PROTECHT

E-BOOK

Managing AI risks: Turning uncertainty into advantage.

A practical guide to embedding AI risk into enterprise governance, risk and compliance with Protecht's AI governance solution.

protechtgroup.com

Executive summary

AI is moving fast; organisations need to act decisively to understand and manage the uncertainty it introduces. *Managing AI risks: Turning uncertainty into advantage* provides a blueprint for embedding AI risk into the heart of enterprise governance, enabling confident innovation without losing control.

Why AI risk matters

AI brings transformative potential but also opacity, unpredictability, and the risk of overconfidence. Without clear guardrails, the same tools that enable efficiency can produce reputational damage, compliance failures, or strategic drift. Risk management isn't about slowing progress: it's about accelerating the right kind.

Understanding the AI risk landscape

AI is not one thing: it's a spectrum of technologies, from predictive models to generative agents, used across every industry. Each implementation reshapes existing risks and creates new ones. Understanding AI risk means understanding how AI interacts with objectives, processes, and people. A structured, objective-led risk view is essential.

The shifting AI regulatory landscape

Regulation is catching up, and fast. The EU AI Act sets a global precedent for risk-based governance, while ISO and NIST frameworks provide practical guidance. Even in jurisdictions without specific AI laws, existing regulations around privacy, discrimination, and misleading conduct still apply.

Managing the risks of AI

AI demands tailored, proactive risk management. That means identifying not just AI-specific risks like hallucinations or model drift, but how AI amplifies broader threats like cybercrime or fraud. Organisations must define clear risk appetite, choose fit-for-purpose treatments, and deploy preventive, detective, and corrective controls that reflect how AI is actually used.

Integrating AI risk into enterprise risk management

AI does not sit on the side: it must be managed within the enterprise risk framework. From governance and risk appetite to operational resilience and third-party oversight, AI should be treated like any other critical system. Standardising assessment, tracking governance indicators, and upskilling staff are vital for sustained assurance.

How Protecht can help

Protecht translates AI governance into real-world practice. The AI governance solution provides a centralised inventory of all AI systems (internal and third-party), capturing risks, controls, datasets, stakeholders, and decision gates across the lifecycle. Automated workflows enable assurance, testing, and reporting, aligned with global best practice. And with Cognita, Protecht's own AI capabilities are developed under the same governance principles, demonstrating safe, transparent, and trustworthy AI in action.

Takeout:

AI introduces uncertainty, but it also creates advantage. With the right risk approach, you can move fast and stay safe.

PROTECHT

Contents

1 Why AI risk matters.

While the term “artificial intelligence” has been around since the 1950s, the explosion of generative AI (and large language models in particular) has put it on the radar of nearly every organisation. The allure of the ‘super employee’ is strong, with common threads of automation, augmentation and acceleration of human capabilities.

This promise is already reshaping industries. In financial services, AI is powering fraud detection and credit decisions. In healthcare, it supports diagnostics and personalised treatments. In retail, it enables tailored customer experiences.

For all its promise, however, AI comes with considerable uncertainty. Many organisations have launched AI pilots with positive results, but struggle to scale. In a survey by KPMG in Q1 2025, only 11% of respondents were actively deploying AI agents. According to McKinsey in Q1 2025, only 1% of leaders call their company “mature” on the AI deployment spectrum.

What gives us confidence that it is safe? What infrastructure and skills do we need? And perhaps most importantly, what does AI-human collaboration look like?

Why we need AI risk management

As AI becomes more embedded in business operations, risk management becomes essential. This includes appropriate awareness and governance – if you don’t define the boundaries within which your people can and cannot use AI, you invite them to make up their own rules. Poorly managed AI can lead to inconsistent outcomes, ethical lapses, reputational damage, regulatory breaches, and even systemic failures. Perhaps worse is fear or lack of knowledge driving blanket bans on use of AI. Risk management becomes an enabler of AI adoption.

What differentiates AI from traditional technologies is its inherent autonomy and opacity. Systems can evolve over time, learn from biased data, or be misused in ways that are not immediately apparent. Without robust risk management, these risks may only be discovered after harm has occurred.

It’s not practical to put your head in the sand if you don’t want to keep up with the dynamic AI landscape. Firstly, existing risks, such as cyber intrusion or fraud are exacerbated by AI use from threat actors. Your people will need to know how to handle these changing threats. Secondly, you might avoid the immediate risks of poor deployment, but you miss the strategic advantages AI can bring – and that your competitors are probably using.

Regulatory drivers for managing AI risk

Across the globe, regulators are moving rapidly to catch up with AI's accelerated development. The European Union's AI Act, the proposed regulations in the United States, and Australia's emerging digital and AI frameworks all signal that organisations must prepare for tighter controls and compliance obligations.

ISO 42001 introduces a management system for AI, complementing ISO 31000's framework for enterprise risk. These standards recognise that AI risk management is not a standalone discipline, but a necessary extension of existing enterprise risk management (ERM) practices.

Regulators are increasingly demanding transparency in model development, fairness in outcomes, traceability of decisions, and accountability for impacts. Failure to comply may lead to legal consequences, financial penalties, and erosion of trust among stakeholders.

The speed of change: Why we need to act now

The velocity at which AI is being adopted and new innovations are being introduced leaves little room for passive observation. Unlike other technologies that follow a more predictable maturity curve, AI capabilities are evolving exponentially. New applications, threats, and dependencies emerge faster than traditional risk and control mechanisms may be used to.

This pace of change introduces a strategic imperative: organisations must develop the capacity to identify, assess, and respond to AI-related risks proactively. This means embedding AI risk management within existing governance and risk frameworks, rather than treating it as a siloed or future-state concern.

AI risk is not a problem to be solved later. Without forethought, AI technical debt can quickly build, with unsafe or insecure systems being built today that may prove challenging to correct tomorrow.

Takeout:

Effective AI risk management will be a key differentiator between organisations that thrive in the era of intelligent automation and those that fall victim to unmanaged uncertainty.

2 Understanding the AI risk landscape.

Artificial intelligence (AI) is not a single technology, but a broad category of systems designed to mimic aspects of human intelligence. These systems include machine learning, natural language processing, computer vision, and more.

Large language models (LLMs), particularly generative AI, represent the latest wave of advancement. These models don't just process data, they generate new content, provide contextual insights, and support human decision-making. In business terms, you can think of AI as an autonomous assistant, capable of augmenting workforces, streamlining operations, and enabling innovation – all at speed and scale.

Use cases across industries

AI's rapid evolution is unlocking applications that were once confined to science fiction:

- **Financial services** are deploying AI co-pilots for risk modelling and compliance, capable of autonomously interpreting regulatory updates and adjusting internal policies in real time.
- **Healthcare** is trialling GenAI models that interpret multimodal patient data (text, image, and genetics) to draft treatment plans, flag anomalies, and support clinical trials with synthetic patient data.
- **Retail** is leveraging AI to create hyper-personalised virtual shopping assistants that engage customers via immersive augmented reality (AR) experiences and simulate future buying behaviour.
- **Manufacturing** is embracing AI-driven design partners that co-create product prototypes, simulate materials under stress, and coordinate with autonomous robotic assembly lines.
- **Public services** are using generative AI for real-time policy drafting, community engagement in native languages, and scenario modelling for disaster preparedness.
- In **governance, risk and compliance** (GRC), AI is being applied to streamline processes and improve consistency. Tools like Protecht's agentic AI assistant Cognita streamline and simplify the GRC process. By embedding domain expertise and safeguards, AI can support more reliable decision-making while helping organisations adapt to evolving compliance demands.

Each of these use cases brings unique opportunities and unique risks.

The future of AI in GRC

AI in GRC is no longer a future aspiration. Many organisations have dabbled in pilot schemes; an analysis tool here, a dashboard there. But the real competitive edge comes from embedding AI into core workflows where risk decisions happen every day.

With Cognita in Protecht ERM, AI stops being a side project and becomes a governance-grade capability. It empowers every decision-maker, strengthens your risk culture, and gives leadership confidence that risk data is timely, consistent, and reliable.

Find out more

What is AI risk?

The ISO 31000 definition of risk is: "the effect of uncertainty on objectives."

The core concept of ERM is managing risks that could impact on those objectives. Consider the below risk and reward framework, and how AI integrates into the definition of risk.

Risk and reward framework for AI.

Objectives layer: Risk only matters if it materially affects objectives. What is material to one organisation may not be material to another. The left side of the pyramid represents operational objectives (running the organisation), while the right represents strategic objectives (changing the organisation).

Processes/projects layer: While objectives headline the business plan, the operating model is how they are delivered. This includes the critical processes and the resources to support them, such as people, systems and technology – including AI. On the right-hand side of the pyramid are strategic projects that deliver change into that operating model. This can include delivering AI initiatives, or project delivery methods enhanced by AI.

Risks layer: This layer considers the risks that can undermine achievement of objectives. The increasing use of AI in the operating model can give rise to novel risks, such as prompt injection or systematic bias. On the strategic side, failing to embrace AI may result in failed strategic objectives. Equally you may stumble on executing your strategic objectives, even if they are chosen wisely.

Enterprise risk management layer: This includes all processes to identify and manage risks. If management of AI and its risks is conducted independently of ERM, you may not have a complete view of risk, inefficient use of resources, or risks not being addressed.

Revisiting our ISO definition, AI risk can be defined as:

- The effect of uncertainty on AI objectives
- The effect of uncertainty created by AI, on objectives

The first relies on how you intend to use AI and introduces new risk. The second relates to how existing risks your organisation faces are evolving, whether you are embracing AI or not.

Defining AI objectives

Before risk can be meaningfully assessed, organisations must define what they want AI to achieve. Is the goal automation of a manual task? Enhancement of human decision-making? Will it interact directly with customers?

At a more granular level, how will that objective be achieved? Different implementations can give rise to different risks. If you intend to use or develop an AI model:

- Is it a third-party model or in-house?
- Is it open source or commercial?
- Is it locally hosted?
- Who will be interacting with the model?
- Who are the beneficiaries? Employees, customers, or other stakeholders?
- Will it be integrated into your products and services?
- Is it integrated with third party systems?

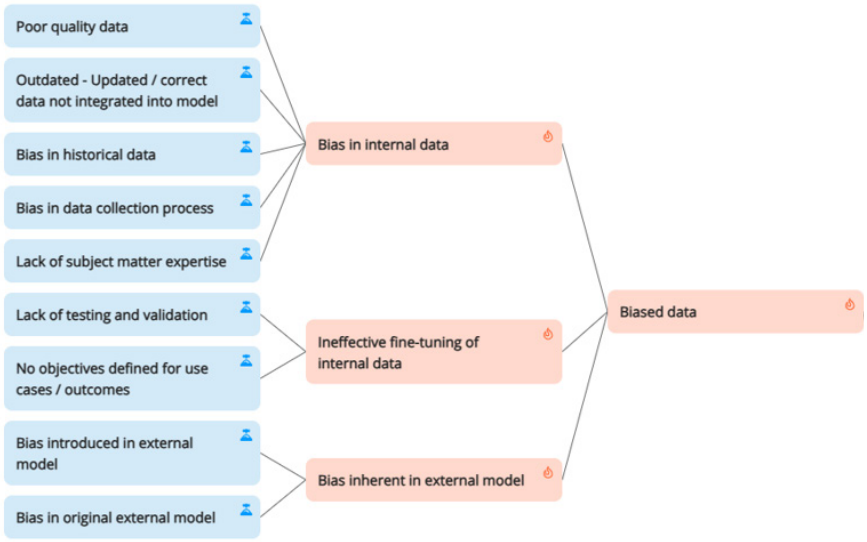
The answers to these questions will shape the specific risks your organisation faces. If you are haphazardly using AI for its own sake, you may end up with misaligned objectives while introducing risks beyond what is acceptable.

The components of risk: a bow tie example

At Protecht, we love risk bow ties as an effective way to both visualise and communicate risk by breaking it down into its key components[1].

- **Impacts:** The impacts on our objectives
- **Risk event:** The risk event that could prevent achievement of objectives. This typically represents the point at which it starts becoming out of your control
- **Causes:** The factors leading up to a risk event

Let's take a common use case of linking a large language model to internal data, creating an interactive knowledge base. For our example use case[2], both employees and customers can utilise the knowledge base to inform their decisions and actions. We define one of the risks as 'Inaccurate AI outputs'. If acted upon, incorrect information may have negative impact on our objectives.



Risk bow tie segment for AI.

Here are some of the ways this risk could unfold:

Biased data: There may be biased, incorrect, or incomplete data in the knowledge base. Users may trust the outputs without realising that bias may have been baked in. If updating processes aren't established, the data may also become outdated.

Data poisoning: Malicious actors could insert harmful or inaccurate materials into the knowledge base, influencing its outputs. That data could be changed through insider threats, compromised credentials enabling access to those data sources, or compromise of the data sources themselves.

[1] Read our Risk bow tie analysis eBook to find out more.
[2] You can download the full-size version of the bow tie example exerpted in this report here.

Prompt injection: Threat actors may trick the model into violating its instructions or bypassing its security features resulting undesirable behaviours such as data corruption or data leakage.

Ineffective prompts: Many LLM models are designed with 'system prompts' which provide overarching instructions to the model, which sit alongside prompts written by individual users. Users may not know how to draft their prompts to complement these system prompts and get the best quality output.

MIT's AI risk repository

One challenge to understanding AI risks is the speed of change, especially over the causal pathways of how risk can unfold. MIT developed the AI risk repository, which at last count housed over 1,600 risks[3].

As a mild critique, the library includes some repetition and calls out different components of risk. It nevertheless remains a great resource and reference point. Start with your AI objectives and use cases, then validate whether components in the repository map to your own context as causes, risks, or impacts. If you use the risk bow tie method above, several risks from the MIT library are likely to exist across a single bow tie.

Risk amplified by AI

AI isn't just about new risks being introduced. The causal pathways for several existing risks are also evolving, especially those with an adversarial component.

Cyber risks are becoming more complex, which can enable phishing attacks at scale, tailored to individual recipients. Many LLMs have coding abilities, and entrepreneurial threat actors can circumvent guardrails to craft novel malware or variants that may not be identified by existing cybersecurity tools.

Fraud may be centuries old, but AI provides new ways for it to be perpetrated.

Could it happen to you?

In a real fraud incident, a Hong Kong employee of UK engineering firm Arup transferred $25M across 5 different bank accounts. What initially sounded suspicious – an email from the UK-based CFO regarding the transfers – progressively appeared less so. The 'CFO' followed up with a request for a video call. You might believe you wouldn't fall for it, but the victim jumped on a call with not only the CFO, but also other colleagues. They all sounded and looked like their real colleagues – but all were fake.

Voice cloning, video deep fakes in real-time, and crafting fake invoices from a simple text prompt are not only possible, but the barrier to entry is rapidly reducing as AI tools become more capable. What worked to manage these risks yesterday may not be sufficient for tomorrow. Even if you are taking a cautious approach to adopting these technologies, your people need to be aware of these technologies, which can include tailored training for key targets for these types of adversarial attacks.

Takeout:

Identifying risks is not the end: it's the beginning. With a clear view of the AI risk landscape, organisations can begin to embed risk identification, assessment, and treatment into their governance structures.

[3] MIT AI Risk Repository

3 The shifting AI regulatory landscape.

If someone asks "what are the risks of AI", an obvious response is "risk to what or who?'. While we covered organisational objectives in the previous section, regulatory bodies have different objectives.

These usually include minimising harm to individuals, to society more broadly, and ensuring fairness. They aim to set minimum standards that organisations need to abide by when implementing or using AI.

European Union

- The EU AI Act, now in force, is the most comprehensive regulatory framework to date, requiring both providers and deployers of AI to adopt principles-based practices over the coming years.
- It introduces a risk-based system of obligations and is already being looked to by other jurisdictions as a model for future regulation.

United States

- The current administration has taken a permissive approach, rolling back earlier executive orders and reducing regulatory hurdles.
- The NIST AI Risk Management Framework provides voluntary guidance, but there is no federal regulation in place.
- Efforts to prevent individual states from imposing their own AI rules have created uncertainty. The federal government even proposed a penalty on states that enforced local AI laws[4], although this was ditched with individual states adopting their own positions, leading to fragmented enforcement.

The challenge for organisations

For businesses operating across multiple jurisdictions, these diverging approaches create uncertainty. The most practical strategy is often to identify the strictest set of requirements and implement governance to meet that standard globally. This reduces compliance risk, avoids duplication, and provides a consistent framework for managing AI safely and effectively.

The EU AI Act: setting the benchmark

The European Union has positioned the AI Act as the global benchmark for comprehensive AI regulation. It defines the responsibilities of providers (those who build AI systems) and deployers (those who use them), recognising that many organisations will be both. Central to the Act is a risk-based approach, classifying AI systems into unacceptable, high, limited, and minimal risk categories.

On the next page, we outline the key obligations under the EU AI Act.

[4] Reuters, AI regulation ban meets opposition from state attorneys general

High-level summary of EU AI Act levels and obligations

Prohibited	High-risk
Manipulative or deceptive AI	Limited biometric identification
AI that exploits vulnerable groups	AI used as a safety product or feature of any products listed in Annex I, including vehicles, machinery or toys
Social scoring models	Critical infrastructure
Crime prediction based on profiling	Education & training (e.g., admissions)
AI to expand facial recognition databases	Employment & worker management
Emotion inference in workplaces or education	Access to essential services
Biometric categorisation	Law enforcement
	Migration, asylum, and border control
	Justice and democratic processes

While not all of these may apply to your organisation, risk managers should scrutinise recruitment practices (employment and worker management) and any AI used for assessing creditworthiness or insurance pricing (captured under access to services). These could fall under "high-risk" applications, demanding additional compliance actions.

For high-risk systems, the obligations differ depending on whether you're a provider or a deployer.

Please note that these tables are simplified, please see Chapter III, Sections 2 and 3 of the EU Act for full details on provider obligations and Articles 26 and 27 for deployer obligations.[5]

Providers	Deployers
Risk management system	Human oversight
Data governance	Ensure data input is relevant and representative
Technical documentation	Use AI in accordance with instructions
Record-keeping	Notify provider of serious incident and cease use
Transparent information to deployers	Create and retain logs
Human oversight	Data protection impact assessment
Accuracy, robustness, and cybersecurity	Fundamental rights impact assessment
Implement a quality management system	
Automatically generate logs	

[5] European Union, Artificial Intelligence Act, Official Journal version of 13 June 2024, available via - EU AI Act Explorer.

Deployers can't simply point to the provider if things go wrong (as in vendor risk management, you can't transfer risk to the vendor). The deployer must demonstrate that they have used it in accordance with instructions (e.g. have not tried to circumvent the provider's safeguards), generate logs of activity and output, and – perhaps most important – include human oversight.

Transparency requirements

You might not have any high-risk systems. However, if you plan to integrate AI into your products or services, transparency requirements may still apply, such as:

- Providers must inform users they are interacting with AI (e.g. not pretending a chat bot is human).
- Providers of systems generating synthetic content must ensure it is machine-readable and detectable as AI-generated.
- Deployers creating synthetic image, audio, or video content (e.g., deepfakes) must disclose that the outputs are artificially created.
- Deployers creating synthetic text or content on public interest matters must disclose it is artificially generated or manipulated.

Even if you aren't regulated by the EU AI Act, some of its requirements simply represent good risk management which you may consider adopting.

Voluntary standards

In the absence of (or to assist you in meeting) formal regulation, AI standards and frameworks have emerged. The two most prominent are ISO 42001 AI Management System, and NIST's AI Risk Management Framework.

ISO 42001: The Global AI Management System Standard

Published in December 2023, ISO 42001 offers a management system standard for AI, mirroring ISO 27001 (information security). It supports organisations in establishing, implementing, maintaining, and continuously improving an AI Management System (AIMS).

Key features include:

- Organisational context and AI governance
- Risk and opportunity assessments specific to AI
- Documented objectives, policies, and controls
- Monitoring and improvement mechanisms

ISO 42001 is particularly valuable for global organisations that need a standardised approach across jurisdictions. It complements the Protecht view of ERM by reinforcing that AI risk management is not separate: it's embedded in the fabric of enterprise risk governance.

NIST AI Risk Management Framework

NIST's AI Risk Management Framework was released in January 2023, shortly after the launch of ChatGPT and the GenAI explosion. In July 2024, the NIST AI RMF was supplemented with the Generative Artificial Intelligence Profile to provide additional actions to address risks introduced or exacerbated by GenAI. The framework includes 4 key domains:

- Govern (embed risk culture)
- Map (to organisational context)
- Measure (risk posed by AI)
- Manage (prioritise and implement risk management activities)

If you are considering adopting a voluntary standard, here are two reminders:

- You don't need to be regulated in order to benefit from effective risk management
- The absence of 'AI' in the title of an Act or regulation doesn't mean your AI use isn't regulated.

AI regulation hiding in plain sight

While regulators and policymakers consider the need for specific AI regulation, don't overlook how existing laws and regulation already cover some of the negative outcomes that AI can produce. Consider regulations related to:

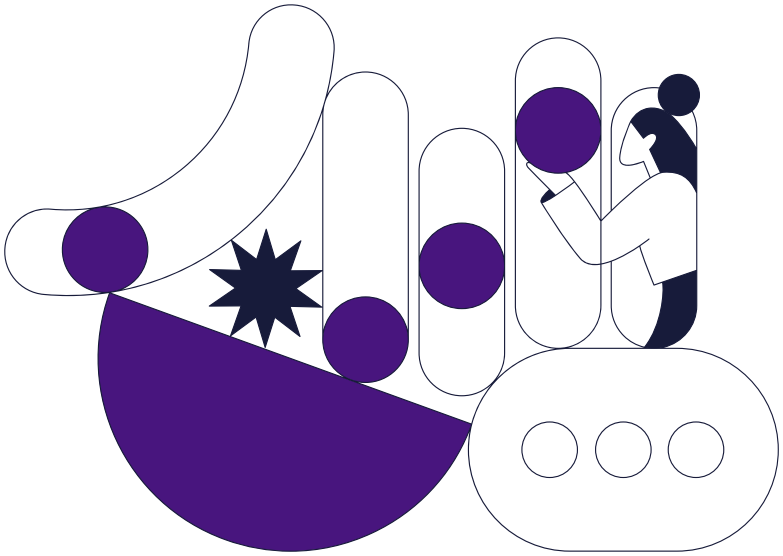
- Data privacy and cyber security
- Misleading conduct
- Discrimination
- Unfair practices
- Safety
- Product Liability
- Directors' fiduciary duties

Some regulators have been vocal and forthright that they will pursue enforcement for ineffective governance or risk management related to AI. ASIC, the corporate regulator responsible for issuing financial services licenses in Australia, released a report in October 2024 *Beware the gap: Governance arrangements in the face of AI innovation*. The message was clear – gaps in governance and risk management were evident amongst regulated entities, and entities needed to consider their existing obligations and duties.

Using AI might be optional to getting caught up in AI-related litigation. AI Washing is the newest flavour of misleading conduct or disclosures, with regulators charging companies who make false claims about their product abilities to leverage AI, riding the AI hype. In April 2025, the US Department of Justice and SEC charged the CEO of Nate Inc, who claimed their e-commerce platform was 'fully automated and scalable', when in fact offshore contractors manually completed the transactions[6].

Takeout:

Regulation is not just a compliance burden: it's a catalyst for better governance and more trustworthy AI. Those who act early will not only avoid penalties but also set the benchmark for ethical and effective AI use.



[6] White & Case, Evolution of AI Washing Enforcement, DOJ Enters the Picture

4 Managing the risks of AI.

Regardless of any regulation, the key is to take a risk-based approach. In this section, we will explore how to operationalise these insights into practical steps for managing AI risks within your organisation.

Identifying and assessing AI risks

Effective AI risk management starts with clear identification and assessment of AI-related risks. Risk identification must be comprehensive and contextual, focusing on how AI modifies both existing risks and introduces new ones.

One of the key aspects of strong governance and risk management is effective stakeholder engagement. AI use is no longer confined to technology teams; it is increasingly integrated into products, services, and processes. This can include shadow AI, where individual users use their favourite chatbot or tools in their daily workflow. Third party tools that are easy to integrate might fly under the radar and not enter a centralised inventory.

The scope of your risk identification may need to include both direct implementations (e.g. customer-facing chatbots) and indirect dependencies (e.g., third-party services leveraging AI tools).

The following are the key steps to identifying and analysing your risks:

1. Identify applicable objectives (organisational, division or process level)
2. Identify critical processes / steps (what must work well in order to achieve those objectives)
3. Identify risks (uncertain events or conditions that would affect those critical processes)
4. Identify causes (the specific causes that could result in the risk occurring)
5. Identify controls (measures already taken to address the likelihood or impact of the risk)

Depending on the scope of your assessment, you may limit this to AI objectives or processes supported by AI, or where AI acts as a cause for other risks.

By using methods like the risk bow tie to get intimate with the risk, it can improve the assessment of the level of the risk – the range of likelihood and potential impacts of the risks.

Defining risk appetite and treatment options for AI

Once you've assessed the level of risk, the next question to ask is "Is that level of risk acceptable?" AI can introduce different types of uncertainties compared to traditional technologies, so organisations may need to refine or supplement their existing risk appetite statements.

A risk appetite statement is intended to provide your people with freedom within boundaries:

- Freedom to take risks
- Freedom to make decisions
- Freedom to act
- Freedom to fail

Define acceptable levels of uncertainty for AI-related activities, such as use of generative models, automation of critical processes, or customer interactions. This can include qualitative statements (such as what types of AI use may or may not be accepted), supported by applicable metrics. Ensure the risk appetite is connected to strategic and operational objectives – if a defined objective is to be on the cutting edge of AI technology while having a 'very low' appetite for AI risk, this is likely to cause conflict.

Risk treatment options

After assessing the level of risk, action may need to be taken. At Protecht we consider seven types of treatment.

Treatment	Definition
Accept	If the risk is within appetite, it is automatically accepted. No formal action needs to be taken.
Process re-engineering	Change the underlying process so that the risk no longer exists. Perhaps the specifications of a particular AI model breach internal policy, and using another model removes that particular risk
Improve existing controls	Existing controls may not have been designed correctly or can be refined to reduce the level of risk. For example, controls to validate outputs may not be robust enough, and need to be enhanced.
Add new controls	Controls gaps may be identified. Perhaps you have strong controls to manage bias, but none to manage prompt injection.
Formal risk acceptance	Risk outside of appetite is accepted (with formal approval) for a finite period. More accurate and up to date information on a customer-facing AI chatbot implementation may increase the expected level of risk, which the board accepts for a 3-month period before a re-deployment can address the issue and ceasing use would cause major services impacts on both customers and the organisation in the interim.
Avoid	Stop the activity altogether that creates the risk. This also avoids the opportunity or upside. AI chatbots might provide efficiency, but if their answers are intermittently misaligned or harmful, avoidance may be the only option if other treatment options won't sufficiently address the risk.
Transfer	This transfers the risk – or at least the impact of the risk – to someone else. Where there is risk insurers see business opportunity, but the AI insurance industry is in its infancy. When it comes to third parties, contractual obligations might limit some of the damage, but rarely all of it. Even if you outsource an AI activity, you are still responsible for managing the risk.

Implementing controls is the common way to treat risks, but don't overlook the other options when they make sense.

A note of caution

There is a lot of (understandable) hype around AI. A common theme of questions in board rooms or executive discussion include 'How are we thinking about AI? How are we building it into our products and services? What are our competitors doing?' This can result in bullish AI pilots and implementations – perhaps without explicit consideration or even wilful ignorance of risk appetite. If you identify a mismatch, address it head on. It's fine to pursue opportunities linked to AI, just make sure you have the risk appetite to match.

Implementing organisational and technical controls over AI

You won't (or shouldn't) implement AI without controls in place. The ISO 31000 definition of control is 'a measure that maintains or modifies risk'. We use an extended definition 'A measure that is aimed at reducing the likelihood of a risk occurring and / or the impact if it were to occur.'

We recognise three main types:

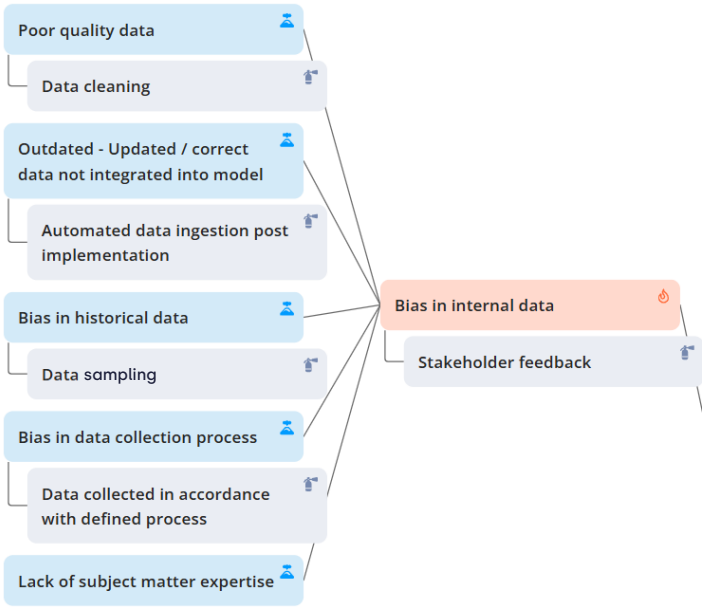
- Preventive – A control that addresses a cause of the risk so that it does not occur.
- Detective – A control that involves the collection and analysis which informs the level of risk. When the information passes a pre-determined threshold, an action needs to occur.
- Corrective – A control that responds to a risk event that has occurred in order to reduce the impact.

Some organisations also recognise directive controls – a control that provides instruction on how people should behave or act.

In some cases, standards or frameworks recommend controls not aligned with our definition. For example, ISO 42001 includes controls to assess the impact of AI systems. However, an impact assessment doesn't modify risk: it is part of a risk assessment. After you've conducted the assessment, you then determine whether you need to treat the risk, including applying controls. You may wish to classify your controls into organisational or governance controls, or a similar approach that separates your AI management system from operational controls over AI.

For governance or directive controls, it can be challenging to demonstrate how they reduce the likelihood or impact of specific risks. While important, they form part of the management system or risk framework, rather than modifying risks directly. When it comes to implementing AI, it will be the specific use cases and the effectiveness of technical controls that are more likely to prevent incidents or poor outcomes.

To consider some of those technical controls, let's return to our use case of creating an interactive knowledge base:



Risk bow tie segment – how bias in internal AI data can occur.

This segment of our risk bow tie breaks down how bias could occur in multiple ways, whether upfront during implementation, or drifting over time. This may warrant a range of specific actions to minimise the likelihood that bias will end up in the model.

It's important to recognise that the appropriate set of controls might, and probably should, differ across use cases or variations in implementation:

- Data ingested from external sources over which you have no direct control should be subject to additional controls or more frequent assurance to ensure they remain appropriate
- If models require personal or sensitive commercial information to be shared, different controls will be needed based on whether the model is locally hosted or using the provider's own servers
- If many employees can change data for the knowledge base, stronger audit controls may be needed to identify when changes are made, alongside regular benchmark assessments to assess for model drift

A comprehensive understanding of the way AI-related risks unfold enables appropriate selection and prioritisation of controls.

Takeout

Effective AI risk management starts with thorough risk identification, including both direct and indirect AI exposures, and continues through clearly defined risk appetite, thoughtful treatment strategies, and the implementation of meaningful controls. Governance, technical rigour, and control effectiveness must align with how AI is actually deployed in your business.

How Protecht approached AI risks in Cognita's design

When building Cognita, Protecht applied the same governance and risk principles that underpin our broader ERM framework. From the outset, we assessed how AI could both create value and introduce new risks — from bias and data security through to accountability and regulatory compliance.

Cognita was therefore developed with:

- **Human-in-the-loop oversight so that AI suggestions remain transparent and subject to sign-off.**
- **Traceable outputs linked to source content, ensuring recommendations can be verified.**
- **Role-based controls and audit trails to maintain accountability within established governance frameworks.**
- **AI knowledge base of Protecht's own risk expertise, avoiding the pitfalls of generic AI tools.**

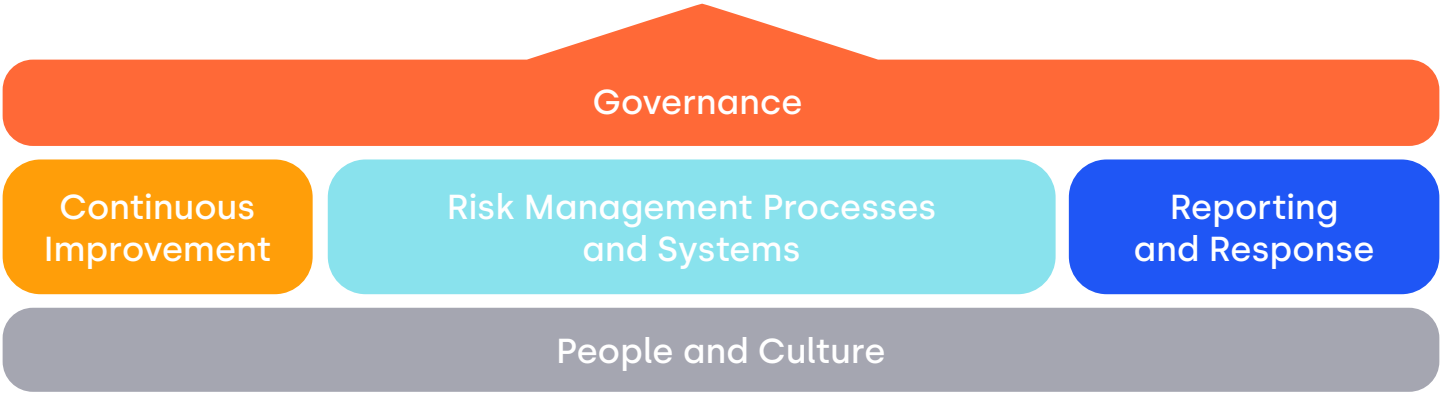
By embedding these safeguards into the product design, Protecht ensured that Cognita strengthens decision-making and risk culture, while remaining aligned with regulatory expectations and customer trust.

5 Integrating AI risk management into an enterprise risk management framework.

AI introduces new risks and modifies existing ones. However, the management of these risks should not be an isolated activity. AI risk management must be fully integrated into your broader enterprise risk management framework (ERMF). Only by embedding AI risks into your existing structures can you ensure a consistent, efficient, and objective-driven approach to managing uncertainty.

The ERMF house: A solid foundation for AI risk management

Protecht's ERMF house model illustrates how AI risk seamlessly fits within a comprehensive risk framework – demonstrating how all risk management activities must support the ultimate aim of achieving objectives. AI risk, like any other type of risk, is integrated into this framework.



The enterprise risk management framework house.

The house is built upon clear layers:

- **Governance:** This is the roof of the house. This includes roles, responsibilities and accountabilities of staff, board, Executive and committees. Governance includes the risk framework and supporting policies, and setting of risk appetite aligned with strategic and operational objectives. AI initiatives should support these objectives, rather than stand alone.
- **People & culture:** People are the foundation of your house. Their daily behaviours collectively make up the culture of your organisation, including their approach to risk management and how and when they utilise AI.
- **Risk processes:** This is the engine room – the daily activities that bring risk management to life. We will look at these through an AI lens in the next section.
- **Reporting and response:** This includes escalations and workflows, reporting over risks and risk management activities, and how that reporting influences decision making.
- **Continuous improvement:** This includes ongoing management and improvement of risk management processes, and development of capability.

Risk processes applied to AI risk

At Protecht, we promote a common set of interconnected risk processes that support effective risk management across all domains, including AI.

Image: Protecht's risk processes diagram.

Risk and control self-assessment

Identify and assess AI-related risks, both those specific to AI systems (such as bias, model drift, adversarial attacks) and those modified by the use of AI (such as cyber risks or fraud). These should use any agreed risk and control taxonomies and libraries for consistency. These are risk bow ties waiting to happen.

Metrics and monitoring

Develop Key Risk Indicators (KRIs) and Key Control Indicators (KCIs) tailored to AI use cases that are mapped across the risk bow tie. For example, monitoring the percentage of AI outputs flagged for rework can serve as an early warning signal before any harm is experienced. These are bow ties that are happening.

Incident management

Incidents involving AI must be captured in your incident management process, whether they arise from AI errors, bias, misuse, security breaches, or other failures. Pay close attention to near misses – with the speed at which AI risks evolve, a near miss today may be a direct hit tomorrow. These are bow ties that have happened.

Controls assurance

Controls over AI should not be set and forget. Conduct design and operating effectiveness assessments of AI-specific controls, acknowledging that they may become ineffective as the underlying technology changes alongside shifting causal pathways.

Issues and actions

Issues may arise from any of the other risk processes. These issues and associated actions should be documented and tracked using centralised ERM tools to ensure accountability for addressing them, with appropriate visibility and reporting.

Compliance management

Your AI-related obligations (whether regulatory or contractual) should form part of your broader compliance and obligation management processes. Compliance obligations represent the risk appetite of society or minimum control standards and should be linked to risks in your risk profile.

Risk in change

Many organisations are currently piloting or implementing significant AI projects. Forging ahead with those projects with little regard for how the operational risk profile will change once they are delivered may result in organisations suddenly finding themselves outside of risk appetite. Or worse, with a string of incidents that undermine their reputation.

These processes are typically part of existing ERM frameworks and can easily be adapted.

When integrating AI into your risk management framework, you also need to consider whether there are any additional processes you have developed for specific risk types that also overlap with AI.

Integrating the AI ecosystem with operational resilience

Operational resilience requires identifying critical operations and ensuring they can continue during disruptions. Use of AI models will increasingly become embedded not just in our personal lives but into our products and services that are delivered to customers and stakeholders.

At a high level, an operational resilience process includes:

1. Identifying critical operations (end to end processes)
2. Defining tolerance levels (at which point harm is intolerable)
3. Identifying sub-processes that make up the critical operation
4. Identifying the resources (people, systems, physical assets etc) that support those processes
5. Identifying vulnerabilities
6. Running scenarios to ensure ability to meet tolerance during disruption

The use of AI would be identified in step four: the AI models that, if they weren't available, would prevent processes from being completed and critical operations being delivered. Where use of Gen AI may differ is in the speed at which these models may change, and the risks that may evolve, and whether organisations consider contingency plans should they fail.

Consider a financial services contact centre that introduces an AI chat model with both voice and text capabilities that reduces their reliance on human-to-human interaction. Imagine the provider suffers from a crippling cyber attack. Does the financial services provider maintain sufficient call centre infrastructure and people capability as a fallback? If the provider is offline for an extended period, is the contact centre's existing AI infrastructure flexible enough to transition quickly to another provider, or does it need to be rebuilt from scratch?

Integrating the AI ecosystem with third party risk management

As the above example demonstrates, AI is rarely a closed-loop internal system. Most organisations rely on external providers for models, datasets, cloud hosting, or AI-enabled services. This extended enterprise introduces third party risks that must be integrated into your risk management process.

The same risk processes mentioned above are all still applicable, however an extra layer of assurance is needed from those third parties:

- What due diligence do you need to conduct before and during onboarding?
- What will happen if they decide to change their models? Do you need contractual clauses around AI model updates, data usage, and explainability?
- What controls do they have in place? What rights do you have to audit them?
- What monitoring do you need over those third parties?
- What is their escalation pathway for incidents? Does it align with internal timeframes?
- Do we need to assess their business continuity capability?

An effective third-party risk management (TPRM) program utilises standardised processes and questionnaires to ensure a consistent approach is adopted across the organisation, and the risk profile introduced by third parties (individually and in the aggregate) is understood and managed.

Breakout: AI in the wild

You don't need to just test AI in the lab: you need to carefully monitor real world implications after deployment, including how they might interact with other AI.

In 2011, dynamic pricing algorithms on Amazon caused two sellers to rapidly increase the price of their copies of a particular book – at one point, the book was listed as $23,698,655[7]. While harmless, it demonstrates how behaviour of even simple algorithms can be hard to predict, let alone the complex AI models we engage with today.

For a more sobering example, in 2010 a trading algorithm placed an unusually high sell order, and competing algorithms joined the fray, buying and selling between each other. The Dow Jones dropped 9% in a matter of minutes. A review identified that it was simply the interaction of the algorithms that caused the spiral – no error could be attributed to a single firm[8].

The human element

While there is debate about whether AI could become conscious in the future, the 'super employee' or agentic workforces aren't human. Your leadership teams and employees remain the most critical part of your organisation. Those humans:

- Design or define use cases for AI
- Identify and assess the risks of AI
- Implement and perform controls over AI
- Remain the 'human in the loop' to validate AI decisions and outputs

The most effective implementations for AI, including agentic AI, are where future AI-human collaboration is deliberately planned or mapped out. Consider new skills your teams might need to develop, identify AI champions, and define AI literacy requirements and training pathways.

The speed of change in AI is outpacing organisations' ability to change, with some pundits suggesting years of innovation to unlock, even if advancements ceased today. The capability and culture of your people are pivotal to how quickly your organisation can adapt. Keep your humans engaged at each part of your AI implementation journey. If your people don't understand AI or your organisations plans (or worse, fear it), they can't or won't speak up if things start to go wrong.

Takeout:

AI risk management should not stand apart: it must be embedded within your enterprise risk management framework. By doing so, you ensure that AI becomes a strategic enabler rather than a source of unmanaged uncertainty.

[7] Wired, How A Book About Flies Came To Be Priced $24 Million On Amazon
[8] Finance Research Letters, Herding and flash events

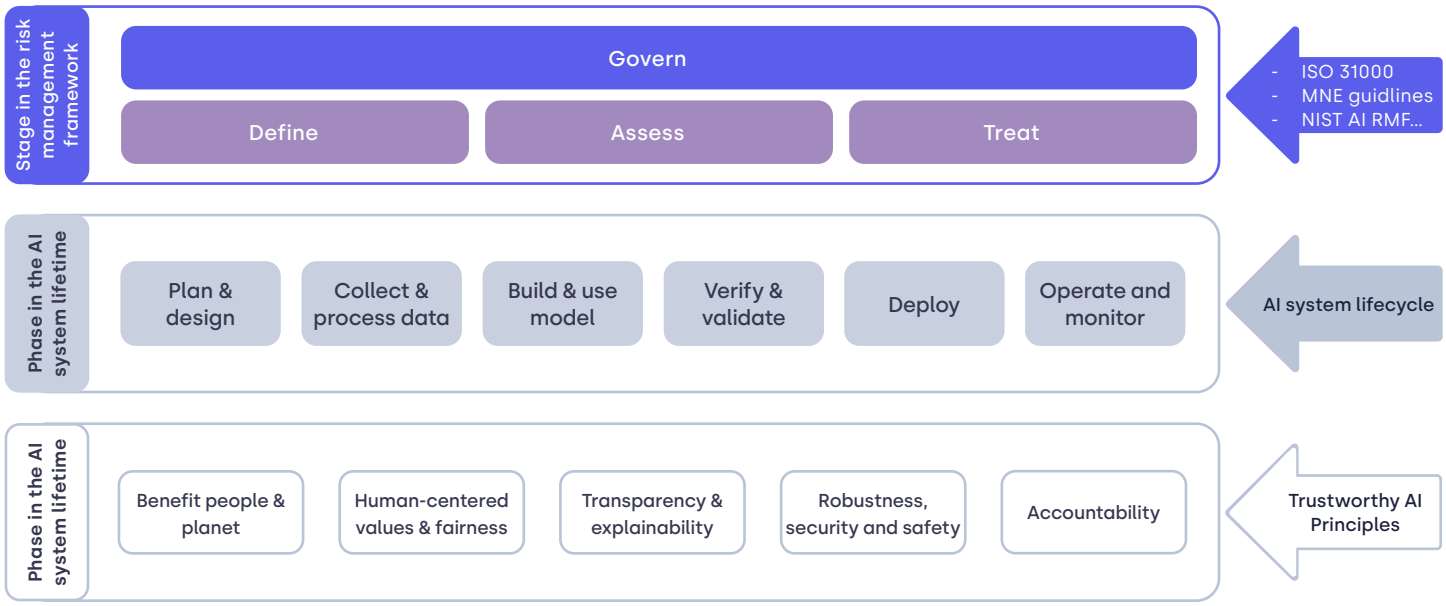
6 Governing AI with Protecht.

Artificial intelligence is no longer an experimental tool deployed on the margins of business. It is now central to customer service, decision-making, and product delivery. With this adoption comes accountability: executives, regulators, and customers all demand evidence that AI systems are transparent, ethical, and clearly delimited.

What does embedded AI governance look like?

Embedded AI governance – as provided in Protecht's AI governance solution – provides a structured way to capture, assess, and monitor every AI system across the organisation, whether developed in-house or procured from third parties. Instead of treating AI risk as an isolated operational issue, this approach integrates AI governance into the enterprise risk framework, ensuring that organisations are both compliant today and future-proofed for tomorrow's regulations.

The OECD AI lifecycle[9] illustrates how governance is applied at every stage of an AI system's journey, from design and development through deployment and ongoing monitoring. We're using this diagram to illustrate the importance of embedding AI governance from the very start of the evolution of an AI system.



OECD AI lifecycle diagram

[9] https://oecd.ai/en/accountability

Step 1: Building a centralised inventory

At the foundation of the package is a single, flexible AI register. Here, organisations log all AI systems in use, from proprietary models to licensed SaaS tools. For each system, users capture key details such as:

- System purpose and intended use cases
- Internal and external stakeholders
- Data sources and privacy/security considerations
- Deployment and configuration decisions, such as environments, thresholds and guardrails

This inventory provides a consolidated, organisation-wide view of AI exposure: something most risk teams previously lacked.

Step 2: Capturing lifecycle governance

AI systems are dynamic, evolving through design, deployment, model drift, data growth, and user expectations. The register reflects this lifecycle by capturing governance checkpoints, including:

- Design and performance choices: model type, intended outcomes, test parameters, and user expectations
- Decision gates: who signs off at each stage, and what change controls are in place
- Transparency and user empowerment: how staff or customers are informed about AI involvement, and how they can challenge or override outputs

Including these checkpoints in the solution helps organisations demonstrate an auditable trail of accountability.

Step 3: Assessing risk and controls

Every AI system carries risks, even when purchased "off the shelf." The solution supports:

- Risk assessment: documenting ethical, compliance, and operational risks such as bias, privacy breaches, or model drift
- Control capture: mapping mitigations to applicable frameworks such as ISO 42001, NIST AI RMF, VAISS
- Due diligence evidence: recording information provided by vendors, even when limited, to show responsible procurement decisions

This structured approach ensures risk managers can answer not just "*what could go wrong?*" but also *"what are we doing about it, and how do we know it works?"*

Step 4: Monitoring and change control

The governance solution recognises that AI does not stand still. It enables:

- Ongoing monitoring of system performance and adverse outcomes
- Change tracking when datasets, models, or parameters are updated

Instead of relying only on point-in-time assessments, the solution ensures continuous oversight aligned with the rapid pace of AI development.

Step 5: Cross-functional oversight and reporting

AI risks are cross-functional by nature. The solution allows organisations to:

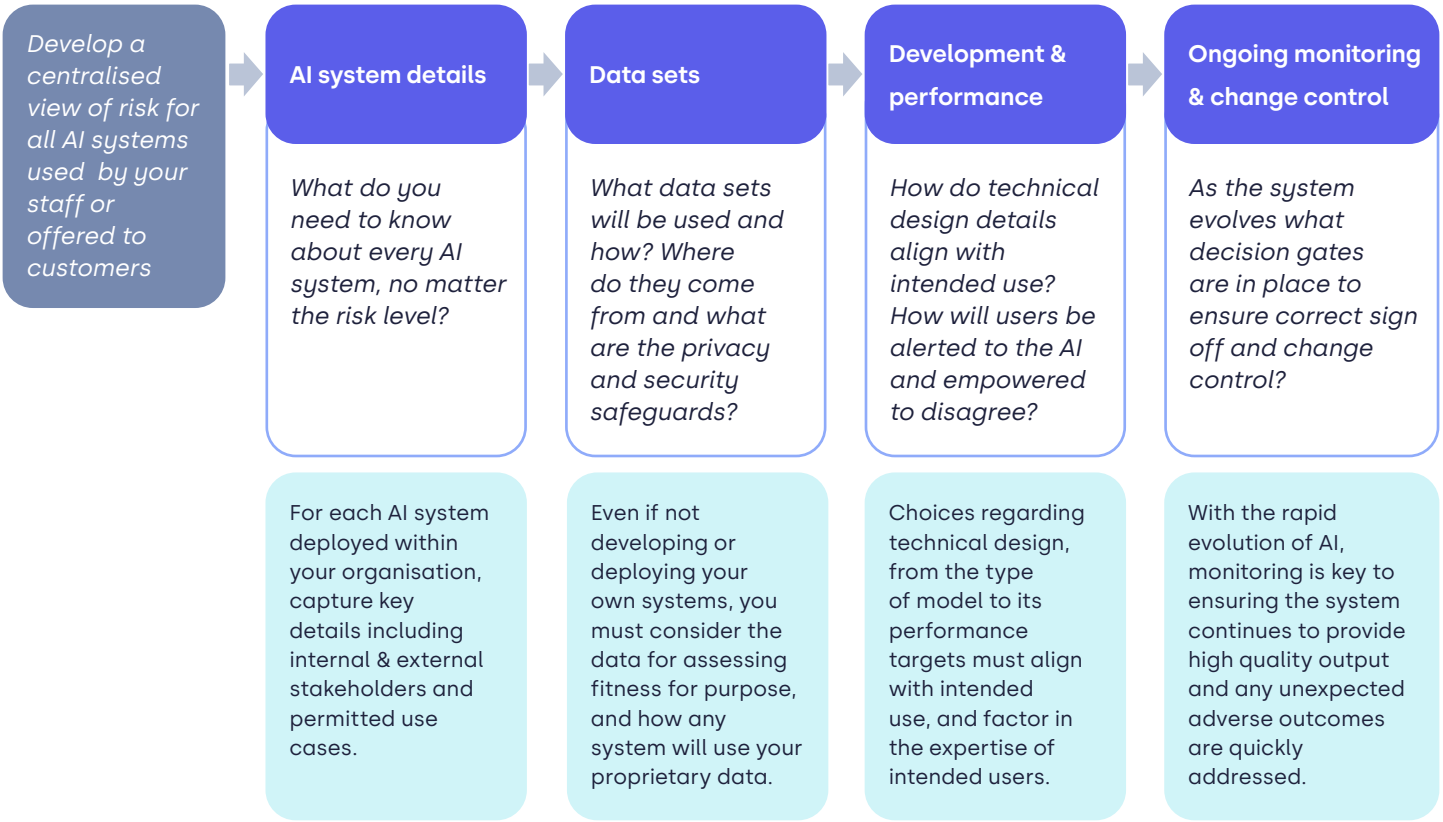
- Assign clear ownership (e.g., CIO, Chief Risk Officer, Data Protection Officer).
- Aggregate data to support the consolidation of risk information across all systems.
- Demonstrate governance, explainability, and compliance to regulators and stakeholders.

This transforms AI risk oversight from reactive troubleshooting to proactive assurance.

The benefits of embedded AI governance

The AI governance solution captures the entire lifecycle of an AI system, from system details and datasets to performance and monitoring, providing a single, centralised view of governance.

How the solution tracks risks and controls



By embedding AI into Protecht ERM's governance framework, organisations gain:

- Centralised visibility of all AI systems, reducing blind spots
- Regulatory alignment with the EU AI Act and global best practices
- Flexibility to support multiple standards without locking into one
- Evidence of due diligence for both in-house and third-party AI
- Continuous monitoring to catch risks early and adapt to changes
- Executive confidence through instant, transparent reporting

Takeaway

The AI governance solution reflects a new reality: managing AI risk is no longer about a single deployment or vendor. It is about governing AI as a portfolio of systems, each with its own lifecycle, risks, and accountabilities.

Protecht gives organisations the tools to capture this complexity in one place, integrate it into enterprise risk management, and move forward with AI adoption confidently and responsibly.

See how Protecht helps you manage AI risk with confidence

From capturing every AI system in a centralised register to embedding governance checkpoints across the lifecycle, Protecht ERM enables you to oversee AI with confidence. Whether you're adopting third-party tools, developing your own models, or preparing for evolving regulations, Protecht gives you a connected framework for transparency, accountability, and assurance.

- Build a complete inventory of all AI systems in use
- Document risks, controls, and due diligence across the lifecycle
- Stay aligned with global best practice including the EU AI Act
- Generate auditable reports and dashboards for boards and regulators

Ready to strengthen your AI risk oversight?

Find out more

Book a demo

About the authors

Michael Howell
Senior Manager, Research and Content

Michael Howell is Protecht's Senior Manager, Research and Content. He is passionate about the field of risk management and related disciplines, with a focus on helping organisations succeed using a 'decisions eyes wide open' approach.

Michael is a Certified Practicing Risk Manager whose curiosity drives his approach to challenge the status quo and look for innovative solutions. Michael harnesses that curiosity in pursuit of risk knowledge, conducting research and developing content to support and advance risk methodology and product design at Protecht.

Michael's industry experience includes managing risk functions, assurance programs, policy management, corporate insurance, and compliance.

David Tattam
Chief Research and Content Officer

David Tattam is the Chief Research and Content Officer and co-founder of the Protecht Group. David's vision is to redefine the way the world thinks about risk and to pioneer the development of risk management to its rightful place as a key driver of value creation in each of Protecht's clients.

David is passionate about risk and risk management and in reaping the value that risk and good risk management can create for any organisation willing to embrace it. He is particularly passionate about risk management research and is prolific in creating a wide range of content delivered in blogs, eBooks, webinars and training courses. He is also the author of A Short Guide to Operational Risk.

Prior to co-founding Protecht, David was the Chief Risk Officer and Head of Operations for the Australian operations of two global banks. He started his career as a Chartered Accountant and Auditor with Grant Thornton and PwC. David is an Associate of the Institute of Chartered Accountants in Australia and New Zealand and a Senior Fellow of the Financial Services Institute of Australia.

ABOUT PROTECHT

Redefining the way the world thinks about risk.

While others fear risk, we embrace it. For over 25 years, Protecht has redefined the way people think about risk management. Through our people, we enable smarter risk taking by our customers to drive their resilience and sustainable success.

We help our customers increase performance and achieve strategic objectives through better understanding, monitoring and management of risk. We provide a complete solution of AI-enabled governance, compliance and risk management software supported by training and advisory services to businesses, regulators and governments across the world.

With our flagship Protecht ERM SaaS platform you can dynamically manage all your risks in a single place: risks, compliance, incidents, KRIs, vendor risk, cyber and IT risk, internal audit, operational resilience, business continuity, workplace safety, and more.

We're with you for your full risk journey. Let's transform the way you understand and manage your risk to create exciting opportunities for growth.

Visit our website:
protechtgroup.com

Email us:
info@protechtgroup.com

EBOOK-AIRISK-GL-2025-1.1

Copyright © 2025 The Protecht Group. All rights reserved. All other registered or unregistered trademarks and service marks are property of their respective companies and should be treated as such. Product images used in this book are illustrated representations only.