of AI solutions is this guide most relevant for?

Due to our focus on the behavioural and organisational risks which could arise when implementing and scaling AI tools within organisations, this guide is most relevant for teams implementing AI tools that are intended to be used directly by people – end users. This is in contrast with more automated, backend AI systems for which, while risk management is still important, typically involve less direct human access.
We have developed this guide based on our experience of rolling out Government Communications' tool Assist, which uses generative AI. However, the risks and mitigation strategies outlined in this guide will likely be applicable to any tool which incorporates wider forms of AI and which are used directly by people.

Is this guide applicable to organisations beyond the public sector?

This guide was designed whilst working on the roll out of an AI tool in UK Government. As a result, readers from the private sector may not recognise all of the examples provided.

However, our engagement with private sector partners suggests that some of the challenges and barriers faced in rolling out new AI tools and services within large public sector organisations are likely to be common to many other organisations, whether public or private. As such, we believe the general framework outlined could be useful to organisations and their teams within the private sector.

How did you develop the six risk categories?

Colleagues within the Applied Data and Insight Team based within Government Communications in the Cabinet Office undertook a comprehensive research programme to support Assist's development, implementation and scaling.

As explored in this guide, this work identified six categories of 'hidden' behavioural and organisational risks which could arise from organisations rolling out AI solutions.

This built on existing AI risk typologies (e.g. by MIT[38]) and was refined through extensive stakeholder engagement across the public, private and third sector.

The themes were developed by analysing a random sample of early user prompts submitted to Government Communications' AI tool, Assist, which involved screening them for potential unintended consequences – for example, how could things go wrong if people use the tool in this way?

These were then further iterated and validated to ensure applicability to wider AI-powered tools, with a focus on risks most relevant to organisational roll-outs of AI tools. This was achieved through extensive stakeholder engagement (recognised in the Acknowledgements) and by building on existing AI risk typologies (e.g. by MIT[39]).

Acknowledgements

This guide was written by Holly Marquez and Dr Moira Nicolson of the Behavioural Science Team in Applied Data and Insight (Government Communications, Cabinet Office).

To make this guide as applicable and useful as possible for people implementing and scaling GenAI-powered tools across the public and private sector, we engaged and consulted a diverse range of stakeholders.

We thank the contributions of these individuals and organisations. Their insight and expertise was instrumental in shaping the Mitigating Hidden AI Risks Toolkit:

Government Communications

Conrad Bird, Director, Strategy and Campaigns

Dr Amanda Svensson, Deputy Director of Applied Data and Insight

Robin Attwood-Martin, Head of Applied Innovation

Marcus Melton, Applied Innovation Lead and Assist Product Manager

Kiran Chahal, Applied Innovation Lead

Dr Ashley Poole, Assist Lead Developer and AI Engineer

Rishi Moulton, Insight and Evaluation Manager

Abby Wade, Applied Innovation Manager

Clement Yeung, Insight and Evaluation Officer

Hayley Higgins, Head of Member Strategy and Services, and the Member Strategy and Services Team

Steven Pirrie, Visual Designer

Kayleigh Lewis, Lead Content Designer

Acknowledgements

Wider Contributors

Professor Susan Michie, Director of the Centre for Behaviour Change, University College London (UCL) and Co-Director of Behavioural Research UK (BR-UK)

Deborah Morgan, PhD Researcher and AI Engagement and Analysis, Technology and Strategy Insights, Government Office for Science

Dr Peter Slattery, Researcher, MIT AI Risk Repository, MIT FutureTech

Ellie Haberlin-Chambers, Implementation Advisor, Department for Health and Social Care (DHSC)

Katie French and Charlotte Ryall, Senior User Researchers, Incubator for Artificial Intelligence (i.AI, Department for Science, Innovation and Technology)

Ann Borda, Ethics Fellow, The Alan Turing Institute

Professor David Leslie, Head of Ethics and Responsible Innovation, The Alan Turing Institute

Antonella Maia Perini, Research Associate, The Alan Turing Institute

Robecca Hogg, Senior Responsible AI Advisor, Defence Science and Technology Laboratory

Ed Butcher, Senior Principal Analyst, AI Concepts and Exploitation Team, Defence Science and Technology Laboratory (DSTL)

Stuart Hossack, Behavioural Insight Lead, HM Courts and Tribunal Service

Rebecca Furlong, Media Specialist, Construct Education

Feedback and collaboration

We welcome you and your teams' thoughts and feedback on this guide, particularly your experience applying this guide to help with scaling and de-risking your own GenAI tools. We also welcome opportunities to collaborate with other teams.

Input from others beyond Government Communications and the Cabinet Office has been and will continue to be invaluable to the project's continuous development.

Get in touch with us by email: gcs@cabinetoffice.gov.uk.

References

1. https://www.instituteforgovernment.org.uk/comment/whitehall-ai-government-pilots
2. https://doi.org/10.1007/s44206-022-00010-6
3. https://doi.org/10.1177/20539517231221758
4. https://doi.org/10.1162/99608f92.ad8ebbd4
5. https://www.bbc.co.uk/news/uk-65746524
6. Shapell et al (2006) Human error and commercial aviation accidents: A comprehensive, fine-grained analysis using HFACS
7. Bruno, Walker and Abujudeh (2015) Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction
8. In healthcare the figure is similar, with many medical tragedies taking place due to human factors. https://www.england.nhs.uk/wp-content/uploads/2013/11/nqb-hum-fact-concord.pdf
9. https://www.penguin.co.uk/books/56495/outliers-by-malcolm-gladwell/9780141036250 https://en.wikipedia.org/wiki/Outliers_(book)
10. Some AI safety research refers to these types of risks using other terms, such as human factors. For example, see: National Institute of Standards and Technology (NIST; 2022) Towards a standard for identifying and managing bias in Artificial Intelligence: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf
11. See for example: https://airisk.mit.edu/.
12. https://plato.stanford.edu/entries/thought-experiment/
13. "AI alignment" has been defined namely as "a field, focused on the technical project of ensuring an AI system acts reliably in accordance with the values of one or more humans"; see Xuan et al (2024). https://arxiv.org/pdf/2408.16984
14. "AI alignment" has been defined namely as "a field, focused on the technical project of ensuring an AI system acts reliably in accordance with the values of one or more humans"; see Xuan et al (2024). https://arxiv.org/pdf/2408.16984
15. Sharma et al. Towards understanding sycophancy in language models. https://arxiv.org/pdf/2310.13548
16. https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input
17. https://www.ibm.com/think/topics/rlhf
18. https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/
19. This approach focuses on putting the necessary human oversight and human in the loop processes in place to validate outputs in situations with high impact or risk#. This seeks to ensure that a person retains a "meaningful" final say over key decisions and maintains the AI as a "co-pilot". Risk ownership can then be placed on the users - or the human overseers - by their organisation. https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems, Roff and Moyes (2016) "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons." Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, Generative AI Framework for HM Government (v1).

References

20. https://www.nber.org/papers/w31747
21. For examples of evidence of low readership of T&Cs, see https://eprints.qut.edu.au/212718/
22. Jailbreaking is a well discussed risk so is not covered here. Jailbreaking occurs when someone circumvents in-built model safeguards, such as through clever prompting techniques. https://arxiv.org/pdf/2401.06373
23. https://www.prompthub.us/blog/a-complete-guide-to-meta-prompting
24. National Institute of Standards and Technology (NIST; 2022) Towards a standard for identifying and managing bias in Artificial Intelligence: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf
25. See page 33 of National Institute of Standards and Technology (NIST; 2022) Towards a standard for identifying and managing bias in Artificial Intelligence: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf
26. https://www.nber.org/papers/w31747
27. For examples of evidence of low readership of T&Cs, see https://eprints.qut.edu.au/212718/
28. https://www.microsoft.com/en-us/research/wp-content/uploads/2025/01/lee_2025_ai_critical_thinking_survey.pdf
29. https://www.science.org/doi/10.1126/sciadv.adn5290
30. https://airisk.mit.edu/
31. https://www.gov.uk/service-manual/service-standard/point-6-have-a-multidisciplinary-team
32. As per the Service Manual: https://www.gov.uk/service-manual/the-team/what-each-role-does-in-service-team
33. https://www.turing.ac.uk/sites/default/files/2024-06/aieg-ati-6-safetyv1.2.pdf
34. https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf
35. Green (2022) The flaws of policies requiring human oversight of government algorithms
36. In line with the Service Standard: https://www.gov.uk/service-manual/service-standard
37. The Alan Turing Institute: AI Safety in Practice Module: https://aiethics.turing.ac.uk/modules/safety/
38. https://airisk.mit.edu/
39. https://airisk.mit.edu/


Cabinet Office