RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

JOINT HIGH-LEVEL RISK ANALYSIS ON AI

Building trust in AI through a cyber risk-based approach

PARIS AI ACTION SUMMIT – TRUST IN AI – CYBERSECURITY

→

This document is published by **the French National Cybersecurity Authority (ANSSI)** and co-signed by the following partners :

- Canada (CCCS)
- Estoni (RIA)
- Finland (NCSC-FI)
- Germany (BSI)
- Greece (NCSA)
- India (CERT-IN)
- Ireland (NCSC-IE)
- Italy (ACN)
- Luxembourg (LHC)
- Malta (MDIA)
- Netherlands (AIVD and NCSC-NL)
- Norway (NSM)
- Poland (NASK)
- United-Kingdom (NCSC-UK)
- Singapour (CSA)
- Slovakia (NCSC-SK)
- Slovenia (URSIV)
- South Korea (NIS)

DISCLAIMER:

The information contained in this document is provided 'as is' and is only intended to contribute to discussions on the risks and opportunities of artificial intelligence. ANSSI and the authoring organisations cannot therefore be held responsible for any loss, injury or damage of any kind caused by its use. The information contained in this document does not constitute or imply the endorsement or recommendation by ANSSI and the authoring organisations of any third party entity, product or service. Links and references to third-party websites and documents are provided for information purposes only and do not imply endorsement or recommendation of these resources over others.

BUILDING TRUST IN AI THROUGH A CYBER RISK-BASED APPROACH

The international agencies and government authorities behind this document advocate for a risk-based approach to support trusted AI systems and for secure AI value chains, and call for the discussion to continue beyond the AI Summit, to equally address opportunities, risks and evolving cyber threat in the context of AI adoption.

AI, a transformative technology under development since the 1950s, now impacts almost every sector from defence to energy, health to finance and many others. Its rapid adoption, including the use of large language models (LLM)[1] and increasing reliance on AI, should encourage stakeholders to assess related risks, including the ones associated to cybersecurity.

Without adequate measures – and given that users still tend to underestimate AI-related cyber risks – malicious actors could exploit vulnerabilities of AI systems[2] and jeopardize the use of AI technology in the future. **It is therefore crucial to understand and mitigate these risks, to foster trusted AI development and fully embrace the opportunities that this technology offers.**

As software systems, AI systems are vulnerable and thus need to be secured by design, based in particular on existing good cybersecurity practices relating to development, deployment, incident management, software supply chain and vulnerability management including both proprietary and open source components. They face the same cyber threats as any other information system (IS), including through their hosting infrastructure, while their interconnection with other systems increases risks of lateralization. There are also AI-specific risks, especially concerning the central role of data in AI systems that poses unique challenges to confidentiality and integrity.

While existing cyber hygiene practices remain widely effective, the emergence of AI-enhanced cybersecurity solutions is an important asset in the face of new and evolving threats. Although they are outside the scope of this document, such solutions already contribute to reinforcing cybersecurity capabilities, and are expected to keep developing, both in terms of monitoring and intrusion detection, but also threat analysis, response automation and digital investigation, automation of security processes, and many others.

In the meantime, the rapidly evolving threat landscape makes it necessary to track malicious AI use, which is expected to grow increasingly sophisticated. AI already amplifies existing attack techniques, by lowering the required expertise to conduct such attacks and also enabling larger scale and efficiency. We see these effects across phishing and social engineering, vulnerability scanning and malicious code development. Advanced generative AI could enable large-scale, cost-effective attacks across the cyber-kill chain.

While the matter of AI-enhanced solutions, whether defensive or offensive, is already well addressed both in academic papers and in various frameworks currently being developed, this document focuses on the cybersecurity of AI systems. It aims to provide a high-level synthetic and comprehensive analysis of related cyber risks and to offer guidance to assess threats and implement adequate security measures building on the Guidelines for Secure AI Systems Development, developed in collaboration with over 20 international organizations and jointly released on November 2023[3].

1. LLM: Large Language Model, a generative AI model used for language processing and natural language generation.

2. An AI system is here defined as a software system that relies on an AI model built through statistical learning on a set of training data

3. Cybersecurity and Infrastructure Security Agency (CISA) and National Cyber Security Centre (NCSC), Secure AI Systems Development Guidelines, November 26, 2023.

CHALLENGES AND SCOPE OF AN AI RISK ANALYSIS

This risk analysis aims to consider not only the vulnerabilities of individual AI components, but also the security of broader AI systems integrating these components. Its purpose is to provide a wide overview of AI-related cyber risks rather than an exhaustive list of vulnerabilities. For further reading see the list of references in Appendix 2

The deployment of AI systems can open new paths of attack for malicious actors if adequate security measures are not implemented. Such a deployment should therefore include a dedicated risk analysis to assess the risks and identify appropriate security measures.

Additionally, understanding AI supply chains is essential to the mitigation of risks associated with the vulnerabilities of suppliers and other stakeholders involved in a given AI system. AI supply chains generally rest on three pillars:

1. Computational capacity;
2. AI models and software libraries;
3. Data.

Each pillar involves distinct, sometimes common, players whose level of cybersecurity maturity may vary considerably.

KEY RISKS AND ATTACK SCENARIOS

The IT infrastructure supporting an AI system faces the same vulnerabilities as any other IT system. An AI system can also be attacked at different stages of its lifecycle, from raw data collection to inference. AI-specific attacks are generally gathered in three categories:

- **poisoning:** altering training data or model parameters to change AI system's response to all inputs or to a specifically crafted input;
- **extraction:** reconstruction or recovery of confidential data such as model parameters, configuration or training data from the AI system or model after the learning phase;
- **evasion:** alteration of input data to change the expected functioning of the AI system.

Such attacks could result in the malfunctioning of an AI system (availability or integrity risks), where the reliability of automated decisions or processes can be compromised, as well as in **sensitive data theft or disclosure (confidentiality risk).**

The opacity of most AI systems today presents additional challenges that users must consider. AI explainability varies widely based on the underlying model: many systems operate as "black boxes", making their decisions difficult to explain or justify. This opacity complicates efforts to secure these systems, as it hinders the ability to find the root cause of errors and other problematic outputs, and makes it more difficult to identify and investigate potential incidents.

The **main risks scenarios** involving an AI system are:

- **Compromising AI hosting and management infrastructure:** malicious actors could impact the confidentiality, integrity, and availability of an AI system by exploiting a wide range of common vulnerabilities, whether technical, organizational, or human. Compromising an AI system's hosting infrastructure is a plausible and critical attack vector, and must be considered throughout the AI system lifecycle.
- **Supply chain attack:** an attacker could exploit a vulnerability in one of the supply chain stakeholders (software libraries, pre-trained model providers, service providers, etc.). For example, open-source

libraries are often used in the development of AI systems, and are often integrated into broader frameworks. An attack on these libraries could jeopardize the entire AI system.

- **Lateralization via interconnections between AI systems and other systems:** AI systems are often interconnected with other ISs for communication and efficient data integration purposes. These interconnections may pose new risks, with, for example, attacks through indirect prompt injection, which exploit LLMs by inputting malicious instructions through external sources controlled by an attacker. Such an attack could be used to extract sensitive information or execute malicious commands remotely. This risk is particularly important if the AI system is interconnected with industrial systems as those systems can directly act on the physical world.

- **Human and organisational failures:** a lack of training can induce an over-reliance on automation and insufficient ability to notice anomalous behaviors of AI systems. In addition, shadow AI[4] can increase risks such as loss of confidential data, regulatory violations, reputational damage to the organization's image, etc. In the long term, the intensive and prolonged use of AI could lead to a risk of technological dependence or "lock-in" where future AI capacities could not be replaced by human action in case of failure. This risk is greater when AI systems are involved in critical activities (e.g. industrial environments), especially where business processes are highly automated.

- **Malfunction in AI system responses:** an attacker could compromise a database used to train an AI model, causing erroneous responses once it is in production. This attack requires significant effort from the attacker as AI model developers' practices tend to improve their resilience to intentional and malicious training data poisoning but can be particularly dangerous when used to categorize data, such as images used in a health or physical security context.

4. Shadow AI is defined as the use of mainstream generative AI solutions without the approval or oversight of the organisation's IT departments.

GUIDELINES FOR AI USERS, OPERATORS AND DEVELOPERS

Analysing the sensitiveness of the use-case should be a first step when considering the use of an AI system. The complexity, the cybersecurity maturity, the auditability, and the explainability of the AI system should correspond with the cybersecurity and data privacy requirements of the given use case.

When a decision is made to develop, to deploy or use an AI solution, in addition to the usual cyber recommendations, the following guidelines constitute good practices for AI users, operators, and developers:

- **Adjusting the autonomy level of the AI system to the risk analysis, the business needs, and the criticality of the actions undertaken.** Human validation should be integrated where necessary into this process, as it will help address both cyber risks and reliability issues inherent to most AI models (e.g., LLM hallucinations);

- **Mapping of the AI supply chain,** including both **AI components** and other hardware and software components, as well as **datasets** (nature, sourcing and processing - in particular to mitigate poisoning and assess the impact of extracting risks);

- **Keeping track of the interconnections** between AI systems and the rest of the information system, making sure each one them is required by the use-case, in order to minimize attack paths;

- **Continuously monitoring and maintaining AI systems,** to ensure that they work as intended, without bias or vulnerability which could impact

cybersecurity, thus mitigating the risks related to the "black box" nature of certain AI systems;

- Implementing a process to **anticipate major technological and regulatory changes** and **identify potential new threats,** to be able to adapt strategies and face future challenges;

- **Training and raising awareness internally** on the challenges and risks of AI, including executives to ensure that high-level decision-making is well informed.

The checklist in appendix 1 could provide AI users, operators and developers with additional measures and recommendations to consider.

GUIDELINES FOR POLICY-MAKERS

Considering regional and national contexts, policy-makers should aim to:

- **Support research relevant to these risks,** including domains such as adversarial machine learning (including AI-specific attacks as well as prevention and detection of such attacks), privacy-preserving computing, emerging offensive uses of AI.

- **Support the development of security evaluation and certification capacity based on shared standards,** to foster trust in AI models, apps, data, and infrastructure.

- **Continue promoting best cybersecurity practices to ensure the secure deployment and hosting of AI systems** with clear guidelines to leverage existing and applicable regulations, adapt security requirements to the level of risk, and by sharing feedback, so that organisations can avoid common mistakes and optimize AI integration into their operations.

- **Foster dialogue between cyber and AI actors,** in particular between cybersecurity agencies and AI Safety Institutes (or similar), while clearly defining respective perimeters and responsibilities, as a way to promote better consideration of the cyber challenges of AI systems. Such collaboration should focus on sharing information on emerging threats and on aligning efforts to protect critical systems.

- **Continue dialogue beyond the AI Summit,** including monitoring the evolving threats to AI systems, and pursue discussions and collaborations at the international level to identify guidelines to better secure the AI value chain and thus foster trust in AI.

→ **APPENDIX 1**

RECOMMENDATIONS FOR THE SECURE IMPLEMENTATION OF AN AI SYSTEM

This is intended as a high-level overview of things to consider and should not be seen as exhaustive. Please refer to appendix 2 for additional framework and guidelines.

1. Recommended self-assessment:

- Have I properly defined and documented the explicit and legitimate purpose(s) of my system, starting from the designing phase if possible?

- Have I properly integrated regulatory aspects into my thought process? Have I verified that the treatment envisaged by the AI system complies with applicable laws and regulations?

- Who has access to the AI system during each different phase of its life cycle?

- Is the principle of least privilege applied, in order to guarantee the security and integrity of the AI system?

- What is the AI system dependency chain?

- What is the reputation of my suppliers and what is their financial health?

- Do my vendors meet cybersecurity standards, whether they are data providers or software component vendors?

- Is it necessary to implement a cloud solution? Have I done a global risk assessment of the possible consequences (data protection, etc.)?

- Do I have a reversibility clause in my service agreement with a provider who can manipulate my data? Is reversibility technically (means or data transfer rate) and chronologically feasible?

- What are the impacts of using AI on the business? Can an AI malfunction endanger my organisation?

- Is there a security foundation at each stage of the AI system life cycle (guides and best practice benchmarks, mapping, etc.)?

- Should my AI models be protected in confidentiality? Are they of significant value to my organisation?

- If relevant, have I properly integrated suitable measures to protect personal data (privacy by design), for both data and metadata, and for the AI system model(s)?

2. Checklist of recommended actions

General recommendations:

☐ limit the use of AI systems for the automation of critical actions on other information systems;

☐ ensure AI is thoughtfully and appropriately integrated into critical processes and provide safeguards;

☐ perform a dedicated risk analysis by integrating the entire organisational context (for instance the impact of an AI system failure should be assessed across the whole organization);

☐ study the security of each stage of the AI system life cycle (from training data collection to inference phase and decommissioning);

☐ conduct a data protection impact assessment if required;

☐ Identify, track and protect AI-related assets;

Infrastructure and architecture recommendations:

- ☐ define the modalities for the use of the AI system and frame its integration into the decision-making process, in particular in the case of automation;
- ☐ apply cloud-specific measures, where appropriate, taking into account applicable regulations and organisational policies;
- ☐ apply the recommendations for outsourcing if applicable;
- ☐ apply secure administration recommendations on the AI system;
- ☐ leverage a controlled access system for critical AI components.

Have a deployment plan

- ☐ design the architecture so that, when scaling occurs, it does not impact negatively the level of security;
- ☐ apply DevSecOps principles across all phases of the project;
- ☐ design the AI system using a privacy by design approach to meet data protection requirements throughout the lifecycle:
  - o take into account data confidentiality issues;
  - o ensure the pseudonymisation or anonymisation of data where necessary;
  - o take the need-to-know issue into account when designing the AI system.

Be vigilant about the resources used

- ☐ use secure formats for obtaining, storing and distributing AI models;
- ☐ implement mechanisms to verify the integrity of model files before loading them;
- ☐ assess the level of trust of libraries and plug-ins used in AI system;
- ☐ ensure the quality and assess the level of confidence of the external data used in the AI system
- ☐ ensure the traceability of the actions carried out on the AI system;
- ☐ ensure that data collection has been carried out in a fair and ethical manner, for those used both for the development and for the use of the system.

Secure and harden the learning process

- ☐ adopt a strict policy on what data is accessed by the AI system, especially sensitive data;
- ☐ secure access and storage of training data;
- ☐ assess the security of the learning and re-learning methods used;
- ☐ implement measures on the extracted data, metadata, annotation and features, and on the AI system model(s) including:
  - o clean up data;
  - o identify relevant and strictly necessary data (in terms of volume, categories, granularity, typology, etc.);
  - o pseudonymise or anonymise data if necessary.

Make the application reliable

- ☐ implement multi-factor authentication for all administration tasks on AI systems;
- ☐ ensure the confidentiality and integrity of inputs and outputs;
- ☐ enforce security filters to detect malicious instructions;
- ☐ ensure that all data, metadata and annotations are kept up to date and accurate (in particular to avoid drift);
- ☐ conduct continuous evaluation of model accuracy and performance.

Thinking an organisational strategy

- ☐ document design choices;
- ☐ supervise the operation of the AI system;
- ☐ identify key individuals and oversee the use of subcontractors;
- ☐ implement a risk management strategy;
- ☐ provide for a degraded mode of operations without AI systems;
- ☐ implement framed generative AI usage policies (depending on the sensitivity of the organisation);
- ☐ establish a process to monitor AI system-specific vulnerabilities;
- ☐ closely monitor technical developments which would, for example, limit the use of personal data;
- ☐ implement a data management system;
- ☐ leverage secure deletion methods for data removal;
- ☐ document datasets used in the product to:
  - facilitate the use of the database;
  - facilitate the monitoring of data over time until their deletion or anonymisation;
  - Reduce the risk of unexpected data use.

Preventive measures

- ☐ regularly train staff on security risks related to AI;
- ☐ carry out regular security audits of the AI system;
- ☐ anticipate as much as possible the problems potentially associated with the exercise of rights (intellectual property and data protection for instance) to training data or to the model itself.

→ **APPENDIX 2**

REFERENCES

AI development

• AIVD. AI systems: develop them securely. 2023. Available at: AI-systems: develop them securely | Publication | AIVD

• G7. Hiroshima Process International, Code of Conduct for Organizations Developing Advanced AI Systems. 2023. Available at: 100573473.pdf

• G7. Hiroshima Process International, Guiding Principles for Organizations Developing Advanced AI Systems. 2023. Available at: 100573471.pdf

• NCSC-UK, CISA. Joint Guidelines for Secure AI System Development. 2023. Available at: Guidelines for secure AI system development - NCSC.GOV.UK

AI use cases

• ANSSI. Security recommendations for a generative AI system. 2024. Available at: https://cyber.gouv.fr/en/publications/security-recommendations-generative-ai-system

BSI, ANSSI. AI Coding Assistants. 2024. Available at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/ANSSI_BSI_AI_Coding_Assistants.pdf?__blob=publicationFile&v=7

• BSI. Generative AI Models: Opportunities and Risks for Industry and Authorities. 2025. Available at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Generative_AI_Models.pdf?__blob=publicationFile&v=6

AI vulnerabilities and security

• CSA Singapore. Guidelines and Companion Guide on Securing AI Systems. 2024. Available at: Guidelines and Companion Guide on Securing AI Systems | Cyber Security Agency of Singapore

• CSA Singapore, Resaro. Securing AI: A Collective Responsibility. 2024. Available at: Discussion Paper on Securing Artificial Intelligence (AI): A Collective Responsibility | Cyber Security Agency of Singapore

• Indian Computer Emergency Response Team (CERT-In) Ministry of Electronics and Information Technology Government of India, Technical Guidelines on SOFTWARE BILL OF MATERIALS (SBOM). Available at: https://www. cert-in.org.in/PDF/SBOM_Guidelines.pdf

• CERT-IN API security - threats, best practices, challenges and way forward using AI. Available at: https://www. cert-in.org.in/s2cMainServlet?pageid=PUBADV01&CACODE=CICA-2023-3248

• Junklewitz, H., Hamon, R., André, A., Evas, T., Soler Garrido, J. and Sanchez Martin, J.I.. Cybersecurity of Artificial

Intelligence in the AI Act. 2023. Available at: JRC Publications Repository - Cybersecurity of Artificial Intelligence in the AI Act

• Kamm L. (Cybernetica AS), Pillmann H. (RIA). Risks and controls for artificial intelligence and machine learning systems. 2024. Available at: Risks-and-controls-for-artificial-intelligence-and-machine-learning-systems.pdf

• MITRE ATLAS: MITRE ATLAS™

• OWASP AI Exchange: https://owaspai.org/ - includes several publications

• Vassilev, A., Oprea, A., Fordyce, A. and Andersen, H. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. 2024. Available at: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations | NIST

Risk management

• NIST AI Risk Management Framework: AI Risk Management Framework | NIST

• OECD. "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI". 2023. Available at: Advancing accountability in AI | OECD

Terminology

• ISO/IEC 22989:2022 - «Information Technology - Artificial Intelligence - Artificial Intelligence Concepts and Terminology». 2022. Available at: https://www.iso.org/standard/74296.html

• OECD. "Defining AI incidents and related terms". 2024. Available at: Defining AI incidents and related terms | OECD

Examples of regulation

• EU Artificial Intelligence Act
Regulation - EU - 2024/1689 - EN - EUR-Lex

• EU Cyber Resilience Act
Regulation - 2024/2847 - EN - EUR-Lex

Version 1.0 – Février 2025 – ISSN en cours

AGENCE NATIONALE DE LA SÉCURITÉ DES SYSTÈMES D'INFORMATION
ANSSI — 51, boulevard de la Tour-Maubourg — 75 700 PARIS 07 SP
www.cyber.gouv.fr