Charter of Trust

Security by Default and Trustworthiness in AI

Publication date: February 12, 2026

Charter of Trust – Principle 3 Working Group

Classification CoT Public

Executive Summary

Artificial Intelligence (AI) is rapidly becoming a cornerstone of economic competitiveness, public service delivery, and national security. At the same time, it introduces new systemic risks to cybersecurity, privacy, and societal trust. This paper, developed under the Charter of Trust's Principle 3 *"Security by Default"*, addresses this dual challenge: securing AI systems throughout their lifecycle while responsibly leveraging AI to strengthen cybersecurity.

Aligned with the Charter of Trust's overarching goals—to protect data, prevent harm to people and infrastructure, and establish a reliable foundation for trust in a digital world—the paper outlines how *Security by Default* can operationalize *Trustworthy AI*. It positions security not as a reactive compliance exercise, but as an inherent, continuously enforced design principle that enables innovation while safeguarding resilience, transparency, and accountability.

Against a backdrop of increasing geopolitical competition, fragmented regulatory regimes, and accelerating AI adoption, the paper highlights the strategic importance of trust as a differentiator for organizations and societies alike. It examines key governance, technical, and regulatory risks surrounding AI, and underscores the need for coherent governance models that integrate cybersecurity, privacy, and ethical considerations from design through deployment and operation.

Building on the Charter of Trust's prior work, the paper provides a high-level framework for embedding Security by Default across the AI lifecycle, aligned with emerging global regulations such as the European Union (EU) AI Act. It also demonstrates how AI, when securely designed and governed, can serve as a powerful enabler of cybersecurity—enhancing threat detection, incident response, and risk management.

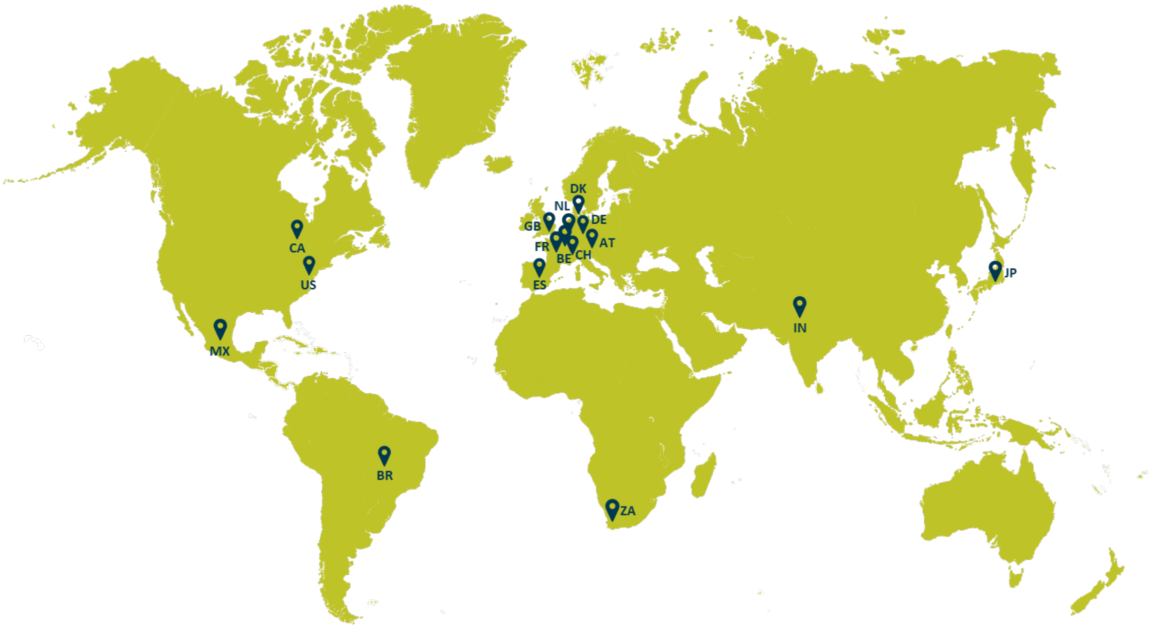
With this, it complements the Charter of Trust's "AI Policy Paper", which offers practical guidance for organizations that develop their own AI systems.

Ultimately, the paper reinforces the Charter of Trust's conviction that trust, security, and innovation must advance together. By embedding Security by Default and Trustworthy AI principles at the core of AI development and use, organizations can strengthen digital trust, improve resilience, and contribute to a safer and more reliable digital future.

Table of Contents

The Charter of Trust: Our Mission

Amidst an increasingly severe and complex threat landscape, the Charter of Trust (CoT) was established at the Munich Security Conference on 16 February 2018 as a non-profit alliance of leading global companies and organizations. Since then, a continuously evolving group of members and partners works together across sectors to strengthen cybersecurity, cultivate digital trust and make the digital world of tomorrow a safer place. Today, our initiative consists of 13 Partners and 17 Associated Partners operating in nearly 170 countries across five continents and representing more than 1.8 million employees.



Countries of establishment of CoT Partners/Associated Partners

On top of their countries of establishment, CoT Partners are also active in close to 170 countries worldwide.

Figure 1: CoT across the globe

All members endorse the **ten fundamental principles of CoT designed** to achieve **three overarching aims:**

- To **protect** the data of individuals and companies;
- To **prevent** damage to people, companies, and infrastructure;
- To **create** a reliable foundation on which confidence in a networked, digital world can take root and grow.

Guided by these principles, the Charter of Trust is working to protect our increasingly digitized world and build a reliable foundation on which trust and digital innovation can flourish. It advances effective cybersecurity policies worldwide and offers expertise in areas such as AI, postquantum-cryptography, security by default, supply-chain protection and education. This publication is issued by the Charter of Trust Working Group on Principle 3 — “Security by Default.”

The Security by Default Working Group: Our Activities

"Security by Default" is a principle emphasizing that security features are designed, implemented and consistently active and functioning throughout the entire lifecycle of a product, service, process, or business model. It represents the third of the Charter of Trust's ten core principles. To advance this principle, a dedicated Working Group—composed of cybersecurity experts from CoT member companies—has pursued a phased program of work.

Phase 1: Security by Default for Products, Functionalities and Technologies

In 2020, the Working Group on Security by Default defined a set of 19 baseline requirements "Phase 1 'Products, Functionalities, Technologies' Baseline Requirements" to drive integration of baseline security mechanisms into the products. Those requirements were followed by the Explanatory Document in 2021, "Achieving Security by Default. An Explanatory Document for the Phase 1 'Products, Functionalities, Technologies' Baseline Requirements", defining the critical cybersecurity requirements to deliver secure products, processes, services and business models.

Phase 2: Security by Default for Processes, Operations and Architectures

In 2021, the Principle 3 Working Group defined a set of 17 baseline security requirements "P3 Phase 2 "Processes, Operations, Architectures" Baseline Requirements" to support organizations in implementing secure development processes and environment. Those requirements were again, in 2022, accompanied by an Explanatory Document "Achieving Security by Default. An Explanatory Document for the Phase 2 "Processes, Operations, Architectures" Baseline Requirements".

Phase 3: Security by Default for Sharing of best practices on Security by Default adoption (Current Phase)

The document on "Secure Development Lifecycle: step-by-step guidelines" from 2023 bridges the two sets of baseline requirements, by showing step-by-step how a product or service can be designed, within a secure development process, and integrates the baseline security mechanisms. Our "Guideline on Cybersecurity Risk Assessment" from 2024 emphasized the importance of caution, proportionality, and due diligence when addressing cyber risks in digitally supported processes and value chains. The guideline offered practical, experience-based advice developed within the Charter of Trust P3 Working Group. Our latest publication from 2025 on "Security by Default in view of major Cybersecurity Regulations" addressed the growing regulatory landscape by providing orientation and guidance for organizations navigating diverse and evolving cybersecurity legislations across different jurisdictions.

The current paper builds on this groundwork, addressing the next level of maturity: ensuring that AI technologies are not only innovative and compliant, but also secure, resilient, and trustworthy throughout their entire lifecycle.

Disclaimer

The following document serves as an overview and general information resource only. It is not intended to provide legal advice or guidance of any kind. While efforts have been made to ensure the accuracy and completeness of the information presented herein, it may not encompass all legal nuances or variations applicable to specific circumstances.

Readers are encouraged to consult with qualified legal professionals or advisors regarding their particular situations or concerns. Reliance solely on the information contained in this document is done at the reader's own risk. The author and publisher disclaim any liability for any loss or damage arising directly or indirectly from the use of or reliance on this document.

Objective

This document explores the dual imperative of securing AI systems and leveraging AI technologies to strengthen cybersecurity. It outlines critical considerations for embedding "Security by Default" into AI design and deployment, while also examining how AI can be harnessed to enhance cyber defense capabilities across domains. By addressing regulatory frameworks, governance models, and practical use cases, the paper aims to guide organizations in building and using AI systems that are both resilient and ethically sound, ensuring trustworthiness in their design and use. It serves as an overarching map to navigate AI as a corporate condition with a pre-plotted destination towards “Trustworthiness” and “Security by Default”. For a more technical guidance on developing own AI systems, please consult the Charter of Trust’s “AI Policy Paper”.

1. Introduction

The Dual Imperative: Securing and Leveraging AI

In the face of rapidly evolving AI, organizations face two critical imperatives: securing AI systems and using AI to strengthen cybersecurity. As AI technologies become integral to both business operations and national security, ensuring that these systems are secure and trustworthy is paramount. At the same time, AI itself holds the potential to enhance cybersecurity capabilities, making it a key tool in the fight against rising cyber threats.

This paper explores these dual imperatives by addressing how "Security by Default" can be embedded into AI design and deployment, and how AI can be leveraged to improve cybersecurity outcomes across different sectors. The goal is to guide organizations in developing and applying AI systems that are resilient, ethical, and secure, fostering trust from the outset.

Why This Matters Now: The Context for AI Security

The urgency of this issue has never been greater. The rapid adoption of AI, combined with increasing cyber risks and the growing momentum of regulatory frameworks such as the EU AI Act, means that organizations must act quickly to ensure their AI systems remain compliant, secure, and trustworthy. Additionally, there is a strong link between technological leadership and trust—especially as AI plays an increasingly central role in geopolitical and economic competition.

As AI becomes more embedded in critical infrastructure, financial systems, healthcare, and even government operations, the stakes are high. Malicious attacks on AI systems could have widespread consequences, including undermining public trust in these technologies. AI systems often rely on vast amounts of sensitive data, operate autonomously, and integrate into critical infrastructure, making them prime targets for cyberattacks.

Without strong cybersecurity measures—such as encryption, access control, threat detection, and secure development practices—organizations risk data breaches, operational disruptions, and reputational damage. As AI continues to evolve, cybersecurity must evolve alongside it, becoming a foundational part of every AI-driven solution.

Defining Key Concepts: “Security by Default” and “Trustworthy AI”

Two foundational concepts are central to this discussion**:** "Security by Default" and "Trustworthy AI"**.**

Security by Default means that security features are built-in, active, and effective from the outset—without requiring user activation—and remain consistently reliable throughout the entire lifecycle of an AI system, product, service, or process. It ensures that protection, privacy, and resilience are integral, automatic, and continuously upheld.

Trustworthy AI, as defined by the EU Guidelines, refers to AI systems that adhere to seven key requirements ensuring ethical, safe, and responsible use. These include:

- **human agency and oversight**, empowering individuals and safeguarding fundamental rights;
- **technical robustness and safety**, guaranteeing resilience, reliability, and fallback mechanisms;
- **privacy and data governance**, ensuring data protection and integrity; transparency, providing clear explanations and traceability;
- **diversity, non-discrimination, and fairness**, avoiding bias and promoting inclusivity;
- **societal and environmental well-being**, fostering sustainability and positive social impact;
- and **accountability**, establishing mechanisms for responsibility, auditability, and redress.

Together, these principles aim to create AI systems that are lawful, ethical, and technically sound.

Embedding Trustworthy AI principles into a Security by Default approach ensures that security becomes a proactive, inherent enabler of trustworthiness rather than an add-on. Security by Default operationalizes key trustworthiness requirements by making them automatic and enforceable. Combined, Trustworthy AI provides the values and objectives, while Security by Default provides the mechanisms that activate and sustain them, resulting in AI systems that are secure, resilient, and aligned with societal expectations from the moment they are deployed.

2. Geopolitical Context of Trustworthy AI

Across organizations, governments, and industry sectors, AI has become a catalyst for innovation and efficiency. But it also introduces new levels of complexity and risk that extend far beyond economics and technology. The ability to develop, secure, and govern AI systems has become a defining factor of national competitiveness and corporate resilience. Thus, before exploring the risks and governance of AI, it is essential to understand the **geopolitical framework** that is shaping its development and regulation.

In today’s global order, the power of nations increasingly depends on their ability to develop, control, scale, and secure critical technologies, and in particular AI. The race for AI leadership, has become a central element of international power dynamics. Technological dominance now serves not only as an economic advantage but as a key instrument of geopolitical influence, fueling a multifaceted, global competition.

Data is the central asset in this contest. This is a source of economic value, predictive power, and strategic control: The fuel of the digital economy, and a critical asset of national and corporate power. Whoever governs data effectively controls the future of AI innovation.

While the **geopolitical race for AI leadership** shapes global strategies and regulations, its real impact is felt within industries themselves. AI technologies are now becoming **tools of geopolitical leverage**, leading to the fragmentation of global technology ecosystems. The result is a landscape defined by competing visions of digital sovereignty, security priorities, and ethical norms. The consequences for global innovation and commerce are significant. Multinational organizations must operate across jurisdictions with diverging rules, standards, and norms—navigating complex web of regulatory expectations and market access requirements. Divergent standards raise compliance costs, disrupt supply chains, and fragment international collaboration.

To navigate this environment, companies must integrate **geopolitical risk assessment** into their innovation and investment strategies, diversify supply chains, and strengthen resilience against technological and regulatory dependencies. Success requires regulatory agility, sustained investment in talent and internal capabilities, and proactive international partnerships that support transparency and interoperability.

In this fragmented landscape, **trust** is a strategic differentiator. Organizations that embed transparency, privacy, fairness, and accountability into their AI products from the outset can more effectively operate across differing jurisdictions and standards. Building **"trust by default"**—rather than "trust by compliance"— is essential for maintaining competitiveness in markets where accountability and ethical integrity are paramount.

3. Major Risks surrounding AI

As organizations scale their use of AI, a broad set of risks emerge that threaten information security, operational stability, and institutional trust. These risks fall into five key categories:

3.1 Governance and Compliance Risks

Regulatory and Compliance Exposure: The AI regulatory environment is evolving rapidly, with new requirements emerging from the EU AI Act, United Kingdom (UK) regulatory frameworks, and sector-specific rules. Uncertainty around compliance obligations increases the likelihood of penalties, reputational harm, and disruption to business operations. Ongoing regulatory tracking and alignment to compliance frameworks are essential.

Legal and Ethical Uncertainty: AI presents unresolved legal challenges around copyright, intellectual property (IP), and ownership of AI-generated content. Disputes over rights, attribution, and permitted usage are increasingly likely. In high-risk domains such as healthcare, financial services, or public safety. AI failures can introduce ethical breaches or severe safety incidents. Dependence on unverified vendor claims further amplifies operational and compliance risk.

3.2 Technical and Operational Risks

Shadow AI: Unapproved or unmanaged AI tools introduce severe data-handling and security risks. Shadow AI bypasses governance controls, potentially exposing sensitive data, expanding the attack surface, and increasing the likelihood of leakage or misuse.

AI Hallucinations: Models can generate confidently incorrect, biased, or misleading outputs. These inaccuracies undermine decision quality, create operational errors, and erode trust in automated tooling.

Model and Data Security Risks: AI systems rely on complex architectures and large datasets, making them susceptible to model poisoning, adversarial inputs, data corruption, and manipulation. Weaknesses in robustness, testing, or resilience can result in systemic failure. Continuous validation, hardening, and security testing are required.

Oversight and Monitoring Failures: Insufficient human supervision during design, deployment, or ongoing use allows issues such as bias, vulnerability, and misuse to persist undetected. Excessive trust in automation can reduce critical scrutiny. Strong oversight and clear accountability are essential.

3.3 Security and Fraud Risks

AI-Enabled Fraud and Attacks: AI now enables more sophisticated phishing, social engineering, malware generation, and disinformation campaigns. High-quality synthetic content makes threats more convincing and difficult to detect. Embedding security-by-design principles throughout the AI lifecycle is crucial.

Autonomy and Control Risks: Highly autonomous AI systems may behave unpredictably or outside intended boundaries. Without enforced constraints, human-in-the-loop controls, and fail-safes, autonomous behaviors can trigger cascading or harmful outcomes.

3.4 Strategic and Financial Risks

Financial Exposure: AI initiatives can be costly, requiring investment in infrastructure, specialized talent, and vendor solutions. Misjudged expectations, cost overruns, or failed deployments can result in material financial losses. Dependence on third-party providers introduces vendor lock-in and pricing volatility.

Long-Term Strategic Risks: AI may contribute to workforce displacement, reduced human oversight, and over-dependency on opaque or unexplainable systems. Poorly anticipating these impacts weakens resilience and undermines trust among employees and the public.

Misalignment Risks: AI systems can optimize objectives that deviate from organizational goals or ethical principles. Misalignment can lead to harmful outcomes, strategic missteps, or unintended behaviors. Strong governance, clear objectives, and constant monitoring are required.

3.5 Human Factor Risks

Overdependence on AI: Relying excessively on automated systems can diminish human judgement, reduce operational adaptability, and increase vulnerability when systems fail or produce flawed outputs. Balanced integration preserves resilience.

Accountability and Responsibility Gaps: Unclear ownership of AI decisions complicates governance, compliance, and incident response. Defined roles, responsibility models, and escalation structures are critical for trustworthy and auditable AI operations.

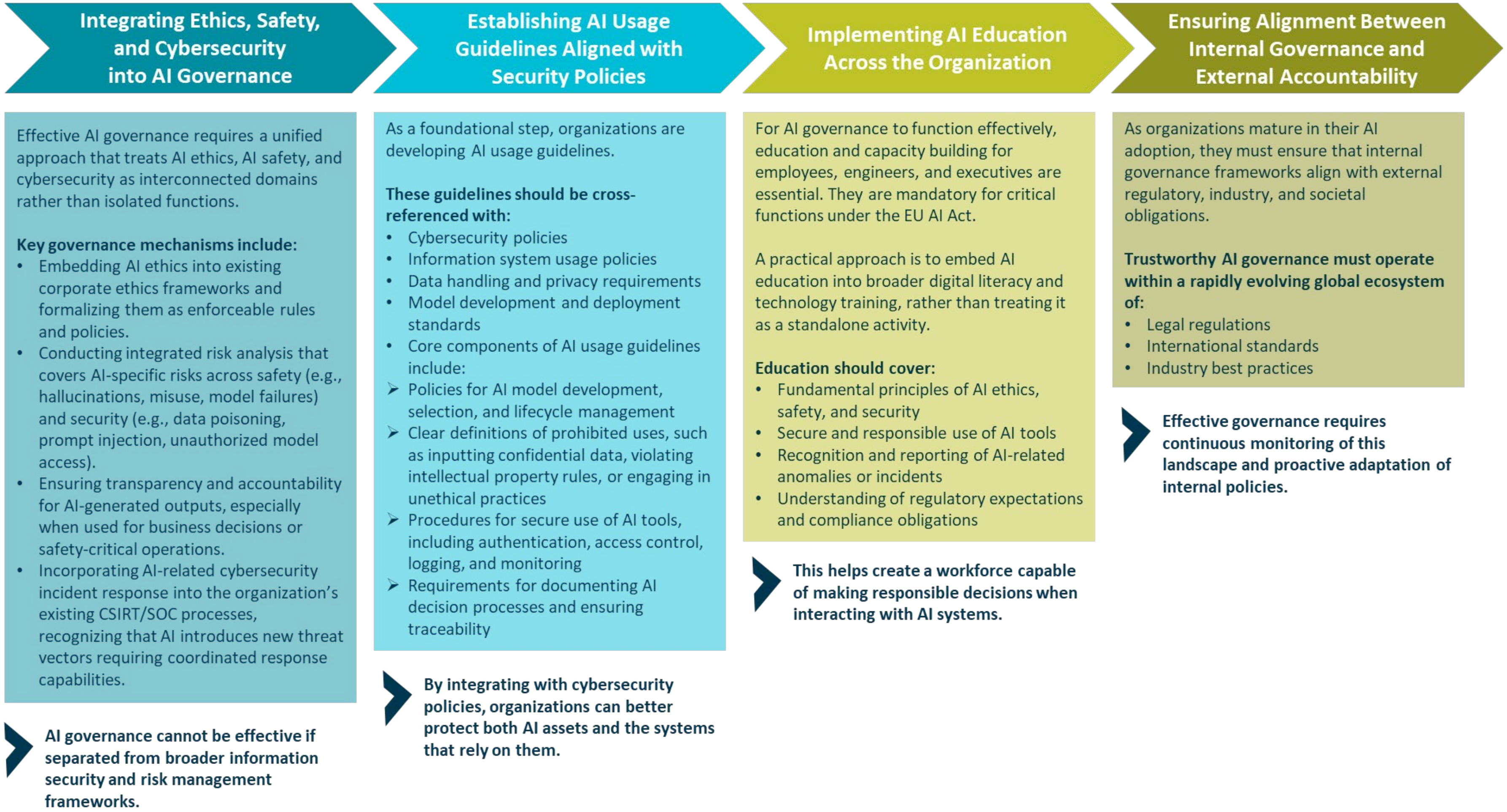
Conclusion

Mitigating these risks requires a comprehensive AI governance framework that promotes ethical, transparent, and secure use of AI technologies. Effective governance includes regulatory monitoring, technical safeguards, security controls, cost-managed investment, and long-term strategic planning to ensure responsible and resilient AI adoption.

4. Development of Governance Models for Implementing Trustworthy AI

As many organizations, Charter of Trust partners are currently assessing and adapting to the multi-dimensional impact of AI and the diverse use-cases of the technology. As such, we are developing answers related to efficient governance and security: Balancing security policy enforcement while utilizing AI technology in evolving and improving products and services.

Development of Governance Models for Implementing Trustworthy AI



Integrating Ethics, Safety, and Cybersecurity into AI Governance

Effective AI governance requires a unified approach that treats AI ethics, AI safety, and cybersecurity as interconnected domains rather than isolated functions.

Key governance mechanisms include:

- Embedding AI ethics into existing corporate ethics frameworks and formalizing them as enforceable rules and policies.
- Conducting integrated risk analysis that covers AI-specific risks across safety (e.g., hallucinations, misuse, model failures) and security (e.g., data poisoning, prompt injection, unauthorized model access).
- Ensuring transparency and accountability for AI-generated outputs, especially when used for business decisions or safety-critical operations.
- Incorporating AI-related cybersecurity incident response into the organization's existing CSIRT/SOC processes, recognizing that AI introduces new threat vectors requiring coordinated response capabilities.

AI governance cannot be effective if separated from broader information security and risk management frameworks.

Establishing AI Usage Guidelines Aligned with Security Policies

As a foundational step, organizations are developing AI usage guidelines.

These guidelines should be cross-referenced with:

- Cybersecurity policies
- Information system usage policies
- Data handling and privacy requirements
- Model development and deployment standards
- Core components of AI usage guidelines include:
  - ➢ Policies for AI model development, selection, and lifecycle management
  - ➢ Clear definitions of prohibited uses, such as inputting confidential data, violating intellectual property rules, or engaging in unethical practices
  - ➢ Procedures for secure use of AI tools, including authentication, access control, logging, and monitoring
  - ➢ Requirements for documenting AI decision processes and ensuring traceability

By integrating with cybersecurity policies, organizations can better protect both AI assets and the systems that rely on them.

Implementing AI Education Across the Organization

For AI governance to function effectively, education and capacity building for employees, engineers, and executives are essential. They are mandatory for critical functions under the EU AI Act.

A practical approach is to embed AI education into broader digital literacy and technology training, rather than treating it as a standalone activity.

Education should cover:

- Fundamental principles of AI ethics, safety, and security
- Secure and responsible use of AI tools
- Recognition and reporting of AI-related anomalies or incidents
- Understanding of regulatory expectations and compliance obligations

This helps create a workforce capable of making responsible decisions when interacting with AI systems.

Ensuring Alignment Between Internal Governance and External Accountability

As organizations mature in their AI adoption, they must ensure that internal governance frameworks align with external regulatory, industry, and societal obligations.

Trustworthy AI governance must operate within a rapidly evolving global ecosystem of:

- Legal regulations
- International standards
- Industry best practices

Effective governance requires continuous monitoring of this landscape and proactive adaptation of internal policies.

Figure 2: Development of Governance Models for Implementing Trustworthy AI

5. Regulatory Aspects

As AI governance becomes more mature, internal policies and controls must increasingly be designed with external obligations in mind. The same mechanisms that integrate ethics, safety, and cybersecurity—risk management, accountability, incident response, and documentation—also determine whether an organization can demonstrate compliance to regulators, customers, and partners. This chapter therefore turns from internal governance to the regulatory landscape shaping AI development and deployment across jurisdictions.

5.1 Global AI Regulation

AI has been subject to ethical, safety, health, and security considerations, but is now emerging as a central **focus of regulation worldwide**. Unlike traditional safety or environmental laws, **AI regulations target a** specific technology, reflecting its strategic and geopolitical significance.

Regulatory approaches vary widely across regions, sometimes even conflicting, creating major challenges for global AI providers, AI developers, and AI users. Exemplary obligations for AI providers usually include an AI lifecycle risk management system, (training) data governance, appropriate levels of accuracy, robustness and cybersecurity, as well as an incident reporting mechanism. AI developers and users must usually fulfill certain obligations regarding AI (input) data and AI system monitoring, incident reporting, human oversight, and transparency. Non-compliance can result in severe penalties and **restricted market access**, while alignment with emerging standards may bring strategic advantages, including funding opportunities or preferred partnerships.

Global competition for AI leadership has also **slowed or stalled some legislative initiatives**, as governments seek to ensure that AI regulation promotes both responsible innovation and safety and control, while preventing monopolistic practices.

In addition, AI is also directly and indirectly affected by several non-technical, but rather human rights, economic policy, or geopolitically motivated regulations for instance regarding AI hardware supply chains, sustainability requirements, usage restrictions, data privacy and copyright issues.

Table 1: Selected global AI regulatory developments as of Q4/2025

Country/Region	Regulation	Comment
European Union	**AI Act** (2024)	EU AI Act regulates AI based on its potential risk to health, safety, and fundamental rights, banning some uses while imposing strict requirements on high-risk systems.
Japan	**AI Act** (2025)	Japan's AI Act primarily sets a framework for future law and policies to regulate AI research, development, and use, while current AI regulation is mainly based on soft laws (e.g., standards, guidelines).

South Korea	**AI Basic Act** (2026)	SK's AI Basic Act with similarities to EU's AI Act has been enacted in 2025 to take effect in 2026.
United Kingdom	**AI Bill** (2025)	The UK's AI Bill (introduced in March 2025), which aims to introduce binding rules for AI regulation, is still in the legislative process as of January 2026 (currently at its 2nd reading). Until it is enacted, AI oversight in the UK continues to rely mainly on soft-law mechanisms, such as standards and regulatory guidelines.
United States	**California AI Bill SB 53** (2025)	California signed AI Bill SB 53, a state law discussed widely as it is the first US law that captures the most advanced AI models. Previous federal legislation on specific issues such as deepfakes or discrimination has been revoked by Trump administration (2025). US AI Training Act (2021) requiring federal agencies to provide AI training for employees in management and acquisition roles.
China	**General Purpose AI (GPAI) Measures** (2023), **AI Labeling** (2025)	China adopted several specific AI regulations primarily on AI use such as GPAI Measures (2023), AI Labeling (2025) including a national AI standards committee (2024).
Canada	Privacy frameworks after **AIDA** pause	After the Canadian Artificial Intelligence and Data Act (AIDA) proposed in 2022 has not been enacted into law, Canada has some smaller scale AI regulation mainly based guidelines and best practice for instance on automated decision-making.

5.2 Interplay of AI Regulations with other Regulations

AI regulation interacts closely with existing legal frameworks, both horizontal (cross-domain) and vertical (domain-specific). Key areas of overlap include:

- **Cybersecurity laws** (e.g., EU Cyber Resilience Act) – ensuring secure AI development, deployment, and operation
- **Data protection** laws (e.g., Data Act, General Data Protection Regulation (GDPR)) – safeguarding privacy, IP, and data access rights
- **Safety regulations** (e.g., Machinery Regulation, Programmable Logic Device (PLD), General Product Safety Regulation (GPSR)) – ensuring AI-enabled solution do not endanger health or safety
- **Fundamental rights frameworks** (e.g., EU Charter of Fundamental Rights, European Convention on Human Rights) – promoting ethical and fair AI use

Bridging the gaps between sector-specific regulations such as the AI Act, Cyber Resilience Act, GDPR, and Data Act, the European Commission recently published the Digital Omnibus Regulation Proposal. Its goal is to establish a comprehensive framework to harmonize and streamline the application of multiple digital regulations within the EU, ensuring regulatory consistency, reducing compliance complexity, and fostering innovation across digital markets. For organizations implementing "Security by Default" and striving for Trustworthy AI, the Digital Omnibus Regulation Proposal offers several key benefits:

- **Regulatory Alignment:** It clarifies the interplay between AI-specific requirements and broader obligations in cybersecurity, data protection, and consumer rights, helping organizations to design AI systems that meet all relevant standards from the outset.
- **Simplified Compliance:** By providing unified procedures and definitions, the proposal reduces the risk of conflicting obligations and supports efficient governance and risk management for AI systems.
- **Enhanced Trust and Transparency:** The proposal encourages the adoption of transparent practices, interoperability, and accountability, which are essential for building user trust in AI technologies.

Integrating the Digital Omnibus Regulation Proposal into AI governance frameworks is a strategic step for organizations seeking to minimize regulatory risks and maximize the trustworthiness of their digital products and services.

5.3 AI System Labels

Inspired by cybersecurity labeling (e.g., US Cyber Trust Mark), several initiatives propose AI system labels to **demonstrate compliance** with technical (e.g., safety, cybersecurity, reliability) and ethical (e.g., fairness, privacy, traceability, explainability, sustainability) obligations and requirements, as defined in frameworks like the EU AI Act. These labels aim to **support regulatory conformity while enhancing user** trust *(*see Figure 3, Figure 4, and Figure 5*).*

Because AI legislation is still evolving (cf. previous section), most proposed AI product labels are in the **very early stages** and should be **considered as proposals** for potential implementation (see Figure 3). While mandatory regulatory compliance with the EU AI Act is integrated into the overarching CE marking, additional dedicated proposals for AI labels exist from organizations such as German VDE, AIGN OS, OECD, IEEE, Artifact Studio, or Center for AI Safety (CAIS).

5.4 GenAI Transparency Labels

A second category of labels focuses on **AI-generated content** to:

a) **Inform users when content was created or modified by AI**
b) Enable other AI systems to **exclude such AI-generated data from training** to avoid model collapse, among others

To be effective, these labels should include both **human-readable notices** and non-removable **machine-readable watermarks, digital signatures**, or similar mechanisms (see Figure 4).

Legislative measures, such as Article 50 of the EU AI Act, also introduce **mandatory transparency labels for generative AI**. These rules require clear identification of AI involvement in the creation of media such as text, images, audio, and video. Transparency labels help **users recognize AI-generated content** (e.g., deepfakes, chatbots), and help **AI developers avoid inadvertently training models on AI-generated material**.

At present, beyond a few simple voluntary text overlays, **no standardized or binding AI transparency labels or machine-readable watermarks** exist. However, a first draft of a Code of Practice with some proposals for the practical implementation of the transparency requirements from the EU AI Act has been available since December 2025.

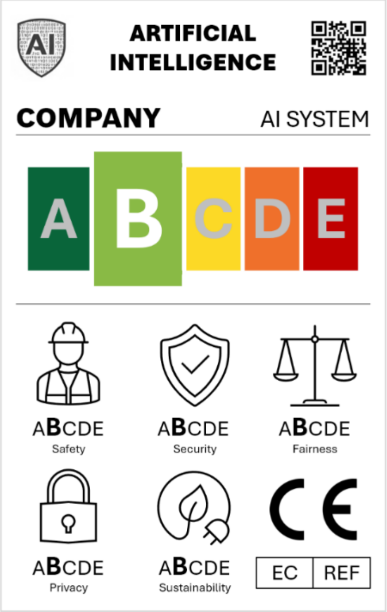


Figure 3: Exemplary AI systems label (AIS label) with exemplary combined rating on AI system safety, cybersecurity, fairness, privacy, and sustainability.

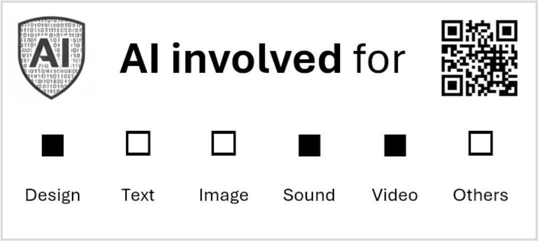


Figure 4: Exemplary human-readable label to mark AI-generated or AI-manipulated media data to avoid misunderstandings.

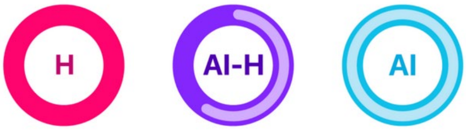


Figure 5: Proposal for a standardized AI label from Artifact Studio for 100% human-generated content (H), for human- and machine-generated content (AI-H) where the ring show the ratio between both, and for 100% machine generated content (AI).

6. Security and Privacy Aspects of AI

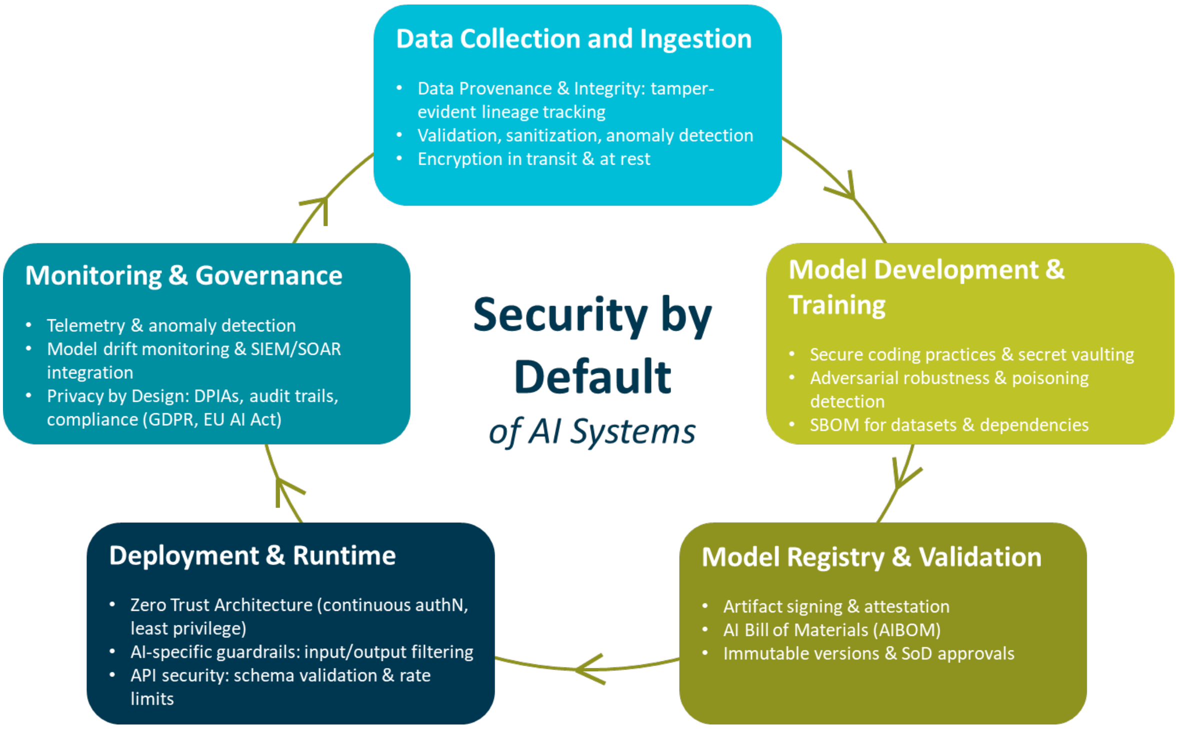

Data Collection and Ingestion
• Data Provenance & Integrity: tamper-evident lineage tracking
• Validation, sanitization, anomaly detection
• Encryption in transit & at rest
Model Development & Training
• Secure coding practices & secret vaulting
• Adversarial robustness & poisoning detection
• SBOM for datasets & dependencies
Model Registry & Validation
• Artifact signing & attestation
• AI Bill of Materials (AIBOM)
• Immutable versions & SoD approvals
Deployment & Runtime
• Zero Trust Architecture (continuous authN, least privilege)
• AI-specific guardrails: input/output filtering
• API security: schema validation & rate limits
Monitoring & Governance
• Telemetry & anomaly detection
• Model drift monitoring & SIEM/SOAR integration
• Privacy by Design: DPIAs, audit trails, compliance (GDPR, EU AI Act)
Security by Default
of AI Systems

Figure 6: Security by Default for AI Systems

6.1 Establishing Security and Privacy by Default: The Foundational Imperative

As organizations accelerate AI adoption to gain competitive advantage, they also expose themselves to new security risks as explained in Section 3.

To operate safely and responsibly, organizations must embed “Security by Default” into every stage of the AI lifecycle. Figure 6 shows some of the key steps that enterprises can adopt to incorporate such an approach seamlessly into their AI initiatives. This diagram illustrates how Security by Default can be embedded across every stage of the AI lifecycle to ensure trustworthy and resilient AI systems. It highlights the key security and privacy controls required during data collection, model development, validation, deployment, and continuous monitoring. By integrating measures such as provenance tracking, adversarial robustness, SBOM validation, and zero-trust runtime protections, it can provide a unified view of the safeguards needed to manage AI risks end-to-end, to enterprise CISOs and Security decision makers. The figure sets the foundation for the detailed security and privacy considerations discussed throughout this section.

This shift – adoption of a “Security by Default” in AI programs - is essential because AI systems are dynamic, probabilistic, and deeply socio-technical. Their behavior can change across contexts, their inputs function like executable instructions, and their dependency on sensitive data intensifies both security and privacy risks.

Technical Risks associated with AI systems

Expanding on the general risks we highlighted in Section 3 of this Paper, Table 2 below summarizes the most common and critical technical security and privacy risks specific to AI systems, which organizations must address to maintain integrity and trustworthiness:

Table 2: Examples of Technical Risks associated with AI systems

Threat Name	Description	Key Areas of Concern
Prompt Injection & Goal Manipulation in Large Language Models' (LLMs)	Exploiting the LLM with malicious inputs to **bypass safety instructions**, reveal sensitive data, or trigger **unauthorized actions** (including agent goal and Instruction manipulation).	A prime target for exploitation, potentially resulting in exposure of **confidential and private data**, system compromise, and unauthorized operations carried out by the agent.
Data Poisoning & Integrity Attacks	Deliberately **corrupting training, fine-tuning, or context data** to embed backdoors, vulnerabilities, or biases that **alter the model's behavior**.	Compromises the foundational learning process. It can cause models to misclassify specific inputs and is critical to defend against due to the risks of **biased or harmful outputs**.
Malicious Content & Deceptive Media	AI creating realistic fake content (e.g., **deepfakes**) for **disinformation, fraud, or impersonation**. This includes AI Agent–driven **social engineering** attacks.	**Deepfake video calls** can impersonate company executives to trick employees into **fraudulent transactions** or phishing. **Voice cloning (vishing)** can lead to substantial financial losses.
Model Theft & IP Extraction	Attackers stealing proprietary **AI models** (weights, architecture, parameters) or **algorithms** via exposed interfaces or repositories.	Leads to the **loss of IP/competitive edge** and violates **copyright risks**. Stolen models can be used to develop more effective adversarial attacks against the agent system.
Sensitive Information Disclosure & Confidentiality	AI models unintentionally **leaking confidential data (Personally Identifiable Information (PII) or trade secrets)** or revealing their **internal**	Includes **Model Inversion Attacks** where output is used to recover PII or training data indicators, breaching **privacy** and regulatory compliance (e.g., **GDPR** or **Health Insurance**

	prompt template/secrets. Confidentiality is paramount.	**Portability and Accountability Act (HIPAA)**). Requires managing the **human factor** risk regarding data access.
Insecure Output Handling	Using **unvalidated AI outputs** in other systems. This occurs when LLM-generated content is passed downstream without adequate **sanitation or validation**.	Can lead to security risks in back-end systems, such as **Cross-Site Scripting (XSS)**, **CSRF**, **privilege escalation**, or **Remote Code Execution (RCE)**. This is exacerbated if the LLM is used to create **Infrastructure as Code (IaC) or Policy as Code (PaC) templates**.
Excessive Privileges	Granting AI systems or their associated plugins permissions beyond what is strictly necessary (**least privilege**).	Amplifies the impact of any exploit, such as prompt injection, potentially leading to the **deletion or exposure of business-critical data** on downstream systems if the **non-human identities (NHIs)** are compromised.
AI Supply Chain Vulnerabilities	Exploiting weaknesses in **third-party AI components**, including unvetted **pre-trained models**, open-source libraries, external data sources, or compromised infrastructure.	Compromised dependencies can **propagate vulnerabilities** across multiple AI systems, potentially leading to **backdoor insertion** or model poisoning. Governance frameworks must address **third-party software and data** risks.
Unwanted / Malicious Use of AI	The intentional misuse of AI's capabilities for **offensive cyber activities** or other harmful, disallowed, or **non-compliant actions**.	Includes the use of specialized AI tools for hacking (**HackGPT**, **PentestGPT**) and general **abuse** or **misuse** of AI platforms for unauthorized purposes (e.g., **resource abuse**).
Model Evasion (Adversarial Attacks)	Crafting subtle, often **imperceptible inputs** that deceive models into making incorrect classifications or decisions. This differs from goal manipulation (Risk#1) which targets **LLMs** instruction-following behavior.	Poses significant **safety risks** in critical systems like autonomous vehicles and can be used to **bypass security mechanisms**. Defenses must accommodate the **dynamic nature** of outputs.

AI Agent & Autonomous System Exploitation	Manipulating self-learning systems that have access to internal tools or external systems to cause harm or leak data.	Exploits unique features like **Memory Poisoning**, **Tool Misuse**, and **Inter-Agent Communication Poisoning**. This is particularly critical in **Multi-Agent Systems (MASs)** where failure can cascade.
Insecure AI System & Component Design	Core flaws in the AI system's architecture, configuration, or security controls, including failures to define the **boundaries** or enforce **secure model requirements**.	Results in broad, systemic weaknesses and increases the attack surface. It includes ignoring **human oversight checkpoints** or failing to address **algorithmic bias**

6.2 Securing the AI Life Cycle

As it is clear from the examples listed in *Table 2*, it is important for enterprises to have a structured approach – like the "Security by Default" approach we are proposing in Figure 6 - to secure the entire lifecycle of their AI programs, applications and workloads. To do so, we recommend a Security by Default principle. Adopting Security by Default ensures robust protection through automated encryption, identity management, and hardened configurations for all the workloads and integrated services in the end-to-end lifecycle of an AI application. With this "Security by Default" principle, enterprise security teams can drive integrity and supply chain assurance by enforcing trusted sources, signed artifacts, and SBOM validation to prevent compromised components. And taking it one step further, for AI development and operations (DevOps) teams, what this translates this to - is adopting the traditional Secure Development Lifecycle (SDLC) and associated development processes - to Machine Learning (ML) and Security Operations needs. This new approach is being called the "Machine Learning Security Operations" (MLSecOps). In Figure 7, we propose an MLSecOps framework, that further strengthens this by unifying DataSecOps and ModelSecOps, combining secure data lifecycle management with secure model DevOps. As shown in Figure 7, this end-to-end approach embeds security controls from data sourcing through deployment and monitoring. And finally, a clear distinction between controls owned internally and those managed by external vendors is also essential, ensuring consistent accountability and reducing supply-chain risks across the AI lifecycle. The subsequent sub-sections explain these concepts in further detail.

A. Integrity and Supply Chain Assurance

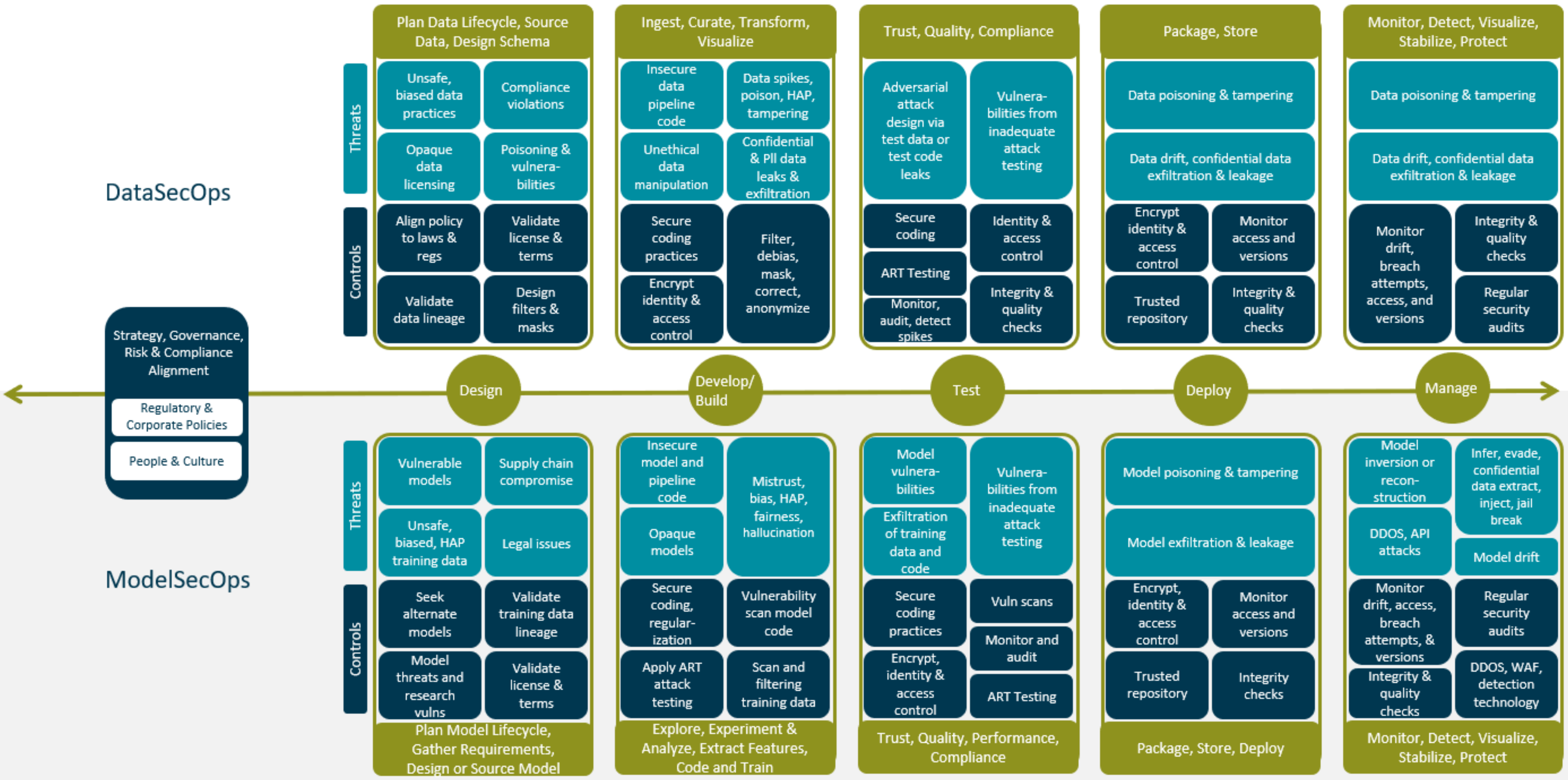
As AI systems become integral to business operations, ensuring their security and trustworthiness is paramount. For this purpose, there are three key areas to address, in a **Pre-deployment phase** of an AI application:

(1) **Data Provenance and Integrity**: Preventing Data Poisoning is foundational. This requires implementing robust data validation, sanitization, and anomaly detection mechanisms and continually assessing data integrity. The entire history of data transformations (lineage) must be tracked and recorded in a non-modifiable, tamper-evident way. Organizations must catalog provenance for all datasets and AI models.

(2) **Model Assurance**: Models, particularly large ones, are opaque (black box) and difficult to inspect. To protect this high-value asset, developers must implement model integrity checks (such as hashing and digital signatures) to prevent unauthorized tampering or modification of model artifacts. This is supported by maintaining an AI Bill of Materials (AIBOM), an inventory of every dataset, dependency, and hardware accelerator used, which aids in governance and vulnerability tracking.

(3) **Supply Chain Risk Management**: Organizations must vet all third-party/open-source components and dependencies thoroughly before use, reviewing licensing and compliance requirements. Governance frameworks, such as the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework (NIST AI RMF), explicitly require policies and procedures to address AI risks arising from third-party software and data and other supply chain issues.

An Illustrative MLSecOps framework



Strategy, Governance, Risk & Compliance Alignment

Regulatory & Corporate Policies

People & Culture

DataSecOps

	Design	Develop/ Build	Test	Deploy	Manage
	Plan Data Lifecycle, Source Data, Design Schema	Ingest, Curate, Transform, Visualize	Trust, Quality, Compliance	Package, Store	Monitor, Detect, Visualize, Stabilize, Protect
Threats	Unsafe, biased data practices Compliance violations Opaque data licensing Poisoning & vulnerabilities	Insecure data pipeline code Data spikes, poison, HAP, tampering Unethical data manipulation Confidential & PII data leaks & exfiltration	Adversarial attack design via test data or test code leaks Vulnerabilities from inadequate attack testing	Data poisoning & tampering Data drift, confidential data exfiltration & leakage	Data poisoning & tampering Data drift, confidential data exfiltration & leakage
Controls	Align policy to laws & regs Validate license & terms Validate data lineage Design filters & masks	Secure coding practices Filter, debias, mask, correct, anonymize Encrypt identity & access control	Secure coding Identity & access control ART Testing Integrity & quality checks Monitor, audit, detect spikes	Encrypt identity & access control Monitor access and versions Trusted repository Integrity & quality checks	Monitor drift, breach attempts, access, and versions Integrity & quality checks Regular security audits

ModelSecOps

	Design	Develop/ Build	Test	Deploy	Manage
Threats	Vulnerable models Supply chain compromise Unsafe, biased, HAP training data Legal issues	Insecure model and pipeline code Mistrust, bias, HAP, fairness, hallucination Opaque models	Model vulnerabilities Vulnerabilities from inadequate attack testing Exfiltration of training data and code	Model poisoning & tampering Model exfiltration & leakage	Model inversion or reconstruction Infer, evade, confidential data extract, inject, jail break DDOS, API attacks Model drift
Controls	Seek alternate models Validate training data lineage Model threats and research vulns Validate license & terms	Secure coding, regularization Vulnerability scan model code Apply ART attack testing Scan and filtering training data	Secure coding practices Vuln scans Monitor and audit Encrypt, identity & access control ART Testing	Encrypt, identity & access control Monitor access and versions Trusted repository Integrity checks	Monitor drift, access, breach attempts, & versions Regular security audits Integrity & quality checks DDOS, WAF, detection technology
	Plan Model Lifecycle, Gather Requirements, Design or Source Model	Explore, Experiment & Analyze, Extract Features, Code and Train	Trust, Quality, Performance, Compliance	Package, Store, Deploy	Monitor, Detect, Visualize, Stabilize, Protect

Figure 7: MLSecOps Framework

The diagram (Figure 7) shows how MLSecOps secures the entire AI/ML lifecycle by combining DataOps and MLOps into a unified security framework. At each stage—data sourcing, ingestion, model training, testing, deployment, and monitoring—it maps common threats to the controls needed to mitigate them.

- **Stage 1 - Data planning & ingestion:** Prevents issues like biased or tampered data using steps such as verifying data licenses and hashing incoming datasets to detect manipulation.
- **Stage 2 - Model development:** Addresses risks like adversarial manipulation through secure coding practices and automated robustness tests (e.g., checking whether small perturbations can fool the model).
- **Stage 3 - Deployment:** Uses trusted registries and signed model artifacts to prevent version tampering or model theft.
- **Stage 4 - Monitoring:** Detects threats such as model inversion or drift by monitoring API behavior and anomaly patterns.

Overall, the framework ensures AI systems remain secure, trustworthy, and resilient through every phase of their lifecycle.

B. Runtime Robustness and Defense Architecture

And once the application is nearing deployment or is in runtime (production), the focus should shift to securing deployed AI systems through zero-trust access, guardrails for safe outputs, and real-time monitoring.

(1) **Zero Trust Architecture (ZTA):** ZTA is the recommended approach, shifting away from network-centric defenses to an asset-centric and data-centric approach. ZTA ensures continuous authentication and dictates that the LLM or agent must operate with least-privilege access (Just-in-Time (JIT)/ Just Enough Administration (JEA)), thereby limiting the potential damage from prompt injection or similar exploits.

(2) **AI-Specific Guardrails**: Because LLMs are non-deterministic, security must be baked into the runtime using purpose-built guardrails. These tools (like LlamaFirewall or NeMo-GuardRails) enforce policy on outputs, sanitize inputs, manage output filtering, and prevent autonomous decisions that violate compliance rules.

(3) **Continuous Monitoring and Incident Response**: Effective defense requires real-time monitoring of AI system behavior. This includes tracking security alerts, logging critical actions, and detecting anomalies or model drift to ensure the model's functionality and trustworthiness remain consistent while in production. Formal AI Security Incident Management Processes must be established and align with enterprise response plans.

6.3 Privacy by Design: Protecting Sensitive Data and Confidentiality

While adopting "Security by default" principles into the design of an AI system, a very important goal for the system design and architecture team, is privacy. AI systems can inadvertently expose PII, confidential business data, or IP through prompts, outputs, or compromised supply chains. Embedding privacy controls from the start reduces these risks and strengthens trust. And as we have covered in the previous sections on Regulations, Global regulations such as GDPR, California Consumer Privacy Act (CCPA), and sector-specific mandates like HIPAA and EU AI Act require organizations to implement privacy-preserving measures by design.

We recommend that Enterprises adopt a "Privacy by design" approach to achieve these goals. That is, to incorporate privacy measures during the design stage of AI systems and applications. Core measures while adopting this approach, could include:

- **Data Minimization:** Collect only necessary data; apply anonymization and pseudonymization.
- **Differential Privacy:** Use privacy-preserving techniques during training and inference.
- **Access Controls:** Enforce Role Based Access Control (RBAC)/ Attribute Based Access Control (ABAC) and least privilege for sensitive data.
- **Encryption Everywhere:** Protect data in transit and at rest; manage keys securely.
- **Compliance Integration:** Bake regulatory requirements into architecture and monitoring workflows.

This approach complements the aspects we covered earlier in **Integrity & Supply Chain Assurance** by ensuring provenance without exposing sensitive data and strengthens **Runtime Robustness** by enforcing privacy guardrails during inference and tool orchestration.

6.4 Protecting AI systems – Architecture Pattern

This reference architecture shown in *Figure 8* illustrates how to secure an agentic AI environment by applying the "Security by default" principles we covered in the previous section —i.e., from user entry to tool execution and enterprise data access—using security-by-default controls, zero-trust principles, and continuous monitoring. The design places enforcement points along the entire interaction path: identity and access management, guarded orchestration, protected model runtime, policy-controlled tool use, and data leakage prevention at enterprise boundaries.

End-to-End Flow & Security Control Points

- **User Entry:** Identity and Access Management (IAM) authenticates users; Guard Monitoring inspects prompts for policy compliance.
- **Agent Orchestration:** Enforces least privilege, validates requests, and applies output filtering at every hop.
- **LLM Runtime:** Hardened environment with guardrails for prompt sanitization and safe outputs.

- **Configuration & Version Control:** Centralized policy management and auditable, signed artifacts.
- **Tool Zone:** Controlled tool invocation with per-tool permissions and monitoring.
- **Enterprise Access:** Agent Access Control and Data Leakage Monitoring protect APIs, apps, and sensitive data.
- **Operations:** Unified telemetry, anomaly detection, alerts, and automated incident response.

To summarize, AI security and privacy are foundational imperatives for secure and trustworthy AI systems, and these systems need a shift to "Secure by default”. Securing AI systems requires addressing technical risks through lifecycle controls—integrity and supply chain assurance, runtime defense, and privacy by design. These measures, driven by global regulations, are reinforced by a zero-trust reference architecture for agentic systems, ensuring resilience, compliance, and trust across data, models, and enterprise integrations.

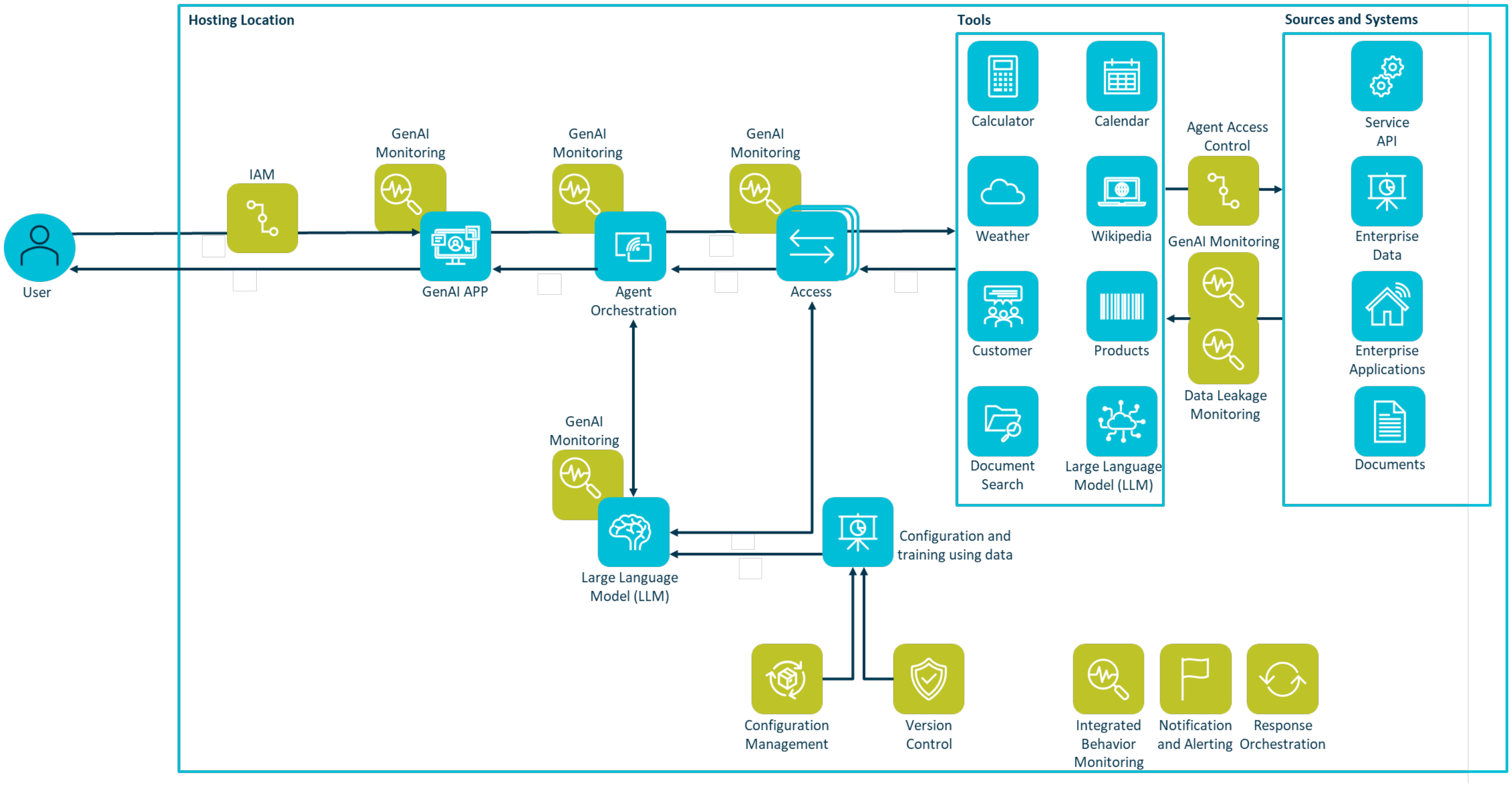


Figure 8: Agentic AI architecture and placement of security components

7. Use Case: AI as a Tool to Enhance Cybersecurity

Beyond the risks, AI can also help implement Security by Default. AI is transforming cybersecurity by enabling automation, rapid threat detection, and smarter incident response. Yet, these advances also introduce new vulnerabilities and ethical concerns. To fully realize AI's potential while minimizing risks, it is essential to embed Security by Default into AI systems for cybersecurity—ensuring that strong safeguards are active from the outset and maintained throughout the AI system's lifecycle. This approach makes security an inherent property of every AI-driven solution, not an afterthought. By combining robust governance with Security by Default, organizations can harness AI's strengths to protect their assets, uphold ethical standards, and build lasting trust.

Table 3: AI usage

AREA	AI USAGE
Threat Detection and Prevention	
AI-Driven Threat Analysis	Rapidly analyzes large-scale data to identify patterns and anomalies indicating cyber threats through automated logfile analysis. This improves both the speed and precision of detection and supports the verification of risk assessments.
AI-Powered Endpoint Detection and Response (EDR)	Establishes baselines for normal activity (e.g., login patterns, process execution) and flags deviations such as lateral movement or misuse of admin tools. AI models continuously monitor endpoints, adapt in real time, reduce false positives, and promote secure behavior as the default user experience.
AI-Insider Threat Detection	Analyzes behavioral patterns to identify potential internal threats, monitoring for data exfiltration, unusual access, and sentiment in communications. Default monitoring policies reduce reliance on manual audits, and risky actions can be automatically escalated or blocked.
AI-Powered Email Threat Prevention	Analyzes headers, content, and sender reputation to block phishing attempts. Every inbound message is analyzed in real time, with links and attachments sandboxed and evaluated. This enhances email security independently of user awareness, minimizing risk by default
Automated Incident Response	AI-powered systems automate responses to cyber incidents (e.g., isolating compromised systems, blocking malicious traffic) and support decisions with real-time logfile analysis. This enables swift containment, reduces response time, and mitigates threats efficiently.

Risk Management, Governance & Compliance	Monitors systems, predicts vulnerabilities, provides mitigation recommendations, and supports self-assessments and compliance tasks. This leads to faster risk detection, fewer exploitable vulnerabilities, and reduced manual effort.
Proactive Defense	Identifies potential threats through anomaly and network detection, supporting red-teaming and penetration testing. This enables earlier threat discovery and proactive defense.
Human Factor	
Enhancing Cybersecurity with AI	Supports human decision-making with real-time insights and automated recommendations, prioritizes alerts, and complements human intuition in threat hunting. This results in faster detection and response, reduced cognitive burden, and more focused attention on critical threats.
AI-Driven Security Awareness Training	Personalizes security training based on employee role, behavior, or past actions. This improves security awareness across the organization, including non-technical staff, without requiring active user initiative.
AI-Based Phishing Simulations	Uses generative AI to create realistic, personalized phishing simulations and automatically generates campaigns. Employees develop instinctive resistance to phishing, and administrators save time on campaign creation.
AI-Enhanced IAM	Automates identity verification and access control, continuously monitors user behavior, and adapts access policies in real time. This improves accuracy, prevents unauthorized access, and enhances protection against insider threats.
Adaptive Access Control (key capabilities)	Enforces least privilege, triggers adaptive authentication for high-risk logins, and applies conditional access rules based on AI-generated risk scores. This enhances security and reduces administrative effort.
AI-Powered Role Mining and Access Reviews	Analyzes access patterns to identify over-provisioned users, recommends least-privilege roles, and generates risk-based access reviews. This reduces excessive permissions and improves security hygiene.
Continuous Authentication via Behavioral Biometrics	Continuously verifies identity using typing patterns, mouse movements, and session behavior, reducing reliance on passwords or static Multi-factor authentication (MFA). This enhances security and protects against account compromise.

AI-Based Identity Risk Scoring	Evaluates risk in real time using factors like impossible travel and unusual access times, enabling adaptive access policies. This improves security through context-aware access control.
JIT Access with AI	Predicts the need for temporary elevated access, grants it when required, and revokes it automatically. This minimizes standing privileges and enhances security.
Secure Configuration Management via CSPM + AI	Continuously audits infrastructure and cloud assets, detects insecure defaults, and enforces secure configurations. This ensures consistent security and reduces manual compliance effort.
Enhanced Patch Management & Vulnerability Prioritization	Assesses vulnerabilities contextually, prioritizes them, and recommends or deploys critical patches. This enables faster, risk-based patching and reduces exposure to vulnerabilities
Automated Security Policy Enforcement via AI (e.g. SOAR)	Automates actions like account lockouts or network quarantines based on detected threats. This speeds up threat response and reduces human error.
Data Loss Prevention (DLP) with AI	Automatically classifies sensitive data and applies policies to block or encrypt it based on context. This ensures consistent protection of sensitive data and reduces manual effort.

8. Conclusion

In conclusion, the rapid adoption of AI presents both significant opportunities and substantial risks. As AI becomes integral to innovation and efficiency across industries, it is essential to prioritize security and privacy throughout the AI lifecycle. The "Security by Default" approach is critical to ensuring AI systems are secure, trustworthy, and resilient to emerging threats.

To achieve this, security and privacy must be embedded into AI governance from the outset, guiding design, deployment, operation, and corporate education. Effective governance frameworks, alongside regulatory measures like the EU AI Act, provide accountability and transparency. It is highly recommendable for organizations to closely follow and adapt to the regulative evolution in this area –not only in the legal departments, but in an overarching capacity. Addressing ethical concerns, including fairness and privacy, is also vital to responsible AI deployment.

Security risks—such as data poisoning, model theft, and malicious AI-generated content—require specialized defenses. AI systems must integrate protective measures like data integrity checks, zero-trust architectures, and continuous monitoring. Privacy-enhancing technologies, such as differential privacy and federated learning, are essential for minimizing data exposure and ensuring compliance with regulations like GDPR.

As AI systems create unique security and privacy risks, organizations must embed Security by Default and privacy by design across the entire AI lifecycle. This requires structured controls for integrity and supply chain assurance (e.g., data provenance, model assurance, third-party risk), runtime robustness and zero-trust defenses (least privilege, guardrails, continuous monitoring), and privacy measures (data minimization, encryption, access control, differential privacy, and built-in compliance). A zero-trust, end-to-end reference architecture with enforcement points from user entry to enterprise data access helps ensure AI systems remain resilient, compliant, and trustworthy in real-world use.

AI’s role in cybersecurity further highlights its dual function as both a tool for securing systems and a potential target. By automating threat detection, incident response, and risk management, AI can help organizations stay ahead of cyber threats, making security a core component of operations. Ultimately, the secure, ethical deployment of AI demands a proactive, holistic approach—one that places security, privacy, and governance at the forefront of every stage of the AI lifecycle. "Security by Default" is not just a best practice; it’s a fundamental standard for ensuring AI systems remain safe, ethical, and trustworthy as they evolve.

9. References and Additional Resources

Publications of national & supranational institutions

Canada

Parliament of Canada (2022): “Digital Charter Implementation Act”, 44th Parliament, 1st session. Bill C-27. URL: https://www.parl.ca/LegisInfo/en/bill/44-1/c-27

China

Cyberspace Administration China (2023): “Interim Measures for the Management of Generative Artificial Intelligence Services”, Office of the Central Cyberspace Affairs Commission. URL: https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

Cyberspace Administration China (2025): “Notice on Issuing the Measures for the Identification of Artificial Intelligence Generated Synthetic Content”, Office of the Central Cyberspace Affairs Commission. URL: https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

The State Council (2024): „China establishes technical committee to further AI standardization”., The Republic of China. URL: https://english.www.gov.cn/news/202412/31/content_WS67736228c6d0868f4e8ee657.html

EU

European Commission (2019): “Ethics and Guidelines for Trustworthy AI”, Independent High-Level Expert Group on Artificial Intelligence. Set up by the European Commission. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission (2025): “AI Act“, Shaping Europe’s digital future. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Official Journal of the European Union (2024): „Regulation (EU) 2024/1689 of the European Parliament and of the Council“. 13.06.2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

Japan

Japan (2025): “Outline of the Act on Promotion of Research and Development, and Utilization of AI-related Technology (AI Act)”. URL: https://www8.cao.go.jp/cstp/ai/ai_hou_gaiyou_en.pdf

NATO

NATO (2024): “Summary of NATO's revised Artificial Intelligence (AI) strategy”. 10.07.2024. URL: https://www.nato.int/cps/en/natohq/official_texts_227237.htm

South Korea

Ministry of Science and ICT (2024): “A new chapter in the Age of AI: Basic Act on AI passed at the National Assembly’s Plenary Session”. Press release. URL:

https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=1071&searchOpt=ALL&searchTxt=

United Kingdom

NCSC (2024): "Machine learning principles", NCSC as part of the Government Communications Headquarters. URL: https://www.ncsc.gov.uk/files/NCSC-Machine-learning-principles.pdf

UK Parliament (2025): "Artificial Intelligence (Regulation) Bill [HL]. Originated in the House of Lords, Session 2024-26", Progress site. URL: https://bills.parliament.uk/bills/3942

United States

California AI Bill SB 53 (2025): "Artificial intelligence models: large developers". URL: https://legiscan.com/CA/text/SB53/id/3271094

Department of Defense (2023): "Data, Analytics, and Artificial Intelligence Adoption Strategy. Accelerating Decision Advantage", Office of Prepublication and Security Review. Department of Defense. United States of America. URL: https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_STRATEGY.PDF

Federal Communications Commission (2025): "U.S. Cyber Trust Mark". URL: https://www.fcc.gov/CyberTrustMark

NIST (2023): "Artificial Intelligence Risk Management Framework (AI RMF 1.0)", National Institute of Standards and Technology. U.S. Department of Commerce. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

NIST (2025): "AI Risk Management Framework", National Institute of Standards and Technology. U.S. Department of Commerce. URL: https://airc.nist.gov/airmf-resources/airmf/

Schwartz, R./ Vassilev, A./ Greene, K./ Perine, L./ Burt, A. (2022): "NIST Special Publication 1270. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.", National Institute of Standards and Technology. U.S. Department of Commerce. URL: https://nvlpubs.nist.gov/NISTpubs/SpecialPublications/NIST.SP.1270.pdf

The White House (2025): "Blueprint for an AI Bill of Rights. Making Automated Systems Work for the American", Office of Science and Technology Policy. Biden White House Archives. URL: https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/

US AI Training Act (2021): "Artificial Intelligence Training for the Acquisition Workforce Act or the AI Training Act". No.117-207. URL: https://www.congress.gov/bill/117th-congress/senate-bill/2551

Joint Publications

ASD’s ACSC (2024): “Engaging with Artificial Intelligence (AI)”, Australian Signals Directorate’s Australian Cyber Security Centre in collaboration with international partners. URL: https://www.cyber.gov.au/sites/default/files/2025-03/Engaging%20with%20artificial%20intelligence%20%28January%202024%29.pdf

Publications of multilateral Organizations

AIGN OS (2025): “AI Governance Trust Label”. URL: https://aign.global/aign-os-the-operating-system-for-responsible-ai-governance/ai-governance-trust-tools-label/

IEEE (2025): “AI Trust Alliance”, jointly operated by IEEE, POSITIVEAI, SystemX and VDE. URL: https://www.trustalliance.ai/

International Organization for Standardization on AI

International Organization for Standardization (2020): “Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence”. ISO/IEC TR 24028.

International Organization for Standardization (2022a): “Information technology — Artificial intelligence — Artificial intelligence concepts and terminology”. ISO/IEC 22989.

International Organization for Standardization (2022b): “Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)”. ISO/IEC 23053.

International Organization for Standardization (2022c): “Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations”. ISO/IEC 38507.

International Organization for Standardization (2023a): “Information technology — Artificial intelligence — Management system”. ISO/IEC 42001.

International Organization for Standardization (2023b): “Information technology — Artificial intelligence (AI) — AI system impact assessment”. ISO/IEC 42005.

International Organization for Standardization (2023c): “Information technology — Artificial intelligence — Guidance on risk management”. ISO/IEC 23894.

International Organization for Standardization (2023d): “Information technology — Artificial intelligence — AI system life cycle processes”. ISO/IEC 5338.

International Organization for Standardization (2023e): “Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems”. ISO/IEC 25059.

International Organization for Standardization (2025): “Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems”. ISO/IEC 42006.

TIC Council

TIC Council (2024): “Trustworthiness of Artificial Intelligence. Recommendations of the TIC sector”, Testing, Inspection and Certification Council. White Paper. URL: https://www.tic-council.org/application/files/9117/1326/9285/White_Paper_on_Trustworthiness_of_Artificial_Intelligence.pdf

OECD

OECD (2021): "Tools for Trustworthy AI. A framework to compare implementation tools for Trustworthy AI Systems", OECD Digital Economy Papers No. 312. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2021/06/tools-for-trustworthy-ai_0e36bb08/008232ec-en.pdf

OECD (2025): "Policies, data and analysis for trustworthy artificial intelligence" URL: https://oecd.ai/en/

OECD (2025): "Accountability (Principle 1.5)", OECD.AI. Policy Observatory. GPAI Global Partnership. URL: https://oecd.ai/en/dashboards/ai-principles/P9

OWASP

OWASP (2025): "OWASP Top 10 for Large Language Model Applications", Open Web Application Security Project. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/

Publications of companies

CoT partner

BSI/ TUV Verband/ Frauenhofer HHI (2021): "Towards Auditable AI Systems Current status and future directions", Federal Office for Information Secruity White Paper. URL: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf?__blob=publicationFile&v=6

IBM (2025): "What is trustworthy AI?". https://www.ibm.com/think/topics/trustworthy-ai

TÜV-Verband (2021): "KI-Studie 2021: Sicherheit und Künstliche Intelligenz", TÜV-Verband e. V. URL: https://www.tuev-verband.de/studien/ki-studie-2021

TÜV AI Lab (2025): "TÜV AI Lab". URL: https://www.tuev-lab.ai/

VDE SPEC (2022): "VCIO based description of systems for AI trustworthiness characterisation", with contributions by Bosch, Siemens and TÜV SÜD. URL: https://www.vde.com/resource/blob/2242194/a24b13db01773747e6b7bba4ce20ea60/vcio-based-description-of-systems-for-ai-trustworthiness-characterisationvde-spec-90012-v1-0--en--data.pdf

Other companies

PWC (2025): „The EU AI Act. Compliance and transformation". URL: https://cee.pwc.com/eu-ai-act-compliance-and-transformation.html

Deloitte (2024): "Trustworthy AI". URL: https://www.deloitte.com/de/de/issues/innovation-ai/trustworthy-ai.html

Publications of Universities and Think Tanks

Díaz-Rodríguez, N./ Del Ser, J., Coeckelbergh/ M., De Prado, M. L./ Herrera-Viedma, E., & Herrera, F. (2023): “Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation”. *Information Fusion*, *99*, 101896. URL: https://www.sciencedirect.com/science/article/pii/S1566253523002129

Frauenhofer IKS (2025): “Trustworthy AI”. URL: https://www.iks.fraunhofer.de/en/research/trustworthy-ai.html

Gillespie, N./ Lockey, S./ Ward, T./ Macdade, A. & Hassed, G. (2025). „Trust, attitudes and use of artificial intelligence: A global study 2025“, The University of Melbourne and KPMG. URL: https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2025/05/trust-attitudes-and-use-of-ai-global-report.pdf

Kaushik, A. (2025): “Leveraging Artificial Intelligence for NATO’s cyber resilience: Preliminary perspectives”, GLOBSEC. 21.02.2025. URL: https://www.globsec.org/what-we-do/publications/leveraging-artificial-intelligence-natos-cyber-resilience-preliminary.

Lahusen, C. / Maggetti, M. /Slavkovik, M. (2024): “Trust, trustworthiness and AI governance”. *ci Rep* **14**, 20752. URL: https://www.nature.com/articles/s41598-024-71761-0

Newman, J. (2023): “A Taxonomy of Trustworthiness for Artificial Intelligence”, UC Berkely & Center for Long-Term Cybersecurity. URL: https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/

Oxford Internet Institute & University of Oyford (2024): “Trustworthiness Auditing for AI”. URL: https://www.oii.ox.ac.uk/research/projects/trustworthiness-auditing-for-ai/#publications

Reynolds, I./Atalan, Y. (2024): “Calibrating NATO’s Vision of AI-Enabled Decision Support”, CSIS. 08.07.2024. URL: https://www.csis.org/analysis/calibrating-natos-vision-ai-enabled-decision-support

10. Acronyms & Glossary

Technical Terms

Acronym	Expansion	Definition
ABAC	Attribute Based Access Control	Authorization model that evaluates attributes (or characteristics), rather than roles, to determine access.
AI	Artificial Intelligence	Technology that enables machines to mimic human intelligence tasks.
AIBOM	AI Bill of Materials	Complete inventory of all the assets in your organization's AI ecosystem.
AIGN OS	Operating System for Responsible AI Governance	Structured, certifiable governance architecture designed to help organizations turn AI- and data-governance principles into measurable, operational practice.
CE	Conformité Européenne	CE Marking is a label indicating that a product complies with applicable EU regulations regarding safety, health, environmental protection, and energy efficiency.
CCPA	California Consumer Privacy Act	Data privacy law that gives California residents rights over their personal information.
CSPM	Cloud Security Posture Management	Set of tools and practices designed to continuously monitor and improve the security of cloud environments.
DevOps	Development and Operations	Culture and set of practices that aim to improve collaboration between software developers and IT operations teams.
DLP	Data Lost Prevention	Set of tools and processes designed to prevent sensitive information from being lost, misused, or accessed by unauthorized users.
EDR	Endpoint Detection and Response	Cybersecurity solutions that collect and analyze data from endpoints to identify suspicious activity, provide

		real-time threat detection, and enable quick investigation and remediation of security incidents.
GenAI	Generative AI	AI systems that create new content such as text, images, or code.
GDPR	General Data Protection Regulation	European Union law that governs how organizations collect, process, and store personal data of EU residents.
GPAI	General Purpose AI	AI systems designed to perform a wide range of tasks across different domains, rather than being specialized for a single application.
GPSR	General Product Safety Regulation	EU regulation that establishes a modernized framework for the safety of consumer products placed on the EU market.
GPT	Generative Pre-trained Transformer	**Hack GPT** refers to AI tools used to develop or automate malicious hacking activities, whereas **Pentest GPT** denotes AI-assisted tools used to support authorized, ethical penetration testing.
HIPAA	Health Insurance Portability and Accountability Act	U.S. law that sets standards for the protection and confidential handling of patients' medical information.
IaC	Infrastructure as Code	Practice in IT and DevOps where infrastructure (servers, networks, databases, etc.) is defined and managed using code rather than manual processes.
IAM	Identity and Access Management	Framework of policies, technologies, and processes that ensures the right individuals have the appropriate access to resources within an organization.
IP	Intellectual Property	Refers to creations of the mind that are **legally protected** to give the creator exclusive rights to use, sell, or license them (such as inventions, literary and artistic works, designs, symbols, names, and images).
JEA	Just Enough Administration	Granting only the minimum required privileges to perform specific tasks.

JIT	Just-in-Time	JIT Access is security practice where users are granted elevated privileges only for the exact time they need them, rather than permanently.
LLM	Large Language Model	AI models trained on vast text data to generate human-like language.
MASs	Multi-Agent Systems	Systems composed of multiple interacting intelligent agents that work together (or compete) to solve complex problems that are difficult for a single agent to handle.
MFA	Multi-factor authentication	Electronic authentication layer requiring two or more credentials to verify identity.
ML	Machine Learning	A subset of AI focused on systems that learn and improve from data.
MLSecOps	Machine Learning Security Operations	Practice of integrating security into the full machine learning lifecycle.
NHI	Non-Human Identities	Machine, application, service account or other non-human entity that needs authentication and access to systems or data.
NIST	National Institute of Standards and Technology	U.S. government agency developing technology and cybersecurity standards.
NIST AI RMF	NIST Artificial Intelligence Risk Management Framework	Guideline developed to help organizations identify, assess, manage, and mitigate risks associated with AI systems.
PaC	Policy as Code	Practice where organizational policies are defined, managed, and enforced using code.
PLD	Programmable Logic Device	Electronic component used to implement digital logic circuits that can be programmed by the user after manufacturing.
PII	Personally Identifiable Information	Set of data that could be used to distinguish a specific individual.

RBAC	Role Based Access Control	Model for authorizing end-user access to systems, applications and data based on a user's predefined role.
RCE	Remote Code Execution	Type of cybersecurity vulnerability that allows an attacker to run arbitrary code on a remote system without authorization.
SDLC	Secure Development Lifecycle	Structured, step-by-step process used by development teams to design, build, test, and deploy high-quality software efficiently
SOAR	Security Orchestration, Automation, and Response	Cybersecurity approach and platform that integrates tools, automates workflows, and coordinates responses to security incidents.
XSS	Cross-Site Scripting	Web security vulnerability that allows an attacker to inject malicious scripts into a trusted website, which then execute in the browsers of visitors.
ZTA	Zero Trust Architecture	Cybersecurity model that assumes no user or device should be trusted by default. Access to resources is granted only after continuous verification.

Organizations and Institutions

Acronym	Expansion
AIGN	The Operating System for AI Governance. Website
CoT	Charter of Trust. Website
EU	European Union. Website
VDE	German Association for Electrical, Electronic & Information Technologies. Website
IEEE	Institute of Electrical and Electronics Engineers. Website
OECD	Organisation for Economic Co-operation and Development. Website

Main Contributors

- **Agora Strategy Group AG**
- **ATOS SE**
- **Danfoss**
- **IBM Corporation**
- **Mitsubishi Heavy Industries, Ltd.**
- **Robert Bosch GmbH**
- **Siemens AG**
- **TÜV SÜD AG**

The Charter of Trust

Protecting the digital world of tomorrow

About the Charter of Trust

The Charter of Trust is a non-profit alliance of leading global companies and organizations working across sectors to make the digital world of tomorrow a safer place. It was founded in 2018 at the Munich Security Conference to enhance cybersecurity efforts and foster digital trust in the face of an increasingly complex and severe cyber threat landscape.

A unique initiative underpinned by 10 principles fundamental to a secure digital world, the Charter of Trust is working to protect our increasingly digitized world and build a reliable foundation on which trust and digital innovation can flourish. It contributes to the development of effective cybersecurity policies that strengthen global cybersecurity posture and provides expertise on topics including AI, Security by Default, supply chain security, and education.

Objectives

The Charter of Trust seeks to harmonize cybersecurity approaches and address cybersecurity challenges from a holistic, ethical and fair perspective. The alliance is collaborating across industries to cultivate, advocate, and enhance global cybersecurity standards. By fostering widespread awareness and sharing expertise, it ensures a cohesive approach to security that enables seamless global interoperability.

Key principles

The work of the Charter of Trust is underpinned by 10 principles fundamental to a secure digital world:

1. Education
2. Cyber-resilience through conformity and certification
3. Transparency and response
4. Regulatory framework
5. Joint initiatives
6. Ownership for cyber and IT security
7. Responsibility throughout the digital supply chain
8. Security by Default
9. User-centricity
10. Innovation and co-creation

Contact

Point of contact: contact@charteroftrust.info