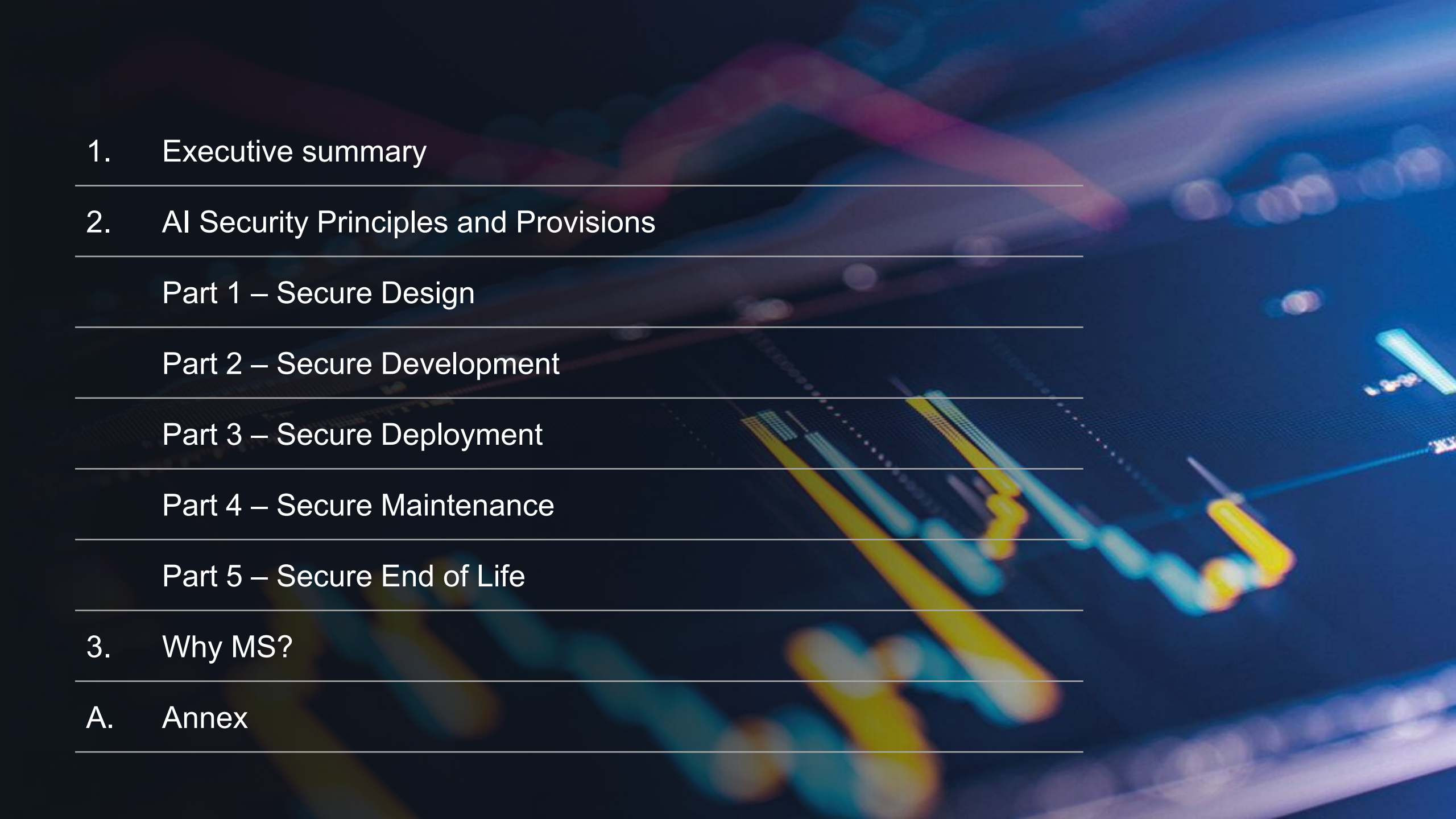
Technical note on

ETSI EN 304 223

Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems



1 Executive summary
General overview

ETSI EN 304 223 defines baseline AI security requirements and principles for Developers, System Operators, and Data Custodians, covering Secure Design, Development, Deployment, Maintenance, and End-of-Life management of AI systems

Background

- The **rapid evolution of AI technologies** and the **associated increase in cyber threats** underscore the need for **cybersecurity standards** to ensure **consistent protection** and **regulatory alignment** across the **EU**.
- **ETSI published** the first globally applicable **European Standard** providing **baseline cybersecurity requirements** for **AI models and systems**.
- The standard sets out a **lifecycle-based framework** with **13 principles** covering **Secure Design**, **Development**, **Deployment**, **Maintenance**, and **End-of-Life**, and is intended to help **stakeholders across the AI supply chain** strengthen **AI security** against emerging threats such as **data poisoning** and **model tampering** (See Annex).

Scope

- Applies to **Developers**, **System Operators**, **Data Custodians**, **End-users**, and **Affected Entities** (See Annex).
- Covers **creation**, **deployment**, **operation**, **use**, and **impact** of **AI systems**, including **open-source** and **proprietary models**.

Next steps

- **31/03/2026**. Deadline for all National Standards Organizations (NSOs) to announce publicly that ETSI EN 304 223 exists.
- **30/09/2026**. Latest date by which National Standards Organizations must publish or formally endorse ETSI EN 304 223.
- **30/09/2026**. Date of withdrawal of any conflicting National Standard

AI Security Principles and Provisions

Secure Design	For **developers**, **system operators**, and **data custodians** and includes provisions such as **AI security training**, **security risk assessment and mitigation**, **documentation of system design and data use**
Secure Development	For **developers**, **system operators**, and **data custodians** and includes provisions such as maintaining **assets inventories**, **access controls**, **monitoring secure processes**, documenting **audit trails**, and **conducting tests and evaluations of AI models**
Secure Deployment	Sets communication requirements for **end users** and **affected entities**, including **clear guidance**, **data-use transparency**, **disclosure of limitations**, **timely security updates** and requires **documented** and **contractually processes**
Secure Maintenance	Requires **developers** and **system operators** to issue **security patches**, test major updates as **new models**, support **change management**, and monitor **logs**, **performance,** and **anomalies** to ensure **ongoing system security**
Secure End of Life	Obliges **developers** and **system operators**, to **securely delete all assets** and **configurations** during **model transfer** or **decommissioning** to prevent **residual security risks**

Access to Document

Developers | System Operators | Data Custodians | End-Users | Affected Entities

2 AI Security principles and provisions

Secure Design

The Secure Design section, composed of four principles, establishes the foundational security measures that organizations must follow to ensure AI systems are safe, reliable, and overseen by trained personnel. It requires organizations to raise awareness, design securely, manage risks continuously, and embed human responsibility throughout the AI lifecycle

Principle 1: Raising awareness of AI security threats and risks

- Integrate regularly updated **AI-specific security content** into organizational **cybersecurity training**, tailored to each role's **responsibilities**.
- Ensure all staff stay informed about emerging **AI-related threats**, **vulnerabilities**, and available **mitigations** through **multi-channel communication**.
- Provide developers with specialized training in **secure AI coding**, **system design**, and techniques to prevent **vulnerabilities** in **models**, **algorithms**, and supporting **software**.

Principle 2: Design the AI system for security as well as functionality and performance

- Conduct and document **requirements**, **AI security risks**, and **mitigation strategies**, involving **Data Custodians** where relevant.
- Design systems resilient to **adversarial attacks**, **unexpected inputs**, and failures, supported by clear **audit trails** for models, datasets, and prompts.
- Perform **AI-specific risk assessments** and **due diligence** for external components/providers, ensuring proper **data sensitivity controls** and **minimal permissions** when interacting with other systems.

Principle 3: Evaluate the threats and manage the risks to the AI system

- Perform continuous **threat modelling** and **risk management**, addressing **AI-specific attacks** and risks from new **configurations** or unnecessary **model functionalities**.
- Apply appropriate **controls** based on **risk tolerance and** communicate unresolved **AI threats** to **System Operators** and **End-users** with clear **impacts** and **recommended actions**.
- Maintain ongoing **monitoring**, ensure external parties can manage **AI security risks**, and recognize that **residual risk** remains even after mitigation.

Principle 4: Enable human responsibility for AI systems

- Ensure **human oversight** is built into AI system design, enabling humans to easily **assess**, **interpret**, and **explain** model outputs.
- Implement and maintain **technical measures** that support oversight as a **risk control and** verify that **Data Custodian–specified security controls** are properly integrated.
- Inform **End-users** about **prohibited use cases** to prevent misuse and reinforce safe, compliant system operation.

2 AI Security principles and provisions

Secure Development

The Secure Development section, which is composed of five principles, establishes the key requirements that govern secure AI development. These principles guide organizations to safeguard assets, harden infrastructure, secure the supply chain, document systems, and conduct thorough testing

Principle 5: Identify, track and protect the assets

- Maintain a comprehensive inventory of all assets **and document their interdependencies** to ensure full visibility and traceability across the AI system.
- Use robust processes and tools to **track, authenticate, and implement version control** for AI-specific assets throughout their lifecycle.
- Protect sensitive data by applying **rigorous validation checks, sanitization measures, and proportionate security controls** appropriate to the level of sensitivity.

Principle 6: Secure the infrastructure

- Strengthen **access control frameworks** for APIs, models, data, and pipelines to ensure that only authorized entities can interact with critical AI components.
- Apply robust **API security controls** to mitigate risks such as reverse engineering and model poisoning.
- Use **dedicated, isolated environments** to reduce the risk of unauthorized actions and lateral movement.

Principle 7: Secure the supply chain

- Follow **secure software supply chain practices** to ensure the integrity and trustworthiness of all components.
- **Justify the use** of any non-documented or unsecured components by conducting formal risk assessments and implementing appropriate mitigating controls.
- **Re-run evaluations** on released models to validate their ongoing security and performance and clearly communicate any changes to end-users.

Principle 8: Document data, models and prompts

- Maintain **comprehensive documentation** and a clear **audit trail** covering system design decisions, development activities, and ongoing maintenance actions.
- Include all **security-relevant information**, such as training data sources, known limitations, guardrails, and any constraints affecting safe operation.
- Release **cryptographic hashes** of model artefacts and document any public data acquisition, including the **source, URL, and date** of retrieval.

Principle 9: Conduct appropriate testing and evaluation

- Conduct **comprehensive security assessment testing** for models, applications, and systems prior to release to ensure they meet required security standards.
- Perform **pre-deployment testing**, preferably with **independent security testers**, to obtain an objective evaluation of potential vulnerabilities.
- Evaluate **model outputs** to identify and prevent risks such as reverse engineering, manipulation, or unintended influence on model behaviour.

Developers | System Operators | Data Custodians | End-Users | Affected Entities

2 AI Security principles and provisions

Secure Deployment, Secure Maintenance and Secure End of Life

These three sections — Secure Deployment, Secure Maintenance, and Secure End-of-Life — establish the requirements for deploying, maintaining and retiring AI systems in a secure and controlled manner. They safeguard data, preserve system integrity, and prevent security risks

Principle 10: Communication and processes associated with End-users and Affected Entities

- Convey to End-users **where and how their data is used, accessed, and stored**, including use for **model retraining** or **human review**.
- Provide **accessible guidance** on system **use, management, integration, and configuration**, highlighting **limitations** and **potential failure modes**, and proactively communicate **security-relevant updates**.
- Support **End-users and Affected Entities** during and after **cyber security incidents**, following a **documented** process agreed in contracts.

Principle 11: Maintain regular security updates, patches and mitigations

- Provide **security updates and patches**, notify System Operators, and ensure updates are delivered to End-users.
- Maintain **mechanisms and contingency plans** to **mitigate security risks** when updates cannot be provided.
- Treat major **AI system updates** as new model versions and perform **security testing and evaluation**.
- Support System Operators in **evaluating and responding to model changes**, including **preview access** via beta-testing or **versioned APIs**.

Principle 12: Monitor the system's behavior

- Log **system and user actions** to support **security compliance**, **incident investigations**, and **vulnerability remediation**.
- Analyze **logs** to detect **anomalies**, **security breaches**, **unexpected behaviour**, or issues such as **data drift** or **data poisoning**.
- Monitor **internal states** of AI systems where this improves the ability to address **security threats** or enables future **security analytics**.
- Monitor **model and system performance** over time to identify **sudden or gradual changes** that may affect security.

Principle 13: Ensure proper data and model disposal

- When transferring or sharing **training data** or a **model**, Developers and System Operators shall involve **Data Custodians** and **securely dispose** of the relevant assets to prevent security issues from propagating between AI system instantiations.
- When **decommissioning** a model and/or system, they shall involve **Data Custodians** and ensure the **secure deletion** of applicable data and configuration details.

3 Why Management Solutions?
Credentials

Management Solutions is experienced in reviewing and developing AI systems across all industries, while ensuring regulatory compliance and meeting supervisors' expectations

Proven Experience in Cybersecurity, ICT Risk and Technological Risk Projects	Value offer in cybersecurity and ICT/IT risk management, combining expertise across technological cybersecurity, ICT/IT risk assessment, and business to create a hybrid profile capable of addressing multiple high-demand areas: • **Governance, Risk and Compliance:** comprehensive risk assessments, definition of risk appetite and digital resilience strategies, organizational design and governance frameworks, development and updating of regulatory and internal policies, establishment of PMOs for ICT/IT risk and cybersecurity programs, support in audit, inspection and certification processes, and the definition and implementation of robust frameworks and methodologies (including scenario-based risk quantification and stress testing). • **Implementation of Controls:** design and deployment of training and awareness programs, crisis and incident management frameworks, access control models, ICT/IT and third-party risk management mechanisms, data security controls, and advisory support in the selection and implementation of appropriate services and technological solutions. • **Execution of Services and Operation of Controls:** operational support to Cybersecurity Offices (CISO/BISO support), ICT/IT 2nd Line of Defense (2LoD) functions, Third-Party ICT Risk Offices, and Digital Operational Resilience Offices, ensuring continuous oversight, monitoring, and enhancement of the control environment.
High-value profiles, expertise and cross vision	• Professionals with strong **understanding, communication, challenge/high-value advice, expertise in ICT risks and their relation to risk management and processes (cross vision).** • Understanding of **business processes**, knowledge of **risk management** methodologies **and strong analytical skills.** • **Detailed knowledge of market regulations, standards and best practices** (COBIT, ISOs, GDPR, NIST, ITIL, SANS, DORA, EBA GL, NIS2, CSF, CSA ...).
Benchmark firm with global capabilities	• **One global Firm, independent and international** (+50 countries), with in-depth knowledge of the businesses in which our clients operate (+2,000 global and local) selecting the most appropriate resources for each project, regardless of where they are located. • **Multidisciplinary team** with strong analytical skills and specialist knowledge. Organized on a matrix basis (customer, industry, competitor and geography). • Consultant accredited by **supervisors and supranational bodies** (ECB, FCA, PRA, BoE, BNH, BNG, BNS, BNM, SBIF, SBS, BCCR, SSN, EIOPA, etc.). • **A strong corporate culture**: commitment, dedication to service and a constant search for excellence. • **A proven track record**, which has resulted in significant organic growth (x50 in 21 years) **benchmarking** capability (presence at top clients in all geographies).
Regulatory and supervisory experience	• We assist supervisors in their on-site inspections in organizations. • We directly support organizations in overcoming on-site monitoring, internal and external audit processes in the area of ICT risks. • Office in Frankfurt as Hub for regulatory analysis and liaison with the supervisor.

A Annex I
Stakeholders

This annex identifies and defines the key stakeholder roles involved in the design, deployment, operation, and use of AI systems, in accordance with the stakeholder model and responsibilities set out in ETSI EN 304 223

Stakeholders	Definitions
Developers 	This encompasses any type of **business**, **organization**, or **individual** across any sector that is responsible for **creating or adapting an AI model and/or system**. This applies to **all AI technologies**, including **proprietary** and **open-source models**. For context, a business or organization that **creates an AI model** and is also responsible for **embedding or deploying** that model/system within its organization is defined in the present document as both a **Developer** and a **System Operator**. Developers can be **AI providers under the EU AI Act** when **putting a system into service** or **placing it on the market**.
System Operators 	This includes any type of **business or organization** across any sector that has responsibility for **embedding or deploying an AI model and system** within its **infrastructure** and/or for its **ongoing maintenance**. This applies to **all AI technologies**, including **proprietary** and **open-source models**. This term also includes businesses that provide a **contractual service** to organizations to **embed or deploy AI models and systems** for **business purposes**. System Operators can be **deployers under the EU AI Act** and may also be **AI providers** if they **make changes to the system**.
Data Custodians 	This includes any type of **business**, **organization**, or **individual** that **controls data permissions** and ensures the **integrity of data** used for an **AI model or system** to function. This stakeholder group also includes entities that define **policies for data use and management** for an AI model and/or system. In the context of an AI system, there may be **multiple Data Custodians**, as data used to create a model may originate from the **deploying organization**, **public databases**, or **other external sources**.
End-users 	This encompasses any **employee** within an organization or business, as well as **consumers**, who **use an AI model or system** for any purpose, including to support **work activities** and **day-to-day tasks**. This applies to **all AI technologies**, including both **proprietary** and **open-source models**. This stakeholder group is defined because the **voluntary Code** places expectations on **Developers**, **System Operators**, and **Data Custodians** to **inform and protect End-users**.
Affected Entities 	This encompasses all **individuals** and **technologies**, such as **applications** and **autonomous systems**, that are **not directly affected** by AI systems or by **decisions based on AI system outputs**. These individuals or entities do **not necessarily interact** with the deployed system or application.

Annex II
Key Terms and Abbreviations

Adversarial attack definition as well as common adversarial techniques used to exploit vulnerabilities in data, models, and input mechanisms of AI systems, as defined by ETSI

Adversarial Attack

Attempt to manipulate an AI model by introducing specially crafted inputs to cause the model to produce errors or unintended outcomes.

Data Poisoning

Type of adversarial attack where malicious data is introduced into training datasets to compromise the AI system's performance or behavior.

Model Tampering

Type of adversarial attack that involves the intentional, unauthorized modification of a machine learning model's internal components.

Meaning of the abbreviations used in this document

Abbreviation	Meaning
ETSI	European Telecommunications Standards Institute
AI	Artificial Intelligence
API	Application Programming Interface
URL	Uniform Resource Locator

Management Solutions
Making things happen

International *One Firm*
Multiscope Team
Best practice *know-how*
Proven Experience
Maximum Commitment

Jorge Monge Alonso
Partner at Management Solutions
jorge.monge.alonso@managementsolutions.com

Marta Hierro
R&D Partner at Management Solutions
marta.hierro@managementsolutions.com