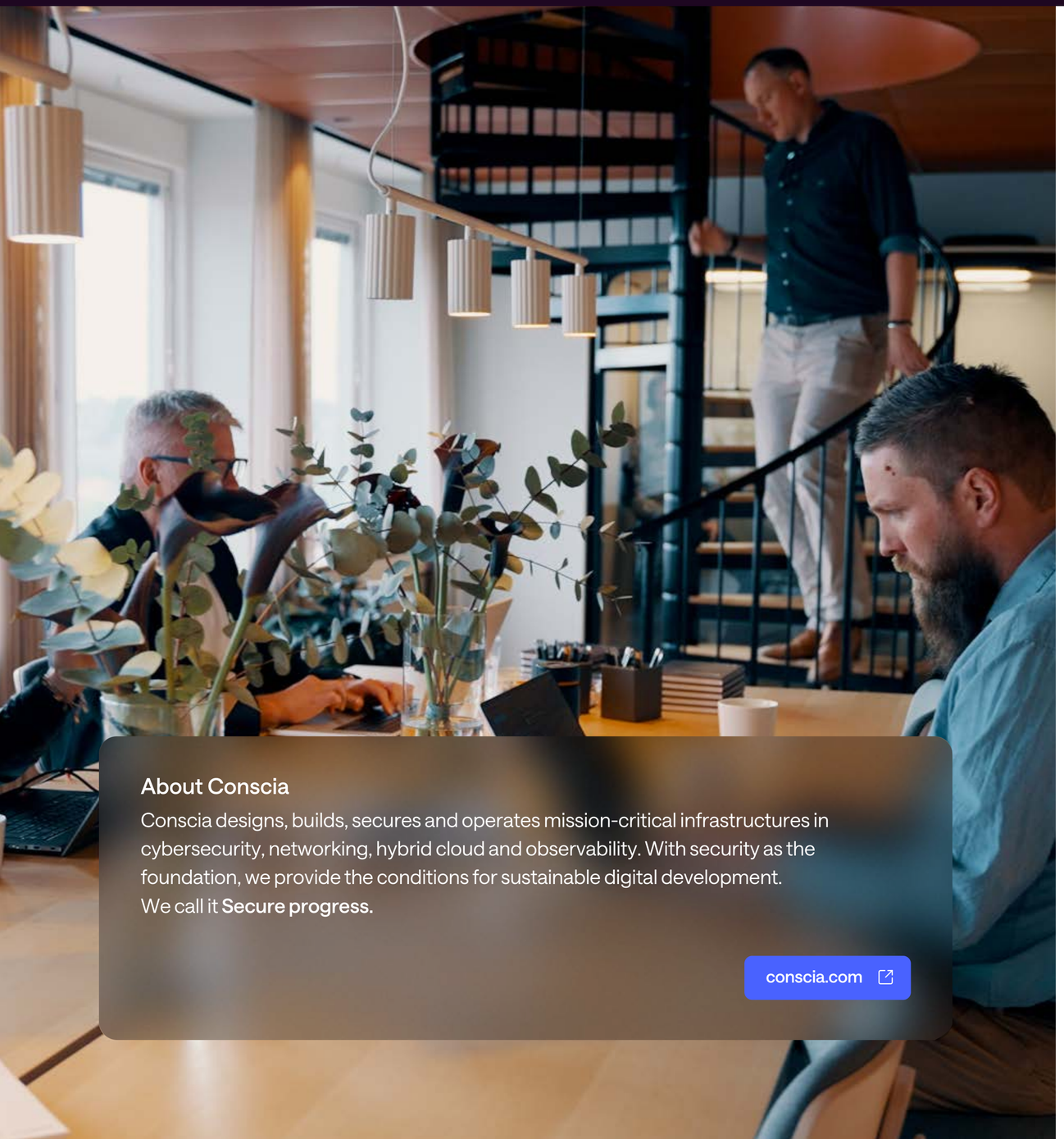
Conscia
Secure progress

Secure AI

Effective & future-proof AI



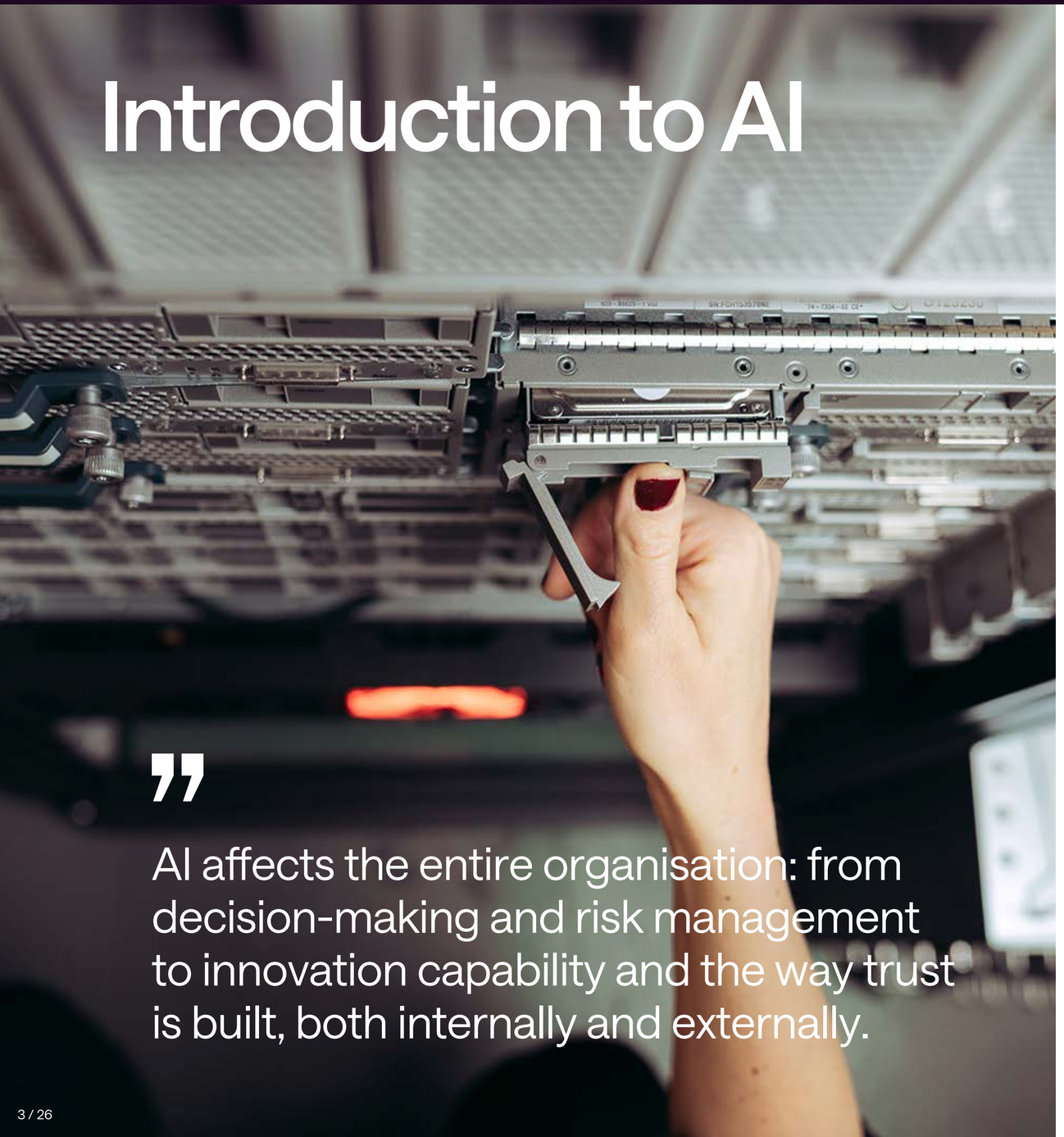
About Conscia

Conscia designs, builds, secures and operates mission-critical infrastructures in cybersecurity, networking, hybrid cloud and observability. With security as the foundation, we provide the conditions for sustainable digital development.
We call it **Secure progress.**

conscia.com

Contents

Introduction to AI



> AI affects the entire organisation: from decision-making and risk management to innovation capability and the way trust is built, both internally and externally.

Safe, efficient & future-proof AI

Artificial intelligence (AI) has in a short time gone from being an experimental tool to becoming a central part of how organisations create value, develop products, and make decisions. AI has left the domain of technology and become a strategic management issue.

The rapid pace of development has fundamentally reshaped the landscape. Questions that once centred on possibilities and potential now concern responsibility, trust, and control to an equal degree. How organisations introduce, govern, and apply the technology will largely determine whether AI becomes a source of long-term value or of uncertainty and unintended consequences.

Many organisations now find themselves moving from curiosity to accountability, and from experimentation to real-world deployment. For most, the question is no longer whether to use AI, but how. How does an organisation ensure its use of AI remains sustainable and secure over time? How can it maintain agility and innovation without losing control? Doing so requires a holistic perspective that brings together technology, governance, capability, and culture.

At the same time, the pace of progress is accelerating at record speed. AI systems are becoming more autonomous, more deeply embedded in core processes, and used by a growing number of people. In parallel, regulatory demands and expectations from customers, partners, and society at large are increasing.

In this environment, technical expertise or isolated initiatives are no longer enough. AI must be managed as the strategic capability it truly is.

About the report

This report is aimed at decision-makers, technical leaders, and business developers who want to understand what AI means for their organisation – beyond the hype and individual technological solutions.

Its starting point is that AI is not primarily a matter of innovation or security, but of governance, responsibility, and long-term capability.

The purpose is to show how organisations can combine rapid technological development with clear control, and how a structured and responsible approach to AI can become a competitive advantage rather than a risk. The report begins with fundamental concepts and the evolution of AI, before moving on to questions of maturity, risk, security, leadership, and safe strategic choices for the future.

Concepts & definitions

In this section, key concepts are defined and explained, including how they relate to business value, risk and governance. What is an AI model? What is meant by an agent? And why is it no longer sufficient to talk only about AI chats and individual tools?

To use AI responsibly and in a controlled way, the technology must be understood as a connected ecosystem, where models, data, agents, integrations and permissions interact – and where each part influences security, compliance and operational impact.

Shared understanding clarifies accountability

AI can mean different things to different people. In many organisations, the term is used for everything from chatbots and assistants to advanced systems that analyse information, make decisions and act directly within operations. Without a shared understanding of what AI actually is, it becomes difficult to manage usage, assess risks and clarify responsibilities.

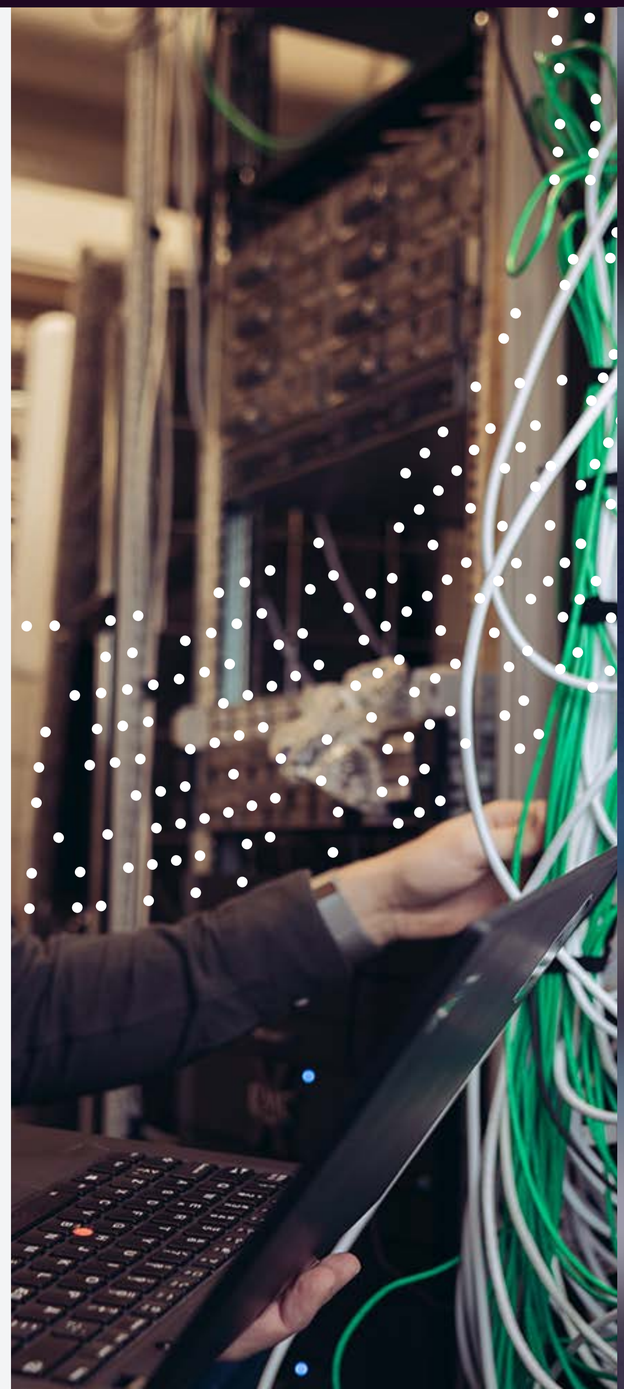
AI has moved from experiments in isolated teams to becoming a natural part of everyday tools – from office support and development environments to analytics and security solutions. At the same time, AI means different things depending on perspective. For some, it is a chatbot; for others, a complex infrastructure of models, agents and integrations deep within core business systems.

Today, it is no longer enough to talk only about prompts and individual AI tools.

To manage, assess and protect AI usage, organisations must view AI as a system of interconnected components. Only when data, identities, models, agents and tools come together do questions of responsibility, control and real impact become clear.

Key terms in brief

- **Artificial Intelligence (AI).** A collective term for technologies that enable computers to analyse, reason, create and make decisions in ways similar to human intelligence.
- **Agent Identity / Non-Human Identity (NHI).** A separate identity for AI agents that enables authentication, permission management and traceable access.
- **Agent Orchestration.** Controls how multiple AI agents cooperate, distribute tasks and coordinate decisions towards shared objectives.
- **AI agent (agentic AI).** Autonomous AI capable of planning, making decisions and acting over multiple steps, often working alongside humans and other systems.
- **AI governance.** Frameworks and roles that govern how AI is developed, used and monitored, focusing on accountability, risk and transparency.
- **AI Gateways.** A central layer for governance, monitoring and policy control of access to AI models and agents.
- **AI Management System (AIMS).** A management system for AI governance aligned with standards such as ISO/IEC 42001, integrating policy, risk and continuous improvement.
- **AI model.** The trained algorithm that interprets data and produces results, such as a language model, image interpreter or prediction engine.
- **AI security.** Technical and organisational safeguards to prevent manipulation, data leakage and malfunction in models and agents.
- **Responsible AI.** AI used ethically, fairly and sustainably, with respect for privacy, human rights and societal impact.
- **Autonomy.** The degree of independence in AI, ranging from human-supervised decisions to fully autonomous agents.
- **AI chatbot.** An AI-driven application that interacts with users via text or speech, answering questions, giving recommendations or performing tasks based on input.
- **Bias.** Systematic distortions in data or models that lead to unfair, inaccurate or biased decisions.
- **Bring Your Own AI (BYOAI).** Employees using their own AI tools and accounts in their work, independently of organisational approval or governance.
- **Data Governance.** Structures and processes ensuring that data is accurate, protected, traceable and used in a controlled way.
- **Deep Learning.** Advanced machine learning using neural networks that process large datasets and identify complex patterns.
- **Generative AI.** AI that creates new content such as text, images, audio or code based on previous data and patterns.
- **Large Language Models (LLM).** Large language models trained on extensive text data that can generate, complete and analyse text.
- **Machine Learning.** A field of AI in which models learn patterns from data and improve over time without being explicitly programmed.
- **Model Context Protocol (MCP).** An open protocol that governs how AI agents access context, tools and resources in a standardised way.
- **Multimodal AI.** Models that can understand and combine multiple types of data – such as text, images and audio – for a comprehensive understanding.
- **Prompt.** An instruction, question or text entered into an AI model to guide its response or generation.
- **Traceability.** The ability to follow the entire chain from data to decision, showing how an AI output was produced.



Shadow AI – a growing reality

Shadow AI refers to the use of AI tools – such as generative services, code assistants and analytics platforms – without the IT department's knowledge or approval. It is AI used out of sight. Not primarily as a rule violation, but as a result of technology becoming accessible and useful before the organisation has created structure.

The phenomenon resembles classic Shadow IT. The difference lies in impact – AI handles data but also shapes analyses, recommendations and actions. When such decisions are made without visibility, consequences may be greater than the organisation can foresee.

Most leadership teams today have limited or no insight into which AI is used across the organisation. This makes it impossible to:

- identify valuable initiatives to support.
- build the right protections around the right data.
- ensure compliance with laws and agreements.
- develop a coherent AI strategy.

To manage AI long-term, more than policies and guidelines are required. It demands a technical and organisational capability built on three connected steps: Discover, Detect and Protect. Together, they form a framework for practical AI governance. (Read more in the chapter AI in the security domain).

AI-development in brief

This chapter provides an overview of how AI has evolved – from early machine-learning approaches to today's multimodal and self-learning agents. What began as a visionary idea of thinking machines in the 1950s has, within just a few decades, grown into one of the most transformative technologies of our time.

The earliest pioneers of AI, Alan Turing (1912–1954) and John McCarthy (1927–2011), were among the first to define concepts for systems capable of reasoning and learning. But the technical capabilities of the time were limited – both in terms of available data and computational power. For many years, the field therefore shifted between periods of optimism and periods of so-called AI winters, when expectations far exceeded what the technology was able to deliver.

The shift of the 2010s

The major turning point arrived in the 2010s. The combination of large-scale data, rising computational power, and breakthroughs in deep neural networks pushed AI from theory into practical use. Suddenly, systems could recognise images, understand language, and predict behaviours.

Within just a few years, AI became a core component across industries – from manufacturing and finance to research and culture. For many people, this was the moment when AI became tangible and genuinely useful in everyday life, thanks to accessible generative tools that demonstrated the technology's real potential.

The current AI era

Today, AI has entered its next phase: the generative and multimodal era. AI systems can create text, code, images, and other forms of content, and can be connected into autonomous agents that learn, plan, and act on their own. At the same time, new phenomena are emerging. One example is BYOAI, where employees use their own AI tools outside of central governance.

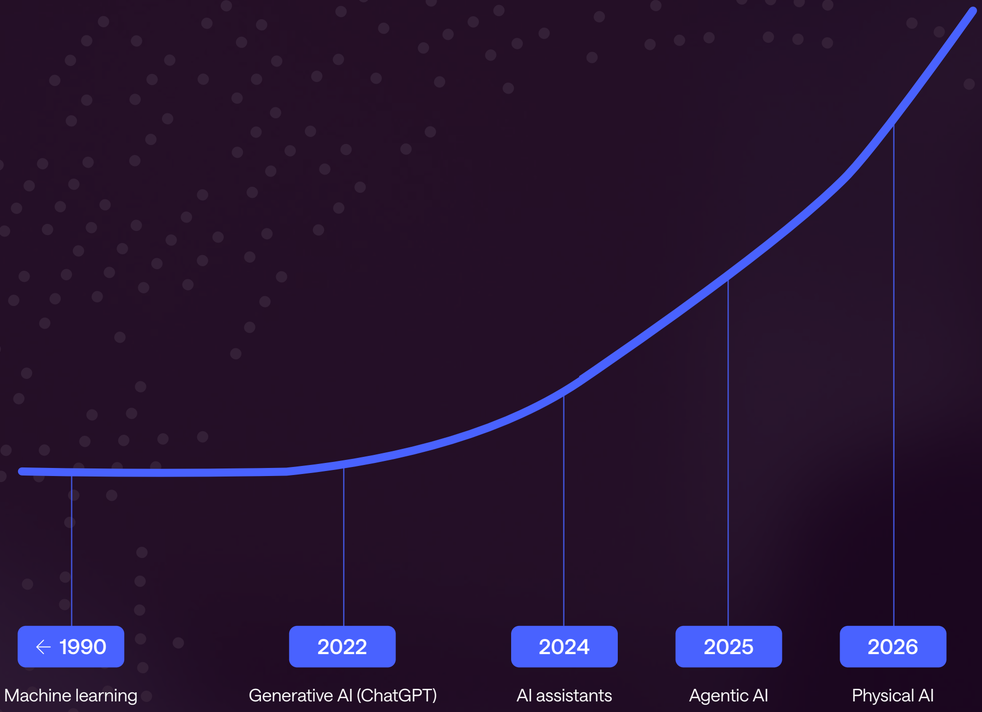
History repeats itself: technology continues to move faster than governance.

This behaviour is a key driver behind what is often referred to as Shadow AI – the growing presence of AI solutions introduced and used outside official structures, policies and security controls.

This is why the following chapters explore how prepared today's organisations really are for this new wave of technology, the risks and threats that emerge, and the role leadership and governance must play as AI becomes a natural part of everyday operations. Today, it is a question of both capability and control – and of how AI can evolve in a way that benefits both the organisation and society at large.

What organisations are facing

The illustration below shows how AI development is accelerating – from machine learning to generative and agent-based AI. Today, the pace of innovation is outstripping organisations' ability to establish the structures needed to manage it, highlighting the growing need for maturity, governance, and deliberate strategic choices.



A technical evolution

AI is often perceived as something sudden – as if the technology jumped from simple chatbots to autonomous agents overnight. In reality, today's AI agents are the result of a long technical evolution, where each step has built on the one before it. To govern, secure, and apply AI responsibly, it is not enough to understand individual concepts. It requires an understanding of how they connect, why they emerged, and what has changed along the way.

Machine learning as a turning point

When AI moved from fixed rules to learning from data.

The first AI systems were built on hard-coded rules and statistical methods. The breakthrough came when machines began learning patterns from data rather than being programmed step by step – a shift that became known as Machine Learning. With Deep Learning, the next leap followed. Neural networks with many layers made it possible to detect complex relationships in large datasets. AI could now recognise images, understand language, and predict behaviours with a level of precision that had previously been out of reach. This is where the AI model was established as the core: a trained algorithm that interprets data and produces results.

Foundation models & generative AI

When broad, general-purpose models made AI accessible across entire organisations.

For a long time, models were trained for one specific purpose at a time. A shift occurred when very large models were trained on broad datasets and then adapted for many different tasks. These became known as foundation models. One particularly important category is LLMs, which work with language and code. They formed the basis for what later became known as generative AI – systems that both analyse and create new content. This was the point where AI became widely accessible. Chat interfaces, coding assistants and writing tools made the technology tangible across entire organisations, not just for development teams. Still, AI remained largely reactive, responding only when prompted.

AI in interaction with systems & data

When models gained context, tools & system connections.

The next step came when models were no longer isolated from their surroundings. They began to receive **context**, access to documents, data, and history, as well as the ability to call tools and APIs. In practice, this meant that AI systems could:

- → Retrieve information
- → Update systems
- → Trigger actions

This was the point where the **AI system** became more important than the model itself. The model was still the brain, but the surrounding system determined the actual impact AI could have on the organisation.

Agentic AI & increased autonomy

When AI systems begin making decisions and initiating actions.

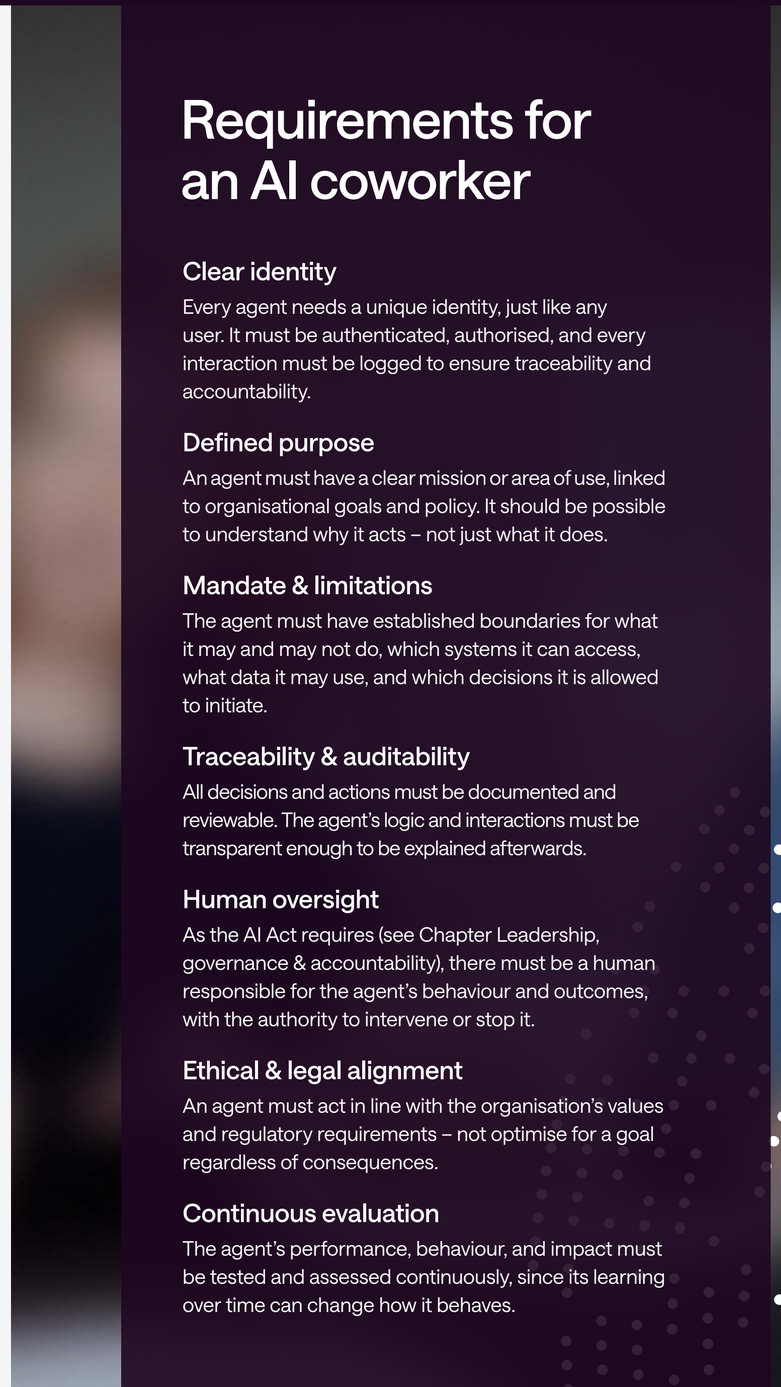
When models are combined with goals, memory, tools, and decision logic, something new emerges: AI agents. An AI agent can break down a task into steps, decide on the next action, and carry it out without a human directing the process. The level of autonomy can vary – from close human oversight to more independent operation. This is the point where AI shifts from being a supportive tool to becoming an operational actor, fundamentally changing the risk landscape.

Agentic AI & the digital coworker

With agentic systems, AI takes on a new role in the organisation – as a digital coworker that can take initiative, make decisions, and interact with other systems. This unlocks significant potential for efficiency and innovation, but it also raises new questions.

When AI takes on an active role in the organisation, it must be governed by the same fundamental principles as human roles: a clear purpose, defined permissions, mandate, and structures for oversight.

Together, these principles form a new way of thinking about AI in the workplace – where AI is seen as a collaborative partner with explicit requirements for governance, follow-up, and ongoing development. Organisations that begin defining role descriptions for their digital coworkers – including mandate, purpose, and oversight – build both safety and trust. This is what lays the foundation for the future collaboration between humans and machines.



Requirements for an AI coworker

Clear identity

Every agent needs a unique identity, just like any user. It must be authenticated, authorised, and every interaction must be logged to ensure traceability and accountability.

Defined purpose

An agent must have a clear mission or area of use, linked to organisational goals and policy. It should be possible to understand why it acts – not just what it does.

Mandate & limitations

The agent must have established boundaries for what it may and may not do, which systems it can access, what data it may use, and which decisions it is allowed to initiate.

Traceability & auditability

All decisions and actions must be documented and reviewable. The agent's logic and interactions must be transparent enough to be explained afterwards.

Human oversight

As the AI Act requires (see Chapter Leadership, governance & accountability), there must be a human responsible for the agent's behaviour and outcomes, with the authority to intervene or stop it.

Ethical & legal alignment

An agent must act in line with the organisation's values and regulatory requirements – not optimise for a goal regardless of consequences.

Continuous evaluation

The agent's performance, behaviour, and impact must be tested and assessed continuously, since its learning over time can change how it behaves.

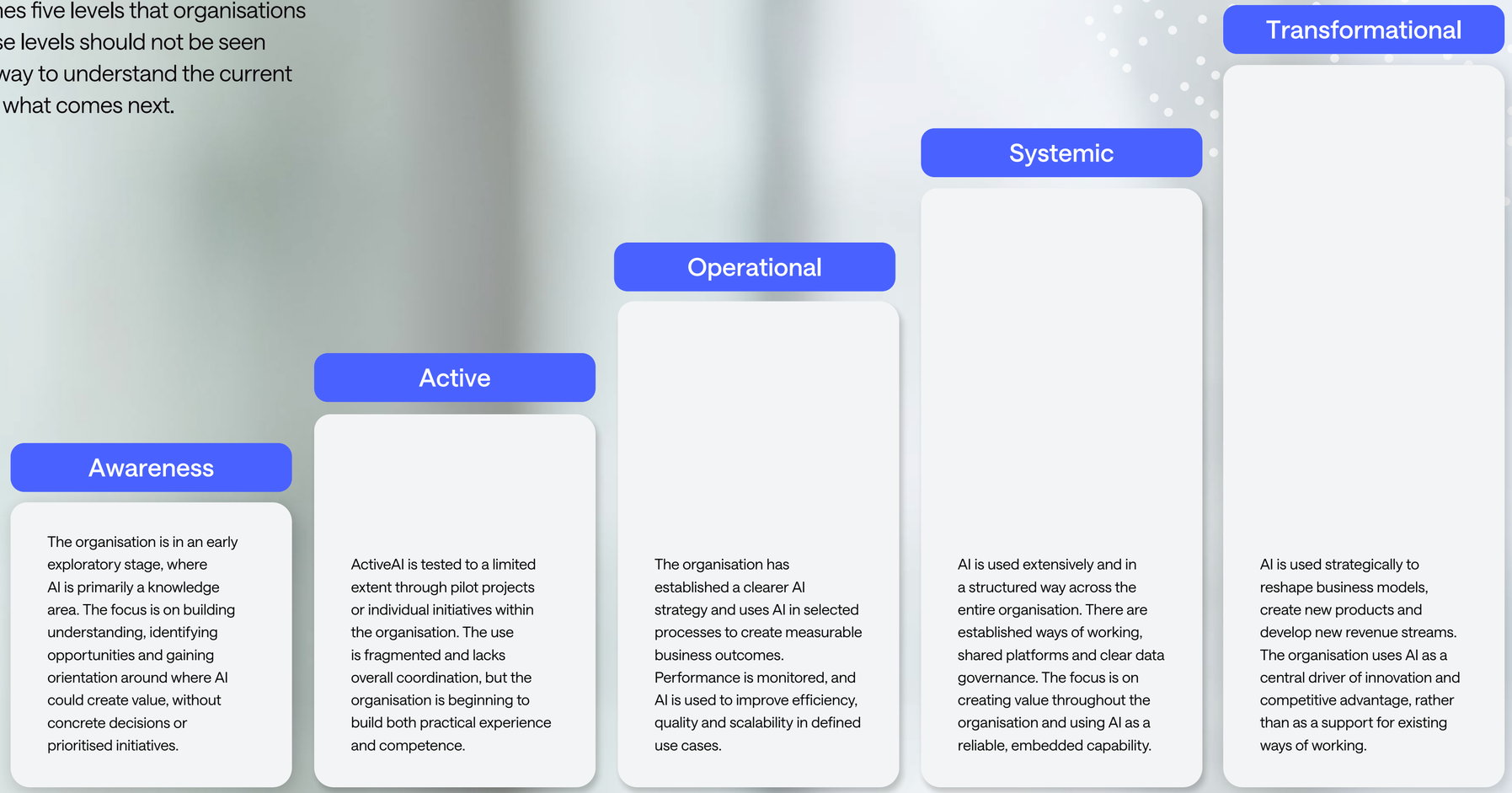
An organisation's AI maturity

The maturity model shown in this chapter outlines five levels that organisations typically move through on their AI journey. These levels should not be seen as fixed stages or goals in themselves, but as a way to understand the current state, identify gaps and build alignment around what comes next.

AI maturity is not about how advanced the technology an organisation uses is. It is about how well AI is integrated, governed and embedded across the business. Many organisations use AI in some form, but the difference between experimenting and creating long-term value is often significant.

A higher level of maturity does not necessarily mean more AI, but a better balance between innovation, governance and accountability. Organisations may be at different levels simultaneously across different parts of the business. The model provides a shared reference point for discussion, prioritisation and leadership.

AI maturity model. From early understanding and experimentation to systematic and transformative use. The model describes how an organisation's focus gradually shifts from learning and testing to governance, scaling and strategic impact. The levels reflect typical patterns in how AI is used, governed and delivers value over time.



Source: AI Maturity Model (Gartner).

Six building blocks of AI maturity

AI maturity is, at its core, about balancing innovation, governance and accountability. Organisations that succeed in establishing AI as a long-term capability do so by building a stable foundation of technology, processes and structure.

Experience from both research and industry shows that six recurring building blocks are present in these organisations: strategy, infrastructure, data, governance, capability and culture. Together, they form a framework for how AI can be implemented, governed and developed in a safe and sustainable way.

The six building blocks of AI maturity

- → Clear AI strategy
- → Scalable, secure infrastructure
- → Centralised, quality-assured data
- → Governance with accountability and transparency
- → Internal AI capability
- → Responsible and learning culture

Strategy

A clear AI strategy links the technology to the organisation's goals and values. It clarifies the purpose of AI use, prioritises initiatives and sets the boundaries for how the technology may be applied. When the strategy is anchored in leadership, it becomes possible to follow up on results and steer innovation in the desired direction.

Infrastructure

AI depends on a robust and reliable infrastructure. Platforms, networks and security architecture must be able to manage large volumes of data, real-time analytics and high demands for availability and integrity. A modern infrastructure enables the development of AI solutions that are scalable and resilient. If technology is the engine, the infrastructure is the road.

Data

Data quality, accessibility and traceability determine how reliable AI models become. Organisations that work systematically with data governance – from collection to use and deletion – can improve the precision of their models and reduce the risk of bias and incorrect decisions. Ultimately, data management is about trust.

Governance

Governance provides the framework for accountability, risk management and decision-making. With clear roles, documented processes and ongoing follow-up, AI can be developed in a controlled manner without limiting innovation. Governance should therefore be viewed as a prerequisite for sustainable AI use.

Capability

AI requires cross-functional capability. Leadership needs to understand how the technology affects business and risk, while technical roles must understand ethical and regulatory requirements. Organisations that invest in internal capability development strengthen both decision-making and innovation capacity.

Culture

A mature AI culture is characterised by curiosity, accountability and transparency. It should be just as natural to discuss risks and limitations as opportunities and business value. It is in the meeting between technology and values that the trust is built which allows AI to become a long-term part of the organisation.

Ready to scale AI?

Many organisations are in the midst of this transition. AI technology has matured rapidly, while governance, accountability and structures are still taking shape. How well these foundations are established will determine whether AI becomes a sustainable part of the business or remains a source of uncertainty and fragmentation.

AI creates new opportunities for analysis, automation and decision support, but the risks increase when the technology is used without clear structures. In organisations that are still exploring AI, usage is often fragmented and lacks transparency. Initiatives are driven locally, responsibilities are unclear, and it becomes difficult to understand how AI actually affects the business. The risk rarely lies in the technology itself, but in the lack of control over data, decisions and usage.

Organisations that are ready to use AI at greater scale have built a stable foundation

With a clear strategy, governance and data foundations, transparency and accountability are created, making it possible to identify risks early and follow up on how AI is used. Control then becomes an enabler of innovation, not a barrier. AI can be scaled into more business-critical contexts without compromising trust or regulatory compliance.

New risks & threats

This chapter highlights how AI introduces new types of threats that differ from traditional IT risks in their nature, scale and impact. The focus is on understanding this shifting risk landscape rather than on how these risks should be managed in practice.

Cybersecurity has traditionally been about protecting systems and data from technical vulnerabilities and breaches. AI fundamentally reshapes this threat landscape. Risks no longer stem solely from isolated flaws – they arise from how AI models and agent-based systems behave over time, how they are influenced by data and context, and how their outputs are used across the organisation. This marks a shift from single-point attacks to more dynamic and harder-to-detect risks. The focus moves from merely preventing intrusions to understanding, monitoring and governing how AI systems influence decisions, processes and trust. The consequences increasingly become operational rather than purely technical.

From vulnerable code to vulnerable models

AI systems are inherently dynamic. They change as they are trained, updated and used, meaning vulnerabilities are no longer static. A model that worked correctly yesterday may behave unpredictably today. As a result, the risk landscape shifts from isolated technical flaws to understanding how models are influenced and evolve over time.

The threats listed to the right illustrate how AI-related risks differ from traditional IT vulnerabilities, and why they are often harder to predict, detect and manage using established security methods.

Prompt injection

Attackers can manipulate an AI system's behaviour through crafted instructions. This may cause the model to reveal information or act in ways that were never intended.

Data poisoning

If training data is tampered with, the model's entire logic can be distorted. This may lead to incorrect decisions, bias, or make the AI vulnerable to attacks once deployed.

Model leakage

Many AI models risk unintentionally disclosing confidential information in their responses, particularly when the same model interacts with multiple users or environments.

Supply chain-risks

AI is rarely built entirely from scratch. Organisations rely on pre-built libraries, open data sources and third-party models. A single insecure dependency can therefore spread risk throughout the entire chain.

Autonomous behaviours

When AI agents can operate independently, new attack surfaces emerge. A misconfigured agent may trigger cascading effects across systems – often faster than a human can intervene.

The 10 most common risks for LLM applications

Source: OWASP LLM Top 10 2025 and Google Forecast 2025

Prompt injection
Manipulation of instructions.

Data leakage
Unintentional exposure of sensitive information.

Supply chain vulnerability
Insecure third-party models and libraries.

Insecure output handling
Models generating harmful content.

Unauthorized code execution
AI triggering dangerous commands.

Overreliance
Blindly trusting AI outputs without verification.

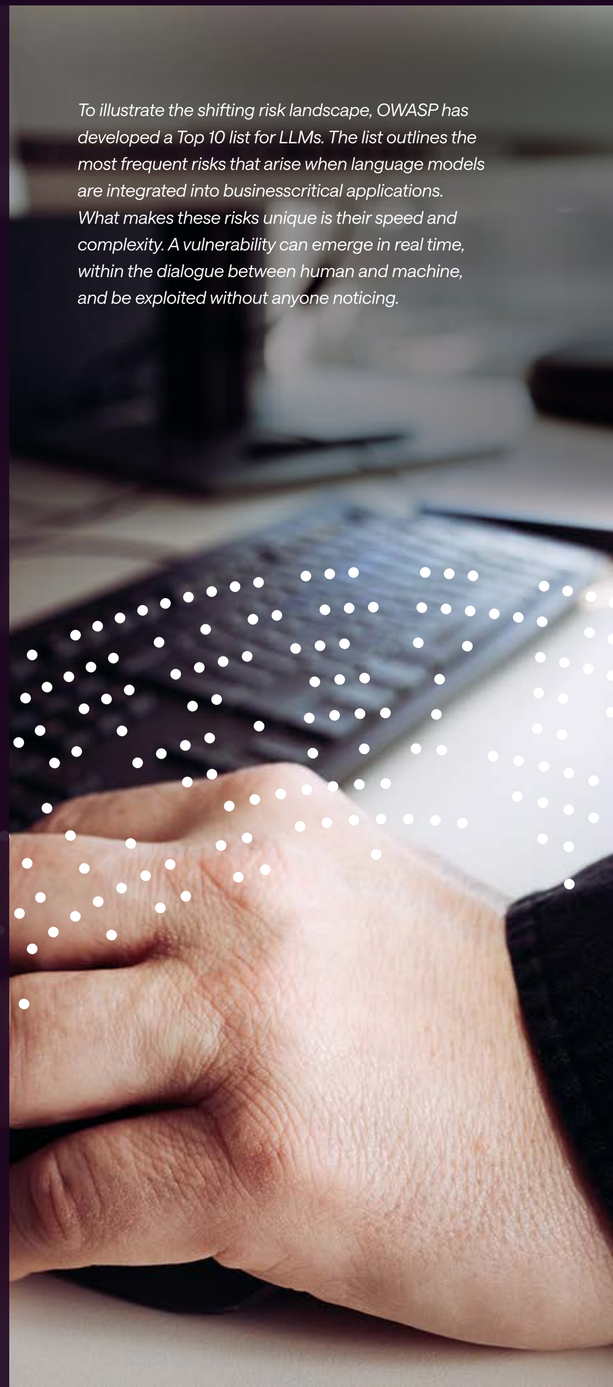
Privacy violations
Insufficient data protection in training data.

Insecure plugin integration
Too many unsecured connections.

Model theft
Theft of training data or model architecture.

Monitoring gaps
Lack of traceability and logging.

To illustrate the shifting risk landscape, OWASP has developed a Top 10 list for LLMs. The list outlines the most frequent risks that arise when language models are integrated into businesscritical applications. What makes these risks unique is their speed and complexity. A vulnerability can emerge in real time, within the dialogue between human and machine, and be exploited without anyone noticing.



When AI is used against us

AI's impact extends far beyond the organisation itself. When generative models are used to produce content at scale, they influence how information spreads, how reality is perceived, and how decisions are made. The information environment becomes more complex – and in some cases deliberately manipulated.

Manipulation & disinformation

AI is also used offensively, from deepfakes to fully generated information campaigns. When the line between what is real and what is fabricated becomes blurred, the consequences become societal. Disinformation can influence public opinion, markets and trust in institutions.

Trust has become the new attack surface.

For organisations, this means that protection is no longer only about data, but about perceptions of who they are and what they stand for.

Faulty decisions & distorted analyses

AI can generate insights that appear logical but are based on inaccurate or biased data. When such analyses are used as decision-making inputs, the consequences can be significant – for example, incorrect assessments of credit risk, production or supply chains. The result may be both financial losses and damage to trust.

Dataexponering & integritet

Generative AI tools used without sufficient control can unintentionally reveal trade secrets, customer data or source code. Incident reports show that accidental data sharing through AI tools is one of the leading causes of information leaks.

Crises of trust

When customers or partners feel that AI decisions are unfair, opaque or incorrect, the damage to the brand can be greater than the incident itself. Trust is often far more difficult to restore than data.

Regulatory consequences

When AI is used in decisions with real consequences, a clear line of responsibility emerges. Regulations such as the EU AI Act require documentation, risk assessments and human oversight in systems classified as high-risk. Organisations that cannot demonstrate how a decision was made risk both sanctions and a loss of trust.

These risks are not only theoretical. Organisations are already being held accountable for AI-driven decisions. This shows that responsibility cannot be deferred to the future but must be built in from the start.

The autonomous shift: when risks become operational

When AI systems move from analysing to acting independently, new possibilities emerge. Agentic systems can perform tasks, make decisions and interact with other systems in real time, creating opportunities for greater efficiency, automation and new business models.

At the same time, this development means that control and decision-making shift increasingly from humans to systems. If an agent receives the wrong instruction, misinterprets an objective or communicates with the wrong system, the consequences can be difficult to predict.

Agentic AI therefore creates a new need for governance.

Each agent needs a defined identity, a role and a purpose, in many ways similar to a human colleague. They must be identifiable, authorised and monitored. In practice, this means that the Zero Trust principle must also apply to autonomous systems and AI agents.

From reaction to resilience

The new risks may seem extensive, yet they follow the same fundamental logic as previous threats: what cannot be seen cannot be protected. To be prepared, organisations need to build resilience rather than reactivity. Four core recurring principles:

- **Visibility.** Knowing where AI is used and how it influences decisions.
- **Understanding.** Being able to explain the models' logic and limitations.
- **Protection.** Putting controls in place for data, access and behaviour.
- **Learning.** Evaluating and improving in line with technological change.

When these principles become part of everyday practice, AI can be managed with the same discipline as other business-critical functions. The goal is not to eliminate risks but to understand and handle them in a structured and professional way.

> Done properly, AI can increase both the speed and quality of security operations and free up time for what humans do best: assessing context, risk and consequence.

AI in the security domain

AI can act as both a safeguard and a threat. This chapter explains how AI strengthens security work, but also how it is used in attacks. From automated detection to manipulation and agent-driven intrusions. The aim is to understand the power of the technology – and the responsibility that comes with it.

AI has rapidly become an integral part of security operations. Models analyse logs, prioritise alerts, and support analysts in complex investigations. At the same time, attackers use the very same technology to scale phishing campaigns, generate deepfakes, and identify new vulnerabilities. Several global reports describe this as a second phase of AI in security: the technology is no longer experimental, but functions as both defence and offence.

On the defensive side, AI is used to streamline and scale security work.

It eases the burden on an overstretched SOC, elevates the quality of decision-making and makes advanced analysis more accessible. At the same time, the shift towards semi-autonomous security operations means systems take on more of the repetitive work while people set the boundaries and make the critical decisions.

Among other things, AI is used to:

- → Filter large volumes of alerts and highlight what genuinely requires action.
- → Detect anomalies in identity and access patterns.
- → Identify deepfakes and manipulated content across communication channels.
- → Provide AIdriven support for everything from threat hunting to report writing.

Done properly, AI can increase both the speed and quality of security operations and free up time for what humans do best: assessing context, risk and consequence.

AI as an attack tool

The same qualities that make AI attractive to defenders make the technology equally appealing to attackers. Google and Microsoft describe how threat actors use generative models to create more convincing phishing, vishing and SMS scams, often tailored to the language, tone and context of each individual recipient. This allows attackers to scale their operations in both volume and precision. Reports highlight several clear patterns:

- **Scaled social engineering.** AI makes it possible to massproduce personalised messages and false profiles, including synthetic voices and video designed to bypass KYC* checks and identity processes.
- **Faster attack chains.** Models are used to write and refine code, conduct vulnerability analysis and automate parts of intrusion and exfiltration flows.
- **Lower barrier to entry.** Access to unguarded LLMs on shady forums enables lessskilled actors to carry out attacks that previously required advanced technical expertise.

The result is a landscape where traditional defences are overwhelmed by more numerous, more credible and faster-moving attacks, while the responsibility for distinguishing real from fabricated content increasingly falls on the recipient.

**KYC (Know Your Customer) is an ongoing process that prevents fraud at every stage of a financial relationship and is based on customer identification, due diligence, and continuous monitoring.*

New attack surfaces in AI-driven environments

AI also introduces its own vulnerabilities. Mapping of cloud environments shows that more than 85% of all organisations already use some form of AI service, often built on young codebases, rapid development cycles and insufficient standards.

This includes exposed AI databases containing logs and keys, vulnerable GPU environments and loosely distributed frameworks where a single flaw can grant full control over the underlying infrastructure.

With agentic AI, an additional layer of risk emerges.

Systems that can independently initiate actions, interact with external systems, and make multi-step decisions can trigger rapid domino effects if granted too much autonomy without sufficient boundaries and oversight – sometimes without an attacker ever needing to gain access in a traditional way.

What separates the mature organisations from the rest?

Several global studies point to the same conclusion: most organisations use AI, but only a small number are truly prepared to do so securely. Accenture shows that around three-quarters lack basic data and AI security frameworks, even as generative AI increases the speed and complexity of threats. Cisco describes how a smaller group of frontrunners combine strategy, controls and technical capability across the entire AI lifecycle.

Organisations that are ahead are characterised by the fact that they:

- Treat AI security as an integrated part of cybersecurity.
- Have visibility into which AI services and models are being used, including BYOAI, which helps minimise Shadow AI.
- Use endtoend encryption, granular access control and monitoring (including for agentic systems).
- Work systematically with AIspecific threat modelling and testing against known LLM risks.

The difference therefore lies not in whether AI is used, but in how consciously it is applied – and in how clearly governance, architecture and capability align.

Protecting AI – technical measures in the right order

When organisations understand where the risks lie, that insight must be translated into practical protective measures. This is where technology and governance meet. Protecting AI is about creating visibility, detecting anomalies and being able to act with precision – not about building higher walls.

AI changes the logic of how protection is built. AI security needs to encompass data, models, decisions and behaviours. The protection must also be dynamic, since AI systems evolve over time. To avoid fragmented and reactive measures, a clear order is required, with each step building on the previous one.

Discover: seeing where AI is actually used

The first step towards control is understanding where AI is actually being used. Without that insight, visibility is impossible to achieve. Most organisations already use AI in some form, but few have a complete picture of where the technology is applied, what data it handles, and which decisions it influences.

Mapping across the organisation needs to cover three levels:

- **System level.** Identifying all AI-related systems and services, both internal and external. This includes standalone AI solutions as well as built-in functionality within existing applications. Many organisations discover that AI is already present in their everyday systems without any formal decision having been made.
- **Data level.** Understanding which data the AI uses, where it comes from and how it moves. Organisations need insight into both training data and data in production, including classification, retention time and access. This is where many of the greatest risks lie – such as bias, exposure and insufficient consent.
- **Organisational level.** Understanding how AI usage affects processes, decisions and business functions. Visibility is partly about technical inventory, and partly about knowing where AI influences the organisation and who carries responsibility.

Traceability is required throughout the entire chain: from data source to model, and onwards to the decision.

Once this overview is in place, visibility becomes part of ongoing governance rather than a one-off exercise. This is where technology, legal frameworks and governance come together. It is the foundation for the next step: being able to detect when AI systems begin to behave in unexpected ways.

Detect: understanding & identifying anomalies

When the map has been drawn, the next step is to understand behaviours. Detecting risks in AI systems differs from traditional monitoring because threats do not always appear in logs or network traffic. Instead, they arise in the model's decisions, in its interpretation of data, or in how AI systems interact over time.

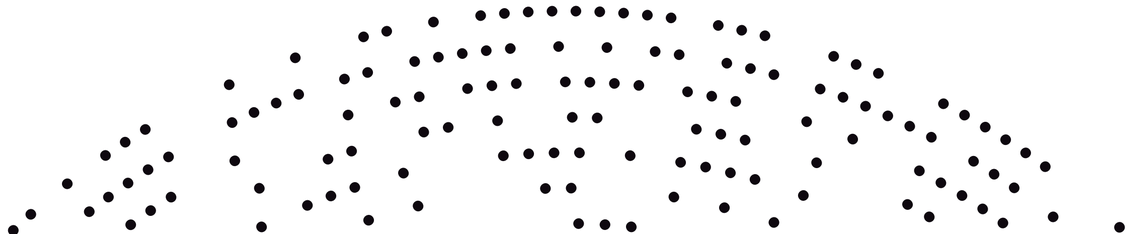
This makes detection a matter of insight rather than warnings and alerts. A model that begins producing unusual outputs may signal poor data quality just as easily as manipulation, and an AI agent may operate within its defined scope yet still cause unintended effects. Identifying such patterns requires both technical tools and analytical understanding.

A modern detection approach is built on three principles:

- **Continuous testing.** AI models need to be tested throughout their entire lifecycle, not only during development. By validating models, organisations can identify anomalies, bias, hallucinations or data poisoning before they cause consequences. Automated test environments are increasingly used to simulate attacks and assess model resilience, for example through methods similar to Tree of Attacks with Pruning (TAP).
- **Behaviour monitoring.** As AI becomes more autonomous, it is no longer enough to monitor systems as a whole. Each agent must be treated as an identity with a defined role, task and level of access. This means the same principles that apply to people – identity, authentication and access control – also apply to software that acts independently. By tracking how an agent communicates, which data it uses and which decisions it makes, anomalies can be identified in real time.
- **Collaboration between technology and the business.** Detection is not purely a technical discipline. Unexpected behaviours are often noticed first by users, analysts or decisionmakers who observe that an AI system's output does not fit the context. This is why observations from the business need to be gathered, reviewed and analysed alongside technical indicators.

An effective detection approach brings together the human and the machine.

As AI systems become more complex, the ability to understand what "normal" looks like becomes crucial. Only then is it possible to react to anomalies – and move on to the next step: building protection with precision.



Protecting with precision

Once the organisation has gained visibility and understanding of its AI use, protection can be built in the right place and in the right order. Protection is not about locking systems behind firewalls, but about creating clear boundaries for how AI is allowed to act – with which data, under what conditions, and with what level of responsibility.

In classic IT security, the focus has been on preventing intrusion. In AI security, it is just as much about preventing misbehaviour. A model trained on sensitive data, an agent communicating with other systems, or an application generating code in production can cause harm without any external attacker ever being involved. Protection must therefore ensure that AI systems understand their limits and remain within them.

Effective protection starts with clarity. Every AI model, service or agent needs a defined identity, purpose and level of access. This is the foundation for applying the same logic as in the wider security architecture: Zero Trust. No model, user or process should have more access than necessary, and all interactions must be verifiable and logged.

In practice, protection is built in several layers, each addressing a specific dimension of risk:

- **Purposebased access.** AI systems should only have access to the data required for their defined task. For example, customer service models should not reach HR data, and analytics tools should not be able to pull raw information from production systems.
- **Segmentation.** AI environments should be separated from the rest of the infrastructure. By isolating training, test and production environments, the effects of errors and attacks can be contained before domino effects spread.
- **Data protection & encryption.** Data used, generated or stored by AI must be encrypted and handled with the same discipline as other businesscritical information. Generative models should be complemented with Data Loss Prevention (DLP) controls to reduce the risk of accidental disclosure of sensitive information. This applies even to internal use – the most common data leak still comes from inside, not outside.
- **Policy enforcement & monitoring.** Protection needs to be dynamic. Policies for AI usage, data sharing and model updates must be translated into technical controls and continuous monitoring, so the system itself can identify when a model acts outside its intended purpose.

  This is where the boundaries between security and governance become clear. When policy, architecture and monitoring align, protection can be built with precision. The goal is to make AI predictable and controllable – not to restrict innovation. When that balance is in place, AI can become an integrated and reliable part of the organisation. A segmented, logged and governed AI environment becomes a controllable asset – part of the organisation's structure rather than an experiment on the side.

From protection to trust

Protecting AI is about creating clearer boundaries, not building higher walls. Organisations that combine visibility with understanding and technical safeguards can manage risks proactively and use AI with confidence.

In the AI era, resilience is as much a question of insight, responsibility and trust as of uptime and availability. When protection is built on understanding rather than fear, the technology becomes an asset rather than a risk. But technical measures alone are not enough. For AI to remain sustainable over time, structure, accountability and culture must evolve together. It is governance that gives direction to the technology – and makes it possible to use AI with both strength and control.

> For AI to become an asset rather than a risk, organisations need to ensure that human judgement remains the final checkpoint.

AI in the security domain is a force multiplier in both directions. It can make security operations more accurate, faster and more accessible, but it can also strengthen threat actors, open new attack paths and deepen existing weaknesses. Global reports are clear: adoption is high, but security and governance have not kept pace.

For AI to become an asset rather than a risk, organisations need to see this dual power as a design requirement: using AI offensively in their own defence, protecting models, data and agentic systems as systematically as any other business-critical asset, and ensuring that human judgement remains the final checkpoint.

Leadership, governance & acountability

When AI becomes part of an organisation's core operations, the demands on governance change. This chapter explores how organisations can lead, steer, and take responsibility for AI – from strategic decisions to legal frameworks, governance, and culture.

As AI moves ever deeper into the heart of an organisation's operations, the responsibilities shift with it. What was once seen as a technical innovation is now integral to how organisations evolve, make decisions, and build relationships. AI drives efficiency and productivity, while also influencing risk management, ethics, and trust. As a result, the conversation has moved beyond development teams and become a strategic concern for the entire organisation.

Decisions, accountability & consequences

Who carries the responsibility when an AI system proposes a decision that affects people or finances? How is the proper use of data ensured? And who reviews the logic behind systems that influence customer experience or risk assessment? These questions are no longer hypothetical – they sit at the heart of how companies lead, grow, and take responsibility when technology can make decisions faster than we can reflect.

When AI demands governance

In recent years, many organisations have focused on experimenting with AI and exploring its potential. But as AI becomes integrated into business processes and decision flows, the expectations change. Innovation must be paired with clear governance. Creativity should not be restricted – but it does need boundaries. Governance is not about resisting new technology, but about ensuring that every initiative proceeds in a controlled manner, where risks are understood and decisions can be traced.

From "Can we?" to "Should we?"

Many organisations still lack a clear overview of where AI is being used. Tools are introduced locally, data is applied widely, and responsibility is spread across functions. The result is a growing shadow of unregistered use – Shadow AI – initiatives driven by good intentions but without visibility, governance, or security safeguards. Moving from fragmentation to coordination requires leadership willing to ask "should we?" rather than "can we?"

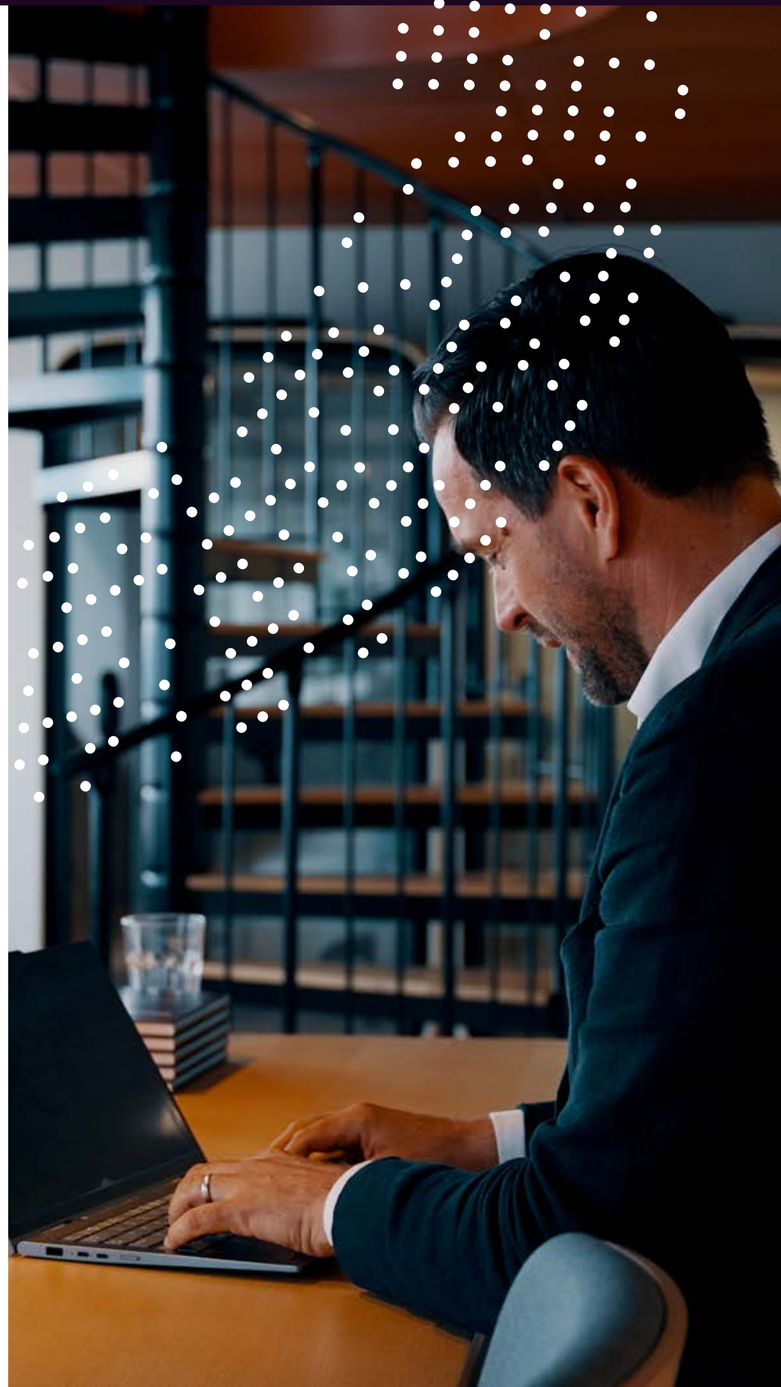
This is where the balance between innovation and control lies – and where the responsibility for the future of AI truly begins.

Organisations that succeed are those that see AI as a shared responsibility – where business, technology, security, and culture are woven into the same conversation. Today's leadership is about guiding the technology, as much as understanding how the technology guides us.

The human ability to guide AI

When AI becomes a significant part of the decision-making machinery, leadership must serve as its compass. This requires an understanding of the technology's possibilities – but also its limitations. Organisations that combine curiosity with responsibility, and technology with trust, will be better equipped to withstand risks while continuing to drive progress.

AI does not only change how we make decisions – it changes who makes them. Ultimately, the future of AI is about humanity's ability to guide the technology in the right direction, rather than the technology's capability on its own.



Human responsibility

AI has taken a clear place in the boardroom. The technology now influences brand, compliance, finance, and trust. It can no longer be treated as an isolated IT matter. Just as cybersecurity and sustainability have become natural board-level priorities, AI is now a central part of strategic governance.

Understanding as a starting point

Responsible leadership begins with understanding – recognising that AI is not a neutral technology. Models are chosen, trained, monitored, and used by people. Every step in that chain shapes decisions, data, and relationships. This is why organisations need to clearly define responsibilities, mandates, and roles, just as they have long done within information security and regulatory compliance.

AI use must also be traceable. Just as in financial reporting, it must be possible to explain how an AI-driven decision was formed: which data informed it, how the model interpreted that data, and which factors were taken into account.

Traceability is a fundamental principle in both good governance and upcoming regulation.

Accountability cannot be delegated

Even if systems can analyse and act, it is always people who carry the responsibility. AI cannot be held legally accountable, yet it can cause consequences for which leadership must answer. This division of responsibility can be described at three levels:

- **Leadership.** Sets the goals, risk levels, and principles for AI use.
- **Business owners.** Ensure that the technology is used in line with purpose, ethics, and legal requirements.
- **Technical functions.** Implement, monitor, and report risks in practice.

At the same time, formal decisions and divisions of responsibility are not enough to manage AI effectively. Leadership needs new capabilities to be able to steer AI: an understanding of how AI influences decision-making and risk, structures that set boundaries and accountability, and a culture that fosters transparency and trust. Only then can responsibility be carried in full.

Regulation as a shared direction

AI has developed faster than any other technological shift in modern times. In just a few years, it has moved from research and experimentation to influencing business decisions, public debate, and government oversight. This rapid evolution has made the need for regulation urgent – not to hinder innovation, but to create trust and predictability.

AI Act – a risk-based framework

With the EU's AI Act, Europe has taken the first step towards a shared model for responsible AI use. The legislation is built on a risk-based logic: the greater the impact a system has on people and society, the higher the requirements for transparency, control, and human oversight.

The purpose is to balance progress and safety, ensuring that AI can be used widely – but not without control.

In 2025, we saw the first major legal cases concerning AI-driven decisions. One AI-based recruitment system was accused of discrimination, another model of issuing incorrect credit assessments. Several courts, both in the EU and the US, confirmed the principle that responsibility always lies with the human, not the algorithm. This is one of the most fundamental principles in both the AI Act and GDPR: automated decisions must always be reviewable, understandable, and open to reassessment by a human.

When law becomes a strategic matter

This will shape the whole of 2026. Ethics, clarity, and accountability are moving up the board agenda as a key part of organisational risk management. Businesses that lack documentation, testing, and human oversight risk fines and crises of trust.

The AI Act will influence everything from product development to data governance and supply chains. The regulation affects both compliance and how organisations build trust. Those who can demonstrate that their AI is safe, traceable, and ethical will gain a strategic advantage – particularly in partnerships and public procurement.

* **The AI Act,** often referred to as the AI Regulation or the AI Law, is the world's first comprehensive AI legislation. Its purpose is to ensure better conditions for the development and use of AI.

Source: European Parliament.

Governance in practice: how AI is managed in everyday operations

When the technical safeguards are in place, the next step is to create long-term stability. Governance is the structure that transforms temporary measures into lasting assurance and makes AI manageable over time – it is the shift from handling AI to steering it.

As AI has become part of the organisation's core operations, the requirements for governance have also changed. Security checks or individual policies are no longer enough. Organisations need a framework that brings together technology, data, people, and decision-making within a single system. The framework needs to be able to answer three main questions.

Who is allowed to use AI, and for what purpose?

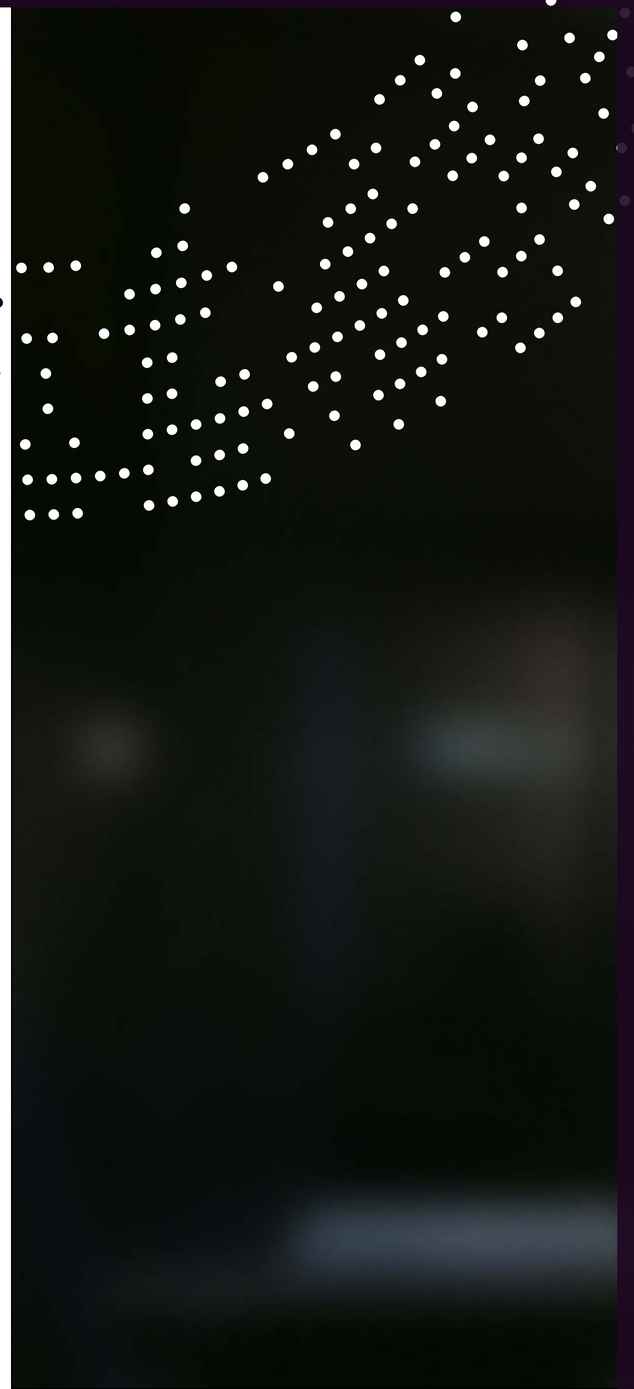
How are risks, decisions, and outcomes monitored?

What do we do when something goes wrong?

Governance ensures that these questions receive clear and consistent answers across the organisation. Organisations that integrate these answers into their strategy will be better prepared than those that simply follow the rulebook. AI governance therefore becomes a way of guiding development, with responsibility as an integrated part of decision-making. In the long run, trust is determined not by the volume of data or the speed of the models, but by how responsibly the technology is used.

A cohesive framework

An effective framework for AI governance needs to be integrated into the organisation. It should be able to grow alongside the organisation and reflect both technological development and the regulatory requirements that follow. International standards such as ISO/IEC 42001 and the AI Act provide a shared foundation for how governance can be structured.



Four fundamental pillars of effective AI governance:

- **Policy & direction.** AI policies should describe purpose, values, and risk levels, as well as clarify responsibility and mandate. A well-written policy serves as a compass for the entire organisation.
- **Roles & responsibilities.** Every function that influences AI – from development and operations to ethics and legal matters – needs a clearly defined mandate. Roles and titles such as AI Governance Lead, AI Security Architect, and Data Steward play an important part here.
- **Risk management & documentation.** All AI use must be traceable and explainable. This applies both to decisions made by models and to the risk assessments carried out before deployment. Traceability is essential for demonstrating accountability.
- **Continuous improvement.** Governance is not a project with an end date. AI systems evolve over time, and the governance framework must be able to evolve with them. Regular reviews, audits, and updates of policies are crucial for maintaining control.

Together, these pillars form a governance system similar to those long used in information security, but with a focus on decision-making rather than data protection alone.

AIMS as part of the management system

AI governance is on its way to becoming formalised as a dedicated management system. International standards introduce the concept of an Artificial Intelligence Management System (AIMS), which serves as a complement to existing frameworks such as the Information Security Management System (ISMS, information security) and the Quality Management System (QMS, quality).

AIMS builds on the same logic: to plan, implement, follow up, and improve, but with a focus on lifecycle, ethics, risk, and transparency within AI. For most organisations, this does not mean starting from scratch, but extending their existing governance to include AI. This is an important distinction. AI governance is a natural step in a development journey that many organisations are already on, moving from information security and data protection towards responsible technology governance. In this way, governance becomes an integrated management system that brings together security, quality, and accountability.

How governance works in practice

In practice, governance means that technology, people, and processes are managed as a whole. A well-functioning framework includes mechanisms for:

- → **AI lifecycle governance**
  Steering AI from development to decommissioning
- → **Incident management & escalation**
  Clear procedures for when something goes wrong
- → **Ethical revie.**
  Independent assessment of models and use cases
- → **Reporting & auditing**
  Regular followup to management and the board

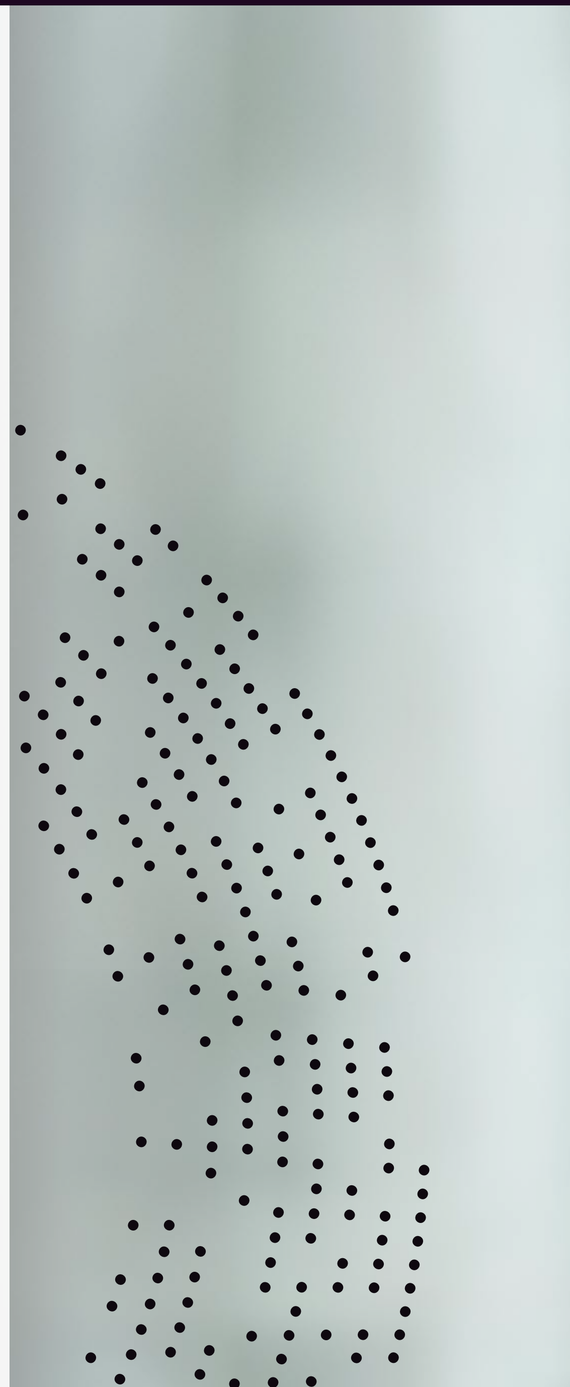
The requirements for transparency, traceability, and accountability have increased as AI has become more widespread, and organisations that already work with ISO 27001, GDPR, or sustainability reporting can often build on their existing structures. Governing AI is essentially about extending the same mindset to a new technological reality.

A link between technology & trust

Governance creates the framework for how AI may be used, how responsibility is distributed, and how value is protected. A strong framework ensures that innovation takes place under control and that risks are managed proactively rather than reactively. Technology can create opportunities, but governance creates trust.

The way forward does not require a fresh start. Many organisations can build on what already exists.

Expanding information security to include AI security, governance to include AI governance, and ethics to include concrete oversight is not a restart. Organisations that establish clear processes for responsibility, testing, and improvement on top of their existing governance structures can continue to grow with AI without losing control.



A shared knowledge base for AI

Leadership in the AI era requires an understanding of how the technology affects people, operations, and decision-making. For AI to be governed effectively, leadership teams and key functions need the knowledge to ask the right questions, assess risks, and make decisions with ethics and business logic in mind.

Boards and executive teams therefore need to build a shared knowledge base around AI – how models are trained, how they can be manipulated, and how decision logic can be influenced. The focus should be on having informed conversations with those who develop and maintain the technology. Many organisations are establishing dedicated roles to coordinate this work:

- → AI Governance Lead coordinates policy, risk, and compliance.
- → AI Security Architect is responsible for identity, access, and data protection.
- → AI Ethics Officer reviews impacts on people and transparency.
- → Data Steward ensures data quality and traceability.

The most important capability, however, is not a title but a mindset: the recognition that AI is something that must be led. Organisations that share this view shift AI from experiment to strategy, from technology to responsibility, and from risk to competitive advantage.

Culture brings governance to life

A framework without culture quickly becomes a paper exercise. Governance works when it is supported by a shared understanding of why it exists. AI touches the entire organisation – technology, legal, communications, HR, and the wider business. This means governance needs to be shared rather than owned by a single function.

Culture is the link between strategy and everyday practice. It determines whether the principles in the framework translate into real-world behaviour. Organisations that build a culture characterised by responsibility and curiosity create the foundation for both safe AI and continued innovation.

Characteristics of a strong AI culture

Transparency

Employees should understand how AI is used, which data underpins decisions, and where they can turn with questions or concerns.

Learning

AI governance requires continuous knowledge-building. New models, tools, and regulations mean that training must be ongoing rather than a one-off effort.

Trust

When governance is perceived as clear and fair, confidence grows. People are motivated to use AI in the right way and feel safe reporting deviations and contributing to improvement.

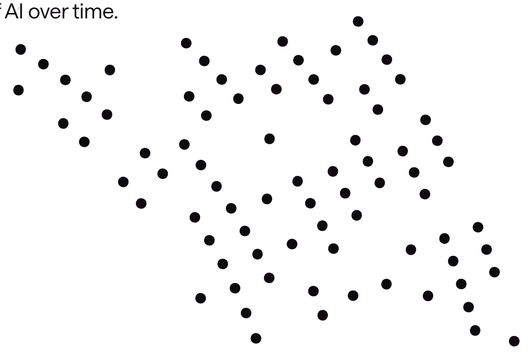
A secure path forward

AI is entering a new phase, where the technology is becoming an increasingly integrated part of organisational operations. The shift towards more agentic systems makes responsibility, transparency, and governance essential. As a result, AI in practice becomes a matter of leadership, sound judgement, and trust.

Regulation is beginning to catch up with technology, even as the technology itself takes on new forms that were previously difficult to predict. In the shift from generative models to agentic systems, AI is assuming a more active role in operations and collaborating with people to an ever-greater extent. This means that responsibility and governance become just as important as innovation. AI therefore touches on issues that reach far beyond technology.

The coming years will be shaped by how organisations build trust around their use of AI.

Organisations that can demonstrate how their AI systems are governed, monitored, and improved are better placed to meet requirements and balance technology with accountability. Success depends on how well organisations understand and lead their use of AI over time.



The future risk landscape

The risks of the future are less about attacks on systems and more about the influence on decisions, data, and trust. As AI becomes more integrated into communication, business logic, and public debate, the boundaries between internal security, information influence, and ethics gradually blur.

Three trends are already clear:

- **Disinformation & fabricated influence.** AI is being used to shape narratives and opinions on a scale that was not previously possible. This makes information influence a strategic risk for organisations as well.
- **AI in the supply chain.** Organisations must examine both their own AI models and those used by suppliers and partners. AI vulnerabilities in thirdparty systems become part of their own risk profile.
- **Regulation as a competitive lever.** Compliance with regulations such as the AI Act and ISO 42001 is becoming part of how organisations build longterm competitiveness. As documentation and traceability become standard, clear governance provides stronger foundations for partnerships and procurement.

From responsibility to action

AI has reshaped how people work, communicate, and make decisions. The greatest shift, however, lies in how the technology is used and integrated in practice. For organisations, the challenge is not to understand what AI can do, but how it should be applied to unite innovation with safety and accountability.

Experiences from recent years show that governance, security, and innovation are not in conflict. They reinforce one another. When structure, trust, and responsibility are built into the technology, AI becomes an enabler rather than a risk. The path forward is defined by three central shifts:

From experiments to ecosystems. AI cannot be run as isolated projects or standalone tools. It must be integrated into an organisation's processes, governance, and culture, in the same way as other strategic initiatives.

From oversight to collaboration. The AI of the future should work in partnership with people, not merely be supervised. By defining roles, mandates, and oversight for digital colleagues, secure autonomy can be achieved.

From regulation to trust. Compliance is the foundation, but not the end goal. Organisations that can show how their AI operates, how decisions can be explained, and how risks are managed will gain something that cannot be legislated into existence: trust.

A shared responsibility

AI is transforming many things, but not the most fundamental one: responsibility always rests with people. The technology can support, suggest, and anticipate, but it is humans who decide how it should be used and which values it should serve. Responsible use of AI is built on sound judgement, in balance with rules and technology. This means organisations need to take a position on difficult questions:

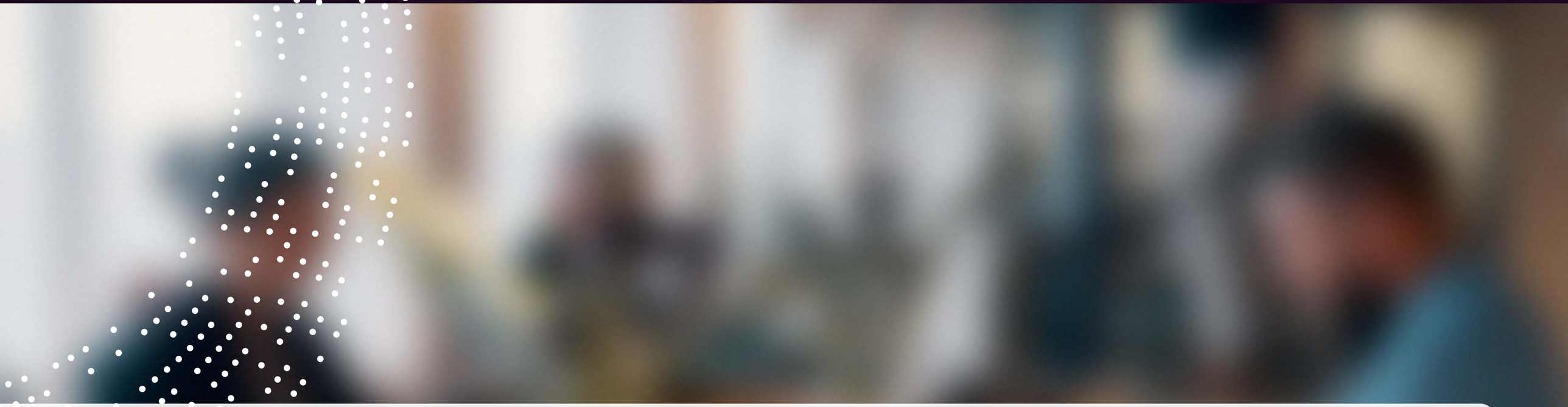
- Which decisions should be automated, and which should always remain human?
- What role should AI play in the organisation, and how can we ensure it delivers more benefit than harm?

Organisations that engage in these discussions create better conditions for managing AI over the long term. They strengthen their technical capability and build the trust required as AI takes on a greater role in operations.

Leading AI over time

All of this marks the beginning of a new era. AI is fundamentally a question of how technology is led and applied, rather than what it may theoretically become. Success will be measured by how wisely organisations are able to guide and take responsibility for AI over the longer term.

Organisations that unite innovation with responsibility, and security with trust, create better conditions for managing future technological shifts – and will also help shape them. Ultimately, it is about making deliberate choices and guiding development with sound judgement. It is not about running the fastest, but about running in the right direction.



Conclusion

Responsibility always rests with people

Artificial intelligence is entering a new phase. What was once handled as tests, pilot projects, and isolated initiatives now shapes how organisations make decisions, run operations, and build trust. As AI evolves from generative models to more autonomous and agentic systems, its presence in operational contexts grows. AI remains technical at its core, but it has become a strategic matter for leadership. This brings higher demands for governance, transparency, and accountability.

At the same time, the risk landscape is changing. Risks arise from technical vulnerabilities, how decisions are influenced, how data is used, and how trust develops over time. Information influence, dependencies in supply chains, and increasing requirements for documentation and traceability mean that security must be viewed in a broader perspective. When AI is used to strengthen security while also enabling more efficient attacks, the ability to understand, monitor, and control its use becomes crucial. In this reality, trust becomes a strategic asset.

To turn ambition into practice, structure is required. AI needs to be integrated into an organisation's processes, governance, and culture in the same way as other businesscritical capabilities.

This demands clear roles, responsibilities, and methods for followup, as well as the competence and judgement needed to determine how the technology should be used. Ultimately, this is a matter of leadership and governance. Organisations that lead AI with sound judgement create the conditions for sustainable development and longterm trust. For the most important insight is that responsibility always rests with people.

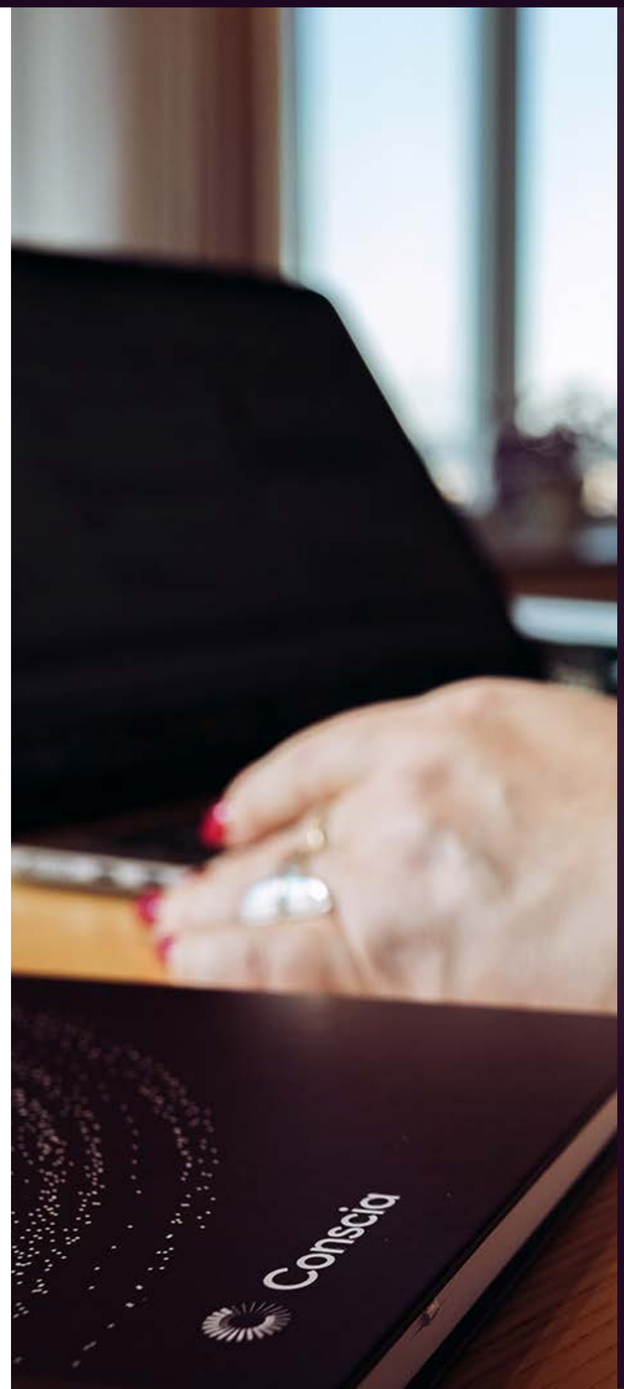
Experts behind the report

Our experts bring extensive experience and deep technical expertise in cybersecurity, networking, hybrid cloud, observability, automation, and artificial intelligence.

Marcus Lind
Business Area Manager Cloud Infrastructure, Observability & Automation
Conscia Sverige

Marcus Nilsson
Consultant Cybersecurity
Conscia Sverige

Octavio Harén
Business Area Manager Cybersecurity & CISO
Conscia Sverige



Sources & references

The report is based on consolidated knowledge and experience, complemented by external reports and data.

Accenture, State of Cybersecurity Resilience, 2025

Cisco, AI Readiness Index, 2025

Europaparlamentet, europarl.europa.eu

Gartner, gartner.com

Google, Cybersecurity Forecast, 2025

Microsoft, Digital Defense Report, 2025

OWASP, owasp.org

Trapets, trapets.com

Wiz, The State of AI in the Cloud, 2025

Conscia

Secure progress