

AI Governance for Business:

Scoping AI Use Cases
and Managing Risks



Innovate
UK

BridgeAI

Table of Contents

Acknowledgements	3
About this project	4
Introduction	5
Research design and methodology	6
How to use the framework	7
Limitations of the framework	9
Part 1: Characterising your AI Use Case	10
Dimension 1: Organisation	10
Business opportunity	10
Use case type	11
Business function	11
Dimension 2: AI system	12
Capabilities	12
Computational infrastructure	13
Deployment environment	13
Technological readiness level	14
Dimension 3: Data	14
Types of input data	14
Processing of personal data	15
Processing of IP-protected data	16
Dimension 4: Economic sector	16
Sector and sub-sectors	17
Part 2: Identifying the risks of your AI use case	19
AI risk management fundamentals	19
Sources of AI risk	21
Data	21
AI model	22
Other AI system components	24
Deployment environment	26
Part 3: Mitigating the risks of your AI use case	30
Product development mitigations	30
Management and oversight mitigations	34
Information and transparency mitigations	38
Education and cultural mitigations	39
Conclusion	41
Glossary	42

Acknowledgements

This project was funded by **Innovate UK BridgeAI**, led by **Arcangelo Leone de Castris** and co-authored by **Dr Florian Ostmann** and **Shakir Laher** at **The Alan Turing Institute**.

The authors would like to acknowledge the valuable contributions of the following individuals and groups and thank them for their dedication and time spent on this work. First and foremost, **Nalanda Sharadjaya** and **Paul Khullar** for supporting us in reviewing the relevant literature, identifying AI use cases, and improving the framework with their critical feedback. We also thank **Prof. Elena Gaura**, **Prof. James Brusey**, **Prof. Keivan Navaie**, **Dr Matthew Foreshaw**, **Prof. Po Yang**, **Prof. Kit Windows-Yule**, **Dr Rachael Stickland**, and **Arielle Bennett-Lovell** for sharing their expertise and providing essential feedback to advance this work. Lastly, we would like to thank the BridgeAI delivery partners and our colleagues at **The Alan Turing Institute**, all of whom allowed us to present our work in multiple fora and collect essential feedback at different stages of the project.

Cite as:

Leone de Castris, A., Ostmann, F, and Laher, S.
(2026). *AI Governance for Businesses: Scoping
AI Use Cases and Managing Risks*. BridgeAI.
<https://doi.org/10.5281/zenodo.14527338>.

Developed by

**The
Alan Turing
Institute**

About the programme

Innovate UK BridgeAI empowers UK businesses in high-growth sectors, driving productivity and economic growth through the adoption of Artificial Intelligence. We bridge the gap between developers and end-users, fostering user-driven AI technologies.

With a focus on ethics, transparency, and data privacy, we aim to build trust and confidence in the development of AI solutions. Strengthening AI leadership, supporting workforces, and promoting responsible innovation, BridgeAI shapes a collaborative and AI-enabled future.

BridgeAI is an Innovate UK-funded programme, delivered by a consortium including Innovate UK, Digital Catapult, The Alan Turing Institute, STFC Hartree Centre and BSI.

www.bridgeai.net

About the project

This project addresses the need for clear, accessible guidance on how organisations can safely integrate AI into their operations. We surveyed the UK business ecosystem to identify the most significant challenges organisations face and how early adopters are addressing them. Based on these insights, we developed a suite of practical resources outlining key requirements, risks, and recommended risk mitigation strategies for responsible AI adoption.

In January 2025, we launched the first version of the AI Use Case Framework—a tool that organisations can use to catalogue their AI solutions. The launch was complemented by four briefings that applied the Framework to real-world AI use cases in the four BridgeAI priority sectors: agriculture, construction, creative industries, and transportation. We then used public feedback and new data collected from the UK business ecosystem to refine and expand the Framework, integrating it with critical guidance on risk identification and mitigation. Published in March 2026, this report presents the final version of the project.

Desired impact

By providing an actionable framework to systematically characterise AI use cases, identify related risks, and mitigate them, we hope to support organisations in identifying promising applications of AI in their area and developing an effective strategy to operationalise them responsibly.

Intended audience

The framework is designed primarily to support business leaders seeking to adopt AI within their organisations. However, the resources developed as part of this project are grounded in widely applicable principles and can benefit non-commercial organisations, government officials and regulators, as well as training officers and advisors working on AI adoption and risk management.

Introduction

Artificial intelligence (AI) is rapidly transforming the global economy, yet many industries are only beginning to explore the opportunities AI technologies offer. Despite adoption being on the rise globally, the technology's potential to improve organisations' productivity and competitiveness remains largely untapped across many sectors. For example, only 20% of small businesses in the UK use AI in their operations, despite 55% stating that it could benefit them.¹

To support organisations on their AI adoption journey, the UK government has launched several high-impact initiatives to "invest and plan for the long-term needs of the [country's] AI ecosystem."² These include, among others, the BridgeAI programme, which aims to promote the development and adoption of AI technologies in sectors with high potential for AI-driven economic transformation.³

Despite important steps forward, several barriers to widespread AI adoption remain. These include high implementation costs, uncertain ROI,⁴ a shortage of digitally skilled personnel,⁵ regulatory ambiguity,⁶ challenges in accessing or collecting high-quality data,⁷ and the difficulty

of integrating AI with legacy systems.⁸ At a more fundamental level, however, businesses cite two key challenges: defining the right AI use cases for specific needs and managing the risks associated with operationalising them.⁹ As highlighted by a recent survey of more than 100 business leaders, companies struggle to define clear use cases due to limited awareness of the full range of practical applications that AI can support.¹⁰ Even when the potential value of AI is well understood, organisations are often held back by the complexity of managing the risks posed by AI systems.¹¹

This report focuses specifically on the last two challenges – identifying the right AI use cases and managing their risks – and provides practical resources to support organisations in navigating them. Part 1 of the report introduces a use case profiling mechanism that organisations can use to think strategically about how AI technologies can be successfully integrated into their operations. Part 2 examines some of the challenges that should be addressed during this process by highlighting the main sources of risk that can emerge across the AI value chain.¹²

1 Russell, C. & E. Quist (2024). *Redefining intelligence: The growth of AI among small businesses*. Federation of Small Businesses. <https://www.fsb.org.uk/resource-report/redefining-intelligence.html>.

2 UK Government (2021). *National AI Strategy*. <https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version>.

3 Innovate UK (2023). *BridgeAI*. <https://iuk.ktn-uk.org/programme/bridgeai/>.

4 CDEI (2021). *UK Business Innovation Survey*. https://assets.publishing.service.gov.uk/media/61bb2e77e90e07044462d8b7/Business_Innovation_Survey_2021.pdf.

5 *AI ecosystem survey informing the National AI Strategy*. The Alan Turing Institute, https://www.turing.ac.uk/sites/default/files/2021-09/ai-strategy-survey_results_020921.pdf.

6 Gillespie, N., S. Lockey, C. Curtis, et al. (2023). *Trust in Artificial Intelligence: A global study*. The University of Queensland and KPMG Australia. https://policy-futures.centre.uq.edu.au/files/16650/Trust%20in%20AI%20Global%20Report_2023_UQ.pdf.

7 IBM Institute for Business Value (2025). *The ingenuity of generative AI. Unlock productivity and innovation at scale*. <https://www.ibm.com/downloads/documents/us-en/10a99803f8afda4c>.

8 Mittal, N., Saif, I. & Ammanath, B. (2022). *Fueling the AI transformation: Four key actions powering widespread value from AI, right now*. Deloitte, <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-ai-institute-state-of-ai-fifth-edition.pdf>.

9 UK Government (2025). *Barriers and Enablers to Advanced Technology Adoption for UK Businesses*. <https://www.gov.uk/government/publications/barriers-and-enablers-to-advanced-technology-adoption-for-uk-businesses/barriers-and-enablers-to-advanced-technology-adoption-for-uk-businesses>.

10 Bicakci, B. et al. (2024). *Leadership in the Age of AI*. Egon Zehnder and Kearney. <https://www.egonzehnder.com/leadership-in-the-age-of-ai>.

11 Mökander, J., et al. (2025). *What the UK Thinks About AI: Building Public Trust to Accelerate Adoption*. The Tony Blair Institute for Global Chance. <https://institute.global/insights/tech-and-digitalisation/what-the-uk-thinks-about-ai-building-public-trust-to-accelerate-adoption>.

12 The objective of Part 2 is not to quantify risk with precision, but to flag key categories of concern and promote structured thinking around mitigation.

Finally, Part 3 offers a structured overview of established AI risk mitigation best practices that can be implemented to address the risks identified in Part 2. Together, they provide essential tools to help innovators leverage AI technologies more confidently and responsibly.

Research design and methodology

The findings presented in the three parts of this report build on existing templates and taxonomies, integrating them with insights collected directly from the UK's business ecosystem.

The development of the use case profiling mechanism described in Part 1 followed an iterative prototyping and refinement methodology. We started by developing a draft classification structure based on existing knowledge of the field and supported by a literature review.¹³ We then refined and validated it through multiple cycles of prototyping, testing on real-world use cases, and collecting feedback from both internal and external stakeholders. For the purposes of trialling and refining the framework, we used 40 AI use cases collected from both primary and secondary sources. These include case study documentation provided by companies, academic papers, and grey literature.

The taxonomy of risk sources and risk mitigation strategies proposed in Parts 2 and 3, instead, is the result of two sequential steps. A first draft taxonomy was developed through a comparative analysis of four AI risk management frameworks: **ISO/IEC 42001:2023 – AI management system** and **ISO/IEC 23894:2023 – Guidance on AI risk management**; the **NIST AI risk management framework**; and the **AI safety governance framework** published by the Standard Administration of China (SAC). The analysis was also informed by the ongoing work of Cen-Cenelec JTC 21 on a new AI risk management standard¹⁴ and by other AI risk taxonomies established in the academic literature.¹⁵ The initial draft taxonomy was then integrated with primary data collected from the business ecosystem through a series of four in-person workshops attended by a total of 190 organisations from across the UK. The result is a taxonomy of 27 sources of risk and 65 recommended risk mitigation strategies.

¹³ Existing frameworks to categorise AI use cases include the *OECD Framework for the Classification of AI Systems*; the *EIT Taxonomy for the European AI Ecosystem*; and *ISO/IEC TR 24030:2024 – AI use cases*. We used these resources to ground and inform our research, and ensure we avoided any unnecessary duplication. At the same time, none of these frameworks covers the full scope of our research, which required the development of the framework we are presenting here. For instance, the OECD Framework focuses on a generic classification of AI systems rather than on specific use cases for business. The taxonomy developed by the EIT aims to map the landscape of AI use cases in the EU but does not provide sufficient detail on some foundational components of AI systems such as the input data, the operational environment of the system, or the readiness level of the technology. ISO/IEC TR 24030:2024 offers a comprehensive template to collect AI use cases but it assumes the availability of information that is usually only available to organisations that already adopted AI and possess detailed insights into the context and the characteristics of the use case. Our framework, while consistent and interoperable with existing taxonomies, is specifically designed to support businesses at the early stage of the AI adoption cycle in identifying and selecting AI opportunities strategically.

¹⁴ CEN/CLC/JTC 21 – Artificial Intelligence. <https://www.cen-cenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/>.

¹⁵ Such as the MIT's AI Risk Repository, the AVID Taxonomy of AI risks, the MITRE ATLAS, the OWASP AI Testing Guide, and IBM's AI Risk Atlas.

How to use the framework

This framework is an actionable tool designed to support the responsible ideation, planning, and management of AI projects within business organisations.

The framework can be applied in a straightforward, iterative manner. It outlines a four-step sequential process that organisations can use to translate high-level commitments to responsible AI into the foundational elements of a comprehensive governance system for implementing AI applications.

Step 1: Define the AI use case

The first step in any new AI project is to define a business problem and/or need, understand what type of solution can address it, and assess its feasibility. Part 1 of this framework offers a blueprinting mechanism that project owners can use to describe the key building blocks of an AI use case in a specific context and embed governance considerations from the start.

Step 2: Diagnose risks

After defining the AI use case, organisations can start mapping its building blocks against the taxonomy of AI risks proposed in Chapter 2. This step involves a systematic assessment of the AI value chain to understand where failure points might occur. The output of this step should be a list of specific vulnerabilities relevant to the specific use case considered – i.e., a risk register.

Step 3: Select appropriate risk mitigation measures

Having diagnosed the risks posed by a specific AI use case, organisations can now define appropriate measures to eliminate or mitigate them. Chapter 3 provides a structured list of established risk mitigation best practices and encourages a holistic approach to risk mitigation by requiring action across four dimensions: a) product development, b) management and oversight, c) information and transparency, and d) education and culture. The output of this step is an AI governance plan that links use-case specific risks to concrete mitigation strategies and that organisations can use to prioritise interventions, assign responsibilities, and define implementation timelines. Such a plan can also concretely support organisations with the development of a formal and comprehensive risk management system.

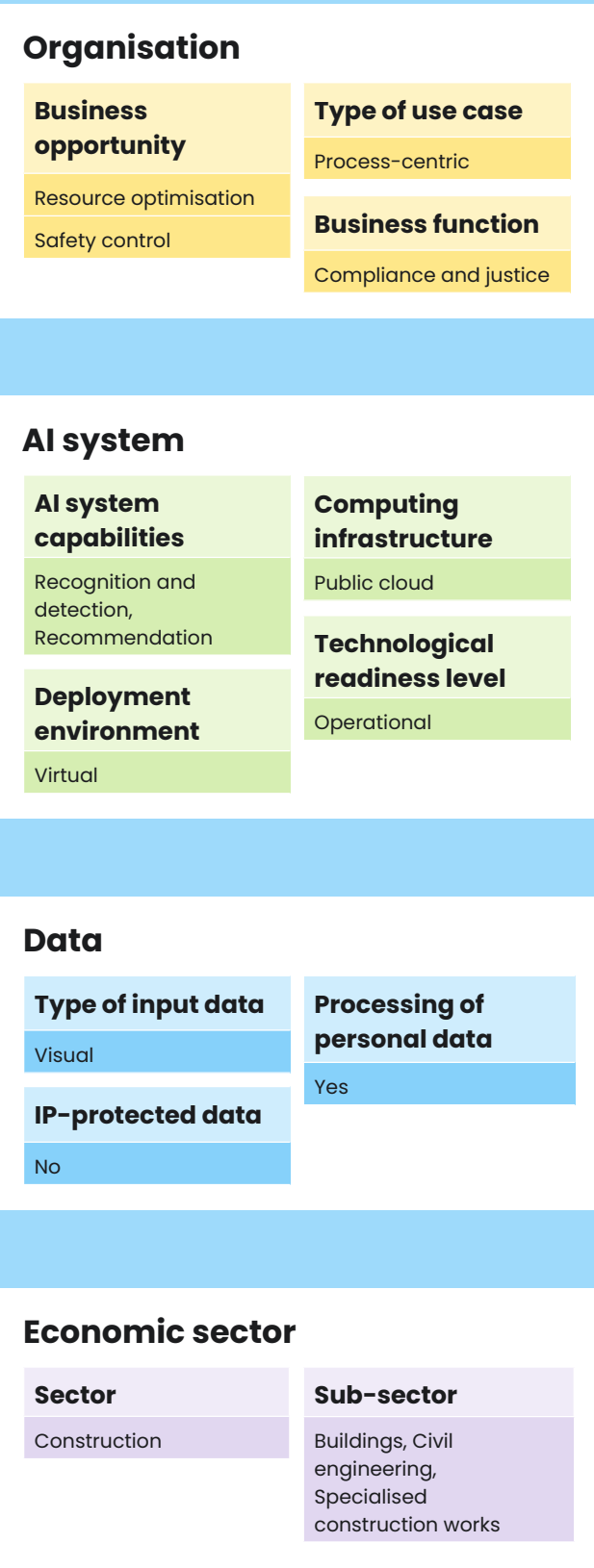
Step 4: Monitor and review

The process described in Steps 1, 2, and 3 should be regularly reviewed and updated during the lifecycle of the AI use case. Considering that most contemporary AI systems are dynamic – they learn, drift, and can change behaviour after deployment – treating this framework as a one-off checkbox could lead organisations to overlook the risks that can emerge after an AI system has operated for some time.

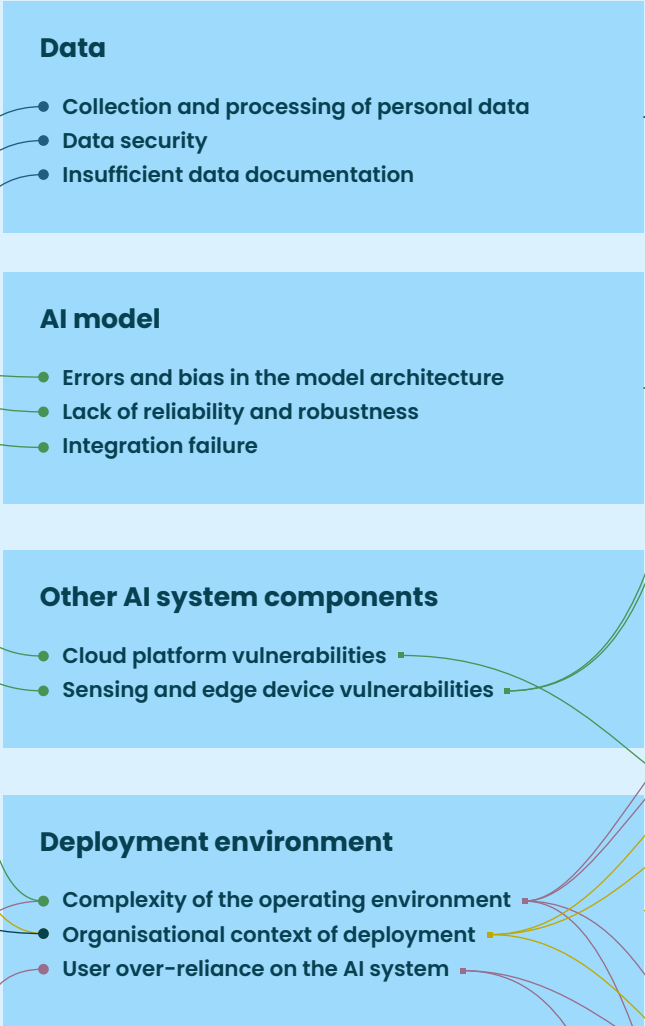
Step 1: Characterise the AI use case

Real-time worksite safety monitoring

Worksite hazards are a major concern in the construction sector. They put workers' lives at risk and can cause costly disruptions. AI technologies can support in detecting these hazards before they escalate, helping safeguard worker safety and minimise losses. For example, AI technologies can be used for real-time worksite monitoring. By processing visual data from cameras placed around the worksite, an AI system can identify and flag instances of unsafe behaviour and other safety hazards. The results of the analysis can then be used to inform the company's worksite safety policies as well as future training and education programmes. This AI solution can be applied to different kinds of worksites and is relevant for all construction sub-sectors. It has the potential to support productivity by reducing worksite incidents and minimising compliance costs for the company deploying it. Similar AI solutions process data containing, among other things, images of workers. For this reason, the companies deploying them should adopt adequate personal data protection measures, such as blurring or masking people's faces before processing the data. Companies deploying these systems should consider possible risks related to the workplace surveillance capabilities of such AI systems and adopt relevant internal policies to address them.



Step 2: Diagnose risks



Step 3: Select appropriate risk mitigations



Step 4: Monitor and review

Limitations of the framework

This framework offers practical resources to support responsible AI adoption. To implement it effectively, however, users should be aware of its limitations:

1. This framework represents a high-level model of a complex reality. Each AI use case is unique and requires evaluating information that is specific to its context of application. The goal is to offer curated, accessible guidance rather than an exhaustive list of every consideration or risk associated with AI adoption. As such, users should treat this framework as a starting point to build on, rather than as a definitive, end-to-end guide.
2. The scope of this framework is limited to governance considerations. It excludes specific economic or operational metrics, such as the cost of investment, potential ROI, or productivity gains related to specific use cases. Business decisions on AI adoption should balance the conclusion supported by this framework with internal economic analysis.
3. Taxonomies are inherently backwards looking. They represent information known at the time of writing. AI risks evolve rapidly, and so do the techniques to address them. Furthermore, our risk and risk mitigation taxonomies do not aim to comprehensively represent every potential risk and solution but focus on the most important ones that businesses should consider. If users apply the taxonomies in Chapters 2 and 3 too rigidly, they might miss emergent or lesser-known risks and mitigation tools. Users should consider the taxonomies proposed in this paper as a baseline reference rather than as the ground truth, and we encourage organisations to regularly update their internal versions of these lists as new risks and solutions emerge on the market.

Part 1: Characterising your AI Use Case

Before implementing an AI use case, it is important to first clearly define its core characteristics—for example, the business value it should unlock, the specific tasks it will perform, the data required for its operation, the readiness level of the technology, and its impact on the operating environment.

Chapter 1 acts as a practical framework for scoping new AI projects. Users can apply it to map out specific characteristics that should be considered for successful and responsible deployment.

A detailed characterisation of AI use cases serves two primary purposes:

1. Clarify the operational requirements that ensure the project is viable within existing technical and resource constraints.
2. Identify high-risk elements early on, enabling the implementation of governance-by-design principles already at the project ideation stage.

Our approach to characterising AI use cases relies on 12 categories of information grouped under four higher-level 'dimensions', depending on whether they capture information about (1) the organisation operating the AI system, (2) the AI system itself, (3) the input data processed by the system, or (4) the economic sector in which the AI system is deployed.

The table below summarises the four dimensions and the categories that can be used to scope the AI use case. In the following subsections, we elaborate on each category of information and its importance for developing a responsible AI adoption strategy.

Dimension 1: Organisation

One of the first steps in developing a targeted strategy for AI adoption is understanding which areas and processes of an organisation's structure can benefit from embedding AI. The first dimension of our framework includes three categories of information that capture important aspects of the role that AI can play within an organisation:

- The **business opportunity** that can be realised with the use case.
- The **type of use case**.
- The **business function** where AI is embedded.

Business opportunity

The first category under the 'Organisation' dimension highlights common business opportunities unlocked by AI. Common examples include:

- **Automation of repetitive tasks** – e.g., customer triaging, invoice processing.
- **Resource optimisation** – e.g., energy grid management, dynamic pricing, supply chain route planning.
- **Quality and safety control** – e.g., inspection of assembly lines, fraud detection.

- **Personalisation and customisation** – e.g., targeted marketing, additive manufacturing.
- **Decision support** – e.g., demand forecasting, customer sentiment analysis.
- **Knowledge extraction** – e.g., search assistance, document summarisation.

AI adoption is still in its early stages, and organisations will keep finding creative ways of applying AI to drive value. As a result, new business opportunities will emerge. Users of this framework are therefore invited to characterise their use cases by identifying the specific challenge they aim to solve, even if it is not included in this list.

Crucially, these examples often overlap. A single AI use will often create multiple business opportunities simultaneously.

Use case type

The second category under the 'Organisation' dimension distinguishes between cases where a company uses AI as a feature of a product or service that it offers and cases where it uses AI internally to support its operations. More specifically, we distinguish between:

- **Product-centric/service-centric** use cases when AI is integrated into a product or service.
- **Process-centric** use cases when AI is used to support the organisation's business functions. Business functions can be fully internal or take place at the interface with partners, clients, or suppliers.

Business function

Focusing solely on process-centric use cases, the third category under the 'Organisation' dimension provides a closer look at which business functions leverage AI. For this research, we understand business functions as the activities carried out by a company to sustain its operations and meet its objectives. We distinguish between 14 business functions:

Accounting – Process of recording, classifying, summarising, and analysing the financial transactions and activities of the company.

Compliance and justice – Process of ensuring that the company's operations comply with relevant laws, regulations, and ethical standards, and that the company is accountable in its internal operations and interactions with stakeholders.

Customer service – Process of assisting customers before, during, and after purchasing a product or service.

Human resource management – Process of recruiting, selecting, training, developing, and managing the company's employees. This includes human resource planning, recruitment, performance management, professional development, compensation and employee relations.

Information Communication Technology management – Process of planning, implementing, and overseeing the information and communication technologies that enable and support the company's operations and objectives.

Logistics – Process of planning, implementing, and managing the flow of goods, services, and information. This includes order processing, transportation, inventory management, warehousing, packaging and labelling, and tracking the status and location of products once shipped.

Management of organisational assets – Process of acquiring, maintaining, upgrading, and disposing of company assets to maximise their value and minimise risks.

Marketing and advertising – Process of promoting the company's products, services, and brand to attract and retain customers. This includes defining and managing the brand, planning and delivering promotional campaigns, promotional content creation, and market research to understand customer needs and preferences.

Planning and budgeting – Process of forecasting the company's future financial position and allocating resources to achieve its goals and objectives.

Production – Process of transforming the company's inputs (such as raw materials, labour, and capital) into products sold to customers. This includes production planning, implementation, and control.

Quality control – Process of measuring, evaluating, and improving the quality and performance of the company's products, services, and processes. This includes quality assurance and control of both products and processes.

Research and development – Process of generating new knowledge, ideas, and

innovations to drive the company's competitive advantage and long-term growth. This includes ideation and exploration, applied research, intellectual property management, etc.

Sales – Process of identifying, engaging, and converting prospective customers into paying clients. This includes processes such as sales prospecting, lead scoring, and negotiation.

Service provision – Process of transforming the company's inputs into services provided to clients. This includes the planning, organisation, and management of resources and operations necessary to provide a specific service.

Dimension 2: AI system

After clarifying the business impact of the AI use case and how it relates to the organisational structure, it is important to consider the main distinctive features of the AI system underpinning that use case. We highlight four categories of information that offer insights into:

- The AI system's **capabilities**.
- The **computing infrastructure** it relies on.
- The **environments impacted by its output**.
- The **technological maturity** of the AI system.

Capabilities

The first category under the 'AI System' dimension identifies the specific tasks that the AI system can perform. This is essential not only for defining implementation objectives and aligning the use case with the company's broader business strategy, but also for identifying potential risks.

Common AI capabilities include:

- **Goal-directed action** – The ability of an AI system to achieve specific goals autonomously, and, where applicable, call on tools at its disposal. Examples include scheduling calls and booking activities based on a given agenda, adjusting inventory levels to ensure effective stock management, or controlling a fruit-harvesting robot.
- **Recognition and detection** – The ability of an AI system to recognise or detect specific events, entities, or behaviours from analogue or digital environments. Examples include monitoring worksites to detect unsafe behaviours or identifying sentiments, opinions or emotions expressed in text, audio, or video data (sentiment analysis).
- **Generation** – The ability of an AI system to generate, transform, and respond to prompts with content such as text, audio, images, video, or code. Examples include summarising and translating text, engaging in conversations with users, producing and editing images and videos, writing code for programming tasks, or creating songs.
- **Optimisation** – The ability of an AI system to identify optimal values for variables, sequences of actions or strategies to achieve goals. Examples include planning navigation routes or scheduling timetables.
- **Prediction and forecasting** – The ability of an AI system to forecast the likelihood of future events or states of the world. Examples include predicting crop yields, the likelihood of machinery failures within a given timeframe, or how customers will behave under certain circumstances.

- **Simulation** – The ability of an AI system to create a digital approximation of an entity, process, environment, or scenario under study. Examples include creating a digital model of a worksite to plan remote interventions or simulating the functioning of an engine to test alternative design specifications.
- **Recommendation** – The ability of an AI system to suggest relevant information based on preferences, needs, and behaviour. Examples include supporting decision-making, recommending content for specific audiences, or enhancing customer experience through chatbots and tailored suggestions.

The capabilities of an AI system have a direct impact on the type and severity of risk it can pose. For example, capabilities that enable an AI system to act autonomously inherently carry higher operational risk, and generative capabilities introduce specific accuracy risks.

AI systems can also exhibit multiple capabilities at once, creating a compounding effect that leads to higher risk levels. For instance, an AI system used for autonomous driving can simultaneously recognise images, predict the likelihood of certain events, and take actions directed towards specific goals. This creates dependency vulnerabilities where a minor error at the recognition stage could lead to actions with potentially catastrophic outcomes.

Computational infrastructure

The second category under the 'AI system' dimension refers to the type of infrastructure required by the AI system. It is important for organisations to think strategically about what computational architecture best matches their requirements and goals based on factors such as cost, privacy, and scalability. Most options fall within one of the following approaches:

- **Public cloud services.**
- **On-premise infrastructure.**
- **Hybrid.**

Public cloud services can provide organisations with access to raw computing power as well as ready-to-use AI platforms, including access to state-of-the-art models and built-in ML Ops tools. This approach is the most common starting point for many organisations because it enables experimentation, prioritising speed over control. However, this approach is also more conducive to vendor lock-in, unforeseen fluctuations in costs, and higher security risk related to the proprietary data uploaded on the cloud.

On-premise infrastructure is more complex and costly to develop, requires a dedicated IT infrastructure team with specialised skills, and can pose procurement challenges due to the current high demand for chips. At the same time, organisations that rely on proprietary computing infrastructure will have more control over the security of their data, will enjoy long-term cost stability, and reduce dependencies on third-party providers. Some jurisdictions require data to stay physically within national borders, and on-premise infrastructure can facilitate compliance with data sovereignty requirements. On-premise infrastructure can also be economically advantageous for organisations that run inference or training continuously.

Considering the pros and cons of different approaches, many organisations end up choosing a hybrid strategy: they start using the cloud for large, one-off workloads and experiments, or to access massive LLMs that would be difficult to host locally, and once a system is stable and expected to run 24/7, move it to on-premise hardware.

Deployment environment

The third category under the 'AI system' dimension focuses on the type of environment influenced by the system's output. Understanding whether an AI system operates in a physical or virtual setting is critical not just for safety and governance, but also for planning investments in hardware and infrastructure.

We distinguish between:

- AI systems that only influence the **virtual environment**. Examples include chatbots, virtual assistants, and gaming applications.
- AI systems that also influence the **physical environment** as a result of being integrated into larger cyber-physical systems. Examples include applications in domains such as robotics, self-driving cars, and smart homes.

Technological readiness level

The fourth and final category under the 'AI system' dimension provides information about how 'ready' for deployment an AI system is relative to a specific use case. We distinguish between the following 'readiness levels':¹⁶

- **Hypothetical** – The development of the AI system has not yet started and has only been theoretically proposed as a conceptual possibility.

- **Early development** – The development of the AI system has started, but has not yet reached the proof-of-concept phase.
- **Proof of concept** – A version of the AI system has been validated in a test or live environment.
- **Operational** – The AI system has been successfully deployed in an operational context.

Considering the technological readiness level of an AI system is critical in determining the feasibility and risk profile of a use case. It can help organisations distinguish between 'possible on paper' or 'in a lab environment' and 'production-ready'. More experimental solutions typically also require higher upfront investment and stricter governance controls.

Dimension 3: Data

Data are essential to support the effective deployment of AI. Sufficient, high-quality data is necessary for determining whether an AI project is feasible and how successful it will be. AI systems can be used to analyse different types of data depending on the tasks they are designed to perform. Understanding what data they process in a specific context has important governance implications. For instance, if an AI system is trained on or takes personal or IP-protected data as input data, it can raise important ethical and compliance questions. Building on these considerations, the third dimension of our framework – 'Data' – includes three categories of information that provide insights relevant to:

- The **types of input data** needed to deploy AI in a specific context.

- Whether **personal data** are processed as part of the system's deployment.
- Where **IP-protected** data are processed as part of the system's deployment.

Types of input data

The first category under the 'Data' dimension provides information about the format of data required to deploy AI in a specific context.

We distinguish between the following types of input data:

Audio – refers to various forms of sound, including the sound of human speech.

Visual – refers to images, videos, or other visual content.

¹⁶ One of the most common frameworks for technological maturity is Technology Readiness Levels (TRLs). The TRL taxonomy, however, is not equally well suited for all types of technology. It is also too granular for the purpose of our research. For this reason, we used TRL as a starting point, simplifying and adapting it to the context of our research while maintaining coherence between the TRL framework and our approach.

Signal – refers to data describing physical phenomena or processes that carry information other than acoustic and visual signals. This includes various types of sensor readings, telecommunication signals, and biomedical signals such as EEG and ECG.

Numerical – refers to data that take the form of continuous numerical values. Examples include lengths, volumes, and time durations.

Categorical – refers to values that can be divided into discrete categories. These categories often represent qualitative characteristics or attributes. Categorical data cover a wide range of domains, including gender, customer satisfaction ratings, and clothing size.

Textual – refers to machine-readable sequences of characters, words, or sentences.

Geospatial – refers to data that describe a location on or near the Earth's surface. Usually, geospatial data combine information about a specific location, information about certain attributes observed in that location (e.g., objects or events), and information about the time at which the attributes are observed.

Organisations considering the feasibility of an AI use case should assess what types of data are required to deploy an AI system in a specific context – knowing this could suggest that the organisation might need to purchase or collect certain types of data it does not already have access to, as well as whether specific controls and audits are required. The different types of input data we identified refer to general, human-interpretable formats in which data can appear at the time of collection. It should be noted that, for this project, we decided to distinguish visual and audio data from signal data, even though they could also be thought of in terms of 'signals'. This choice is motivated by the fact that computer vision and audition systems are increasingly common in industry applications and present distinctive

characteristics and challenges compared to AI systems that process other types of signals.

Processing of personal data

Across the different types of input data mentioned above, the data processed by AI systems may or may not represent personal data. In the second category of the 'Data' dimension, we therefore distinguish between AI systems that:

- **Process personal data.**
- **Do not process personal data.**

Developers and deployers of AI systems that process personal data – i.e., any information that relates to an identified or identifiable individual¹⁷ – are required to comply with the legal requirements set by relevant data protection regulations and implement specific measures based on the circumstances in which the data processing takes place – e.g., based on the nature, scope, context, and purposes of the processing, and on the risks this poses to individuals' rights and freedoms.

In the UK, the legal regime relevant to data protection is laid out by the Data Protection Act (DPA) and the UK General Data Protection Regulation (GDPR). Both regulations adopt a risk-based approach to data protection insofar as they require actors processing personal data to identify the risks that their processing activities pose to the data protection rights of affected individuals. For instance, if an organisation's processing of personal data is "likely to result in a high risk" to individuals' rights and freedoms, the UK GDPR and DPA require that organisation to perform a data protection impact assessment (DPIA) – i.e., a process to support the organisation understand the risks connected to its personal data processing and implement appropriate measures to mitigate those risks. Based on the ICO's Guidance on AI and data protection,¹⁸ the use of AI often involves a type of processing that is likely to result in a high risk to individuals' rights and freedoms and will, therefore, activate the legal requirement to perform a DPIA.

¹⁷ Examples of personal data include names, addresses, dates of birth; web-based data such as user location, IP addresses, cookies; Radio Frequency Identification tags; health and genetic data; biometric data; data on political opinions, sexual orientation, and racial and ethnic data.

¹⁸ ICO (2023). *Guidance on AI and data protection*. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/>.

Provided that the data protection implications of AI shall be assessed based on the contextual characteristics of each use case, knowing whether personal data are included in the data processed by a specific AI system is an important first step to assess whether the organisation deploying AI in that way is data protection compliant.

Finally, it is important to consider that not all data collected from individuals are automatically considered personal data. If they are fully anonymised and can no longer identify the individuals they originally referred to, their treatment will be exempt from the requirements of the DPA and the UK GDPR. Companies should exercise caution when anonymising data, however. First, the process of anonymisation is still considered processing of personal data. Second, if the individuals to which the data refer can be re-identified through any 'reasonably available means', then the data are not truly anonymised, and data protection requirements will apply.¹⁹

Processing of IP-protected data

The third and final category of the 'Data' dimension distinguishes between AI systems that:

- **Process IP-protected data.**
- **Do not process IP-protected data.**

Developers and deployers of AI systems that process IP-protected data – such as data protected by copyright, dataset rights, and trade secrets – should consider both the risks related to the data fed into the system (input) and the legal status of the data that the system produces (output).

While many jurisdictions and regulators are still developing approaches to manage the challenges that AI poses to copyright law, organisations should be aware that training or fine tuning an AI system using copyright-protected data can create significant legal risks depending on the circumstances in which the data usage takes place – e.g., based on the source, licensing terms, commercial intent, and the nature of the output generated. Organisations that use third-party AI systems should also consider that inputting proprietary data or trade secrets into those systems may cause that information to be incorporated into the training dataset, which could even lead to losing trade secret protections associated with it.

Regarding risks related to the AI system's output, organisations may be liable for copyright infringement if their AI system reproduces a piece of copyrighted content present in its training dataset.

Dimension 4: Economic sector

The fourth dimension of our framework captures information related to the economic sector in which the AI use case is deployed. Even for AI systems that are technically applicable across sectors, the specific socio-economic context in which a system is deployed will at least partially determine the regulatory landscape, safety standards, and operational norms to which it is subject. For example, an AI system applied in healthcare can be subject to specific medical

device regulations or patient privacy laws, whereas the same AI architecture applied in retail might only trigger standard consumer protection laws. Different sectors and subsectors also have different risk tolerance standards. A 5% error rate in a 'Creative Industries' recommendation engine can have fundamentally different implications than a 5% error rate in a system controlling an autonomous vehicle.

¹⁹ ICO (2022), *What is personal data*, <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-is-personal-data/>.

Sector and sub-sectors

Due to the vast diversity of industry standards and the rapidly evolving regulatory landscape, an exhaustive mapping of all sector-specific norms is beyond the scope of this framework. Instead, users should leverage the 'Economic sector' dimension as a prompting mechanism to identify and map specific sector best practices and regulatory requirements relevant to each of their use cases.

Below, we provide a sub-sector breakdown of four sectors selected as examples:

- **Agriculture**
- **Construction**
- **Creative industries**
- **Transport**

A more detailed discussion of AI use cases in these four sectors and their respective sub-sectors can be found in the **sector-specific briefs**.

Organisation

Business opportunity	Type of use case	Business function		
Automation of repetitive tasks	Product-centric / Service-centric Process-centric	HR management	Accounting	Logistics
Resource optimisation		Research and development	Planning and budgeting	Compliance and justice
Quality and safety control		Management of organisational assets	Sales	ICT management
Personalisation and customisation		Customer service	Marketing and advertisement	Service provision
Decision support		Quality control	Production	
Knowledge extraction				

AI system

AI system capabilities		Computing infrastructure	Deployment environment	Technological readiness level
Goal-directed action	Prediction and forecasting	Public cloud services	Physical	Hypothetical
Recognition and detection	Recommendation	On-premise infrastructure	Virtual	Early development
Generation	Simulation	Hybrid		Proof of concept
Optimisation				Operational

Data

Types of input data				Processing of personal data		IP-protected data
Audio	Signal	Categorical	Geospatial	Yes	No	Process IP-protected data
Visual	Numerical	Textual				Do not process IP-protected data

Economic sector

Sector	Subsector			
Agriculture, forestry and fishing	Crop	Animal production and hunting	Forestry and logging	Fishing and aquaculture
Construction	Buildings	Civil engineering	Specialised construction works	
Creative industries	Advertising and marketing	Architecture	Crafts	
	Design	Film, TV, radio and photography	IT, software and computer services	
	Museums, galleries and libraries	Music, performing and visual arts	Publishing	
Transportation and storage	Air	Rail	Road	
	Space	Water	Warehousing and support activities	

Part 2: Identifying the risks of your AI use case

Once organisations develop a clear understanding of the organisational, technical, and regulatory profile of an AI use case, they can use that information to start identifying potential sources of risk and developing targeted controls. Part 2 of this report provides the essential taxonomy for this phase, outlining the most important sources of risk that organisations should consider as part of their use case implementation strategy.

While some risks are specific to particular use cases, many AI applications share overlapping risk characteristics. Developing a systematic

understanding of these common vulnerabilities can significantly improve an organisation's ability to manage them. By mapping the use case profile against the risk factors outlined in this chapter, users can identify the specific elements in their AI use cases that could lead to harm. This process lays the foundation for Part 3, which presents a curated set of established industry best practices for AI risk mitigation. By combining these diagnostic insights with actional control strategies, this framework aims to offer a practical resource for developing effective, robust AI risk management protocols.

AI risk management fundamentals

This section offers a high-level overview of the main elements to consider in AI risk management. While this framework is not a step-by-step guide on setting up a risk management system, introducing basic risk management notions helps contextualise the concepts discussed in the following sections.

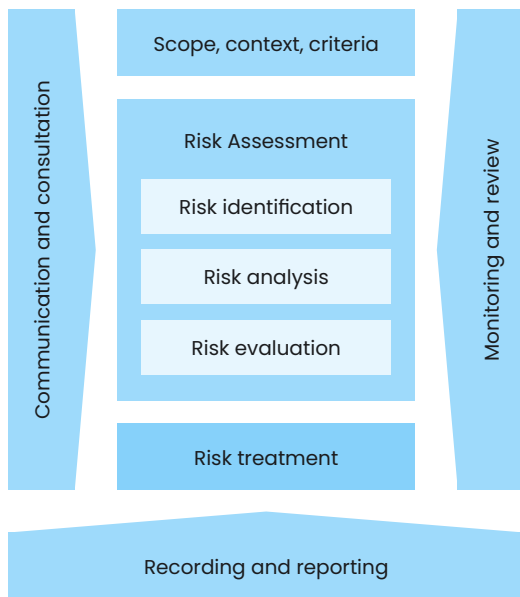
Defining AI risk management

AI risk management is the systematic and continuous application of management policies, processes and practices to the tasks of analysing, evaluating, controlling and monitoring risks throughout the entire lifecycle of an AI system.

AI risk management may be approached as the adaptation of traditional risk management processes to the specific context of AI. They broadly share the same objective – to reduce risk as effectively and early as possible in the lifecycle of a product or service – and structure. AI risk management, like traditional risk management processes, consists of four fundamental steps:

- Identifying sources of risks.
- Estimating the risks posed by each source and evaluating if they are acceptable or require mitigation.
- Adopting the necessary risk mitigation measures to bring all unacceptable risks under the acceptability threshold.
- Monitoring and reviewing the effectiveness of the risk mitigation measures.

Risk management process



Source: ISO/IEC 31000:2018

Despite significant overlap with traditional risk management, AI risk management also presents distinctive features:

- First, the risks posed by AI are often dynamic, subject to distribution shift, and poorly described by historical data.
- Second, they can be harder to predict and quantify when the model behaviour is non-linear and opaque.
- Third, they require continuous monitoring and oversight because, especially with adaptive AI models, they can evolve rapidly.
- Fourth, in more severe instances, they could have a systemic impact, causing harm at scale.
- Finally, effective AI risk management requires considering both safety and security risks. Traditionally, safety and security risk assessments are distinct processes characterised by bespoke methodologies and different

skillsets. Safety assessments tend to focus on unintended events such as human errors and system failures, while security assessments focus on intentional attacks. In the case of AI, these two domains are likely to intersect (e.g., prompt injections, data exfiltration via model outputs, data poisoning). As such, it is important to adopt a broad and integrated approach to AI risk management.

Therefore, while traditional risk management systems remain largely relevant in the context of AI by providing strong foundations to build on, it is important for organisations to consider how to adapt and integrate those systems to account for the new challenges brought about by AI technologies.²⁰

The importance of AI risk management in businesses

AI risk management is essential for organisations seeking to adopt AI safely and drive meaningful and lasting transformation. In addition to reducing the likelihood and impact of adverse events, it also creates business opportunities. By proactively managing risks, organisations can make more informed decisions, become more resilient against uncertainty and hazards, and signal their trustworthiness to investors, customers, and other stakeholders. In some jurisdictions, setting up a risk management system is mandatory for certain AI applications. Beyond mere compliance, however, robust risk management anchors AI initiatives in ethical and safety best practices, empowering enterprises to scale AI technologies across the organisation with confidence.

²⁰ For additional information and practical support with setting up an AI risk management system, we refer to the following resources: ISO 31000:2018, ISO/IEC 42001/2023, ISO/IEC 23894/2023, and the NIST AI RMF.

Sources of AI risk

As innovation advances and AI technologies become more sophisticated and complex to manage, organisations interested in making the most of their AI systems must account for and manage an increasingly broad range of risks.

AI risks can emerge at every stage of the product lifecycle and are driven by a combination of technical and societal factors. Common risk drivers include the quality of the data used for training, validation, and operation of the system, the system design, the type of infrastructure on which it relies, and its interaction with the environment after deployment – e.g., how it is used or misused, who operates it and for what purpose, whether it interacts with other AI systems, its degree of autonomy, and the context in which it is deployed.

To navigate the wide array of potential vulnerabilities, this framework maps risks to four categories that reflect core features of an AI use case:

- **Data**
- **AI model**
- **Other AI system components**
- **Deployment environment**

Data

Data are the foundational input used to train, test, and operate AI models. They constitute the source from which an AI model learns patterns, correlations, and relationships, and will determine the model's initial capabilities and logic.

The most common sources of risk related to the data used to train, test and operate an AI system include:

Collection, processing, and storage of personal data

Organisations have a duty to protect the personal data they control and process from unauthorised or unlawful data practices and accidental loss, destruction, or damage. Failure

to adequately protect personal data can result in violations of data protection law and harm to data subjects. While data protection risks are common to all organisations controlling and processing personal data, AI systems can exacerbate them and make it more complex to identify and manage them. For instance, the practice of collecting large amounts of data to train advanced AI systems can lead to the inclusion of personal information in training datasets that is not necessary or directly relevant to the purpose of the system, potentially violating the legal principle of data minimisation. AI systems also challenge the storage limitation principle.²¹ As AI systems 'learn' mathematical representations of the data they are trained on, deleting personal data after the necessary retention period or based on specific user requests can be technically challenging, detrimental to the performance of the system, and difficult to demonstrate. Specific risks also exist when personal data are processed for profiling and automated decision-making purposes. Depending on the specific use case, these practices might be prohibited or warrant the organisation carrying them out to adopt additional transparency and safety measures. Finally, the highly complex and layered nature of AI system supply chains can make it difficult to assign data protection roles and responsibilities to the different organisations involved in their development and operation. These are just a few examples of issues to consider when thinking about the interaction between AI and data protection. The field evolves rapidly, and data protection regulators are still grappling with some of the challenges that AI systems raise.

Collection and processing of IP-protected data

The collection and processing of data could potentially violate intellectual property rights when the data involves copyright-protected work. If an AI system processes only non-

²¹ Art. 5(1)(e) of the UK GDPR requires organisations to keep personal data only as long as necessary for the purpose it was collected, and to delete or anonymise it afterward.

copyrightable operational data such as customer orders, sensor data, and financial transactions, safeguards aren't usually required. But if an organisation trains or tunes an AI system using copyright-protected works, measures such as licensing, filtering, and provenance tracking might be necessary to avoid copyright liability.

Data quality

Data quality refers to how well a dataset meets the quality standards required by a specific AI application, including criteria such as completeness, representativeness, consistency, uniqueness, accuracy, and timeliness. Poor quality data used to train, test, fine-tune, and operate an AI system, including issues such as missing values, outliers, formatting errors, and duplicates in the dataset, increases the risk that a system will produce inaccurate, discriminatory, or even illegal outputs. Assessing the quality of the data used by an AI system is especially important in cases where multiple data sources are integrated and involve real-time data collection and processing. Data quality assessments should also include data labels in cases of supervised and semi-supervised learning, where labels are usually produced by human annotators.

Data security

Data security refers to the measures adopted by an organisation to protect sensitive data from unauthorised access, manipulation, or theft. Data security risks include data theft, unauthorised data disclosure, and data poisoning.²² Data security breaches can harm both the organisation and the subjects whose data were compromised, possibly also leading to the violation of data protection obligations. While data security risks are shared by all organisations controlling data, AI systems can add new layers of complexity and create novel vulnerabilities. For instance, datasets used to train AI systems are especially valuable due to the quantity of information they contain and can be specifically targeted by malicious actors. AI systems also present distinctive vulnerabilities which can be

exploited to access confidential data, and their performance can be affected when the integrity of the data they are trained on or that they process is compromised. For more information on security vulnerabilities specific to AI systems, see the sections on *AI model* and *Other AI system components* later in this document.

Data bias

Datasets are biased when they represent certain groups or individuals more than others, resulting in the underperformance of the AI system for less represented categories. An AI system trained or tuned on biased data may produce unfair or discriminatory results. It will generate content that misrepresents certain groups or individuals and make predictions or take decisions that unfairly penalise them. A system producing unfair outputs can cause harm to individuals and communities by amplifying existing discriminatory dynamics and can lead to legal liability under anti-discrimination law, such as the Equality Act (2010) in the UK.

Insufficient data documentation

Data documentation refers to the practice of recording information about the data used to train and tune an AI model. This includes information on the collection and preparation of the data, data provenance, lineage, and ownership, and how datasets are treated to mitigate risks. Data documentation is necessary to understand model behaviour, assess its trustworthiness, and identify potential risks. In some cases, the lack of sufficient documentation could also lead to legal and compliance breaches.

AI model

AI models are mathematical representations of patterns and relationships in the data they are trained on that are optimised to achieve predefined goals. They are the core component of AI systems. In most cases, AI models are trained to make predictions or recommendations, recognise patterns, or generate content.

²² An adversarial attack where false or incorrect data are injected into the training or fine-tuning datasets. Through data poisoning, malicious actors can alter the behaviour of an AI model for their own benefit.

Some of the main sources of risk to consider when assessing AI models include:

Errors and bias in the model architecture

AI model development involves a multitude of human choices (e.g., how to preprocess the data, what family of models to use, which goals to optimise for, and which evaluation metrics to test against). Each of these choices reflects assumptions, trade-offs, and omissions that developers make, which are then baked into the model. These are not coding bugs, but structural design choices. If these choices are the product of judgment mistakes or unconscious bias, they can lead to the AI model behaving in unexpected ways, including producing inaccurate outputs and performing poorly on the tasks it was originally designed for.

Goal misspecification

One of the most fundamental sources of risk in AI is the misalignment between the goals an AI model is optimised for and the real-world objectives it is intended to pursue. Goal misalignment can occur when the objectives, reward signals, instructions, or performance metrics used to train and evaluate an AI model do not fully capture the intentions and goals of the humans who operate it. Misalignment between the behaviour an AI model is optimised for and the objectives of the organisation deploying it can lead to unexpected and even harmful model behaviour.

Lack of model interpretability and explainability

A model is interpretable and explainable when it is possible to understand its internal mechanics and describe how it produced a given output in ways that humans can comprehend. Explainability techniques are especially valuable for complex models whose mechanics are objectively impossible to inspect and interpret comprehensively – i.e., black box models. Interpretability and explainability can be critical to establishing the trustworthiness of a model, to debug issues, address unexpected behaviours, and ensure accountability. Low degrees of interpretability and explainability may result in the inability to fully control the AI model and account for its

outputs. It is worth noting that interpretability and explainability can take on slightly different forms depending on the primary audience they are directed to. For instance, if they are directed to developers and engineers, they will involve technical information describing the different components of the architecture of the system and how several factors contributed to the final output. If, instead, they are directed to the public, they will provide a more generic explanation of the system's decision-making process to support informed use and trust.

Lack of reliability and robustness

Reliability is the ability of an AI model to perform as expected under normal circumstances, while robustness refers to the ability to maintain a consistent performance when faced with noisy, adversarial, or out-of-distribution inputs. Many factors can contribute to an AI model being unreliable or non-robust. In addition to data quality and security issues, factors that can negatively affect the performance of a model include overfitting, security vulnerabilities affecting the output of the model, model drift, and constraints on the computational resources used in the deployment phase. An AI model that behaves unreliably or that is not robust to unusual operational conditions is more prone to failures.

Model security

Another major source of AI risk is the potential for an AI model's behaviour or output to be altered by adversarial actors. Attacks such as model extraction and inversion, prompt injection, prompt leakage, and jailbreaking can cause AI models to expose sensitive data, proprietary algorithms, and other confidential information through their output. This can result in unauthorised data access, privacy violations, intellectual property breaches, and the generation of harmful or illegal content. These attacks could also modify a model's decision-making processes, causing it to execute arbitrary commands. This is especially risky in cases of cyber-physical systems and where multiple models interact with each other.

Insufficient model documentation

Model documentation refers to the availability of information on the model design, development, and evaluation processes. The absence of sufficient information about an AI model can lead to the inability to identify and manage some of the risks it poses.

Integration failure

Integration failures refer to cases where AI models that worked well in a lab or controlled environment fail to interact as intended with the surrounding software, hardware, and processes that characterise the system in which it is embedded for a specific and concrete application. This could be caused, among other issues, by interoperability and scalability challenges, dependencies between the model and certain components of the system that may change or be updated over time, or assumptions made at the development stage that do not hold during deployment.

Other AI system components

In addition to one or more AI models providing core capabilities and the data they were trained on, AI systems comprise compute infrastructure, hardware, and software components that enable operations and context-specific workflows.

When assessing the risk profile of an AI system, in addition to considering data- and model-related vulnerabilities, organisations should also consider the following sources of risk:

Cloud platforms

Cloud platforms are core enablers of AI systems. Their function is to provide the infrastructure, services, and tooling that let organisations develop and deploy AI applications without having to build and maintain in-house compute infrastructure. Cloud platforms offer several important benefits for organisations, such as faster time to market, scalability of the AI workload, and cost-effectiveness. However, the cloud layer could also represent a critical dependency and expose the organisation relying on it to a specific set of risks. Aspects to consider in this respect include dependencies between the availability and reliability of the service and

the operations of the AI system. Disruptions in the cloud service or even in the internet network on which the system relies might slow down or halt the operations of the system. This is especially relevant for latency-sensitive applications. Depending on a single cloud provider might also limit the possibility for a business to migrate its data, AI models, or modify workflows further down the line, exposing an organisation to a risk of vendor lock-in. Another critical dependency relates to the security of the cloud infrastructure. Organisations could suffer leaks of sensitive information due to provider-side breaches, including data and model theft.

On-premise servers

Organisations that train and/or run AI systems on proprietary servers need to consider specific sources of risk that can affect both the technical performance and the security of those systems. From a technical point of view, considering that AI training and inference require large computational resources, it is important to assess if the server capabilities available are sufficient for the workload required by the AI system. Insufficient capabilities could slow down or prevent the system from operating. Servers also require continuous maintenance to prevent system failures and downtime. From a security perspective, servers are vulnerable to theft, tampering, and environmental hazards such as fire or flooding. Damaged or compromised servers can lead to system failures, data loss, or model loss. Finally, the architecture of servers optimised to handle AI workloads introduces specific vulnerabilities that can be exploited by side-channel attacks. These attacks leverage the analysis of GPU vectors such as timing, power, and electromagnetic emissions to extract sensitive data or cryptographic keys.

User interfaces

User interfaces (UI) are the tools through which humans and other systems can interact with an AI system. In practice, organisations can choose different types of interfaces depending on the users who interact with the system, the context of application, and the tasks that the system is expected to perform. The main types of user interfaces for AI include

application programming interfaces (APIs), text-based conversational interfaces, interactive dashboards and portals hosted on websites or apps, voice and speech interfaces, and robotic interfaces. Each type of user interface presents a different risk profile, and a comprehensive discussion would require a deeper level of analysis than this report allows for. However, some of the most common UI risks relate to how adversarial actors can leverage the interface to attack the system, stealing valuable information or manipulating its behaviour. For example, conversational interfaces are conducive to prompt-based system manipulation and sensitive data disclosure by unaware users. Voice and speech interfaces are vulnerable to voice spoofing or misrecognition and can inadvertently collect sensitive audio data from the environment. Lastly, APIs can be targeted to gain unauthorised access to the model or reveal sensitive information and are vulnerable to cyberattacks aimed at shutting down the availability of the service.

Sensing and edge devices

AI systems can collect data from the physical world through sensors that convert environmental signals into digital information. The main categories of sensors include audio, visual, motion, position, temperature, pressure, touch, proximity, and biometric sensing. In some cases, sensors are integrated with computing devices that can preprocess the data locally and share summaries or analyses with the central system. Computing devices located close to the data sources are generally referred to as edge devices. These devices are vulnerable to both technical malfunctions and various types of adversarial attacks and tampering. From a technical point of view, they present a risk of containing defective components, degrading over time, being sensitive to stressful environmental conditions, and sensor miscalibration that can lead to bias in the data collected. In some cases, electricity and network connectivity might also be factors to consider when assessing potential causes of system failure. From a security perspective, sensing and edge devices can be stolen and physically manipulated or damaged to

compromise data capture and negatively affect the system. These devices can be targeted by cyberattacks to either damage their software components or intercept the stream of data they share with the central system. Additional cybersecurity vulnerabilities can also arise from their integration with communication protocols such as Wi-Fi or Bluetooth, as well as APIs and user interfaces. Finally, for devices that can capture personal data – e.g., cameras, biometric sensors, fitness and health devices – organisations should also address considerations related to compliance with relevant privacy and data protection regulations.

Physical actuators

In addition to sensing the world, AI systems can also act on it through the integration with physical actuators. Actuators are physical components that convert the output of AI systems into physical actions. Examples of physical actuators include electric, pneumatic, and hydraulic motors, steering and braking components, manipulation and handling devices, exoskeletons, valve and pump actuators, lighting controls, and printers. Considering that actuators can enable AI systems to directly impact the physical environment, the systems controlling them could cause serious safety incidents, including physical harm, if they fail to function as they should. Potential sources of risk to consider in relation to actuators encompass both their technical performance and their security. Factors that can affect the technical performance of actuators include the durability of their components, their robustness to harsh environmental conditions, their calibration, precision, response time, latency between the decision taken by the central system and the response of the actuator, including potential connectivity issues, and issues related to the integration of software and hardware components, including errors in the communication between the central system and the actuators. From a security point of view, potential sources of risk to consider include both cybersecurity vulnerabilities – such as the risk of firmware manipulation and denial of service attacks – and physical safety risks, especially in cases where malfunctions in the system

could harm humans interacting with it or the surrounding environment.²³

Deployment environment

Finally, another fundamental set of risks arises from how the AI system interacts with its deployment environment – the specific domain in which it operates. Identifying these risks implies analysing the characteristics of that domain and the entities within it. For example, if the system interacts with humans, organisations should consider whether they are vulnerable subjects or whether their fundamental rights could be compromised. Furthermore, organisations should consider whether the system could damage protected property, critical infrastructure, or natural resources. Lastly, it is crucial to evaluate how various actors might foreseeably use—or misuse—the system and the potential consequences of those actions.

This category outlines the primary sources of risk associated with common deployment environments. However, it is important to always assume that specific applications will introduce unique risks beyond the general challenges highlighted here.

Complexity of the deployment environment

An AI system's deployment environment includes the combination of hardware, software, human operators, and contextual factors within which the system is deployed and performs its tasks. The greater the complexity of an AI system's deployment environment, the higher the likelihood and potential impact of risks associated with its operation. The complexity of the deployment environment can be determined by factors including:

- The nature and structural characteristics of the environment, such as whether the deployment environment is fully or only partially observable, deterministic or stochastic, static or dynamic, discrete or continuous, digital or physical, and if it

works following known or unknown rules.


- The sensitivity of the functions in which the AI system is embedded.
- Whether the AI system is integrated into other systems or operates independently.
- The type of entities that can interact with the system in each environment.
- The regulatory and compliance landscape of the environment.
- The environment's security threat profile.

For example, a system deployed across jurisdictions, in environments regulated in different ways, can increase the risk of non-compliance; a high-value system operating in an open environment will be more vulnerable to adversarial attacks; a system deployed in a highly dynamic environment and integrated into several other systems – such as an object recognition system for autonomous driving – will have a more complex risk profile compared to an AI system used to detect and filter spam emails. In some cases, such as under the EU AI Act, systems deployed in specific contexts are identified as especially risky and can be subject to more stringent safety requirements (e.g., critical infrastructure, employment, law enforcement).

Organisational context

As part of assessing the complexity of the deployment environment of AI systems, organisations should consider potential sources of risk related to the organisational context in which the system is deployed. Some risks may depend on the business function the system supports and the role it plays within it. Some business functions, such as finance and accounting, are highly regulated. Deploying a recommendation system in that context can pose compliance risks that would not exist if the same system were deployed in Marketing and Sales. Other business functions, like Customer Service and HR, are characterised by data-related sensitivities. Deploying an

²³ For a comprehensive list of safety considerations relevant for industrial robot applications, see, for example, ISO 10218-1:2025 Robotics – Safety requirements – Part 1: Industrial robots and ISO 10218-2:2025 Robotics – Safety requirements – Part 2: Industrial robot applications and robot cells.



AI system in those contexts requires setting up stronger data protection safeguards to account for a higher risk to information rights.

Social context

Another important set of considerations relates to the social context with which AI systems interact. Identical systems can have different risk profiles depending on the people, communities, and institutions they interact with. Vulnerable groups such as medical patients, elderly people, immigrants, and ethnic minorities are more affected by harms than general users. Risks like bias, system failure, and misinformation carry higher weight in sensitive social contexts. Furthermore, different cultural communities have different standards and expectations of fairness, privacy, and safety. In this sense, social contexts might have an impact on whether a certain AI product is welcome on the market. More generally, organisations should consider questions of social justice when identifying and mitigating risks. When AI systems are deployed in contexts where structural power dynamics exist – such as between government agencies or law enforcement and citizens, employers and employees, and educators and students – they have the potential to amplify existing inequalities.

Level of autonomy of the AI system

AI systems can be granted varying degrees of autonomy by their developers. Basic AI systems perform predefined tasks under human supervision and in a controlled environment. More advanced ones, however, can be programmed to take actions with minimal to no human supervision through the ability to call functions or leverage tools to interact with other systems in response to a prompt. These systems sense and learn from the environment, make real-time decisions, and initiate actions accordingly. Between these two extremes, there is a spectrum of degrees of autonomy driven by core product design choices. Higher levels of autonomy translate to a higher risk of harm in case of system malfunction, and each organisation is responsible for deciding how to navigate the trade-off between automation

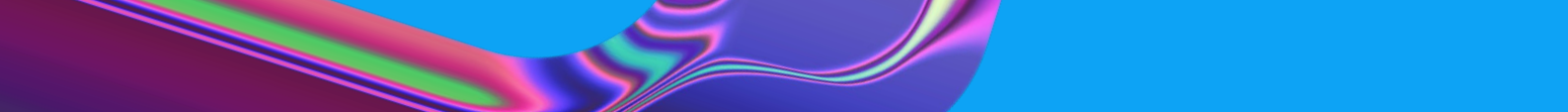
and safety. Autonomous systems can amplify an organisation's impact but can also cause incidents severe enough to put it out of business.

Interaction between AI systems

AI-to-AI interactions are another potential source of risks. When multiple AI systems operate in the same virtual environment (e.g., cloud APIs, financial trading platforms, content moderation systems) or the same physical environment (e.g., autonomous vehicles, collaborative robots, IoT-enabled factories), they can introduce new categories of risk that would not exist if they operated in isolation. These include instances where AI systems pursue conflicting goals, compete in ways that can create instabilities and unpredictable dynamics (e.g., in financial trading), create dependencies due to which failures in one system can cascade to others, and establish feedback loops where the output of a system becomes the input of another, leading to the reinforcement of errors and biases. Multiple systems interacting with each other could also complicate or prevent human oversight due to the lower predictability, transparency, and interpretability of compounded behaviours. Similarly, when harm occurs, it might be unclear which system was responsible, complicating the enforcement of liability regimes and the pursuit of redress by affected individuals and organisations. Safety risks increase exponentially when AI systems have the capacity to operate in physical environments. Multiple autonomous machines operating in the same space greatly increase the risk of accidents.

Misuse of the AI system

Organisations should also be mindful of the risks that can arise when AI systems are used in ways inconsistent with their intended purpose because of human error or malicious intent. For instance, employees who lack adequate training or the organisational incentives to prioritise workplace safety could inadvertently use AI systems in harmful ways. If, as a result, a system causes harm to third parties, the organisation could be vicariously liable for the employee's actions. In highly regulated sectors such as



healthcare, financial services, or insurance, failing to use the AI system as intended could also result in regulatory sanctions. Other examples of improper use could include feeding sensitive data into an AI system that is not designed to handle it securely or misinterpreting the output of the system, leading to overconfidence or misinformed decisions. AI deployers should also account for the risk of intentional misuse of their systems. Possible instances of such misuse include:

- using AI to create misleading content, such as artificial images, videos, audio messages, and fake news.
- monitoring employees beyond what's lawful or profiling individuals in unauthorised ways.
- obtaining sensitive information without authorisation.

Human oversight and validation errors

Humans might be involved at different stages of an AI system's workflow to improve its accuracy, transparency, and accountability. For example, a human-in-the-loop approach can provide a formal record of why a decision was implemented or overturned, leaving an audit trail that can be evaluated when needed. However, the added layer of human intervention can pose challenges. Human reviewers can introduce unintended bias or errors in the decision-making process if not adequately trained to interpret and review an AI system's output. In addition to the risk of introducing their personal bias, human reviewers are vulnerable to automation bias – a phenomenon where humans choose to favour automated recommendations over their own judgement, even when those recommendations are, in fact, wrong. Different human reviewers may also interpret the same output differently in cases where there is no clear right or wrong, and can get tired, distracted, or confused.

User over-reliance on the AI system

In AI-assisted decision-making tasks, over-reliance refers to a misplaced trust in the model's output when this is, in fact, incorrect. The risks of blindly trusting an AI system's

recommendation can affect both the organisation using the system and third parties interacting with it. For example, over-relying on the recommendation of an AI system used for cancer detection could lead to devastating consequences for the patient. Over-reliance risks can also be more subtle and unrelated to the accuracy of the system's output. When organisations or individuals excessively rely on AI systems to perform important tasks, potentially dangerous technological dependencies can emerge. Risks include:

- A decline in know-how.
- Skills degradation.
- Propagation of biases.
- Poor decision-making.

Environmental impact

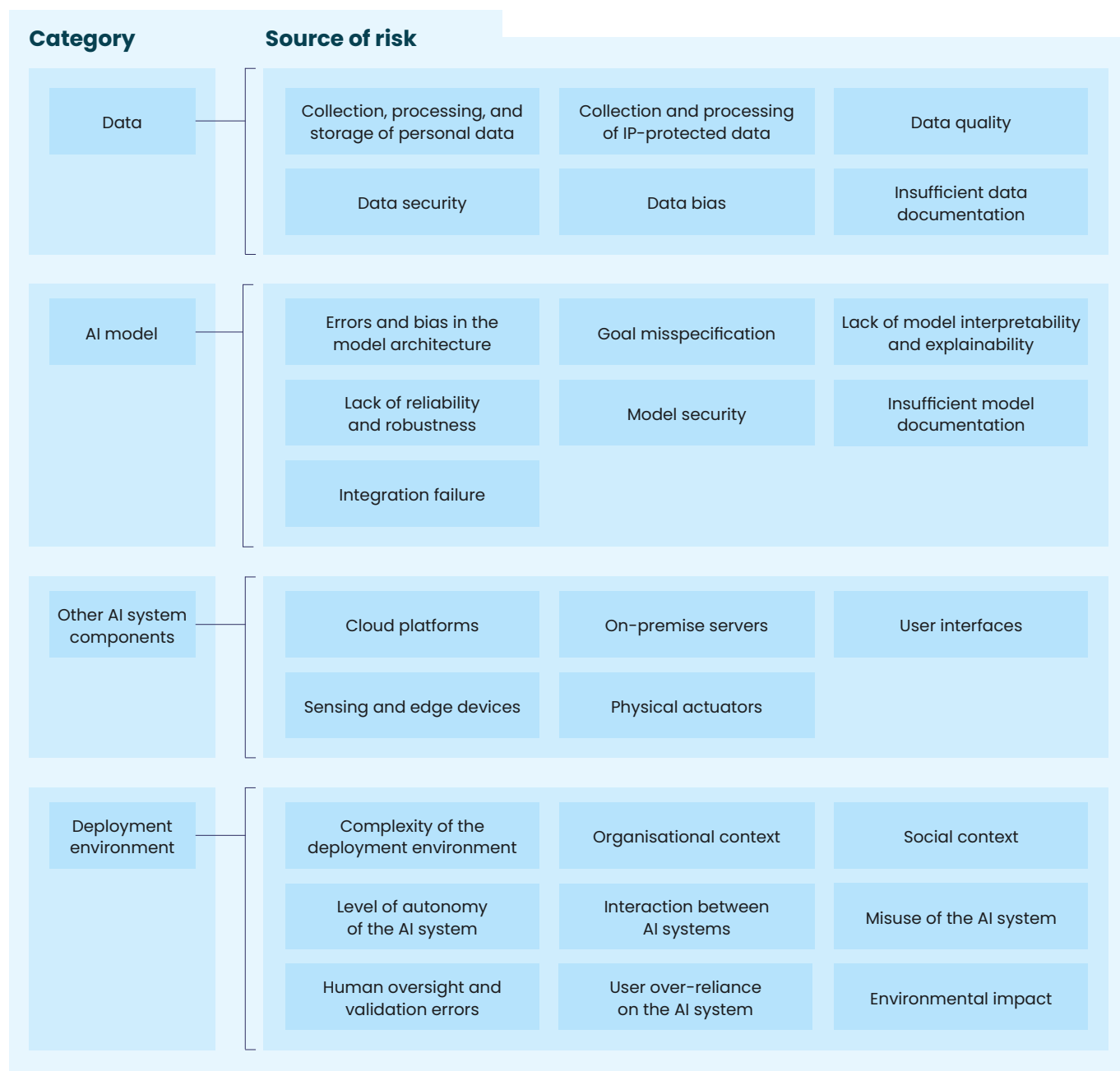
AI systems can have a negative impact on the natural environment due to the resource-intensive processes required for their development and operation. Factors determining the environmental impact of an AI system include:

- The computational intensity of the AI model.
- The energy sources used to develop and operate the AI system.
- The energy efficiency of the AI system's infrastructure.
- The volume of electronic waste generated when hardware becomes obsolete.
- Whether the AI system is used to optimise resource usage (e.g., smart grids) or triggers higher energy consumption behaviours (e.g., increased digital traffic).

The energy consumption of AI systems is increasingly becoming a regulatory and reputational risk for organisations. Under frameworks such as the EU's Corporate Sustainability Reporting Directive and its AI Act, or California's SB 253, some companies are required to report the emissions of their AI systems, including those generated by software and services that they buy. Failure to report these emissions could lead to penalties or prevent businesses from accessing certain

portions of the market where clients require vendors to meet certain sustainability standards. As investors and consumers become more aware of the environmental risks posed by AI systems and the data centres powering them, 'green AI' is becoming a brand differentiator.

Summary of AI risk sources



Part 3: Mitigating the risks of your AI use case

Once organisations identify potential sources of risk in their AI systems, they should assess whether these risks are serious enough to warrant mitigation. In risk management, this is known as the risk analysis and evaluation stage. If the risks identified are deemed unacceptable, organisations should implement measures to eliminate or reduce their likelihood and/or impact.

Considering the complexity and diversity of AI risks, managing them effectively will typically require a combination of different measures. To develop a comprehensive risk mitigation approach, it is useful to distinguish between mitigations across four different levels of control:

- 1. Product development mitigations** are technical and process-based measures integrated directly into the design, development, and deployment of AI systems to improve systems' technical accuracy, robustness, safety, and security.
- 2. Management and oversight mitigations** are measures that establish structures, processes, and rules throughout the organisation to ensure human and institutional control, accountability, and oversight across the AI lifecycle.
- 3. Information and transparency mitigations** are disclosures and other mechanisms that promote internal and external transparency around AI systems, their risks, and limitations.
- 4. Educational and cultural mitigations** are to training and value-alignment interventions to strengthen the organisational understanding of AI risks and embed responsible AI competencies into the organisation's culture and decision-making processes.

Each control level includes mitigations that can be used to address risks related to the four categories of risk sources discussed in the previous section: data, AI model, AI system, and the interaction between the system and its deployment environment. The following sections provide an overview of the most important risk mitigations for each level of control.

Product development mitigations

Product development mitigations are the first line of defence against technical and operational risks such as bias, inaccuracy, security vulnerabilities, and privacy breaches. They operationalise responsible AI principles through design and engineering choices, implementing preventive or detecting controls that address risks at their source.

Important product development mitigations include:

Data profiling and quality assessments

Data profiling is a technical analysis of the content, structure, and quality of a dataset. Assessing the quality of a dataset that will be used to test and train an AI model – including the integrity, accuracy, completeness, consistency, and relevance of the dataset – is especially important to address potential sources of risk that could compromise a model's performance, robustness, and security.

Data cleaning

Organisations can use various techniques to filter out unwanted information from datasets, including duplicates, wrongly formatted data, missing values, and other anomalies. These techniques are employed to ensure appropriate data quality and, as a result, higher accuracy and reliability of the AI models that are trained or tuned on such data. Ensuring datasets are clean also helps prevent the accidental retention of sensitive data, supporting compliance with data protection regulations and reducing security risks.

Data debiasing

Organisations can apply various techniques to reduce bias in the datasets they use to train and tune AI models. Some of the most common include:

- **Data augmentation** refers to methods used to supplement incomplete or imbalanced datasets with missing real-world data. Augmented datasets tend to be larger and more diverse than the datasets they are derived from and can be used in AI model training to reduce overfitting and improve robustness.
- **Random sampling** involves selecting a random subset of data points from an input dataset, where each data point has an equal probability of being selected, to create balanced training and testing datasets.
- **Synthetic data** refers to artificially generated data that can be used to train AI and test models as an alternative or complement to natural data. The use of synthetic data has different benefits, including giving developers more control over the statistical properties of the dataset, balancing out bias in the natural data, and avoiding using personal information of real individuals and organisations. However, if not properly generated or validated, synthetic datasets can still reveal training patterns that can compromise privacy.

Data security

Data security is the practice of maintaining the confidentiality, integrity, and availability of an organisation's data. Ensuring the security of personal data is also a legal obligation under both the UK's and the EU's GDPRs. Some of the most common data security interventions include:

- **Access controls and authentication protocols** ensure that only authorised users can access certain data.
- **Data backup and recovery** protocols support an organisation's ability to recover its data after incidents like hardware failures, disruptive cyberattacks, and natural disasters.
- **Data encryption** is the process of encoding information through a secret key so that only those who possess that key can access it. Organisations use encryption to protect sensitive data from unauthorised access and mitigate the severity of data breaches.
- **Data erasure** is a technique that allows full overwriting of data, making them irrecoverable after disposal and preventing unauthorised access or data theft.
- **Data versioning** is the process of storing different versions of datasets used during the AI lifecycle. It lowers the cost of errors and incidents by giving organisations the option to revert to previous data versions.
- **Differential privacy** is the method of adding calculated noise to a dataset to achieve a mathematical guarantee of privacy. By adding noise to the data, differential privacy makes it difficult for potential attackers to identify sensitive information in the dataset. Too much noise, however, can affect the accuracy and performance of the system.
- **Federated learning** is a decentralised approach to training AI where, instead of uploading data collected by a device to a central cloud server, an AI model is sent down to individual devices (e.g., smartphones), analyses the data locally and sends back only a mathematical summary of the analysis, while the raw data remain on the device. Often, data sent back from multiple devices are aggregated in a single average before being processed to anonymise the contributions of each source of data – a privacy-enhancing technique known as 'secure aggregation'.
- **Trusted execution environment (TEE)** are secure areas within a computer's processor that isolate sensitive information and code. TEEs rely on hardware-level isolation to ensure that even if the main operating system is infected by a virus, the secrets inside the enclave remain secure. These environments use techniques like strict access controls, memory encryption, and remote attestation to verify their integrity.

Adversarial training

Adversarial training is a technique to make AI models more robust to adversarial conditions. By training AI models to give the correct output for simulated adversarial inputs, adversarial training directly addresses vulnerabilities that make AI models fail or misbehave when targeted by adversaries or facing harsh environmental conditions.

Model debiasing

In addition to data-focused interventions, AI model bias can also be reduced during pre-training and fine-tuning by negatively reinforcing biased outputs. Debiasing methods include adjusting model weights to give underrepresented groups more influence on the model's output, introducing fairness constraints in the loss function, removing bias-related directions from embeddings (hard debiasing), and introducing structured knowledge bases in the dataset, such as explicit equality statements.

Fine-tuning

Fine-tuning is the process of optimising a pre-trained AI model for a specific application by leveraging the existing knowledge of a model and refining it with an additional level of training based on a narrower, application-specific dataset. In addition to customising the performance of AI models, fine-tuning can be used to mitigate several risks, including:

- Bias in the pre-trained model.
- Domain mismatch.
- Lack of robustness.
- Goal misspecification.
- Misuse.
- Lack of explainability.

Retraining

Whereas fine-tuning leverages additional data to optimise an existing model, full retraining means starting again by training the base architecture of a model with new or partially modified data, objectives, or constraints. Relative to other risk mitigation measures, retraining is a very costly intervention, but in some cases, it might be the only possible solution. For instance, retraining might be required for cases of data poisoning, discovery of systemic bias in the

training dataset, structural flaws in the model architecture or objective function, and severe model drift.

Functional safeguards

Functional safeguards are controls integrated into AI systems to prevent or mitigate unintended behaviour. Functional safeguards can focus either on mitigating harmful behaviour when an AI system interacts with a user, or on controlling who can access and interact with the system. Examples of popular functional safeguards include:

- **Content filters** that scan inputs and outputs for prohibited content (e.g., copyrighted content, sensitive data, hate speech, potentially dangerous information) and block or redact the interaction with the user. Content filters tend to work well in blocking obviously harmful content, but their effectiveness drops when users actively try to jailbreak them.
- **Hardcoded responses** for pre-defined sensitive topics or categories of conversation.
- **Circuit breakers** to limit how many interactions users can have with a system in a given timeframe.
- **Input validation** to detect instances of prompt injection and other adversarial interactions.
- **Role-based access control**, based on which users are granted the minimal set of access permissions they need for their role.
- **Network segmentation** to separate the networks where the AI model runs from those used by other devices to better control traffic and improve security.
- **Safety red teaming**, where specialised teams stress-test AI systems to evaluate their behaviour and alignment with guidelines and policies, even when it is encouraged to violate them. Safety red teaming is useful to mitigate the generation of potentially harmful content and hallucinations.

Security testing

Security testing refers to a set of methods and techniques that organisations can use

to identify and address vulnerabilities in their AI systems before adversaries exploit them. Security testing encompasses both the data, model, and system levels, and includes techniques such as:

- **Access control and infrastructure security testing**, where specialised teams evaluate potential gaps in the system infrastructure (e.g., servers, cloud services, APIs, edge devices) that can be exploited to gain access to sensitive information or modify the behaviour of the AI model. Some of the most popular tests falling under this category are penetration tests, secret and key management tests, and privilege escalation tests.
- **Data poisoning and integrity testing**, to verify that the data pipeline of the AI model is resilient against malicious data injections that can modify the model's behaviour.
- **Model theft and inversion testing**, where an AI system is tested to assess the extent to which attackers can extract sensitive information from the system, including proprietary information about the AI model, trade secrets, and sensitive training data.
- **Security red teaming**, where specialised teams conduct attack simulations to test how AI systems respond to realistic adversarial threats, including interactions with other AI and IT systems.

Monitoring and evaluation

By regularly monitoring and testing AI models, organisations can detect and address sources of risk early, implement timely and targeted interventions, and reduce the potential negative impact that those risks could cause. AI models should be tested and monitored both pre- and post-deployment.

- **Pre-deployment evaluation** is where organisations verify that an AI system is fit for purpose before it interacts with its deployment environment. It can be done in a simulated or sandboxed environment and is essential to gather empirical evidence of how an AI model behaves when applied to real-world tasks. Key features tested before deployment include performance,

robustness and security, fairness, explainability and transparency, and data integrity.

- **Post-deployment monitoring and evaluation**, instead, is where organisations verify that AI systems remain fit for purpose once deployed in the real world. Post-deployment monitoring and evaluation must be performed continuously while an AI system is operational and is often done through tools such as:
 - » User interaction tracking systems.
 - » Out-of-distribution detection.
 - » Misuse detection.
 - » Usage pattern analysis.

The insights collected through post-deployment monitoring and evaluation are essential to detect, among others, early signs of model drift and degradation, integration failures, emergent capabilities, or malicious misuse, and can be used to activate the implementation of bespoke safeguards.

Edge computing

Central to the concept of edge computing is the idea of moving inference workloads closer to the source of data. This can be useful to mitigate specific security and performance risks. For example, processing data locally without transferring them to a centralised cloud can reduce the risk of interception or leakage of sensitive information. Additionally, in applications where real-time response is critical, edge computing can help mitigate the risk of delays due to network latency and the risk of service disruption due to outages or poor connectivity.

Fail-safe design and calibration of system components

When AI systems rely on sensors and actuators to perceive and interact with the physical environment, ensuring they follow the principles of fail-safe design and are properly calibrated is essential to minimise risks. A fail-safe design is a strategy where a system defaults to a known safe state in case of failure. For example, an actuator can be designed with a mechanism that forces it back to a safe position in case of power loss. Calibration, instead, is the process of ensuring that the digital model of a component

matches the physical reality. For instance, sensor readings can drift over time even as inputs stay the same. This is due to factors like challenging environmental conditions, contamination, and ageing components, and can lead to a deterioration of the system's performance. Similarly, miscalibrated robotic arms risk harming humans, damaging equipment, or causing defects in production. In general, recalibration interventions should be periodically planned in all cases where AI systems that interact with the physical world require precision and repeatability.

Shutdown mechanisms

A system shutdown mechanism – sometimes called a kill switch – is designed to stop an AI system's operation when it behaves unpredictably or dangerously. Shutdowns may be costly for organisations, but they can be necessary as a last-resort measure to avoid the escalation of potential damages. For example, in cases where an AI system has been hacked by adversarial actors, a forced system shutdown could limit the leverage of attackers.

AI redundancy

A common method to improve the resilience and reliability of AI systems, especially in high-stakes or mission-critical applications, is to exploit redundancies to manage failures and ensure the desired level of functionality. Redundancies can be built into both hardware, software, and the computational infrastructure. Organisations can even deploy two or more systems that run in parallel so that if one fails or behaves unpredictably, the other takes over to ensure continuity of operations.

Carbon footprint reduction techniques

Organisations can implement several strategies to reduce the carbon footprint of their AI systems. These include:

- Right-sizing AI models to tasks. Organisations can often save resources by choosing smaller, purpose-built solutions over very large, general-purpose models with outsized capabilities relative to the task they need to perform.
- Prioritising fine-tuning existing models over developing new models from scratch.
- Scheduling high-intensity compute workloads to run during times of the day when the local energy grid is powered by renewables rather than coal or gas.
- Leveraging model optimisation techniques – such as quantisation (reducing the precision of the numbers used to represent the model's parameters and computations) and distillation (teaching a smaller 'student' model to mimic a larger 'teacher' model) – to improve their energy efficiency.

Continuous maintenance

After deploying an AI system, organisations should undertake the necessary activities to ensure that both its software and hardware components function correctly, efficiently, and securely. Continuous maintenance is necessary to correct faults, improve performance, or adapt a product to a modified environment to prevent downtime and failures in the system.

Management and oversight mitigations

Management and oversight mitigations provide the organisational scaffolding that ensures AI systems are not only technically sound but also embedded in a framework of human and institutional controls to monitor, detect, and address unforeseen and potentially harmful events across the AI lifecycle.

Important management and oversight mitigations include:

AI management system

Organisations should develop, implement, and continually review an AI management system – i.e., a structured framework of policies, processes, and controls designed to oversee the responsible development and use of AI systems. A well-structured AI management system should include:

- An organisation-wide policy specifying what the objectives of the organisation's AI projects are, how they link to broader organisational

objectives, and how its approach to managing them contributes to responsibly advancing them.

- A detailed description of the roles responsible for managing AI projects within the organisation.
- Documentation relative to the AI systems developed and/or used by the organisation, including their intended purpose, context of application, and potential sources of risk.
- An AI risk management framework to identify and treat risks posed by AI systems.
- An assessment of the potential impact of AI systems on individuals, groups, society, and the natural environment.
- An assessment of the competencies of the persons doing work that affects the organisation's AI projects and, where needed, a plan of action to secure them.

Organisations should conduct regular internal audits to assess whether the AI management system is effectively implemented, maintained, and updated as needed.

AI impact assessment

AI impact assessments are processes designed to identify, analyse, and document the impact of an AI system on the organisations that develop, provide, and operate it as well as on individuals, society, and the natural environment. AI impact assessments should cover the entire lifecycle of AI systems, from the design and development stage until after their deployment and evaluate both direct and reasonably foreseeable impacts. AI impact assessments provide a systematic and auditable way to manage ethical, social, and environmental risks of AI and should be seen as complementary to the risk management and broader AI governance approach of an organisation. Impact assessments are evidence of due diligence, they support regulatory compliance and signal trustworthiness.

Data protection impact assessment (DPIA)

Data protection impact assessments are processes that can be used to identify, minimise, and document data protection risks related to the development and operation of an AI system. Under the UK GDPR, DPIAs are mandatory for

projects that are likely to result in a high risk for individuals. However, even when not mandated by law, DPIAs can be valuable risk mitigation tools. They help organisations map data flows, carefully consider trade-offs between privacy risks and commercial benefits, and implement measures to reduce those risks where appropriate. The goal of DPIAs is not to eliminate all risk, but to demonstrate that the organisation recognised them and took reasonable steps to manage them. They are typically implemented before organisations start processing data.

IP safeguards

Intellectual property and copyright violations can occur when copyrighted information is used to develop or operate an AI model. While the legal implications of using copyright-protected work for training or operating AI models in most jurisdictions are still being defined, there is a growing corpus of cases where organisations have been sanctioned for copyright-related violations related to their AI practices. To avoid financial and reputational drawbacks, organisations can implement a series of safeguards to reduce the risk of violating IP rights. These include:

- Conducting due diligence on data vendors and scraping tools, and negotiating contractual clauses requiring data vendors to warrant that their datasets have been obtained in lawful ways.
- Implementing usage restrictions if the model's outputs are generated using work that could infringe on protected IP rights.
- Integrating IP compliance checkpoints into the AI model lifecycle (before training, before and after deployment).
- Licensing copyrighted data from copyright holders.
- Tracking the lineage of datasets by recording how each dataset was obtained, whether it contained copyrighted work, what transformations were applied to the data, and how the data was used in model training.
- Vetting data sources and filtering out copyrighted information.

Data audit

Data audits are structured reviews and evaluations of the data used to develop and operate an AI model, including how the data is collected, stored, used, and protected throughout its lifecycle. Data audits are performed after the deployment of an AI system and aim to evaluate the data governance of an organisation and ensure compliance with relevant policies and regulations. Data audits can be performed by internal teams or by specialised third parties and have become important tools to identify and mitigate risks such as poor data quality and bias, security vulnerabilities, privacy breaches, and other compliance risks.

AI audit

AI audits are structured evaluations performed to verify whether an AI system adheres to internal organisational policies, applicable laws, and standards governing its development and operation. AI audits are comprehensive reviews that combine technical, procedural, and legal checks to ensure the AI system does not introduce unacceptable risks. They can be conducted by dedicated internal teams or by third-party organisations and are valuable instruments to identify and address risks related to regulatory compliance, security, and performance, including robustness, accuracy, and fairness issues. By regularly auditing their AI systems, organisations can identify risks earlier, reducing their potential impact.

Third-party assurance and certification

Organisations can assure the quality and security of their systems through independent evaluations performed by specialised third parties. Third-party evaluations can cover different aspects of an AI system, including its training data, the performance of its model, its overall safety, and the security of its various components. Following the evaluation, providers of assurance services often issue a certification which the organisations that submitted their systems to evaluation can use as a stamp of approval signalling due diligence and trustworthiness.

Human oversight

Organisations can enable humans to play strategic roles in the operational management of AI systems. Embedding humans into AI workflows can be necessary to improve the accuracy, safety, transparency, and accountability of AI systems. In some cases, ensuring human oversight of AI systems can also be a legal requirement. For example, Art. 14 of the EU AI Act requires that AI systems used in high-risk applications are effectively overseen by natural persons during the period in which they are in use. There are several ways in which humans can be integrated into the workflow of AI systems, including:

- By verifying the correctness or fairness of the AI system's outputs. Humans can validate or override the system's output before it is used to trigger actions or decisions, or in real time, while the decision is being taken. In the latter case, humans won't be able to prevent a potentially harmful decision or action.
- By providing feedback to help the AI systems improve and adapt to a changing environment.
- By complementing the AI system's decision-making capabilities in especially complex or sensitive scenarios, where humans may have a better understanding of cultural norms and ethical grey areas.

Effective human oversight requires competency, authority to intervene and eventually override the AI system output, real-time visibility into the system's operations, and clear escalation paths. Considering the potentially high volume of outputs of certain AI systems, human reviewers can struggle to validate every decision taken by the systems they are responsible for. Popular approaches to dealing with large output volumes include reviewing random output samples or target outputs that fall outside a certain confidence threshold. In all cases, it is important to record and document all human interventions to enable the creation of audit trails that can be reviewed as necessary, ensuring transparency and accountability.

Incident monitoring and response

Assuming that AI systems might fail at some point, organisations can develop plans to prepare for scenarios in which they might have to respond to incidents caused by such failures. A comprehensive incident response plan identifies the information that should be recorded about AI incidents, the stakeholders that should receive such information, the processes and measures that would be implemented in each foreseen scenario to minimise the impact of incidents, and the responsibilities for the execution of those processes. Having a structured and comprehensive incident management policy helps organisations control AI risks, improve their accountability and transparency, and position themselves as responsible and trustworthy market actors. Additionally, recording information about AI incidents can help better understand the causes of an incident. Such knowledge can then be leveraged to improve products and governance processes.

Accountability and redress frameworks

Accountability is a cornerstone of good governance. Organisations should be able to take responsibility for the decisions and actions of their AI systems, explain them and offer a remedy in case of harm. In this sense, organisations should identify the individuals responsible for the design, deployment, and outcomes of their AI systems, ensuring there is a clear and traceable chain of responsibilities for everything the AI system does. Organisations should also adopt mechanisms that allow individuals and other affected parties to seek remedies for the harm caused by their AI systems. This includes complaint procedures, internal investigation pathways, system-level corrections, and compensation schemes.

AI ethics committees and boards

AI ethics committees and boards are formal governance bodies tasked with the function of providing guiding principles for the ethical development and use of AI systems and resolving questions related to the ethical and societal implications that they pose. Typical duties of AI ethics committees and

boards include developing AI ethics policies, auditing the use or proposed use of AI systems, developing AI ethics training programmes for the organisation's employees, and examining proposals for AI research projects. AI ethics committees and boards typically bring together senior employees with diverse backgrounds and from different teams, including product development, legal, policy, human resources, external communications, and operations. In some instances, an organisation could decide to appoint an AI ethics board comprised of external members to ensure an even higher degree of independence and accountability.

Trustworthy AI procurement clauses

For organisations that do not build their entire AI workflow internally, procurement is the first line of defence in AI risk management. When procuring AI models, integrated systems, data services, or cloud infrastructure on the market, the trustworthiness of the vendor directly determines the risk exposure of their customers. Defining procurement clauses that promote contracts with trustworthy AI vendors is, therefore, a critical risk mitigation mechanism. By implementing trustworthy procurement clauses, organisations can ensure that the technology they purchase meets minimum safety, fairness, sustainability, and compliance standards; that liability if things go wrong is clearly allocated between vendor and buyer; and incentivise vendors to disclose information necessary to perform due diligence checks. Examples of such clauses include requesting:

- Model/system cards.
- Reports from third-party audits and security testing.
- The right to audit before concluding the transfer.
- Indemnity from potential IP violations caused by data practices implemented by the provider.
- The integration of the International Data Transfer Agreement framework for cross-border data transfers.
- Sustainability metrics (e.g., Power Usage Effectiveness ratings).

Information-sharing and transparency mitigations

Information-sharing and transparency mitigations aim to foster accountability, shared understanding, and informed decision-making by internal and external stakeholders. Internally, they enable different teams to make coordinated and informed decisions; externally, they facilitate assurance, trust, and compliance with transparency obligations and best practices.

Important information and transparency mitigations include:

AI documentation

AI documentation is the practice of recording information on how AI systems are designed, trained, and deployed. By maintaining comprehensive AI documentation practices, organisations can support the responsible development and use of their AI systems, including by enabling the identification and mitigation of downstream risk. AI documentation also improves the interpretability of AI systems and ensures accountability by facilitating regulatory oversight and legal redress in case of harm. Additionally, documentation practices can also benefit developers by incentivising scientific rigour and supporting reproducibility. In some jurisdictions, documenting AI systems can be a legal obligation. This is true, for example, for high-risk AI systems and general-purpose AI models under Articles 11 and 53 of the EU AI Act.


Key elements to consider in AI documentation include:

- **Data documentation** (e.g., via datasheets) provides information on all movements and storage of datasets, including data provenance and collection methodologies, permissible uses, annotation procedures, processing history, and ownership. Maintaining comprehensive data documentation has several important advantages. It helps data teams perform root cause analysis in case of unexpected results and identify ethical and security issues in the datasets, supports the protection of personal data, reduces the risk of IP and copyright violations, and facilitates data audits and quality assessments.
- **Model documentation** provides information on how a model is built, trained, and fine-tuned, how it is evaluated, and in what context it is intended to be used. Model cards are a common format for sharing information about AI models.
- In addition to information about the underlying AI model(s), **AI system documentation** should also include information about the system architecture, the compute resources needed to operate it, safety and security evaluations performed, its intended uses, risks, and performance. Increasingly, AI system documentation also includes carbon footprint data. Like AI models, the key features of an AI system are often summarised through templates called 'system cards'.

It is important to consider that AI documentation can serve different purposes depending on the audience it is intended for. AI documentation can be designed for both internal and external stakeholders. Internal documentation tends to be more technical and comprehensive, including proprietary and other sensitive information. It is generally aimed at enabling internal understanding of AI systems, control, evaluation, maintenance, and risk management. External documentation is primarily intended for stakeholders and regulators. It is designed to show compliance, trustworthiness, and accountability. It must balance transparency with confidentiality, and tends to be more legible and abstracted from technical detail.

Risk transparency

Risk transparency is the practice of providing visibility into the features of an AI system that could pose a risk for the organisation deploying it, its users, and other third parties that might be indirectly affected. Informing relevant stakeholders of the risks posed by an AI system helps promote a culture of risk awareness and supports informed and responsible use of AI systems, contributing to mitigating risks related to unintentional misuse. Information about AI risks is generally included in broader AI documentation practices. However, risk transparency goes a step



further. It involves communicating information about risks and safeguards in a targeted way to different stakeholders to mitigate the impact or likelihood of specific risks.

Disclosure of AI usage

Disclosing the use of AI in products and services promotes transparency, accountability, and risk mitigation. Clear disclosure – such as notifying users that they are interacting with an AI system or labelling AI-generated content – helps consumers make informed decisions and reduces the risk of misunderstanding and misuse. It also supports compliance with emerging regulations (e.g., the EU AI Act) and industry standards. However, excessive or poorly designed disclosures can overwhelm users and lose impact. Organisations should therefore develop a policy that defines when disclosure is necessary (e.g., where AI materially affects individuals or decisions), what information to include (e.g., purpose, limitations, human oversight), and how to present it (e.g., clear notices in interfaces).

Education and cultural mitigations

Education and cultural mitigations promote ensuring staff understand AI risks, limitations, and ethical implications, and contribute to the effectiveness of the other three control levels. Policies, process, and tools have greater impact if people understand them and internalise their purpose.

Important education and cultural mitigations include:

AI training

Staff skills, knowledge, and understanding of AI systems are integral to effective AI risk management. Without proper staff training, even technically sound AI systems can cause failures, compliance breaches, and ethical violations. To mitigate these risks, organisations should ensure that staff responsible for managing or using AI systems understand how they work, how to interpret their outputs, the risks they pose, the legal regime they are subject to, and how to behave in case of incidents or failures. Considering that organisations cannot train everyone on everything, AI training measures

should consider the technical knowledge, experience, and education of staff as well as the context in which the AI system is used. The more structured and role-specific training programmes are, the better they will support staff members with identifying and managing AI risks. A common approach is to develop increasingly detailed training modules for different categories of staff: a general module for all employees covering basic awareness, company policies, and prohibited practices; a user-focused module on prompt engineering, recognising errors and hallucinations, and incident reporting; a governance module on compliance, risk management, and procurement; and a technical module on privacy-preserving techniques and security. In all cases, however, organisations should design AI upskilling programmes that focus as much on efficient use as on safety, ethics, and regulatory compliance.

Responsible user guides

Responsible user guides are guidelines provided to users or operators of AI systems that explain how to use the system appropriately, safely, and ethically, including its limitations, risks, and safeguards. They are a more structured and comprehensive approach to risk transparency.

Embedding a safety-centric culture

Organisations can foster a positive and risk-aware culture by demonstrating a commitment to safety, security, privacy, accountability, and ethical values. In the context of AI, a safety culture means, for example, that a junior data scientist feels safe delaying a product launch because they found a bias in the training data, and they are rewarded, rather than punished, for doing so. Concrete mechanisms that organisations can use to foster this culture include:

- Implementing a blameless incident reporting policy: when accidents occur, the focus is on process failures rather than personal failures.
- Promoting the idea that safety is everyone's responsibility.
- Giving specific roles (e.g., the AI governance lead) the power to halt deployment if a risk threshold is breached or vulnerabilities are detected.

- Including safety and sustainability metrics as KPIs in performance reviews, rather than evaluating success based solely on model performance or speed of delivery.

Summary of AI risk mitigations

Product development mitigations

- **Data profiling and quality assessment**
- **Data cleaning**
- **Data debiasing**
 - » Data augmentation
 - » Random sampling
 - » Synthetic data
- **Data security**
 - » Access controls and authentication
 - » Data backup and recovery
 - » Data encryption
 - » Data erasure
 - » Data versioning
 - » Differential privacy
 - » Federated learning
- **Adversarial training**
- **Model debiasing**
- **Fine-tuning**
- **Retraining**
- **Functional safeguards**
 - » Content filters
 - » Hardcoded responses
 - » Circuit breakers
 - » Input validation
 - » Role-based access control
 - » Network segmentation
 - » Safety red teaming
- **Security testing**
 - » Access control and infrastructure security testing
 - » Data poisoning and integrity testing
 - » Model theft and inversion testing
 - » Security red teaming
- **Monitoring and evaluation**
 - » Pre-deployment evaluation
 - » Post-deployment monitoring and evaluation
- **Edge computing**
- **Fail-safe and calibration of system components**
- **Shutdown mechanisms**
- **AI redundancy**
- **Carbon footprint reduction techniques**
- **Continuous maintenance**

Management and oversight mitigations

- **AI management system**
- **AI impact assessment**
- **Data protection impact assessment (DPIA)**
- **IP safeguards**
 - » Data vendor due diligence
 - » Usage restrictions
 - » IP compliance checkpoints
 - » Licensing
 - » Data lineage tracking
 - » Vetting data sources
- **Data audit**
- **AI audit**
- **Third party assurance and certification**
- **Human oversight**
- **Incident monitoring and response**
- **Accountability and redress frameworks**
- **AI ethics committees and boards**
- **Trustworthy AI procurement clauses**

Information-sharing and transparency mitigations

- **AI documentation**
 - » Data documentation
 - » Model documentation
 - » System documentation
- **Risk transparency**
- **Disclosure of AI usage**

Education and cultural controls

- **AI training**
- **Responsible user guides**
- **Embedding a safety-centric culture**

Conclusion

With AI set to enable new ways of generating value in nearly every industry, organisations should think strategically about how to integrate this technology into their operating model. To optimise the benefits of enterprise AI, it is imperative to move beyond hype and identify solutions that closely align with specific objectives and can drive measurable impact. This can pose challenges, especially for businesses at the early stages of their AI adoption journey. One of the major barriers preventing AI solutions from scaling across organisations is the difficulty of effectively controlling the technology. AI applications require carefully designed controls to ensure they are safe and reliable.

To help overcome these barriers, this paper proposes a framework designed to support organisations in laying out in a clear and structured format the key information to consider when assessing potential AI use cases, identifying the sources of risk that these may pose, and selecting the right risk mitigation strategies to control them.

Ultimately, we view this framework not as a rigid rulebook, but as a starting point that organisations can build upon and adapt to their specific needs, ensuring that innovation and safety advance in parallel.

Glossary

AI governance: A set of structures, policies, tools, and processes that organisations use to ensure AI systems are effectively controlled and aligned with strategic goals throughout their entire lifecycle.

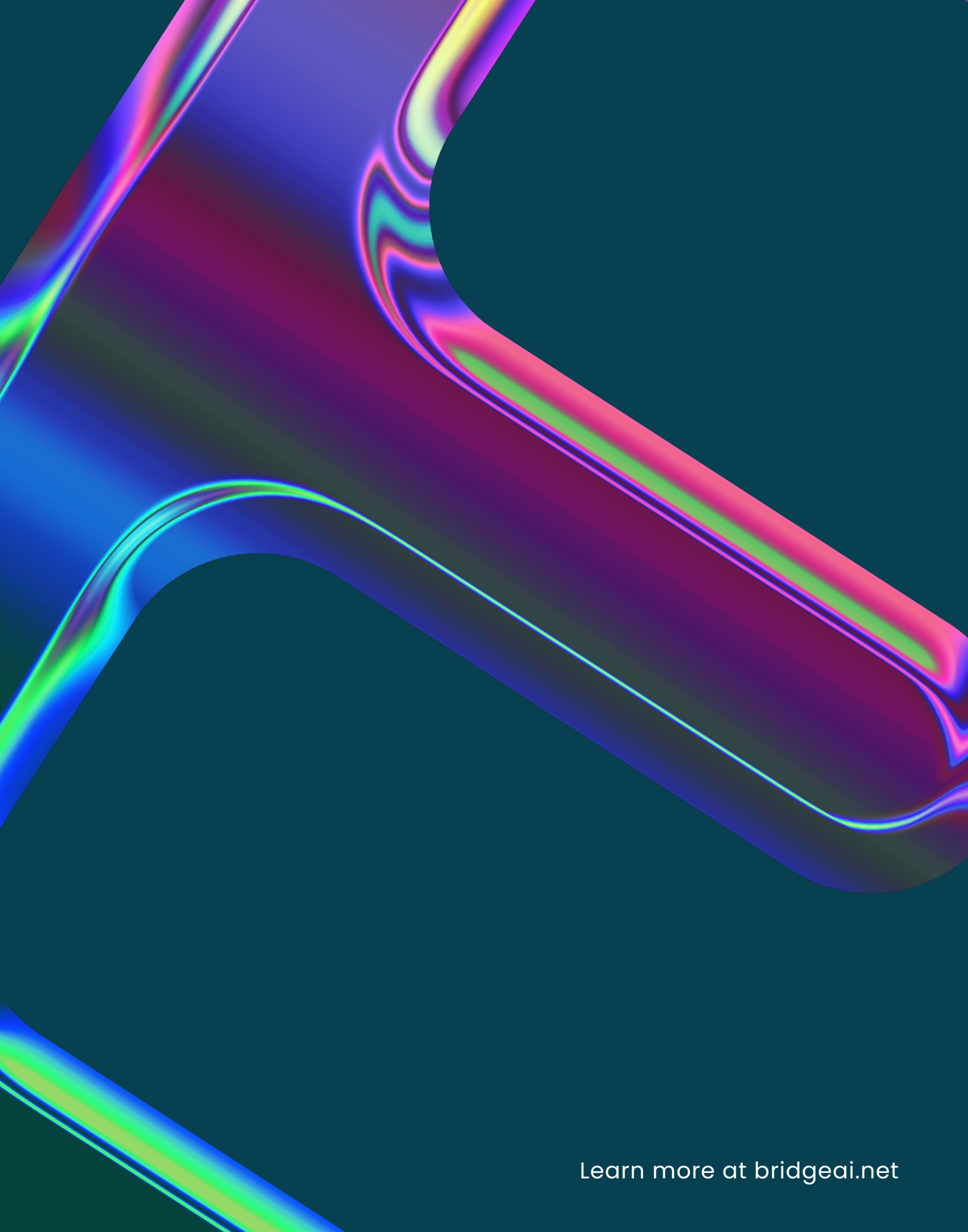
AI use case: A specific, practical application of an AI system.

AI system: An integrated set of components, including AI models, data pipelines, software, and hardware, designed to perform a set of given tasks.

AI model: A mathematical construct obtained by training a machine learning algorithm on a given dataset. An AI model will typically generate an inference or prediction based on input data.

Risk: the combination of the probability of an occurrence of harm and the severity of that harm.

AI risk management: A continuous, lifecycle-wide process for identifying, assessing, mitigating, and monitoring AI-related risks.



Learn more at bridgeai.net