



Department for
Science, Innovation
& Technology

Guidance

AI Risk Management Toolkit: guidance

Published 8 September 2026

Contents

Purpose of the toolkit

About AI risk assessment tools

Using these tools

AI risk identification

AI risk appetite

Quantifying AI risks

Quantifying risk likelihood

AI risks and treatment options

Additional resources

Appendix 1: Risk Identification Questions

Appendix 2: Risk Appetite Level Definition

Appendix 3: List of potential risk impacts

Appendix 4: Risk treatment suggestions



This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit nationalarchives.gov.uk/doc/open-government-licence/version/3 or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or email: psi@nationalarchives.gov.uk.

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.

This publication is available at <https://www.gov.uk/government/publications/ai-risk-management-toolkit/ai-risk-management-toolkit-guidance>

Purpose of the toolkit

The Artificial Intelligence (AI) Risk Management Toolkit is intended to support anyone involved in the design, operation, procurement and delivery of products empowered by or enabled by AI. The toolkit will benefit multi-disciplinary project teams when navigating AI adoption, which may include data science, engineering, project delivery, IT, change management and communications professionals.

A clear understanding and justification of AI risks and treatment strategies is key to using AI responsibly and ethically. The scope of AI projects can be wide and may include:

- integrating commercial solutions
- driving adoption across wide user bases
- building models in-house
- supporting internal and external operations

The AI Risk Management Toolkit is intended as a starting point for teams looking to implement good risk management practices.

About AI risk assessment tools

Our inability to clearly identify and assess risk, particularly the fear of unknown risks, can be a significant barrier to the adoption of emerging technologies. With AI, the potential risks to individuals and society are regularly highlighted in public discourse, so for the UK public sector to succeed in its ambition to realise significant benefits from AI, understanding and addressing these risks is critical, while taking a good risk in pursuit of benefit.

This Toolkit has been created to assist departments with identifying and managing risks associated with the development, deployment, procurement and use of AI solutions. The Toolkit aligns with the Risk Management toolkit (RMF) found in the Orange book (<http://www.gov.uk/government/publications/orange-book>), which establishes risk management principles and processes for government departments and other UK public sector organisations in their routine operations. The Toolkit is designed to implement Section D: Risk Management Processes (<https://www.gov.uk/government/publications/orange-book/the-orange-book-management-of-risk-principles-and-concepts#section-d-risk-management-processes>) of the Orange Book. These include:

- risk identification and assessment
- risk treatment
- risk monitoring
- risk reporting

Additionally, these tools are designed to work together with other established toolkits, including the Cyber Assessment toolkit (CAF) (<https://www.security.gov.uk/policy-and-guidance/introduction-to-the-cyber-assessment-framework-caf/>). This Toolkit is relevant throughout an AI system's lifecycle, from use case identification to retirement, to its use and adoption.

While AI is a relatively new technology, effective risk management shares principles with best practices for software development and data science risk management. An emphasis on data quality, model understanding and validation ensures effective management of arising risks. Teams should aim to highlight the commonalities by incorporating existing governance structures, and take advantage of existing expertise. The fear of unknown risks should be tackled by taking a proactive approach to risk where ongoing risk management is carried out during the lifecycle of an AI system.

Included within the Toolkit is:

- a guide to AI risk assessment which sets out best practice and builds upon the principles set out in the Orange book (<http://www.gov.uk/government/publications/orange-book>)
- a set of critical questions that will stimulate thorough investigation around areas of potential risk to expose where AI risks are hiding for a specific AI solution
- a workbook to record identified risks, their assessment and associated treatment actions, although an existing risk register may be used if sufficient information is tracked to ensure risks, mitigations, and owners can be observed and updated.
- an AI Risk Monitoring Dashboard which provides an overall perspective on the risk profile of the AI solution, as well as the chances of different degrees of success and failure

Using these tools

AI systems change and evolve during their lifetime, so there is no 'final version' to validate against. Ongoing risk management is required throughout the AI lifecycle, from use case identification to retirement. This is because:

- the ability to effectively address risks is maximised if they are identified at the outset, and a legal, ethical and robust by design approach is adopted
- the legal, regulatory and ethical landscape, and technical capabilities are continually evolving, so risks and their management will change over time.
- it's important to continuously re-assess risks and treatments as new risks may arise, or treatments may need changing as their efficacy decreases

A multi-disciplinary team, referred to as the AI risk management team, is required to effectively drive active risk identification and management activities. Your team should ideally be led by an identifiable individual known as an AI governance officer. An individual or a smaller team may hold multiple responsibilities within your risk management team to provide additional input to drive good risk management. These individuals include:

- AI governance officers, who are responsible for the implementation and oversight of processes, activities and policies relating to the use of AI within the organisation, directorate or team
- senior leaders, who are responsible for steering the direction and success of AI solutions through defining and inspiring a long-term vision and setting risk appetite and tolerance levels
- the data teams, who are responsible for managing the environment and process through which AI solutions are developed, produced and deployed, and manage the data architecture, flows and lineage to support data collection, preparation and processing
- AI practitioners, such as data scientists, software engineers, machine learning and AI engineers, who are responsible for the production of AI solutions which automate tasks and solve key problems that have been communicated from the organisation
- the security team, who are responsible for ensuring that AI solutions and their associated infrastructure and architecture function as safely as possible by monitoring and minimising the threat of cyber-attacks and ensuring compliance with requisite laws, regulations and policies
- legal and compliance professionals, who are responsible for ensuring that AI solutions are operating in alignment with applicable laws, regulations and organisational policies, including the mitigation of future legal risks as legal and regulatory landscapes evolve
- business domain experts, who are responsible for providing insights and expertise in the production and implementation of AI solutions within a specific context
- end users, who are both direct and indirect stakeholders who engage with and utilise AI solutions, such as colleagues who use AI solutions to support their role and responsibilities, external stakeholders who use AI products to access goods or services

The Government Digital Service (GDS) encourages departments to keep a central log of AI risks, and to share these with GDS and the Department of Science, Innovation and Technology's (DSIT) Central AI Risk Toolkit Team (CAIRF) team. This will help to prioritise risks where investment in causal mapping and designing treatment options would provide maximum benefit to departments.

AI risk identification

A risk is defined as a potential future event that affects your objectives and impacts the outcomes. Your multi-disciplinary AI risk management team should aim to identify as many potential scenarios as possible, however unlikely they may seem. A multi-

disciplinary approach is crucial for thorough and ongoing risk identification. AI systems involve various dimensions, including technical, ethical, legal, and social aspects. Your AI risk management team, as set out above, should include the relevant expertise required to identify risks collectively.

AI systems evolve over time, and new risks can emerge as the system is developed, deployed, and iterated upon. Initial risk assessments provide a foundational understanding, but the landscape of potential issues can and will shift. Changes in data inputs, model updates, shifts in usage context, or evolving external factors can introduce unforeseen risks. Therefore, risk identification should not be treated as a one-time task but rather integrated into the continuous management of the AI tool. This ensures that emerging risks are identified and addressed promptly to maintain the system's safety, reliability, legal and ethical integrity.

Reassessment of risks should occur as required. For example, you may consider reassessing risks when a machine learning model has sufficient monitoring to capture model drift and the underlying model is updated or after Alpha and Beta releases. Upon creating a live service, you should continuously monitor AI technical performance and further evaluate its ability through stress testing. Use real-world Alpha and Beta scenarios to monitor potentially unforeseen failures and analyse past incidents to predict future risks.

A list of risk identification questions is set out in [Appendix 1: Risk Identification questions](#) of the AI Risk Management Toolkit. This list is not exhaustive, as it aims to provide a starting point from which your AI risk management team can build upon using the principles set out in the [AI Playbook for government](#) (<https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government/artificial-intelligence-playbook-for-the-uk-government-html>).

In the AI Risk Management Toolkit, various categories are set out to direct activities and ensure a thorough investigation of potential risks associated with AI systems. These categories are:

- financial risks that could lead to financial losses for the organisation, including increased operational costs, expenses related to fixing or maintaining AI solutions, and potential financial implications from automated decisions
- legal and regulatory compliance risks related to failing to meet legal toolkits, regulatory regimes, and guidance relevant to the deployment of AI solutions, considering issues like data protection, equality law, and compliance with evolving AI-specific laws
- appropriate transparency and explainability risks related to the AI solution's lack of transparency or explainability, which considers whether users and those impacted by the AI solution understand how it works, its decision-making processes, and if they are aware that they are interacting with AI
- fairness risks of the AI solution being unfair, biased, or discriminatory towards specific groups, which considers whether the AI solution could negatively impact an individual's rights or create unfair market outcomes, including compliance with equality laws and guidelines
- accountability and governance risks associated with a lack of clear accountability and effective governance of AI solutions, which address issues around identifying who is responsible for the AI solution and its outputs and actions, establishing risk

management processes, and ensuring clear communication and roles throughout the AI lifecycle

- contestability and redress risks related to the ability of a user or affected parties to contest an AI solution's output or seek redress if they believe they have been harmed or wronged, including whether there are accessible and transparent redress mechanisms and procedures for remediating harm
- technical robustness risks related to the AI solution's ability to function reliably and maintain performance as intended throughout its lifecycle, including assessing data quality, model reliability, and the solution's ability to perform under unexpected situations or data
- security risks related to potential security threats arising from the AI solution's deployment, such as data poisoning, leakage, or cyber-attacks, which consider whether the organisation is more vulnerable to attacks due to the AI solution and if security measures are in place to protect against these risks
- risks to people and the environment concerning the potential impact of the AI solution on an individual's physical and mental well-being, and the safety of critical infrastructure and the environment, which assesses the likelihood of safety-related incidents and whether steps are taken to mitigate these risks

AI risk appetite

Before you plan risk treatments and mitigation options, the severity and likelihood of the risk need to be compared to the risk appetite. This risk appetite should be set out at an organisational level if possible, or at least across a wider team to ensure a consistent approach. Your risk appetite for AI should then ideally be aligned to your department's wider risk appetite and signed off at the departmental board level. Defining your risk appetite is considered to be the most challenging aspect of any risk toolkit, especially in AI's evolving field.

Without clearly defined and measurable levels of risk appetite, your users will invariably encounter a number of challenges and blockers when trying to manage risks for new AI tooling and solutions, which will eventually stifle innovation. Alternatively, a properly communicated and appropriate risk appetite can actively enable your organisation to achieve its targets.

The risk appetite set out in Appendix 2: Risk Appetite has been designed to align with the Orange Book (<http://www.gov.uk/government/publications/orange-book>). You should adapt and replace these examples when defining their level for each section.

Quantifying AI risks

Once you have identified your risks, their impact and likelihood must be scored between 1 and 5. For likelihood, a score of 1 means that the risk is highly unlikely to occur, and 5 means that it's highly likely to occur. For scoring impact, 1 would have a negligible impact, and 5 would have a catastrophic impact.

Quantifying risks can be complex, and the guidance below aims to support your team in giving a fair assessment of each identified risk.

Inputting the likelihood and impact scores into the workbook will provide the corresponding risk score. The risk score is the product of the likelihood and impact scores. Depending on the risk appetite of your organisation, your team may decide to apply a weighting according to its relative importance. For example, an internal-facing system may be less concerned about financial risks but more concerned about contestability, redress, or transparency risks.

Quantifying risk likelihood

Due to the novelty of AI projects, your department should combine qualitative and quantitative approaches throughout the AI development lifecycle to assess risk likelihood across the different risk categories.

All risks should be assessed against a standardised likelihood scale, such as given below.

Likelihood Level	Description	Probability Range
1 - Rare	Highly unlikely to occur	< 5%
2 - Unlikely	Unlikely but possible	5% - 20%
3 - Possible	Possible to occur at some time	20% - 50%
4 - Likely	Likely to occur in many circumstances	50% - 80%
5 - Almost Certain	Almost certain to occur	> 80%

The novelty of AI can bring challenges in estimating the likelihood of a risk occurring. There is a lack of historical data to derive assumptions from, so your AI project teams should use a combination of the below quantitative and qualitative analyses methods to estimate the risk likelihood.

Historical data analysis

This method gathers relevant historical data related to the service or product and analyses patterns and trends to understand the frequency and circumstances under which a risk might occur. Consider product and service specific metrics such as service uptime, frequency of use, historical user errors and bug frequency.

Model analysis

This method applies techniques to assess the bias, accuracy, and quality of AI models during the model development stage, feeding in assumptions about the risk likelihood. These include accuracy, precision and recall of a classification model, precision for a recommendation engine, and BERTscore for a language model.

Expert judgement

Engage with AI specialists, data scientists, risk management and legal experts to assess the likelihood of identified risks.

Experimentation and monitoring

In the Alpha and Beta stages, you should experiment and conduct user research in a safe environment to understand the frequency of risks manifesting, and iterate your assumptions as required.

Quantifying risk impact

When quantifying the impact of a risk, first identify if the impact is quantifiable or not. The table in [Appendix 3: List of potential risk impacts](#) sets out a list of example risks and whether they are quantifiable. If they are not quantifiable, you should consult with relevant stakeholders and experts to assign a sensible impact score to the risk. The list of potential risk impacts uses quantifiable and non-quantifiable categories to help you identify and evaluate the various implications of risks associated with AI technologies. Quantifiable impacts include measurable effects such as financial loss, operational downtime, and complaint volumes, while non-quantifiable impacts cover areas like strategic setbacks, morale, and public trust.

Included in [Appendix 3: List of potential risk impacts](#) is a table setting out impact levels for quantifiable risks. When a risk may result in multiple outcomes, you should assign weightings to each potential outcome to calculate an overall risk impact score, or use the highest impact score when calculating the risk score.

AI risks and treatment options

Treatment options aim to manage an identified risk to bring it within risk appetite or tolerance limits. Your treatment options should adhere to institutional standards

wherever possible, and may change and evolve during the lifetime of an AI system. Risk treatment options fall within 4 categories, with varying levels of impact and cost. These are:

- avoidance against any exposure to risk, which is usually suitable when both the potential impact and cost of mitigating are too high
- limiting risk exposure by taking some action to maintain risk levels within an acceptable range, and employing a blend of acceptance and avoidance
- transference, where all or a portion of risk liability and responsibility are shifted to the third-parties best placed to manage it, which may be appropriate if addressing a risk requires expertise or resources that are better provided externally
- acceptance, where risks are acknowledged, and potential impacts are accepted without taking further actions to mitigate or eliminate them, which is appropriate when an identified risk is within your organisation's risk tolerance

Selecting the appropriate treatment option for a risk is a highly technical activity. The approach will be very dependent on the risk itself and the context in which it has arisen, which is why a multi-disciplinary approach is so essential. There may be multiple options available that need to be weighed up against their potential impacts on the end users and costs. An important but often overlooked option is to have a clear response plan if a risk does occur. A quick response once a risk has been identified as occurring could significantly reduce the impact of the risk.

An inexhaustive and illustrative list of general risk treatment options for common AI risks is set out in [Appendix 4: Risk treatment suggestions](#). This advice is general and without context, so not all risk treatment options will be appropriate to all situations, and there may be context-specific options missing. Additionally, the AI landscape is evolving, and so are the methods and metrics for treating AI risks. You should actively monitor the latest available methods and techniques for treating risks, as it is crucial to ensure industry-standard best practice approaches are adopted.

Additional resources

AI Standards and the AI Standards Hub

Standards Development Organisations (SDOs) are actively developing standards in relation to AI technologies. These include the International Organisation for Standardisation (ISO), International Electrotechnical Commission (IEC), and Institute for Electrical and Electronics Engineers Standards Association (IEEE SA). Many of these standards are designed to address known risks in relation to AI technologies. The [AI Standards Hub \(https://aistandardshub.org/the-ai-standards-hub/\)](https://aistandardshub.org/the-ai-standards-hub/) has a standards database which covers relevant standards that are being developed or have been published by a range of prominent SDOs. The AI Standards Hub also has a range of e-learning training covering topics such as assessing and mitigating bias and discrimination in AI.

DSIT's CAIRF team are considering high-level AI risks by preparing common scenarios and risks. They are also involved in mapping out the causes of these risks to facilitate teams in better understanding how they can treat the underlying cause of the risks.

Appendix 1: Risk Identification Questions

Risk Category: Financial

Risk Identification Questions

- Are there any financial implications of using the AI tool that need to be considered? For example, have the increased operational costs or costs to scale the solution post-piloting been considered?
- Have the financial risks from unstable AI systems and harmful results that lead to errors (which require costly refurbishment and replacement costs) been considered?
- Is there a need for additional resources to fix and maintain the AI solution, and have the financial implications of this been considered?
- Have potential financial implications related to non-compliance with legal provisions, including fines and damage to the reputation of the department, been considered?
- What (i) legal toolkits and (ii) regulators and associated regulatory regimes and guidance are relevant given the context that the AI solution is going to be or has been deployed in?
- Have the following legal toolkits been considered at a minimum:
 - (i) Human Rights Law
 - (ii) Equality Law (including the Public Sector Equality Duty (PSED))
 - (iii) Data Protection Law (e.g. UK GDPR)
 - (iv) Intellectual Property
 - (v) Social Value Act
 - (vi) applicable AI Laws (E.g. EU AI Act)
 - (vii) etc?
- Have the following regulatory regimes and guidance been considered at a minimum: (i) EHRC and its guidance on meeting the Public Sector Equality Duty (PSED) (ii) ICO and its guidance on AI and data protection?
- How will the AI solution be reviewed to ensure compliance against the legal toolkits and regulatory regimes identified?
- How will a regular review of the legal and regulatory landscape be conducted, noting that it is currently evolving?
- What other legal or regulatory toolkits are relevant with respect to the data collected for and used or generated by the AI solution, such as the Data

Risk Category: Legal and regulatory compliance

Risk Identification Questions

- What legal or regulatory toolkits are relevant with respect to the model selection, training, deployment and scaling for the AI solution that allows us to continue to export to our trading partners?
- How is the model trained? Does it involve using data in open-source environments or anywhere else in which there may be a risk of failing to comply with relevant laws and regulations, such as the Data Protection Law?
- How is legal and regulatory compliance ensured during deployment? When and how regularly is this monitored?
- What are the policies, processes and procedures for ensuring legal and regulatory compliance of the AI solution throughout the AI lifecycle?
- What processes are in place if issues are discovered? Are these sufficient to react and address the issue in an appropriate timeframe given the context the AI solution is going to be or has been deployed in?
- Do you have the requisite expertise internally to fully understand the legal and regulatory implications of the AI solution in the context that it is going to be or has been deployed in? What training do these individuals and teams receive to remain up to date? Is this training mandatory?
- What knowledge and skills do others in the organisation need about the legal and regulatory requirements to ensure the AI solution is compliant? What training do these individuals and teams get to ensure they have the necessary knowledge and skills?
- What third party data, software, hardware or other infrastructure are deployed in the AI solution?
- What are the terms of the contract or licence? How are they being complied with?
- How do you ensure that any third party data, software, hardware or other infrastructure complies with the legal toolkits and regulatory regimes and guidance that are relevant given the context that the AI solution is going to be or has been deployed in?
- Who are the intended users of the AI solution? Do they need to know that they are interacting with an AI solution? What are their expectations to know they are interacting with an AI solution? Do they need an explanation for a specific output from an AI solution? If yes, what does a good explanation look like?
- Who is impacted by the output of the AI solution and how? Do they need to know that they are impacted by the output of an AI solution? What are their expectations to know they are impacted by the output of an AI solution? Do they need an explanation for a specific output from the AI solution? If yes, what does a good explanation look like?
- Have you complied with the GDS and RTA's Algorithmic Transparency Recording Standard?

- Are there any legal or regulatory requirements that require minimum standards with respect to transparency and explainability?

Risk Category: Appropriate transparency and explainability

Risk Identification Questions

- What action has been taken to address the appropriate level of transparency and explainability that has been identified? How is this going to be monitored?
- What action has been taken to maintain a regime that allows us to continue to export to our trading partners, such as consideration given to the EU AI Act and other appropriate AI legislation.
- Can you provide details on what data is collected and what features are selected to train the AI model and why? Can it be explained why some data are selected whilst others are not?
- Are users reliably aware about the data obtained and how it is used? Can individuals access their own data? Can users consent or object to the use of their data?
- Does the way you collect and process data sufficiently enable the type of explanation required?
- Can you provide details on what AI model has been selected and why this one was chosen over others? Can you provide details on how the AI model was trained?
- Can you explain what the AI model does and how it makes decisions? What are the main contributors that influence the AI model's output?
- Does the model that is chosen and the details about how it was chosen and trained sufficiently enable the type of explanation required? Does the way the AI solution is deployed and scaled enable the type of explanation required?
- Is it necessary to be able to explain why the AI model has made a particular recommendation?
- Does the way the AI solution is deployed and scaled enable the organisation to be transparent about when people are interacting with the AI solution? How is this presented to the user (such as explanations and outputs of the model).
- Does the organisation and its policies, processes and procedures promote transparency about when AI solutions are used and the outputs that these generate?
- Does the organisation and its policies, processes and procedures enable the type of explanation required? Could test explanation methods and the resulting explanations be shared with relevant AI actors, end users and potentially impacted individuals prior to deployment to gain feedback?
- Are outputs from the AI solution and any requested explanations clearly and accurately communicated in a timely fashion?
- Does the organisation have the relevant internal expertise to enable the type of explanation required?

- Who needs to be able to explain the AI solution's outputs? Can these stakeholders explain the AI solution's outputs? If not, then what is communicated? What skills, knowledge and training are required to enable these stakeholders to give meaningful explanations?
- Are the explanations and outputs clearly and transparently communicated to your end users?
- Are there any requirements from third parties for transparency and explainability? How are these being complied with?
- If any third party data, software, hardware or other infrastructure is being used, how do you ensure that it can enable transparency and the type of explanation required?
- What does fairness look like in the context that the AI solution is going to be or has been deployed in?
- Could the AI solution impact an individual's or organisation's legal rights?
- Who are the intended users of the AI solution? Does the AI solution unfairly advantage or disadvantage a particular group? This could be by the way it is accessed and the output from it.
- Who is impacted by the outputs of the AI solution? Does the AI give biased or discriminatory outputs?
- Are there any legal or regulatory requirements that require minimum standards with respect to fairness and eliminating unlawful discrimination and bias? These include the (i) Equality Law, PSED and EHRC Guidance, (ii) Data Protection Law and ICO Guidance.
- How has fairness, discrimination and bias of the AI solution been tested? How is it going to be monitored?
- Have you considered areas that are difficult to assess across the protected characteristics, and how would you manage this if a particular risk was to arise?

Risk Category: Fairness

Risk Identification Questions

- Have you considered the fairness of the AI solution across multiple protected combined characteristics, such as intersectionality?
- Are any of the variables in the data set likely to be a proxy for protected characteristics, and are you able to test for proxies with protected data characteristics, or could any other variables in the data set introduce bias against those with protected characteristics?
- How is it ensured that selected data for modelling is a fair representation of the relevant population? How do you evaluate and monitor this?
- How is it ensured that selected data for modelling does not misrepresent or underrepresent individuals based on their protected characteristics?
- Is the data set to be used representative? Do you measure if there are any disparities/inequalities in the raw data selected for AI models? How is this

measured? And what is done to remediate any identified issues?

- How do you ensure the AI solution maintains fairness? How do you measure and test this? What monitoring and controls do you have in place?
- Does the AI solution display bias toward certain groups? Is differential treatment of groups legal and justified by the context the AI solution is deployed in? How do you test and monitor this?
- How is it ensured that the AI solution does not misrepresent or underrepresent individuals based on their protected characteristics?
- Do you measure if there are any disparities or inequalities in model predictions or outcomes? What is the remediation if the model provides worse results for a particular group?
- What is the risk for the model to create toxic outputs?
- Does the way the AI solution is deployed and scaled ensure that the AI solution maintains fairness?
- How is fairness ensured during deployment? How do you measure and test this? When and how regularly is this tested?
- Is the service fair to those that object against automated decision making?
- Does the organisation and its policies, processes and procedures promote fairness with respect to how AI solutions are developed and used, and the outputs that these generate throughout the AI lifecycle?
- What processes are in place if issues are discovered? Are these sufficient to react and address the issue in an appropriate timeframe given the context the AI solution is going to be or has been deployed in?
- Does the organisation have the relevant internal expertise to identify and address unfair and unlawful discriminatory outcomes in the AI solution and monitor for these throughout the AI lifecycle?
- Who needs to be able to identify unfair and unlawful discriminatory outcomes in the AI solution? What skills, knowledge and training are required to enable these individuals and teams to do this?
- Are there any requirements from third parties for fairness and elimination of bias? How are these being complied with?
- If any third party data, software, hardware or other infrastructure is being used, how do you ensure that it doesn't contain or display discriminatory bias or produce unfair outcomes?
- Has an individual been identified who is ultimately accountable for the AI solution and its outputs? - Who is responsible for AI risk management efforts? Do they have the right level of authority, human resource and budget?
- Is there a diverse team engaged in AI risk management efforts to effectively identify, evaluate and manage AI risks?
- Are there any legal or regulatory requirements with respect to accountability and governance given the context the AI solution is going to be or has been deployed in?
- Who is responsible for the data collected and selected for the AI solution and other data processing activities related to the AI solution? How does this person report to the individual with overall responsibility?

- Are there established data governance practices to oversee data compliance activities? How do these fit together with data monitoring and compliance activities for the AI solution? - Who is responsible for model selection and training activities? How does this person report to the individual with overall responsibility?

Risk Category: Accountability and governance

Risk Identification Questions

- What processes are followed to review model selection and training activities?
- Who is responsible for the way the AI solution is deployed and scaled? Who is responsible for ongoing monitoring activities? How does this individual or individuals report to the individual with overall responsibility?
- What existing organisational governance and risk controls are in place, and how does AI risk management connect into these?
- Are roles, responsibilities and lines of communication related to AI risk management documented and clear to individuals and teams throughout the organisation? - Are there clear policies, processes and procedures for monitoring and mitigating AI risk across the whole AI lifecycle?
- Are there whistleblower policies in place to facilitate reporting of serious concerns with respect to the AI solution? Are stakeholders encouraged to raise concerns? How are concerns treated?
- Is there an incident response plan?
- Do the individuals who are responsible for the AI solution or aspects of it have the requisite skills and knowledge to understand the risks and make decisions? What additional training or support do they need?
- Do the individuals who are responsible for the AI solution or aspects of it have the requisite authority to access the information they need to effect change?
- Are there any requirements from third parties for accountability and governance? How are these being complied with?
- Who are the types of users of the AI solution and what are their expectations to be able to contest an output from an AI solution and seek redress if they think they have been harmed or wronged?
- Who is impacted by an output of the AI solution and what are their expectations to be able to contest an output from an AI solution and seek redress if they think they have been harmed or wronged?
- Are there any legal or regulatory requirements with respect to contestability and redress for people and organisations who have been harmed or wronged by an output from the AI solution?
- Has GDPR's right to rectification been considered as part of any contestability and redress risks?

Risk Category: Contestability and redress

Risk Identification Questions

- Are there defined procedures for remediating any harm or wrong caused by the AI solution? How are those procedures formed, reviewed and updated?
- Are there any redress mechanisms? How is it ensured that they are accessible and transparent to all, including those with vulnerabilities?
- Who is responsible for procedures and decisions with respect to addressing harms or wrongs caused by the AI solution? Do they have the requisite skills, knowledge and authority to do this effectively?
- What are the mechanisms to address disputes if there is an issue with any third party data, software, hardware or other infrastructure being used in the AI solution? Is this meaningful and a way to pass risk onto the third party?
- What is the level of performance the AI solution must be able to maintain given the context in which the AI solution is going to be or has been deployed in?
- Are there any legal or regulatory requirements with respect to technical robustness given the context the AI solution is going to be or has been deployed in?
- How is the level of performance going to be measured and monitored? What metrics or data are going to be used?
- Does the user have a role to play in technical performance, such as human-in-the-loop?
- What data is collected and selected to train the AI model? How is this data assessed for being suitable to train the model given the context in which it is going to be or has been deployed?
- How could the performance and robustness of the AI solution be affected by data quality and data pipeline issues?
- Will the data system be supplemented with data across the AI lifecycle? How do you monitor and control the data set in live deployment?

Risk Category: Technical robustness

Risk Identification Questions

- Has the AI solution been trained on data sets from other AI models? If so, how will data quality and data collapse risks be managed?
- How is the AI model assessed for being suitable in terms of performance and reliability given the context in which it is going to be or has been deployed?
- How could the AI solution performance and robustness be affected by how the AI model is designed and developed?
- Is the model trained under conditions that mimic adverse conditions and natural drift?

- Does the AI solution perform as intended when deployed and scaled? For example, does it perform as intended in less than ideal conditions and when encountering unexpected situations and data?
- How is technical robustness ensured during deployment? How do you measure and test this? When and how regularly is this tested for, such as drift, misinformation and disinformation?
- What are the policies, processes and procedures for ensuring technical robustness of the AI solution throughout the AI lifecycle? Is a robust by design approach adopted?
- What processes are in place if issues are discovered? Are these sufficient to react and address the issue in an appropriate timeframe given the context the AI solution is going to be or has been deployed in?
- Does the organisation have the relevant internal expertise to be able to design and deploy a technically robust AI solution and monitor it throughout the AI lifecycle?
- If the user has a role to play in technical performance, such as human-in-the-loop, then does this user have the requisite expertise and information to do this effectively? What training and additional resources might these users need?
- Are there any requirements from third parties for technical robustness? How are these being complied with?
- If any third party data, software, hardware or other infrastructure is being used how do you ensure that it is technically robust given the context in which the AI solution is going to be or has been deployed in?
- What contingency plans are in place to handle failures or incidents in third-party data, software, hardware and other infrastructure?
- What security threats arise by virtue of the context in which the AI solution is going to be or has been deployed in?
- Has your organisation created new security risks by deploying the AI solution? Are there new ways that your organisation is vulnerable to attack, both externally and internally? Who are the new threat actors? If you have completed Section A2 (Risk Management) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.
- Are there any legal or regulatory requirements with respect to security given the context the AI solution is going to be or has been deployed in?
- Have the National Cyber Security Centre's guidance on principles for the security of machine learning (<https://www.ncsc.gov.uk/collection/machine-learning-principles>) been considered and followed?
- Do you have an overarching AI security strategy supported by clear policy, procedures and processes? If you have completed Section A1 (Board Direction) and B1 (Service Protection Policies and Processes) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.

Risk Category: Security

Risk Identification Questions

- Could the way the data is collected, selected and processed cause any security issues? If you have completed Section B3 (Data Security) in the Cyber Assessment toolkit, you can use this to provide evidence here instead (assuming your response has been expanded to include use of AI)
- Is data selected from trusted sources and is a list of trusted sources kept up to date, with exceptions to collecting untrusted data requiring management approval? If you have completed Section B3 (Data Security) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.
- Are data sets appropriately tracked and verified via cryptographic hash before use? - Are data sets uniquely identified so that any unauthorised changes to an approved data set would cause a review of the data set?
- What type of data will the AI model ingest? Does it also interact with a business critical function or support critical infrastructure?
- How could security be affected by data quality and data pipeline issues?
- Does your organisation have robust data security policies and procedures in place relating to the collection and storage of data from all sources that are used for AI? If you have completed Section B3 (Data Security) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.
- Could the model selected for the AI solution cause any security issues?
- Could how the AI model is designed and developed embed security vulnerabilities that need to be addressed?
- Are you following any policies or processes relating to the training of models and algorithms? What are they?
- Is the model training code reviewed by a responsible party? Who?
- Is the model trained under conditions that mimic adversarial conditions?
- Does the AI solution cause any security issues when deployed and scaled? For example, could new security issues arise when the AI solution encounters unexpected situations and data?
- How is security ensured during deployment? For example, are models adequately tested for vulnerabilities prior to deployment? How do you test this? When and how regularly is this tested?
- What are the policies, processes and procedures for ensuring the AI solution is secure throughout the AI lifecycle? Is a secure by design approach adopted? If you have completed Section B1 (Service Protection Policies and Processes) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.
- What processes are in place if issues are discovered? Are these sufficient to react and address the issue in an appropriate timeframe given the context the AI solution is going to be or has been deployed in? Are disaster recovery and business continuity procedures in place? If you have completed Objective D (Minimising the Impact of Cyber Security Incidents) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.

- Does the organisation have the relevant internal expertise to be able to identify and address security issues caused by the AI solution and monitor these throughout the AI lifecycle?
- Who needs to be able to identify security issues with the AI solution? What skills, knowledge and training are required to enable these individuals and teams to do this? If you have completed Objective C (Detecting Cyber Security Events) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.
- What security education and awareness activities are available for the users and super users of the AI solution? If you have completed Section B6 (Staff Awareness and Training) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.
- Are there any requirements from third parties with respect to security? How are these being complied with? If you have completed Section A4 (Supply Chain) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.
- If any third party data, software, hardware or other infrastructure is being used, how do you ensure that it's secure given the context in which the AI solution is going to be or has been deployed in? Could this be a source of security vulnerabilities? If you have completed Section A4 (Supply Chain) in the Cyber Assessment toolkit, you can use this to provide evidence here instead, assuming your response has been expanded to include use of AI.
- Could the AI solution impact an individual's physical or mental well-being?
- Could the AI solution impact the safety of infrastructure or the environment?
- How are you going to protect vulnerable users?
- Are there any legal or regulatory requirements with respect to safety given the context the AI solution is going to be or has been deployed in?
- How are any safety risks going to be measured and monitored?
- Could the way the data is collected and selected cause any safety issues?

Risk Category: Environment and People

Risk Identification Questions

- How could safety issues be affected by data quality and data pipeline issues?
- Could the model selected for the AI solution cause any safety issues?
- How could safety issues be affected by AI model design and development?
- Does the AI solution cause any safety issues when deployed and scaled? For example, could new safety issues arise when the AI solution encounters unexpected situations and data?
- How is safety ensured during deployment? How do you measure and test this? When and how regularly is this tested?

- What are the policies, processes and procedures for ensuring the AI solution is safe throughout the AI lifecycle? Is a safe by design approach adopted?
- What processes are in place if issues are discovered? Are these sufficient to react to and address the issue in an appropriate timeframe given the context the AI solution is going to be or has been deployed in?
- Does the organisation promote a culture of safeguarding the wellbeing of end users of its AI solution, particularly those with vulnerabilities?
- Does the organisation have the relevant internal expertise to be able to identify and address safety issues caused by the AI solution, and does it monitor these throughout the AI lifecycle?
- Who needs to be able to identify safety issues with the AI solution? What skills, knowledge and training are required to enable these individuals and teams to do this?
- Are there any requirements from third parties with respect to safety? How are these being complied with?
- If any third party data, software, hardware or other infrastructure is being used, how do you ensure that it's safe given the context in which the AI solution is going to be or has been deployed in?

Appendix 2: Risk Appetite Level Definition

	Averse	Minimal	Cautious	Open
Legal, regulatory and compliance	Zero appetite for any AI project with the chance of legal challenge, even if highly unlikely to be successful.	Appetite for risk taking limited to those AI projects with no chance of any significant legal challenge even if unlikely to be successful.	Appetite for risk taking limited to those AI projects with little chance of a successful legal challenge.	Appetite for AI projects with the potential to expose the department to legal challenge, but only where steps are taken to mitigate the legal challenge so that you are likely to win, and the gain will outweigh the adverse impact.
Accountability and	Avoid AI projects with	Willing to consider AI	Willing to consider AI	Ready to take on AI projects

	Averse	Minimal	Cautious	Open
governance	associated accountability and governance risk. No decisions are taken outside of AI processes, and oversight and monitoring arrangements. Organisational AI controls minimise accountability and governance risk, with significant levels of resource focused on detection and prevention.	projects with low accountability and governance risk, which support delivery of priorities and objectives. AI processes, and oversight and monitoring arrangements enable limited risk taking. Organisational AI controls maximise accountability and governance through robust controls and sanctions.	projects where benefits outweigh accountability and governance risks. AI processes, and oversight and monitoring arrangements enable cautious risk taking. Controls enable accountability and governance by maintaining appropriate controls and sanctions.	with some accountability and governance risk when benefits outweigh risks. AI processes, and oversight and monitoring arrangements enable considered risk taking. Levels of accountability and governance controls are varied to reflect the scale of risks with costs.
Contestability and redress	Zero appetite for any AI project where there is a possibility of impacted third parties and actors in the AI life cycle contesting an AI decision or outcome.	Appetite for risk taking limited to those AI projects where there is a very small likelihood of any impacted third parties and actors in the AI life cycle contesting an AI decision or outcome.	Appetite for risk taking limited to those AI projects where there is little chance of any third parties and actors in the AI life cycle contesting an AI decision or outcome, and where appropriate steps are taken to minimise likelihood.	Appetite for AI projects with the potential for third parties and actors in the AI life cycle to contest an AI decision or outcome, and where appropriate steps are taken to minimise likelihood.
Appropriate transparency and explainability	Avoid AI projects with associated transparency and	Willing to consider AI projects with low transparency	Willing to consider AI projects where benefits outweigh	Ready to take on AI projects with some transparency and

	Averse	Minimal	Cautious	Open
	explainability risks.	and explainability risk which support the delivery of priorities and objectives. Organisational controls minimise risk of a lack of transparency and explainability, with significant levels of resource focused on open communication related to AI systems to relevant people where they are able to access, interpret and understand the decision-making process of an AI system.	transparency and explainability risk. Appropriate processes are in place to ensure communication of appropriate information about the AI solution to relevant people, supported by procedures for them to access, interpret and understand the decision-making processes used by AI. These processes enable cautious risk taking.	explainability risk when benefits outweigh these risks. AI processes, and oversight and monitoring arrangements enable considered risk taking. Levels of transparency and explainability controls are varied to reflect scale of risks with costs and trade-off between accuracy and explainability.
Fairness	Zero appetite for any AI projects that may impact the rights of individuals or organisations, discriminate unfairly against individuals, or create unfair market outcomes.	Appetite for risk taking limited to those AI projects where there is low chance of any repercussion to the rights of individuals, unfair discrimination against individuals, or creating unfair market conditions, and only if essential to	Appetite for risk taking limited to those AI projects where there is little chance of any significant impact to the rights of individuals, unfair discrimination against individuals, or creating unfair market conditions, and only if benefits	Appetite for AI projects with the potential for the rights of individuals or organisations to be impacted, unfair discrimination against individuals, or the creation of unfair market conditions, where the benefits

	Averse	Minimal	Cautious	Open
		delivery and significant steps are taken to eliminate the risk, such as human-in-the-loop.	outweigh risk and steps are taken to mitigate the risk.	outweigh the risks and where appropriate controls are in place to minimise the risk. Controls are varied to reflect the scale of risks with costs.
Technical robustness	Zero appetite for AI systems and technology that are unable to reliably function as intended throughout the AI lifecycle.	Only essential AI systems and technology developments to protect current operations.	Accept the need for technical robustness of AI technologies with risk mitigated through stress testing, performance KPIs and continuous improvement.	AI systems and technology robustness are considered to enable improved delivery. Agile principles may be followed.

	Averse	Minimal	Cautious	Open
Environment and people	Zero appetite for any AI projects with a high chance of repercussions to individual's physical and mental well-being, and the safety of critical infrastructure and the environment.	Appetite for risk taking is limited to those AI projects where there is low chance of any repercussion to an individual's physical and mental well-being, and the safety of critical infrastructure and the environment, and only if essential to delivery and significant steps are taken to eliminate the risk, such as human-in-the-loop.	Appetite for risk taking is limited to those AI projects where there is little chance of any significant impact to an individual's physical and mental well-being, and the safety of critical infrastructure and the environment, and only if benefits outweigh risk and steps are taken to mitigate the risk.	Appetite for AI projects with the potential to impact an individual's physical and mental well-being, or expose the safety of critical infrastructure and the environment, where the benefits outweigh the risks and where appropriate controls are in place to minimise the risk. Controls are varied to reflect the scale of risks with costs.
Security	No tolerance for security risks such as data poisoning and leakage, perturbation attacks, prompt injection threats arising from AI systems and technology, causing loss or damage to HMG property, assets,	Minimal risk appetite for security risks such as data poisoning and leakage, perturbation attacks, prompt injection threats arising from AI systems and technology causing loss or damage to HMG property, assets, information or	Limited security risks such as data poisoning and leakage, perturbation attacks and prompt injection threats arising from AI systems and technology are accepted to support business needs. Appropriate security	Cautious acceptance of security risks such as data poisoning and leakage, perturbation attacks and prompt injection threats arising from AI systems and technology to support business needs.

	Averse	Minimal	Cautious	Open
	information or people.	people. Stringent security measures are in place to protect against these risks.	measures in place to protect against these risks. Steps taken are varied to reflect the scale of risks with costs.	Appropriate security measures are in place to protect against these risks. Steps taken are varied to reflect the scale of risks with costs.
Financial	No appetite for any financial impact or loss resulting from the use of the AI solution.	The department is willing to accept minimal financial impact if essential to delivery.	The project seeks safe AI delivery options and is cautious towards little residual financial loss resulting from use of the tool, and only if it could provide some benefits to the user or organisation.	Prepared to invest in AI for productivity and efficiency benefits, minimising the possibility of financial losses by managing the risks to an acceptable level.

Appendix 3: List of potential risk impacts

Impact Categories

	Business	Staff	Service User/Citizen	Reputation
Quantifiable	- Operational downtime - Financial loss - Service delivery delays - Productivity	- Employee turnover - Absenteeism rate - Training costs, with the	- Service access delays, with additional days added to the processing time for public service	- Media coverage - Social media sentiment - Public

	Business	Staff	Service User/Citizen	Reputation
	gains, with an increase in productivity in sectors/ and departments adopting AI compared to those that do not - Cost of implementation, with capital expenditure on AI technologies as a percentage of total capital investment - Public sector spending and revenue, with an increase in unemployment benefits due to job displacement vs tax revenue from AI-driven economic growth.	training hours needed per employee, and the cost required to adapt to new AI systems - Job displacement, with the reduction in workforce in departments where AI is implemented, by percentage - Job transformation, with the number of tasks automated within a specific job role - Employee stress and mental health, with the number of stress-related absenteeism cases or mental health support requests	applications due to AI errors - Complaint volume and user satisfaction, with an increase in the number of complaints filed due to AI-driven service disruptions - Cost to citizens, with the GBP amount citizens have to spend additionally due to delays or inaccuracies caused by AI systems - Service quality and reliability, with the percentage of interactions resulting in errors or failures - Data privacy and security, with the increase in reported data privacy incidents involving customer information - Service personalisation, with the percentage of users dissatisfied with personalised recommendations or decisions - User retention and churn, with changes in the number of users	enquiries and investigations - Complaint volume, with the increase in customer complaints regarding AI-driven decisions or services - Ethical concerns, with the number of reported ethical breaches or controversies involving AI continuing to use the service over a given period
Non-quantifiable	- Strategic setback - Resource	- Morale and job satisfaction - Workplace	- Public trust and confidence - Citizen	- Loss of stakeholder trust due to

Business	Staff	Service User/Citizen	Reputation
diversion - Innovation stagnation - Loss of human skills and knowledge, with long-term dependency on AI systems and potential difficulties in reverting to manual processes if needed - Dependence on technology, with economic instability and disruption if AI systems experience significant outages, errors or cyber-attacks	culture - Talent attraction - Ethical dilemmas, with moral distress and potential conflict with organisational policies used in AI decision-making - Adaptation challenges, including frustration, decreased productivity, and reluctance to adopt new technologies	satisfaction - Equity and fairness - Citizen displacement and social isolation, with social and psychological consequences such as decreased mental well-being and increased societal fragmentation - Loss of privacy with public resistance to AI technologies and potential backlash aimed at protecting individual privacy rights	AI system issues affecting collaboration and partnerships - Brand perception, with long-term damage to the government's reputation for reliability and competence in using AI - Public perception of competence, with public doubt about the government's ability to manage AI systems effectively and responsibly - Decision-making transparency, where lack of transparency can erode trust among stakeholders and complicate accountability for AI-driven decisions

Table setting out impact scores for each quantified impact.

Impact Categories

Impact Level	Business	Staff	Service User/Citizen	Reputation
1. Negligible	- Operational downtime is less than 1 hour - Financial loss is less than £10,000 - Service delivery delays are less than 1 hour - Strategic setback: minimal - Resource diversion: minimal - Innovation stagnation: minimal	- Employee turnover is less than 1% - Absenteeism rate is less than 1% - Training costs are less than £1,000 - Morale: no change - Job satisfaction: no change - Workplace culture: no change	- Service access delays are less than 1 hour - Complaint volume increases by less than 1% - Cost to citizens is less than £1,000 - Public trust and confidence: no change - Citizen satisfaction: no change - Equity and fairness: no change	- Media coverage: negligible - Social media sentiment: minimal - Public enquiries: none - Stakeholder trust: no change - Brand perception: no change - Public perception of competence: no change
2. Minor	- Operational downtime is between 1 to 4 hours - Financial loss between £10,000 to £50,000 - Service delivery delays between 1 to 4 hours - Strategic setback: minor - Resource diversion: minor - Innovation stagnation: minor	- Employee turnover between 1% to 2% - Absenteeism rate between 1% to 2% - Training costs between £1,000 to £5,000 - Morale: slight decrease - Job satisfaction: slight decrease - Workplace culture: slight negative impact	- Service access delays between 1 to 4 hours - Complaint volume increase of between 1% to -2% - Cost to citizens between £1,000 to £5,000 - Public trust and confidence: slight decrease - Citizen satisfaction: slight decrease - Equity and fairness: slight negative impact	- Media coverage: minor - Social media sentiment: minor - Public enquiries: minor - Stakeholder trust: slight decrease - Brand perception: slight negative impact - Public perception of competence: slight impact
3. Moderate	- Operational downtime	- Employee turnover	- Service access delays	- Media coverage:

Impact Level	Business	Staff	Service User/Citizen	Reputation
	between 4 to 8 hours - Financial loss between £50,000 to £200,000 - Service delivery delays between 4 to 8 hours - Strategic setback: moderate - Resource diversion: moderate - Innovation stagnation: moderate	between 2% to 4% - Absenteeism rate between 2% to 4% - Training costs between £5,000 to £20,000 - Morale: noticeable decrease - Job satisfaction: noticeable decrease - Workplace culture: noticeable negative impact	between 4 to 8 hours - Complaint volume increase between 2% to 4% - Cost to citizens between £5,000 to £20,000 - Public trust and confidence: noticeable decrease - Citizen satisfaction: noticeable decrease - Equity and fairness: noticeable negative impact	moderate - Social media sentiment: moderate - Public enquiries: moderate - Stakeholder trust: noticeable decrease - Brand perception: moderate negative impact - Public perception of competence: moderate impact
4. Major	- Operational downtime between 8 to 24 hours - Financial loss between £200,000 to £1 million - Service delivery delays between 8 to 24 hours - Strategic setback: significant - Resource diversion: significant - Innovation stagnation: significant	- Employee turnover between 4% to 8% - Absenteeism rate between 4% to 8% - Training costs between £20,000 to £100,000 - Morale: significant decrease - Job satisfaction: significant decrease - Workplace culture: significant negative impact	- Service access delays between 8 to 24 hours - Complaint volume increase between 4% to 8% - Cost to citizens between £20,000 to £100,000 - Public trust and confidence: significant decrease - Citizen satisfaction: significant decrease - Equity and	- Media coverage: major - Social media sentiment: significant - Public enquiries: significant - Stakeholder trust: significant decrease - Brand perception: major negative impact - Public perception of competence: significant impact

Impact Level	Business	Staff	Service User/Citizen	Reputation
			fairness: significant negative impact	
5. Catastrophic	- Operational downtime is more than 24 hours - Financial loss is more than £1 million - Service delivery delays of more than 24 hours - Strategic setback: critical - Resource diversion: critical - Innovation halt	- Employee turnover is more than 8% - Absenteeism rate is more than 8% - Training costs is more than £100,000 - Morale: severe decrease - Job satisfaction: severe decrease - Workplace culture: severe negative impact	- Service access delays are more than 24 hours - Complaint volume increase is more than 8% - Cost to citizens is more than £100,000 - Public trust and confidence: severe decrease - Citizen satisfaction: severe decrease - Equity and fairness: severe negative impact	- Media coverage: severe - Social media sentiment: severe - Public enquiries: numerous and severe - Stakeholder trust: severe decrease - Brand perception: severe negative impact - Public perception of competence: severe impact

Appendix 4: Risk treatment suggestions

AI Risk	Risk Treatment Options	Risk Treatment Category
Poor accuracy or performance	Establish policies setting out the required level of human involvement in AI-augmented decision-making.	Limitation
Poor accuracy or performance	Establish policies and testing procedures to enable regular and consistent monitoring and testing practices throughout the AI lifecycle. These should include metrics to demonstrate whether or not the system is fit for purpose	Limitation

AI Risk	Risk Treatment Options	Risk Treatment Category
	and functioning as claimed, as well as defined acceptable limits for performance.	
Poor accuracy or performance	Establish incident response procedures for when the AI solution performs beyond acceptable limits. Put alerts in place for when performance indicators observed in production differ from pre-deployment test outcomes, or when anomalies are detected. Depending on the context, this could include real-time flagging of potential incidents and human review of outputs and decisions.	Avoidance
Poor accuracy or performance	Establish and regularly review procedures to bypass or deactivate the AI solution, including plans for redundant or backup systems to ensure continuity of operational functionality and regularly reviewing incident thresholds for activating bypass or deactivation responses.	Avoidance
Poor accuracy or performance	Red-teaming or adversarial testing of AI systems under stress conditions to seek out failure modes or vulnerabilities in the system by external experts independent from the AI project team.	Limitation
Poor accuracy or performance	Make end users clearly aware of the limitations of the system, that they use the outputs at their own risk, and are responsible for verifying the outputs.	Transference
Poor accuracy or performance	Establish proficiency standards required for end users of the AI solution. Define and develop training material and keep it under review to ensure it is up to date. Ensure end users are properly trained to interpret the AI solution output and detect errors and incidents.	Limitation
Accountability	Senior leaders and executives must own responsibility for the organisational risk appetite for AI risks and overall management of those risks.	Acceptance
Accountability	Identify a specific team or individual who is responsible for AI risk management efforts across the organisation.	Acceptance
Accountability	Establish accountability metrics to determine whether the AI project team and other	Limitation

AI Risk	Risk Treatment Options	Risk Treatment Category
	accountable owners maintain clear and transparent lines of responsibility, and are open to inquiries.	
Accountability	Establish confidential feedback mechanisms and whistleblowing policies for internal and external stakeholders to raise issues with AI solutions.	Limitation
Bias and fairness	Ensure compliance with relevant law, including Data Protection Law, Equality Law and the Public Sector Equality Duty, and follow guidance from relevant regulators including the ICO and EHRC. Complete a Data Protection Impact Assessment (DPIA).	Limitation
Bias and fairness	Establish policies around AI project teams being composed of a diverse group of individuals who reflect a range of backgrounds, perspectives and experience. Establish policies around using participatory design approaches for AI solutions.	Limitation
Bias and fairness	Establish policies, procedures and metrics around testing and verifying the data sets and models used in AI solutions, including those from third parties and external sources.	Limitation
Bias and fairness	Have separate individuals and teams who are independent from the AI project team carry out tests and evaluation tasks to help counter implicit biases, groupthink and sunk cost fallacy.	Limitation
Transparency	Establish standardised AI solution documentation, and regularly review documentation policies that capture information related to the AI solution, including information related to responsible owners and their contact information.	Limitation
Transparency	Establish policies for an AI model inventory system, and regularly review its completeness, usability, and efficacy.	Limitation
Transparency	Document and review the use and efficacy of different types of transparency tools. Follow industry best practice and standards at the time a model is in use.	Limitation

AI Risk	Risk Treatment Options	Risk Treatment Category
Transparency	Establish policies and processes regarding public disclosure of the use of AI and risk management material such as impact assessments, audits, model documentation and validation and testing results. Identify transparency metrics to assess whether stakeholders have access to necessary information.	Limitation
Transparency	Establish stakeholder engagement plans and follow up on feedback.	Limitation
Transparency	Establish policies and processes for transparency around the decommissioning and retirement of AI solutions.	Limitation
Explainability, or how a prediction or decision was made, and interpretability, or why a prediction or decision was made	Establish policies and procedures around developing explainable and interpretable AI solutions. These could include policies around: when possible, approaches should be utilised that are inherently explainable, such as decision trees; explaining systems using a variety of methods, such as visualisation, model extraction and feature importance; testing explanation methods and resulting explanations prior to deployment to gain feedback from relevant stakeholders and potentially impacted individuals or groups about whether explanations are accurate, clear, and understandable.	Limitation
Explainability, or how a prediction or decision was made, and interpretability, or why a prediction or decision was made	Document and review the use and efficacy of different types of explainability and interpretability tools. Follow industry best practice and standards at the time a model is in use.	Limitation
Explainability, or how a prediction or decision was made, and interpretability, or why a prediction or decision was made	Establish mechanisms that facilitate the auditability of the AI solution, including development process traceability, sources of training data, logging of the AI solution's processes, outcomes, and impacts both positive and negative.	Limitation
Privacy	Ensure compliance with Data Protection Law and follow guidance from the ICO. Complete a	Limitation

AI Risk	Risk Treatment Options	Risk Treatment Category
	Data Protection Impact Assessment (DPIA).	
Privacy	Where possible, utilise privacy-enhancing technologies (PETs) for AI, as well as data minimising methods such as de-identification and aggregation.	Avoidance
Security	Adopt machine learning and end-point security countermeasures, such as robust models, differential privacy, authentication and throttling.	Limitation
Hallucinations and misinformation	Establish policies and procedures that AI generated content should be checked for accuracy by end users.	Transference
Legal compliance	Consider legal compliance from the outset. This includes working closely with legal teams to identify and understand the legal and regulatory requirements specific to your use case and business purpose, as well as the context of the deployed AI solution. Translate these into the way the AI solution is designed, trained and deployed to ensure the technology is aligned with the legal requirements identified. Follow regulatory guidance from relevant regulators such as the EHRC, ICO and relevant sector specific regulators.	Limitation
Legal compliance	Establish ongoing monitoring, and review for compliance against applicable laws and regulations, including preserving materials for forensic, regulatory, and legal review as required.	Limitation
Legal compliance	Maintain policies for training and re-training staff about legal and regulatory considerations that may impact AI-related design, development and deployment activities.	Limitation
Does not deliver expected benefits	Establish policies to determine if AI is the right tool to deploy when weighing up negative risks against its benefits.	Limitation
Does not deliver expected benefits	Define KPIs and success metrics of the AI solution to enable testing throughout the AI lifecycle to ensure it's delivering the expected benefits and impact.	Limitation

AI Risk	Risk Treatment Options	Risk Treatment Category
Does not deliver expected benefits	Establish policies and procedures around participatory stakeholder engagement throughout the AI lifecycle to help project teams design and develop the AI solution that will deliver impact, and decide if the AI solution should be pursued at all. Document and action feedback.	Limitation
Does not deliver expected benefits	Keep a narrow application scope as the benefits and risks will be easier to map. Decisions made about acceptable tradeoffs can decide whether to pursue the AI solution or not.	Limitation
Failure of third party elements of the AI solution	Establish policies and procedures to test, evaluate and validate third party elements, whether data, software or hardware systems, to ensure legal and regulatory compliance as well as compliance with other organisational policies and procedures.	Limitation
Failure of third party elements of the AI solution	Establish processes for third party suppliers to report known or potential issues and vulnerabilities in supplied resources.	Limitation
Failure of third party elements of the AI solution	Establish policies and procedures that include redundancies for covering third-party functions.	Limitation
Failure of third party elements of the AI solution	Establish and regularly review procedures to bypass the AI solution, including plans for redundant or backup systems to ensure continuity of operational functionality. Ensure contracts and licences have robust warranty and indemnity provisions.	Avoidance, Transference



