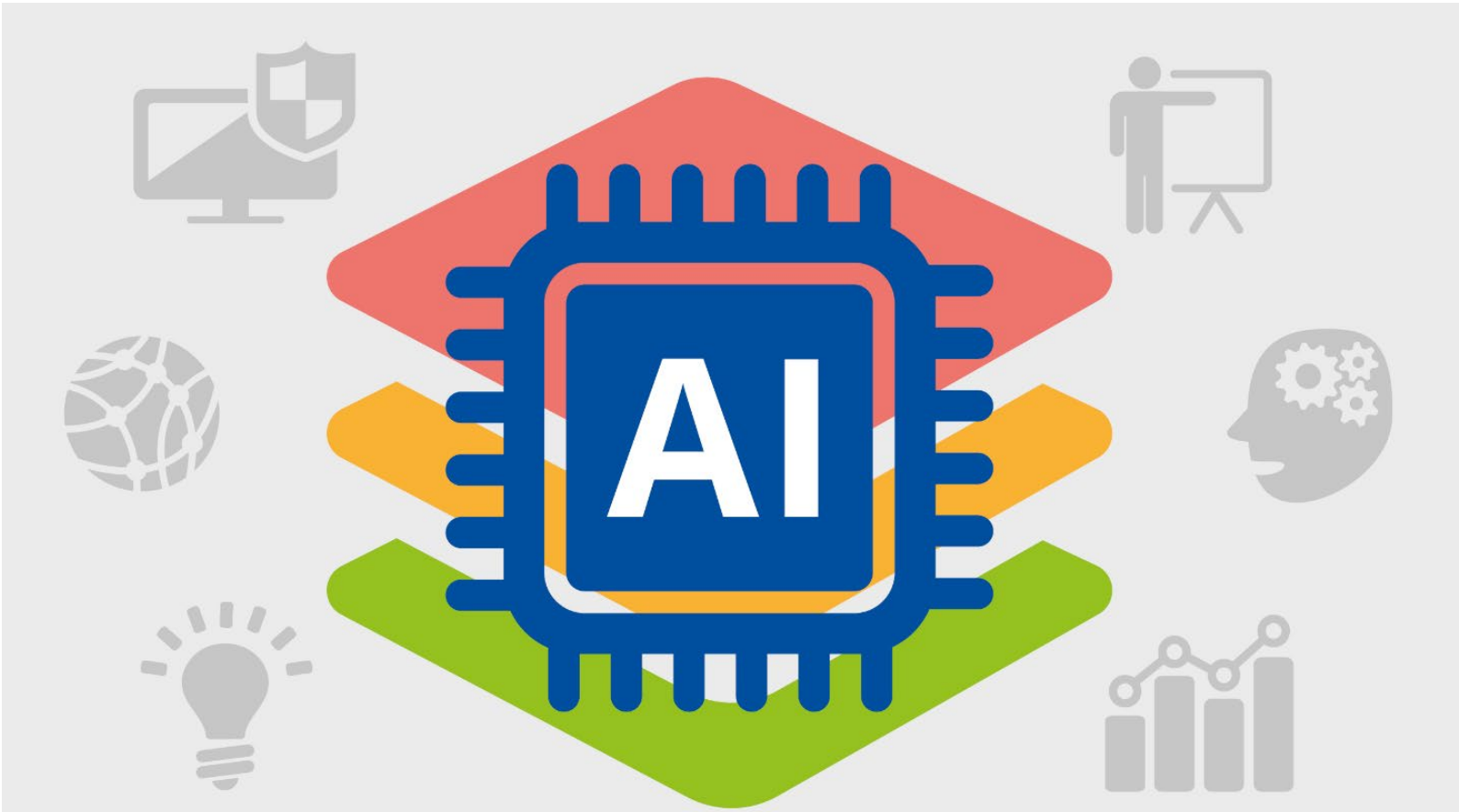
EUROPEAN UNION AGENCY
FOR CYBERSECURITY



MULTILAYER FRAMEWORK FOR GOOD CYBERSECURITY PRACTICES FOR AI

JUNE 2023

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the EU agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, ENISA contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the EU's infrastructure and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found here: www.enisa.europa.eu.

CONTACT
To contact the authors, use info@enisa.europa.eu
For media enquiries about this paper, use press@enisa.europa.eu

AUTHORS
Nineta Polemi, Isabel Praça

EDITORS
Monika Adamczyk, Konstantinos Moulinos (ENISA)

ISBN 978-92-9204-619-4 doi:10.2824/588830 TP-04-23-025-EN-N

TABLE OF CONTENTS

1. INTRODUCTION 4

1.1. AIMS AND OBJECTIVES 4

1.2. BENEFITS AND BENEFICIARIES 4

1.3. METHODOLOGY 5

2. FRAMEWORK FOR GOOD CYBERSECURITY PRACTICES FOR AI 6

2.1. LAYER I – CYBERSECURITY FOUNDATIONS 7

2.2. LAYER II – AI FUNDAMENTALS AND CYBERSECURITY 13

2.3. LAYER III – SECTOR-SPECIFIC CYBERSECURITY GOOD PRACTICES 23

3. SURVEY ANALYSIS 27

3.1. METHODOLOGY 27

3.2. SURVEY ANALYSIS 28

3.3. SURVEY CONCLUSIONS 34

4. CONCLUSIONS AND THE WAY FORWARD 36

ANNEX I: QUESTIONNAIRE 38

ANNEX II: AI-RELATED STANDARDS 42

A.1 AI SECURITY-RELATED STANDARDS 42

A.2 DESIGN-RELATED STANDARDS 42

ANNEX III: LIST OF ABBREVIATIONS 44

EXECUTIVE SUMMARY

In April 2021, the European Commission published a proposal for an artificial intelligence (AI) regulation ([1]). The proposal focuses on high-risk AI systems, for which requirements include adequate levels for robustness, accuracy and cybersecurity. The proposed regulation requires the use of technical standards during the design and development of high-risk AI systems to ensure a consistent and high level of protection of public interests, such as health, safety and fundamental rights. Work on the AI-related standards has already begun, however standards development takes a long time, so they most likely will not be ready before the regulation enters into force. Until then, a collection of good practices would be beneficial for the AI ecosystem stakeholders.

To this end, ENISA has published already two studies on cybersecurity for AI: *AI Cybersecurity Challenges – Threat landscape for artificial intelligence* ([2]) and *Securing Machine Learning Algorithms* ([3]), which provide guidance for cybersecurity within the AI machine learning (ML) life cycle. However, these studies do not fully cover the entire AI life cycle (from concept to decommissioning), the associated infrastructure and all the elements of the AI supply chain.

The importance of identifying good cybersecurity practices for AI, which go beyond ML, has also been noticed by the Commission, which requested ENISA assistance in identifying not only existing cybersecurity practices for AI, but also in gathering information on the current state of cybersecurity requirements for AI on the EU and national levels, along with the monitoring and enforcement of these requirements by national competent authorities (NCAs).

In this report, we present a scalable framework to guide NCAs and AI stakeholders on the steps they need to follow to secure their AI systems, operations and processes by using existing knowledge and best practices and identifying missing elements. The framework consists of **three layers (cybersecurity foundations, AI-specific cybersecurity and sector-specific cybersecurity for AI)** and aims to provide a step-by-step approach on following good cybersecurity practices in order to build trustworthiness in their AI activities.

We gathered information through a survey, which is based on the framework presented in this report and the main principles of the proposed Artificial Intelligence Act (AI Act) and the coordinated plan on AI ([4]) from the NCAs (AI-specific or cybersecurity-related). **We analysed the current state of cybersecurity requirements** and **monitoring and enforcement practices that the NCAs have adopted** (or plan to develop) to ensure that the national AI stakeholders address cybersecurity requirements. **The survey results revealed that the readiness level of NCAs is low and that further measures are needed**. The report also points out additional research efforts needed for the development of these additional cybersecurity AI practices.

The main recommendation is to treat the cybersecurity of AI systems as an additional effort to existing practices for the security of organisations' information and communications technology (ICT). The existing cybersecurity practices need to be complemented with AI-specific practices, which address, among other things, their dynamic socio-technical nature. Examples of additional practices include dynamic, measurable risk assessments of AI technical (e.g., poisoning data) and social threats (e.g., bias, lack of fairness) and continuous risk management (RM) during the AI system life cycle. The operational environment (e.g., energy sector) and usage (e.g., monitoring the smart meters) of the AI system need to be considered for the realistic and accurate mitigation of sectoral threats.

[1] Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.
[2] ENISA, *AI Cybersecurity Challenges – Threat landscape for artificial intelligence*, 2020, https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.
[3] ENISA, *Securing Machine Learning Algorithms*, 2021, https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms.
[4] Annexes to the Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions – Fostering a European approach to Artificial Intelligence – Coordinated plan on artificial intelligence 2021 review, COM(2021) 205, https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review.

1. INTRODUCTION

In recent years, AI systems and related technologies (e.g., robots, biometrics, surveillance cameras, IoTs, drones) have been increasingly deployed and used by all economic sectors (e.g., health, energy, telecom, financial) in their daily business activities. However, to unveil the full business potential of AI – and also to serve European values and the democratic rights of Europeans – adequate level of cybersecurity and privacy need to be ensured in these systems. EU is already working on the necessary legal cybersecurity instruments to protect the AI developments towards serving EU citizens and has invested highly in the trustworthiness of AI through Horizon 2020, Horizon Europe and the Digital Europe Programme. At the same time, the majority of Member States (MS) have put in place national AI strategies where the NCAs play a crucial role in the effort to build innovative human-centric AI ICT products in Europe.

Various standards are being developed by numerous bodies – such as the European Telecommunications Standards Institute (ETSI), the European Committee for Standardization, the International Organization for Standardization (ISO), the National Institute for Standards and Technology (NIST), the Institute of Electrical and Electronics Engineers (IEEE) and the Open Web Application Security Project – along with recommendations and white papers by cybersecurity organisations – such as ENISA, the European Union Agency for Law Enforcement Cooperation, the European Defence Agency, the European Cyber Security Organisation and the Centre for European Policy Studies (CEPS) – and international bodies such as the OECD and the UN. These provide policy and technical measures (design principles, integration platforms, test cases), practices and further research needed to secure AI products (e.g., software, hardware, systems, services) and ensure their human-centric design.

Numerous existing traditional cybersecurity practices and solutions (methodologies, tools, recommendations) can be used to guide AI stakeholders in undertaking appropriate traditional controls. These good practices are presented in different documents which address various layers of the ICT environments (e.g., physical, network, informatics, data, services) that host AI products, making it difficult for the AI stakeholders to determine the ones appropriate for their environment. Furthermore, the dynamic nature of AI imposes some open issues and additional cybersecurity measures required in undertaking additional effective measures. Additional cybersecurity practices are also needed when AI products target a specific economic sector (e.g. health, energy, automotive) to meet sectoral security requirements. Finally, further research activities are needed to enforce the resilience and security of the AI-based products.

1.1. AIMS AND OBJECTIVES

The objectives of this study are:

- to develop a **framework for AI good cybersecurity practices (FAICP)** necessary for securing the ICT infrastructures and the hosted AI, taking into account the AI life cycle which goes beyond ML (from system concept to decommissioning) and all elements of the AI supply chain, associated actors, processes and technologies;
- to collect **information from EU NCAs** about **the national cybersecurity requirements for AI** and how **compliance with these requirements is monitored and enforced** nationally;
- to **identify challenges and gaps in the existing cybersecurity practices for AI** that will help AI stakeholders from all sectors bring trustworthiness to their AI-related operations and business.

1.2. BENEFITS AND BENEFICIARIES

The target beneficiaries of this study are as follows.

- **AI stakeholders.** Designers, developers, integrators, manufacturers, operators, service providers, supply chain business partners, auditors, legislators, policymakers and professional users.
- **National authorities.** Authorities, agencies and competence centres for monitoring and assessing cybersecurity or AI activities.

1.3. METHODOLOGY

In this study, we treat AI systems as cyber assets within an ICT infrastructure. In particular, we identify their main components: data sources, data, algorithms, training models, implementation/data management/testing processes and users. These components of the AI systems belong in the layers of an ICT infrastructure within an enterprise. The report takes this view in order to develop a framework that can easily group the cybersecurity best practices in multiple layers.

Because AI systems are part of the ICT infrastructure, not only AI specific cybersecurity practices must be applied, but also those that protect the ICT encompassing the AI elements. In order to achieve this, we conducted a literature review to identify the main cybersecurity challenges, standards and best practices that contribute towards addressing these challenges. ENISA's previous work on cybersecurity of AI[5] was also extensively used, along with best practices for AI published by various organisations.

Using the cybersecurity concepts described in the FAICP framework and the main principles (related to cybersecurity) of the AI Act and the Coordinated Plan on AI, a survey was developed and conducted with NCAs (AI-specific or cybersecurity-related), to identify the current level of MS preparedness in the monitoring and enforcement of cybersecurity requirements for AI systems.

The steps we followed can be summarised as follows.

- We used definitions used in relevant standards on various cybersecurity AI-related concepts and in the European taxonomy proposed by the Joint Research Centre (JRC)[6].
- We reviewed the interrelations of the cybersecurity concepts in the various standards (e.g., ISO2700x, ISO15408, ETSI SAI, ISO/IEC 24368:2022, ISO/IEC 22989:2022).
- We reviewed the relevant cybersecurity legislation, i.e. NIS[7], NIS 2[8], the proposed AI Act and the proposed Cyber Resilience Act.
- We analysed the current state-of-the-art for cybersecurity AI-related standards from various organisations (e.g., ETSI, European Committee for Standardization, ISO, IEEE, NIST), best practices published (e.g., Organization for Economic Co-operation and Development or OECD, ENISA, JRC, European Cyber Security Organisation, CEPS, BSA, ARM) and recommendations.
- We identified various tools that can be used for the development of trustworthy AI (e.g., OECD AI Policy Observatory[9], MITRE ATLAS[10]).
- We searched for best practices in the uptake of AI in various critical (based on NIS and NIS 2) sectors (e.g., automotive, energy, finance, health, industry, telecoms).
- We reviewed national AI strategies and the assessment of the maturity of the implementation of the strategies by JRC and ENISA's cybersecurity review.
- We developed a questionnaire to assess the current state of policies for cybersecurity enforcement of AI.
- We identified the open issues and additional cybersecurity practices that need to be developed due to the dynamic and socio-technical nature of AI systems.

[5] https://www.enisa.europa.eu/topics/iot-and-smart-infrastructures/artificial_intelligence.
[6] See: Nai Fovino, I., Neisse, R., Hernandez Ramos, J., Polemi, N., Ruzzante, G., Figwer, M. and Lazari, A., *A Proposal for a European Cybersecurity Taxonomy*, JRC Technical Reports, Publications Office of the European Union, Luxembourg, 2019, https://publications.jrc.ec.europa.eu/repository/handle/JRC118089.
[7] Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union
[8] Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive)
[9] https://oecd.ai/en/
[10] https://atlas.mitre.org/

2. FRAMEWORK FOR GOOD CYBERSECURITY PRACTICES FOR AI

OVERVIEW OF THE FRAMEWORK

The proposed **FAICP framework** is a simple approach to guide NCAs, individual AI stakeholders and the research community on how they can use the existing cybersecurity practices, what additional cybersecurity activities are needed to address the specificities of AI and the additional practices required when AI systems are employed in specific sectors (e.g., health, energy, telecom).

The framework was developed using the following principles.

- **Inclusive.** Uses past experience and builds upon it.
- **Holistic.** Considers the AI systems within the ICT infrastructure and embraces all cybersecurity practices needed around and within the AI systems and their individual components.
- **Expandable.** Its generic and yet embracing structure can include future developments in all three layers.
- **Multi-use.** Useful to AI stakeholders independently of the sector.
- **International.** Includes European and international efforts, standards and recommendations.

The FAICP is a scalable 3-layered framework:

Figure 1: FAICP – A scalable framework for AI-related cybersecurity good practices

- **Layer I (cybersecurity foundations).** The basic cybersecurity knowledge and practices that need to be applied to all ICT environments that host/operate/develop/integrate/maintain/supply/provide AI systems. Existing cybersecurity good practices presented in this layer can be used to ensure the security of the ICT environment that hosts the AI systems.

- **Layer II (AI-specific).** Cybersecurity practices needed for addressing the specificities of the AI components with a view on their life cycle, properties, threats and security controls, which would be applicable regardless of the industry sector.
- **Layer III (Sectoral AI).** Various best practices that can be used by the sectoral stakeholders to secure their AI systems. High-risk AI systems (i.e. those that process personal data) have been identified in the AI Act and they are listed in this layer to raise the awareness of operators to adopt good cybersecurity practices.

2.1. LAYER I – CYBERSECURITY FOUNDATIONS

AI systems are hosted in ICT infrastructures and in this first layer of the proposed framework, we emphasise the need to start by securing the ICT-hosted ecosystem as a whole using basic cybersecurity practices. We present the basic cybersecurity principles and procedures as described in various standards, methods and best practices that need to be applied by AI stakeholders. However, due to the dynamic, constantly evolving nature of AI systems, the cybersecurity foundations built in this layer leave some additional open issues that will be outlined and further analysed in Layer II, where additional cybersecurity practices will accompany the basic ones described in this layer.

The key elements of this layer are:

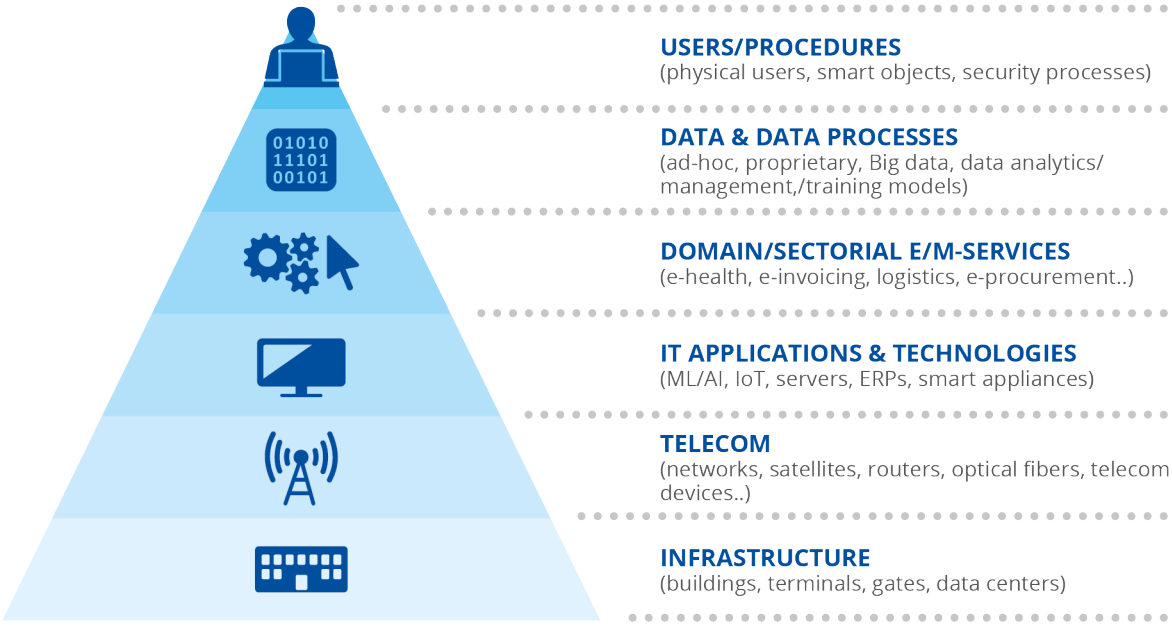
- security management of the ICT infrastructure hosting AI systems;
- security management;
- cybersecurity certification;
- cybersecurity legislation and policies that affect AI systems.

Securing the ICT infrastructure hosting AI systems

ICT encompasses the infrastructure and assets that enable digital computing. All organisations rely on the secure operations of ICT for their business/digital activities, regardless of whether the ICT is hosted in-house or owned by a third party (cloud provider, supply chain business partner).

The components of any ICT infrastructure can be viewed as a scalable pyramid of six building blocks[11]:

Figure 2: AI systems are hosted in the ICT infrastructure



The first basic building block (Infrastructure) consists of all physical assets, used in the 2nd building block (Telecom) where all types of networks and telecom equipment are placed. These are necessary for the 3rd block (IT applications and technologies), which also contains assets related to AI systems. The 4th block (Domain/sectoral e/m-services) includes all digital services, while the 5th block (Data and data processes) includes all of the types of data used in the

[11] Polemi, N., *Port Cybersecurity – Securing critical information infrastructures and supply chains*, 1st edition, Elsevier, 2017.

previous blocks. Finally, the 6th block (Users/procedures) includes all users that interact with all components from the previous blocks, i.e. internal and external physical entities (e.g. persons, enterprises), smart objects (e.g. IoT) and operational procedures.

Any ICT system is a cyber-physical system, since the first and last blocks (Users and Infrastructure) of the ICT are the physical layers, whereas the four intermediate blocks are the cyber layers. Cybersecurity of an ICT infrastructure should cover the following dimensions (also known as 'CIA'): confidentiality, integrity/authenticity and availability/non-repudiation (Figure 3) for all six blocks and all assets within the layers of the ICT infrastructure.

Figure 3: Information aspects protected according to ISO 27001

Security management

Risk management is the basic cybersecurity practice for ensuring that an enterprise is secure, by identifying and evaluating threats and vulnerabilities, potential impacts and by measuring risks. According to the NIS and NIS 2 directives, all essential entities important for the functioning of society **need to assess and mitigate their risks**. Therefore, the first step in the security of AI systems and the security of their life cycle is **to operate in a secure environment, i.e. to secure the ICT infrastructure that hosts the AI systems.**

The various types of threats to ICT infrastructures are listed below.

- **Adversarial threats.** These pose malicious intentions (e.g. denial of service attacks, non-authorised access, masquerading of identity) to individuals, groups, organisations or nations.
- **Accidental threats.** These are caused accidentally or through legitimate components. Human errors are a typical accidental threat. Usually, they occur during the configuration or operation of devices or information systems, or the execution of processes.
- **Environmental threats.** These include natural disasters (floods, earthquakes), human-caused disasters (fire, explosions) and failures of supporting infrastructures (power outage, communication loss).
- **Vulnerability.** This is an existing weakness that might be exploited by an attacker.

For the identification of general cybersecurity threats, AI stakeholders wishing to secure their ICT infrastructure can use the annual *ENISA Threat Landscape* [12] report on the state of the cybersecurity threat landscape, or similar reports such as the annual technical threat reports published by other organisations (e.g. the Open Web Application Security Project or OWASP)[13].

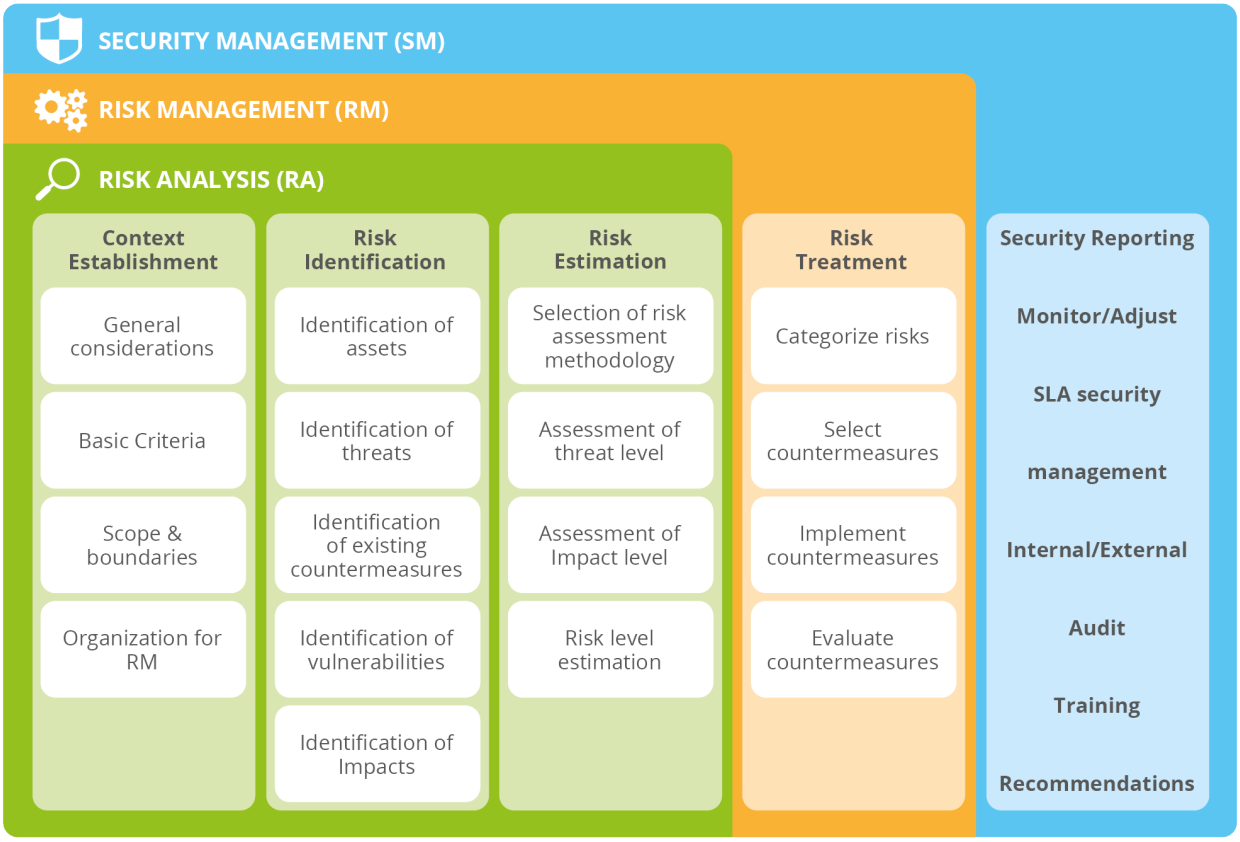
Security management[14] includes two main phases.

- **Risk analysis.** Threat/vulnerability/impact analyses and risk estimations are conducted on all assets within the perimeter of the assessment (e.g. components of medical devices, cyber assets within a hospital's infrastructure).

[12] See https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends.
[13] See the OWASP Top 10: https://owasp.org/www-project-top-ten/.
[14] NIST Cybersecurity Framework, https://www.nist.gov/cyberframework.

- **Risk management.** Risks are treated by selecting and implementing appropriate countermeasures (Figure 4). The appropriate selection of controls requires a cost-benefit analysis, to determine the risks the manufacturer is willing to accept and compare the costs of those risks against the benefits.

Figure 4: Phases of security management based on NIST



The major focus lies on enterprises and the identification, analysis and evaluation of threats and vulnerabilities, along with the estimation of risk levels to the respective enterprise assets. The outcome of a risk analysis is **a list of threats to all assets** of the enterprise ICT system, together **with the corresponding risk levels of these threats to all assets**.

Since its creation, ENISA has worked on RM and has produced several methodologies and best practices (see Table 1) that can be used to conduct RM and can be used by AI stakeholders to secure the ICT infrastructure that host their AI systems.

Table 1: ENISA publications on risk assessment

Publication name
ENISA, *Methodology for Sectoral Cybersecurity Assessments*, 2021, https://www.enisa.europa.eu/publications/methodology-for-a-sectoral-cybersecurity-assessment
ENISA, *National-level Risk Assessments: An analysis report*, 2013, https://www.enisa.europa.eu/publications/nlra-analysis-report
ENISA, *Consumerization of IT: Final report on risk mitigation strategies and good practices*, 2012,

https://www.enisa.europa.eu/publications/COIT_Mitigation_Strategies_Final_Report
ENISA, 'Inventory of Risk Management / Risk Assessment Methods and Tools', n.d., https://www.enisa.europa.eu/topics/risk-management/current-risk/risk-management-inventory/inventory-of-risk-management-risk-assessment-methods-and-tools?v2=1&tab=details
ENISA, *Risk Assessment – Guidelines for trust service providers, part 2*, 2013, https://www.enisa.europa.eu/publications/tsp2-risk
ENISA, *Cloud Computing – Benefits, risks and recommendations for information security*, 2009, https://www.enisa.europa.eu/publications/cloud-computing-risk-assessment
ENISA, *Methodology for Sectoral Cybersecurity Assessments*, 2021, https://www.enisa.europa.eu/publications/methodology-for-a-sectoral-cybersecurity-assessment

The mitigation of risks found in an ICT infrastructure requires a selection of countermeasures (soft measures, e.g. procedures or processes and hard measures, e.g. technical controls). The AI stakeholders can use ISO 27002[15] for the implementation and management of technical controls and also technical controls proposed by international organisations (e.g. SANS Top 20[16], UCI[17], CIS Critical Security Controls[18]).

Apart from these guidelines, a number of EU research projects related to RM, where innovative security management tools have been developed, can be useful to AI stakeholders[19].

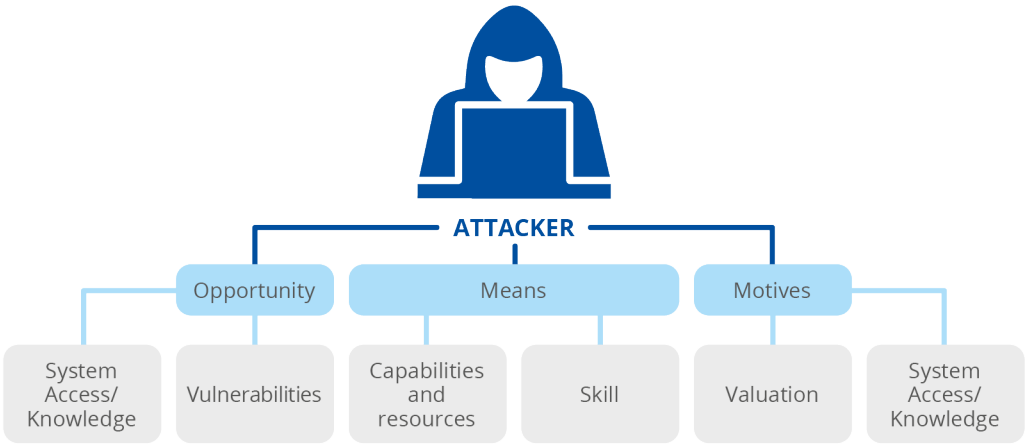
Threat agents and attackers' profiles in AI ecosystems

AI stakeholders need to be aware of their adversaries in the operational environment. Three key components characterise potential adversaries: means, motive and opportunity. An attack occurs if the attacker has the means to execute it, the opportunity to do so and exploit a vulnerability, and a motive to target the victim in question.

AI stakeholders and operators need to analyse potential attackers in order to estimate their risk levels more realistically and accurately and to undertake appropriate countermeasures[20].

[15] https://www.iso.org/standard/54533.html
[16] https://www.sans.org/critical-security-controls/
[17] https://security.uci.edu/security-plan/plan-controls.html
[18] https://www.cisecurity.org/controls/cis-controls-list/
[19] For a list of relevant EU projects, see the CORDIS website: https://cordis.europa.eu
[20] Kioskli, K., Polemi, N., 'Estimating attackers' profiles results in more realistic vulnerability severity scores', in Ahram, T. and Karwowski, W. (eds), Human Factors in Cybersecurity, AHFE (2022) International Conference, AHFE Open Access, Vol. 53, AHFE International, 2022, http://doi.org/10.54941/ahfe1002211

Figure 5: Adversary characterisation



Adversarial threats[21] are mainly caused by people who have a deliberate intention to cause harm. Typically, these threat actors are referred to as attackers or adversaries. In the literature, cyber threat actor lists and taxonomies are still being developed, and most of these lists identify the following **intentional threat actors**: insider attackers, cyber terrorists, hacktivists / civil activists, organised cybercriminals, script kiddies, state-sponsored attackers, commercial industrial espionage agents, cyberwarriors / individual cyber fighters, cyber vandals and black hat hackers.

Today there is no universally accepted standard for an attackers' taxonomy and new definitions and proposals for taxonomies are still emerging; 11 attacker types were defined by ENISA in 2021[22] by consolidating, refining and improving previous taxonomies, which reflect the current threat landscape and can be mapped to other taxonomies in use by MS and EU bodies. The attackers target ICT infrastructures hosting AI systems/products or AI systems at any stage of their life cycle.

Cybersecurity certification

Cybersecurity certification under the EU's Cybersecurity Act (CSA)[23] is intended to increase trust and security for European consumers and businesses of ICT products (including the ones using AI technologies).
The main standard for certification is ISO/IEC 15408[24] (particularly the Common Criteria – CC[25]) that establishes the principles for ICT security assessment, while ISO/IEC 18045[26] provides a methodology to help an evaluator conduct a CC evaluation by defining the minimum actions.

These standards have been implemented in various methodologies (e.g. ETSI-TVRA[27], ENISA-RCA[28], CYRENE[29]) that AI stakeholders can use to evaluate ICT products. These methodologies can be used to evaluate the security of ICT assets hosting AI components, such as a server that hosts AI models or a supply chain service in which AI assets participate during the provision of the service.

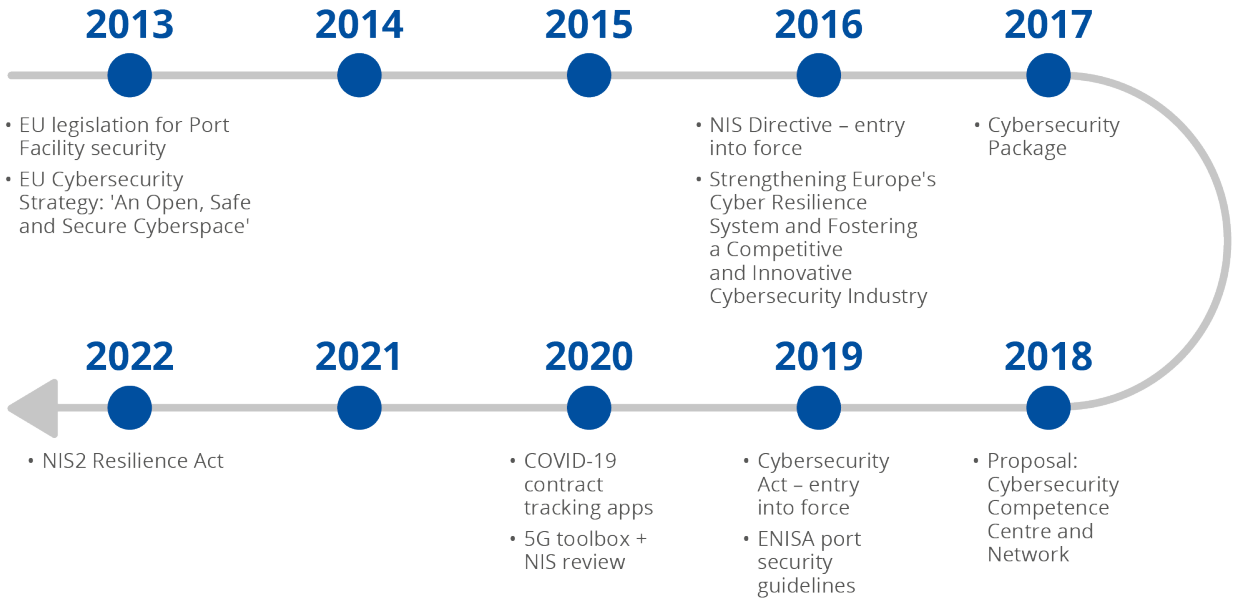
Evaluation methodologies to identify security requirements for development of certification schemes for AI products are not available yet. Additional research efforts are needed to evaluate the security of AI systems, due to their dynamic nature.

Cybersecurity legislation and policies

[21] Source: ENISA, *Methodology for Sectoral Cybersecurity Assessments*, 2021, https://www.enisa.europa.eu/publications/methodology-for-a-sectoral-cybersecurity-assessment.
[22] See footnote ([17]).
[23] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act), https://eur-lex.europa.eu/eli/reg/2019/881/oj.
[24] https://www.iso.org/standard/72891.html.
[25] https://www.commoncriteriaportal.org/.
[26] https://www.iso.org/standard/72889.html.
[27] https://www.etsi.org/deliver/etsi_ts/102100_102199/10216501/05.02.03_60/ts_10216501v050203p.pdf.
[28] https://www.enisa.europa.eu/publications/methodology-for-a-sectoral-cybersecurity-assessment.
[29] CYRENE was an EU Horizon 2020 project, see https://www.cyrene.eu, accessed on 14 May 2021.

The operators of ICT infrastructures need to be aware of and comply with all EU legislation, recommendations and directives, from the cybersecurity strategy in 2013 to the NIS 2 directive and the Cybersecurity Resilience Act in 2022.

Figure 6: Cybersecurity legal/policy EU instruments



Several pieces of legislation and policies have been developed to ensure the most effective responses and the ICT infrastructure needed to comply with these policies. **NIS 2**[30] **and the CSA**[31] **are considered to be Europe's two most important and far-reaching pieces of cybersecurity legislation**[32] **and the general data protection regulation (GDPR)**[33] **is the key personal data protection act**, emphasising supply chain security and privacy respectively, which are **most relevant for the life cycle of the AI systems** as well.

The EU's common security and defence policy (CSDP)[34] is another important element, since it is the main instrument of the EU for dealing with new and unconventional security threats and serves to prepare a possible common European defence of the EU. Since AI is considered a technology that will play a crucial role for defending the EU, **it is also important that this policy is considered**.

The CSA[35] establishes a **cybersecurity certification framework** for products and services. This framework provides EU-wide certification schemes as a comprehensive set of rules, technical requirements, standards and procedures. This way it is possible to ensure the general public trust in the cybersecurity of IT products and services. It is important that we can see that a product has been checked and certified to conform to high cybersecurity standards. AI-related products will gain trustworthiness if they are certified and, in the years to come, various cybersecurity schemes will be developed for AI products to specify the security requirements.

Another important initiative is the **European Cybersecurity Competence Centre**[36], which aims to increase Europe's cybersecurity capacities and competitiveness, working together with a Network of National Coordination Centres[37] to build a strong cybersecurity community. Also, the establishment of national **computer security incident response**

[30] Revised Directive on Security of Network and Information Systems (NIS 2), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2020:823:FIN.
[31] https://eur-lex.europa.eu/eli/reg/2019/881/oj.
[32] https://digital-strategy.ec.europa.eu/en/policies/nis-directive.
[33] The GDPR applies to the processing of personal data regardless of the means by which personal data is processed and thus applies to AI systems that process personal data. However, a number of AI-related data protection issues are not explicitly answered in the GDPR and need to be specified. For additional information on this topic, see: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf.
[34] https://www.eeas.europa.eu/eeas/common-security-and-defence-policy_en.
[35] https://eur-lex.europa.eu/eli/reg/2019/881/oj.
[36] https://cybersecurity-centre.europa.eu/index_en.
[37] https://cybersecurity-centre.europa.eu/nccs_en.

teams (CSIRTs) is an essential step to facilitate the building of cyber capacity both within and across nations and to make it more effective[38]. The European Cybersecurity Competence Centre will guide AI stakeholders in enhancing the cybersecurity of their products and advance their research efforts and developments. CSIRTs will gain the necessary capabilities to guide stakeholders to respond to AI attacks or using AI technologies to defend their infrastructures. These are just some of the instruments developed under the **EU cybersecurity strategy**[39], which aims to build resilience to cyber threats and ensure that citizens and businesses benefit from trustworthy digital technologies.

In addition, the **new legislative framework (NLF)**[40] improves market surveillance, introduces rules to better protect both consumers and professionals from unsafe products (of EU or non-EU origin), sets rules for accreditation and establishes a common legal framework for industrial products. The NLF will enhance the security of AI-based products.

The **European Chips Act**[41] is relevant to AI security because semiconductors are the elements of platform technology of the 21st century that will be used for AI developments and for embedding strong security measures. The EU globalised semiconductor industry will be supported by this proposed act.

The **Cyber Resilience Act**[42] will set new cybersecurity rules for digital products and ancillary services. This initiative will also promote the security of AI products, since it aims to address market needs and protect consumers from insecure products by introducing common cybersecurity rules for manufacturers and vendors of tangible and intangible digital products.

The EU legislative instruments and policies are mature and embrace AI system trustworthiness. The upcoming challenge is upscaling and embracing the legal and policy requirements to technical requirements, design specifications and concrete testing and assessment of AI systems.

New challenges

The common cybersecurity practices need to be embraced with additional practices that will meet the security requirements of AI systems. Due to the dynamic and multifaceted nature of these systems, the following additional challenges need to be addressed.

- AI risk assessments should be dynamic and combined with anomaly detection approaches, as for ICT systems in general.
- Measuring AI threats and evaluating AI risks require the development of a widely accepted scaling system that can meet common social and ethical values.
- A taxonomy of AI attackers needs to advance the existing taxonomies, in order to better understand the motives, capabilities, objectives and psychological profiles of the AI adversaries.
- Evaluation of an AI product against a static set of requirements can quickly become outdated, therefore dynamic RM and conformity assessment throughout the entire AI life cycle are required.
- No new standards or legislative instruments are needed, but there is a need for targeted guidelines, best practices and tools that will help the evaluation of AI-cybersecurity and trustworthiness.

2.2. LAYER II – AI FUNDAMENTALS AND CYBERSECURITY

In the previous section we addressed the various blocks within an ICT infrastructure and discussed the characteristics of the first blocks and the related tools and legislation. AI systems are part of the 3rd block, see Figure 2 in Section 2.3. In this chapter, we assume that AI systems are supported by a trusted hardware infrastructure and focus on the particularities of these types of systems, their properties, threats, risks and related tools and legislation.

The key elements of this layer are:

- AI legislation
- Types of AI

[38] ENISA, *ENISA CSIRT Maturity Framework – Updated & improved*, 2022, https://www.enisa.europa.eu/publications/enisa-csirt-maturity-framework.
[39] https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-strategy.
[40] https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en.
[41] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-chips-act_en.
[42] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13410-Cyber-resilience-act-new-cybersecurity-rules-for-digital-products-and-ancillary-services_en.

- AI assets and procedures
- AI threat assessment
- AI security management
- AI-related standards
- Ethical and trustworthy AI
- Tools
- Networks and initiatives

AI legislation

The cybersecurity legislation presented in the first layer is complemented with AI-specific legislative efforts. The most important Commission proposal is the AI Act[43], which puts forward the proposed regulatory framework on AI with the following specific aims:

- ensure that AI systems placed on the EU market or put into service are safe and respect existing law on fundamental rights and EU values;
- ensure legal certainty to facilitate investment and innovation in AI;
- enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems;
- facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.

In addition to the AI Act proposal, the Commission has published a proposal for an AI liability directive, whose purpose is to 'improve the functioning of the internal market by laying down uniform rules for certain aspects of non-contractual civil liability for damage caused with the involvement of AI systems'[44].

Types of AI

According to the OECD[45], 'An AI system is a machine-based system that can influence the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to: (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g. with ML) or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy.'

AI is a broad topic which can be further dissected into multiple subfields, which in turn are often mentioned interchangeably. Some of these are described below.

- **Computer vision.** This is related to the automatic processing of visually rich data such as images or videos. Some of the main tasks under this domain are object detection, facial recognition, action/activity recognition and human pose estimation.
- **Expert systems.** Expert systems are highly interpretable white-box programs that use a knowledge-based approach, where domain information provided by experts in the field is used by a knowledge engineer to populate a knowledge base (e.g. a set of if–then rules). At the inference phase, the content of the knowledge base is used by an inference engine to derive new conclusions for a given set of observed facts.
- **Machine learning.** ML is arguably the most disruptive subfield of AI, introducing a new paradigm for the design of intelligent systems. ML algorithms can learn predictive rules from hidden patterns in labelled/unlabelled data on their own, without needing to be explicitly programmed for a specific task. Furthermore, **deep learning (DL)**, which mimics the structure and way of working of the human brain, is currently the most promising branch of ML, benefiting from large amounts of available data.
- **Multi-agent systems.** These are part of distributed AI and address the interaction between several autonomous entities designated as agents. Agents can perceive their surrounding environment on their own, and collaborate or negotiate with other agents to interact with them in a beneficial manner.
- **Natural language processing.** This makes use of computational techniques to learn, understand and produce content in human language with respect to several levels of

[43] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX %3A52021PC0206.
[44] https://ec.europa.eu/info/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en.
[45] https://www.oecd.org/digital/artificial-intelligence/.

linguistic analysis.

- **Robotics.** Robotics is related to the development of physical machines with variable degrees of autonomy. These are able to continuously adapt to their ever-changing environments by several loops of actions such as perceiving, planning and executing.
- **Speech recognition.** The speech recognition domain encompasses methods for processing speech automatically, providing better ways of interfacing with computers.

ML and DL undoubtedly pose the main challenges to security, as grey-box and black-box models dominate the field and imply a dynamic analysis of the threats, not just along the life cycle, but also in the interrelations within other blocks of an ICT infrastructure. The following sections discuss many of the threats related to this subfield.

No-code AI reduces the time to build AI models to minutes, enabling companies to easily adopt ML models in their processes. **No-code AI solutions are focused on helping non-technical users build ML models without getting into the details of every step in the process of building the model.** This makes them easy to use but harder to customise. Multiple no-code AI platforms, i.e. software that allows people without specialised skills to build algorithms, are proliferating rapidly. In the future, people will not just want to deploy different models, but potentially thousands of pieces of AI software. They will be able to design and create their own algorithms.

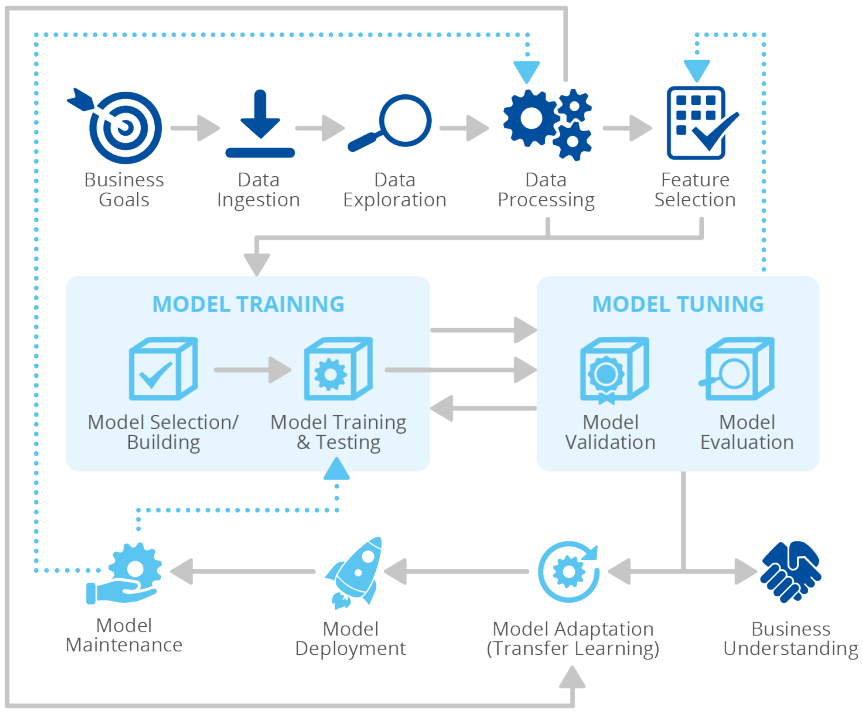
Empowering every employee to build and train AI algorithms will make it impossible to assess the trustworthiness of these algorithms in terms of transparency, ethical, data privacy, non-bias or governance pitfalls. The rise of no-code AI makes it imperative to develop strong auditing tools and policies around the use of AI and have systems in place to ensure that everyone using the no-code software understands and abides by these policies. **Advanced tools are needed to audit how these no-code AI models have been trained, in order to secure them by design**.

AI assets and procedures

The AI domain is broad and therefore requires a structured and methodical approach to understand its different facets. ENISA has proposed a generic reference model for a functional overview of typical AI systems[46]. However, due to the vast number of technologies, techniques and algorithms involved in these systems, mapping them all in a single life cycle would be too ambitious. ENISA then proposed a life cycle[47], illustrated in Figure 7, that is based on ML, as the particularities of the many subfields of AI – namely natural language processing, computer vision, robotics, etc. – make use of ML that has been spearheading the explosion of AI usage in different domains.

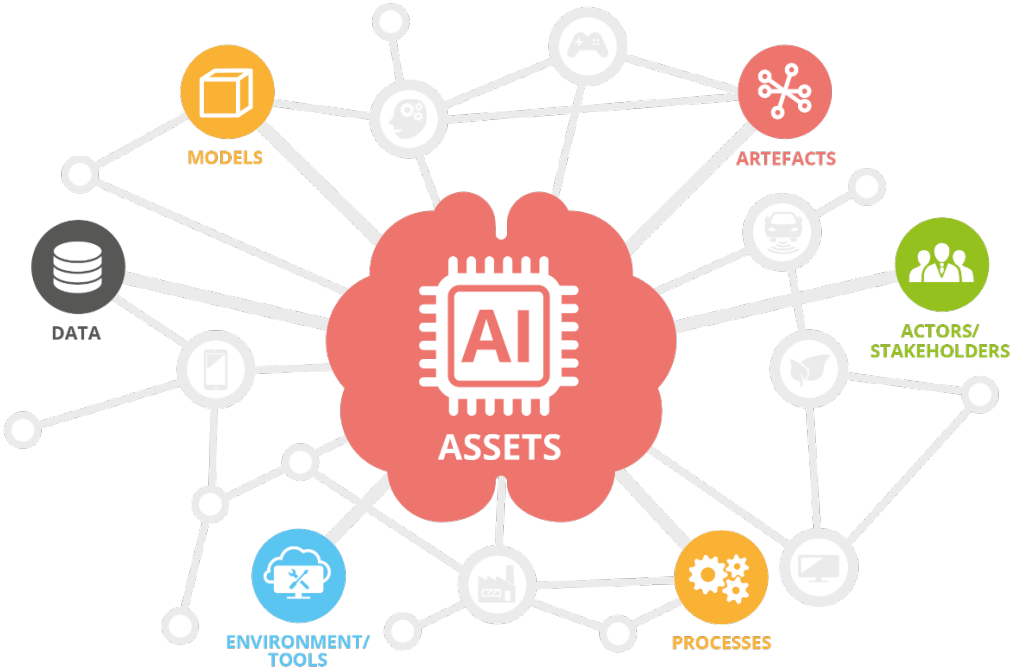
[46] ENISA, *AI Cybersecurity Challenges – Threat landscape for artificial intelligence*, 2020, https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.
[47] See footnote [43].

Figure 7: ENISA AI life cycle



In the same report, ENISA identified the most relevant assets[48], based on the functional description of specific stages and, in order to reflect AI components, also including assets that support the development and deployment of AI systems.

Figure 8: AI assets



- **Data.** Raw data, public data sets, training data, testing data, etc.
- **Models.** Algorithms, models, model parameters, hyper-parameters, etc.

[48] See footnote [43]

- **Artefacts.** Data governance policies, descriptive statistical parameters, model frameworks, etc.
- **Actors/stakeholders.** Data owners, data scientists, data engineers, model providers, etc.
- **Processes.** Data ingestion, data pre-processing, data collection, data augmentation, feature selection, training, tuning, etc.
- **Environment/tools.** Algorithm libraries, ML platforms, optimisation techniques, integrated development environments, etc.

AI threat assessment

AI systems greatly contribute to automate and enhance decision-making in a wide variety of day-to-day tasks, enhancing business processes all over the world. Nonetheless, as with any other ICT system, AI-powered ones can also be victims of cybercriminals and multiple cybersecurity threats (see Section 2.2) with the objective of hijacking their normal functioning for malicious purposes.

The additional required risk assessment efforts that are specific to AI must:

- include not only technical and physical threats, but also threats mentioned in the EU AI Act, such as loss of transparency, loss of interpretability, loss of managing bias and loss of accountability;
- enhance the types of impact factors, such as robustness, resilience, fairness and explainability;
- be dynamic and combined with anomaly detection approaches, as for ICT systems in general.

ETSI has published an AI threat ontology ([49]) to define what would be considered an AI threat and how it might differ from threats to traditional systems.

As explained in the NIST AI Risk Management Framework[50], AI systems are socio-technical in nature, meaning that the threats are not only technical, legal or environmental (as in typical ICT systems), but social as well. For example, social threats – such as bias, lack of fairness, lack of interpretability/explainability/equality – are directly connected to societal dynamics and human behaviour in all technical components of an AI system, and they can change during its life cycle. How these societal threats can impact individuals with different psychological profiles, groups, communities, organisations, democracies and society as a whole need to be analysed and measured before we estimate the risks. Actually, events that can compromise the characteristics of AI systems, as described in Figure 9 in the next section, are specific threats for AI systems which are social, policy and technical AI threats.

For example, bias is a new threat targeting the AI system and the different stages of the AI life cycle (design, development, deploying, monitoring and iteration), as analysed in the BSA framework[51]. The CEPS *Artificial Intelligence and Cybersecurity – Technology, governance and policy challenges* report[52] also provides an overview of the current threat landscape of AI, ethical implications and recommendations. The ARM framework ([53]) provides a simple interactive approach to explain the various principles of trustworthy AI. Additional AI-specific threats are described in more detail in Section 3.3 of this report.

The AI threats themselves can be of several types and affect all AI subfields. These can be mapped into a high-level categorisation of threats based on ENISA's threat ([54]) taxonomy, comprising:

- nefarious activity/abuse
- eavesdropping/intercept/hijacking
- physical attacks
- unintentional damage

[49] ETSI, *Securing Artificial Intelligence (SAI) – AI threat ontology*, Group report, DGR/SAI-001, 2022, https://www.etsi.org/deliver/etsi_gr/SAI/001_099/001/01.01.01_60/gr_SAI001v010101p.pdf.
[50] Tabassi, E., *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, Gaithersburg, MD, 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.
[51] https://www.bsa.org/reports/confronting-bias-bsas-framework-to-build-trust-in-ai.
[52] https://www.ceps.eu/wp-content/uploads/2021/05/CEPS-TFR-Artificial-Intelligence-and-Cybersecurity.pdf.
[53] https://interactive.arm.com/story/building-trustworthy-ai/page/3?utm_source=linkedin&utm_medium=social&utm_campaign=2022_client_mk04_arm_na_na_awa&utm_content=whitepaper.
[54] https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends/enisa-threat-landscape/threat-taxonomy/view.

- failures or malfunctions
- outages
- disaster
- legal.

On the other hand, ML-related threats can affect different steps of the ML life cycle. The most important high-level ML threats can be described as follows[55].

- **Evasion.** Evasion is a type of attack in which the attacker works with the ML algorithm input to find small perturbations which can be used to exploit the algorithm's output. The generated input perturbations are designated as adversarial examples.
- **Poisoning.** In a poisoning attack, the attacker alters the data or the model to modify the ML algorithm's behaviour in a chosen direction (e.g. to sabotage its results or to insert a back door) according to its own motivations.
- **Model or data disclosure.** This threat is related to the possible leaks of all or partial information about the model, such as its configuration, parameters and training data.
- **Compromise of ML application components.** This threat refers to the possible compromise of an ML component, for example by exploiting vulnerabilities in the open-source libraries used by the developers to implement the algorithm.
- **Failure or malfunction of an ML application.** This threat is related to the failure of the ML application. It can be caused by denial of service due to a bad input or by the occurrence of an untreated handling error.

All of these threats can be mapped to multiple vulnerabilities, such as lack of training based on adversarial attacks, poor control over which information is retrieved by the model, lack of sufficient data to withstand poisoning, poor access rights management, usage of vulnerable components and missing integration with the cyber resilience strategy.

In a report of a quantitative study with 139 industrial ML practitioners[56], despite most attacks being identified as related to the ICT infrastructure, some ML-related attacks were also identified. The number of reported AI threats was marginal, with 2.1 % of evasion attacks and 1.4 % of poisoning attacks recognised by the organisations.

AI security management

The RM conducted for an entire infrastructure (see Section 2.2) will need to be complemented with conducting RM in all AI systems hosted in the ICT infrastructure.

This section introduces AI properties and the security controls that can be employed to minimise the impact of AI threats aimed at compromising AI trustworthiness. The ISO 2700x[57] standards, the NIST AI framework[58] and ENISA's best practices can all be used for AI RM and it is strongly recommended that they be followed when implementing more general-purpose security controls.

AI trustworthiness

In order to understand the concepts and risks associated with the usage of AI, it is important to start by analysing the level of trustworthiness and the desirable properties to consider. We define AI trustworthiness as the confidence that AI systems will behave within specified norms, as a function of some characteristics such as: accountability, accuracy, explainability, fairness, privacy, reliability, resiliency/security, robustness, safety and transparency.

In this section, an overview of these characteristics is provided, along with their relationships with the risk assessment framework based on NIST[59].

- **Accountability.** Ensures responsibility for AI, which in turn implies explanation

[55] For additional information on ML-specific threats and security controls, see: ENISA, *Securing Machine Learning Algorithms*, 2021, https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms.
[56] Grosse, K. et al., "'Why do so?" – A practical perspective on machine learning security', Cornell University, 2022, arXiv:2207.05164.
[57] https://www.iso.org/search.html?q=27000.
[58] https://www.nist.gov/itl/ai-risk-management-framework.
[59] See footnote [58]

and justification; humans and organisations should be able to answer and be held accountable for the outcomes of AI systems, particularly adverse impacts stemming from risks.

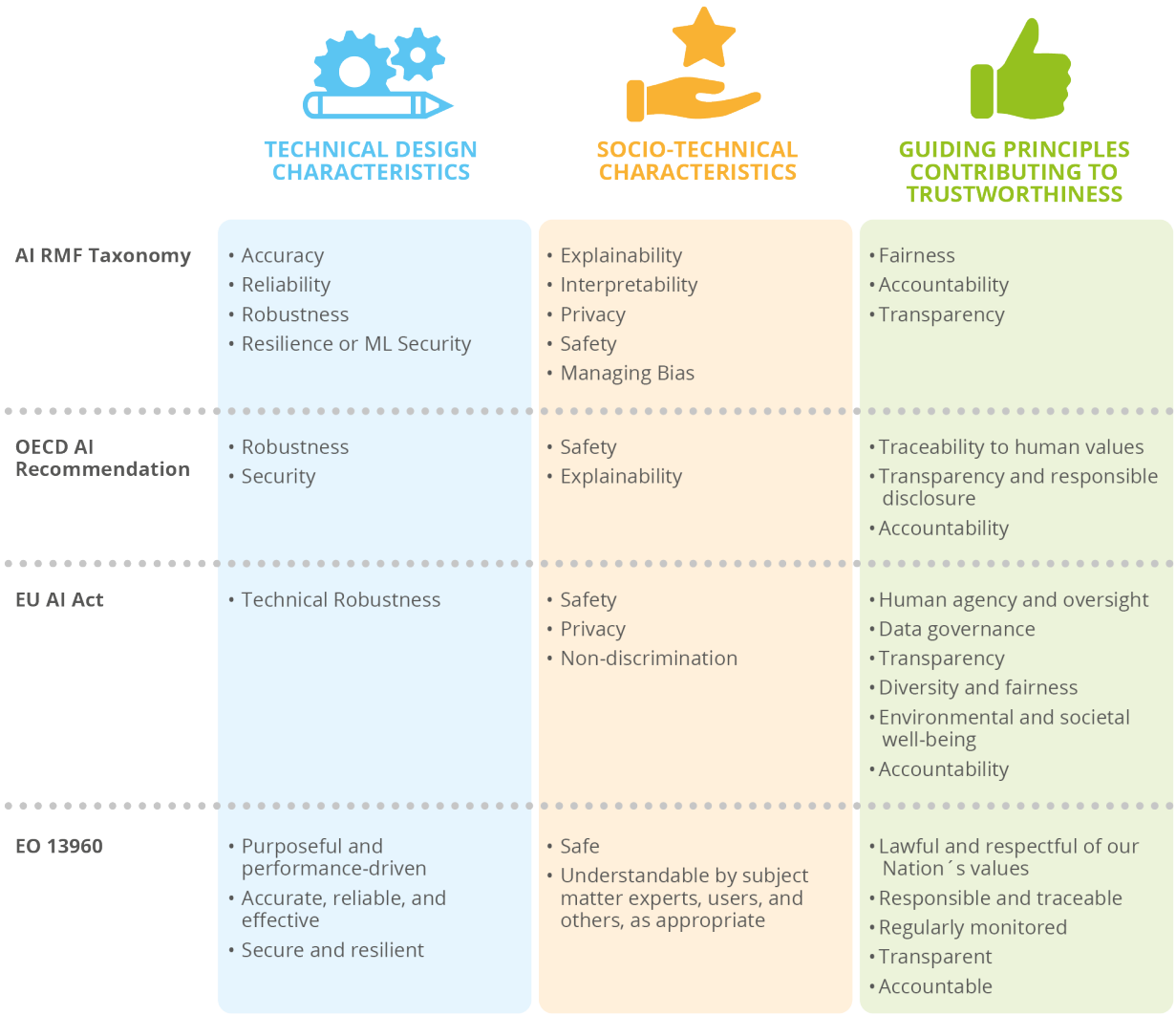
- **Accuracy.** Correctness of output compared with reality; RM processes should consider the potential risks that might arise if the underlying causal relationship inferred by an AI model is not valid.
- **Explainability.** Provides a description of the conclusion/decision made in a way that can be understood by a human; risks due to explainability may arise for many reasons, including for example a lack of fidelity or consistency in explanation methodologies, or if humans incorrectly infer a model's operation, or the model is not operating as expected.
- **Fairness.** Neutrality of evidence, not biased by personal preferences, emotions or other limitations introduced by the context, equality (of gender and opportunity). Fairness is a concept that is distinct from but related to bias. According to ISO/IEC TR 24027:2021, bias can influence fairness. Biases can be societal or statistical, can be reflected in or arise from different system components and can be introduced or propagated at different stages of the AI development and deployment life cycle.
- **Privacy.** Secure management (process, analysis, storage, transport, communication) of personal data and training models; ability to operate without disclosing information (data, model); identifying the impact of risks associated with privacy-related problems is contextual and varies among cultures and individuals.
- **Reliability.** Ability to maintain a minimum performance level and consistently generate the same results within the bounds of acceptable statistical errors; may give insights about the risks related to decontextualisation.
- **Resiliency.** Ability to minimise impact, restore safe operating conditions and come out hardened from an adversarial attack.
- **Robustness.** Ability of an AI system to maintain a previously agreed minimum level of performance under any circumstances; this contributes to sensitivity analysis in the AI RM process.
- **Safety.** Preventing unintended or harmful behaviour of the system to humans or society; safety is highly correlated to risks.
- **Security.** Ability to prevent deviations from safe operating conditions when undesirable events occur; ability to resist attacks; ensures confidentiality, integrity, authenticity, non-repudiation, availability of data, processes, services and models.
- **Transparency.** Ability to foster a general understanding of AI systems, make stakeholders aware of their interactions with AI systems and allow those affected by an AI system to understand the outcome. It also enables those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

The NIST AI framework organises these characteristics in **three classes** (**technical, socio-technical and guiding principles)** and provides a **mapping of the taxonomy to AI policy documents**[60], as can be seen in Figure 9. The technical characteristics in the framework taxonomy refer to factors that are under the direct control of AI system designers and developers and which may be measured using standard evaluation criteria.

At this level, properties like accuracy, reliability, robustness and security are referred to in most of the documents. Socio-technical characteristics in the taxonomy refer to how AI systems are used and perceived in individual, group and societal contexts. At this level, the focus is on safety, explainability and privacy. The guiding principles in the taxonomy refer to broader societal norms and values that indicate societal priorities, where fairness, accountability, transparency and traceability are the most highlighted.

[60] See footnote [59]

Figure 9: AI characteristics mapping to policy documents



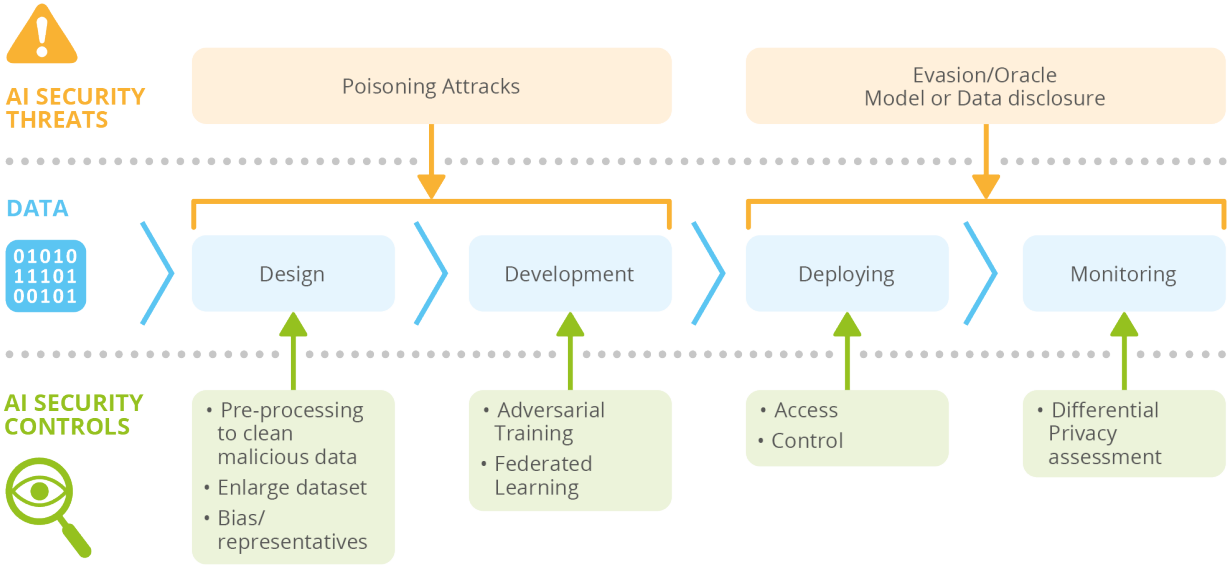
Security controls

On the other hand, specific ML security controls can be mapped for the introduced threats to provide efficient ways of prevention and mitigation. For **evasion**, tools can be implemented to detect whether a given input is an adversarial example, adversarial training can be used to make the model more robust, and models that are less easily transferable can be used to significantly decrease the ability of a given attacker to properly study the algorithm that works underneath the system.

Similarly, for **poisoning** attacks, processes that maintain the security levels of ML components over time should be implemented, the exposure level of the used model should be assessed, the training data set should be enlarged as much as possible to reduce its susceptibility to malicious samples, and pre-processing steps that clean the training data from such malicious samples must also be considered.

Model or data disclosure can be protected by applying proper access control and federated learning to minimise the risk of data breaches. Similarly, to reduce the level of compromise of ML application components, these should be compliant with protection policies, fully integrated to existing security operations and asset management processes, and evaluated according to the level of security of their foundation blocks (e.g. libraries that are responsible for the algorithm implementation). Finally, to prevent failure or malfunction of ML applications, employed algorithms should have their bias reduced, should be properly evaluated to ensure that they are resilient to the environment in which they will operate and should encompass explainability strategies.

Figure 10: The relation between AI threats and security controls



Security testing

In 2022, the ETSI working group on AI published draft of *Security Testing of AI*[61]. This report identifies methods and techniques that are appropriate for security testing of ML-based components. The scope of this report covers the following elements.

- Security testing approaches for AI (used to generate test cases that are executed against the ML component).
- Security test oracles for AI (enable the calculation of a test verdict to determine whether a test case has passed, i.e. no vulnerability has been detected, or failed, i.e. a vulnerability has been identified).
- Definitions of test adequacy criteria for security testing (used to determine the overall progress and can be employed to specify a stop condition for security testing).

According to the report, security testing of AI does not end at the component level. As for testing of traditional software, its integration with other components of a system needs to be tested as well.

Security testing of AI has some commonalities with security testing of traditional systems, but also provides new challenges and requires different approaches, due to:

- significant differences between subsymbolic AI and traditional systems that have strong implications on their security and on how to test their security properties;
- non-determinism that may result from self-learning, i.e. AI-based systems may evolve over time and as a consequence, security properties may degrade;
- the test oracle problem, where assigning a test verdict is different and more difficult for AI-based systems, since not all expected results are known a priori;
- data-driven algorithms, where in contrast to traditional systems, (training) data forms the behaviour of subsymbolic AI.

AI-related standards

Several initiatives are underway to provide standards and specific guidelines for AI security and trustworthiness: ISO/IEC is working on RM, trustworthiness and management systems; ETSI provides an AI threat ontology and data supply chain security, among other items; and IEEE is working on AI explainability. In this section, a list of the current available standards and initiatives is presented. The reader can find a list with AI-related standards in Annex II.

Ethical and trustworthy AI

Besides the standardisation organisations, other groups are also working on guidelines for ethical and trustworthy AI. The following list shows some examples that we identified during our desktop research:

[61] https://portal.etsi.org/webapp/WorkProgram/Report_WorkItem.asp?WKI_ID=58860

- *Ethics Guidelines for Trustworthy Artificial Intelligence*[62];
- Data Ethics of Power. A Human Approach in the Big Data and AI Era[63];
- White Paper on Data Ethics in Public Procurement of AI based Services and Solutions[64];
- 5 things lawyers should know about artificial intelligence[65];
- How brain-inspired technologies can support ethical AI[66].

Tools

In addition to legislation and standards, we have identified other initiatives and tools that focus on a more practical approach to assess and guide the achievement of AI security and risk assessment.

- ***The Assessment List for Trustworthy Artificial Intelligence***[67] is a practical tool that helps businesses and organisations to self-assess the trustworthiness of their AI systems under development. It was developed by the High-Level Expert Group on Artificial Intelligence in the *Ethics Guidelines for Trustworthy Artificial Intelligence* report, which provides a detailed assessment list.
- The **OECD**[68] provides a classification of AI systems and tools for developing trustworthy AI systems.
- **MITRE ATLAS**[69] is a knowledge base of adversary tactics, techniques and case studies for ML systems based on real-world observations, demonstrations from ML red teams and security groups, and the state of the possible from academic research. ATLAS is modelled after the MITRE ATT&CK framework and its tactics and techniques are complementary to those in ATT&CK.
- **AI security risk assessment**[70]. Counterfit is an automation tool for security testing AI systems as an open-source project. Counterfit helps organisations conduct AI security risk assessments to ensure that the algorithms used in their businesses are robust, reliable, and trustworthy. This tool was provided by Microsoft as a means to automate techniques in MITRE's Adversarial ML Threat Matrix.
- **GuardAI**[71] is a platform for evaluating AI model robustness against adversarial attacks and natural noises. The goal of GuardAI is to simulate adversarial and malicious inputs, which can fool AI models and force AI applications to make wrong predictions. GuardAI supports different techniques of adversarial attacks, noise, domain adaptation simulation, popular ML frameworks and main computer vision tasks.

Networks and initiatives

Apart from the above, various Commission research projects related to AI can also be useful as they promote AI values, such as trustworthiness and responsibility, and bring together different AI stakeholders from research to business. Below are examples of the initiatives and networks, we have identified during our market analysis, however this is by no means a complete list as new ones are created on ongoing basis.

- **The European AI on demand** platform[72] is a facilitator of knowledge transfer from research to multiple business domains. The platform serves as a catalyst to aid AI-based innovation, resulting in new products, services and solutions to benefit European industry, commerce and society. The platform aims to create value, growth, and jobs in Europe through an ecosystem and a collaborative platform that unites the AI community, promotes European values and supports research on human-centred

[62] European Commission, *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, Brussels, 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
[63] Hasselbalch, G., Data *E*thics of *P*ower – *A human approach in the big data and AI era*, Edward Elgar Publishing, Cheltenham, 2021, https://www.e-elgar.com/shop/gbp/data-ethics-of-power-9781802203103.html.
[64] Hasselbalch, G., Kofod Olsen, B. and Tranberg, P., *White paper on data ethics in public procurement of AI-based services and solutions*, DataEthics.eu, Denmark, 2020, https://dataethics.eu/wp-content/uploads/dataethics-whitepaper-april-2020.pdf.
[65] Leong, B., Hall, P., '5 things lawyers should know about artificial intelligence', Mind Your Business column, ABA Journal, 2021, https://www.abajournal.com/columns/article/5-things-lawyers-should-know-about-artificial-intelligence.
[66] Shea, T., 'How brain-inspired technologies can support ethical AI', LinkedIn post, 2021, https://www.linkedin.com/pulse/how-brain-inspired-technologies-can-support-ethical-ai-tim-shea.
[67] https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence.
[68] OECD, *OECD Framework for the Classification of AI Systems*, OECD digital economy papers, No 323, 2022, https://oecd.ai/en/classification.
[69] https://atlas.mitre.org/.
[70] https://github.com/Azure/counterfit.
[71] https://www.navinfo.eu/services/cybersecurity/guardai/.
[72] https://www.ai4europe.eu/.

and trustworthy AI.

- **ALLAI**[73] is an independent Dutch organisation dedicated to drive and foster responsible AI. ALLAI's vision is responsible AI for a world where AI is developed, deployed and used responsibly, i.e. in a safe and sustainable manner and in line with ethical principles, societal values, existing and new laws and regulations, human rights, democracy and the rule of law.
- The **Confederation of Laboratories for Artificial Intelligence Research in Europe**[74] seeks to strengthen European excellence in AI research and innovation. The network forms a pan-European Confederation of Laboratories for Artificial Intelligence Research in Europe. Its member groups and organisations are committed to work together towards realising the vision of CLAIRE: European excellence across all of AI, for all of Europe, with a human-centred focus.
- **The European Network of Human-centred AI**[75] aims to facilitate a European brand of trustworthy, ethical AI that enhances human capabilities and empowers citizens and society to effectively deal with the challenges of an interconnected globalized world.
- The **D-seal** initiative[76] is the new labelling programme for IT security and responsible use of data in Denmark. It provides interesting guidelines and criteria on how to combine IT security and responsible use of data in the same label, with AI as one of the criteria.
- **AI testing and experimentation facilities** have been included by the Commission in the Digital Europe Programme. These are meant to be large-scale reference sites for 'testing state-of-the-art AI-based soft and hardware solutions and products'[77].

New challenges

The security of AI should be considered at all stages of its life cycle, taking into account the following elements.

- AI systems are multi-disciplinary socio-technical systems and their threats are technical, societal, ethical and legal. Collaboration between cybersecurity experts, data scientists, social scientists, psychologists and legal experts is needed in order to identify the continuous evolving AI threat landscape and develop corresponding countermeasures.
- Among the different types of AI described in Section 2.3.3, ML and DL undoubtedly pose the main challenges to security and imply a dynamic analysis of the threats, both along the life cycle and in the interrelations within other blocks of an ICT infrastructure.
- AI-specific risk assessment efforts need to consider their unique properties and enhance their robustness, resilience, fairness and explainability, along with preventing loss of transparency, loss of managing bias and loss of accountability.
- Assigning a test verdict is different and more difficult for AI-based systems, since not all of the expected results are known a priori.

2.3. LAYER III – SECTOR-SPECIFIC CYBERSECURITY GOOD PRACTICES

AI is a technology that has entered all economic sectors (e.g. automotive, health, maritime, finance). The third layer of the FAICP framework provides additional recommendations and best practices available in order to address cybersecurity issues in the AI systems used in some of these sectors.

While almost every economic sector already relies on AI systems, we have identified below only those sectors for which we managed to find relevant cybersecurity guidelines. Additionally, ENISA's reports can be used to identify sectoral threats (e.g. 5G, AI, supply chain).

Energy

[73] https://allai.nl/.
[74] https://claire-ai.org/.
[75] https://www.humane-ai.eu/.
[76] https://d-seal.eu/.
[77] https://digital-strategy.ec.europa.eu/en/activities/testing-and-experimentation-facilities.

The emergence of industrial automation and control systems, AI, smart grids and autonomous devices have made the energy sector a target for cyberattacks, while the existing interconnectivity and the rapidly complexity of the underlying infrastructures increase the security threats and their cascading effects.

The energy sector uses different IoT devices (e.g. miniaturised sensors to monitor transmission pipelines); drilling rigs and robots to inspect and repair infrastructure; virtual power plants, microgrids or cloud management services for solar, building automation; new applications with close integration of demand and response providing unparalleled flexibility; expanded telecommunication infrastructures and networks with increased usage of mobile devices. However, all of these technologies (mostly from foreign manufacturers) have many vulnerabilities and a high number of potential attack points, increasing the cybersecurity challenge in the clean energy industry, as described in the 2019 ENISA report *Industry 4.0 Cybersecurity: Challenges and recommendations*[78].

Therefore, operators, stakeholders and networks must urgently focus on security as part of their ICT and IT infrastructure, in order to enhance their information security and privacy practices and address the origins of their main security problems. These may include remote work during operations and maintenance, using technologies with known vulnerabilities, new highly-interconnected services, a limited cybersecurity culture among vendors, suppliers and contractors, data networks between on- and offshore facilities and outdated control systems in facilities.

The following best practices can provide guidance to AI stakeholders in the energy sector:

- *Cybersecurity in the energy sector*[79];
- *Transforming the energy industry with AI*[80];
- *Artificial Intelligence for Energy Systems Cybersecurity* (The National Renewable Energy Laboratory report)[81];
- ENISA report *EU Cybersecurity Market Analysis – IoT in distribution grids*[82].

Health

Many medical devices – from glucose meters, insulin pumps, virtual home assistants and cardioverter defibrillators to smart wearable devices, sophisticated software and hospital equipment, along with medical services and applications – are connected over the network and often use AI technologies. Although new connected medical devices help in fighting the increasing cost of healthcare – by reducing the need for hospitalisation, developing personalised therapies and creating intelligent point-of-care diagnostic tools – they also introduce new cybersecurity risks and their interoperability, security and resilience levels are considered to be low. Recently, an attack crippled more than 400 hospitals across Puerto Rico, the United Kingdom and the United States ([83]).

There are three **primary attack vectors through which connected medical devices might be compromised.**

- **Devices.** Cybercriminals exploit device vulnerabilities that exist in their memory, firmware, physical interface, web interface or network services. Other aspects such as unsecure default settings, outdated components and unsecure update mechanisms can also be exploited. Outdated legacy devices are the main targets, due to their unpatched implemented vulnerabilities.
- **Communication channels.** A device can be compromised by attacking the channels used to connect it with another device. In this vector, spoofing and denial of service attacks are common. Conventional wireless sensor networks consist of wireless nodes equipped with antennas, which broadcast radio signals in all directions and are consequently prone to eavesdropping attacks. An attacker can use this data to introduce themselves as an authorised member to launch an impersonation attack. Thus, eavesdropping is very simple for an attacker while the patient data is transmitting

[78] ENISA, Industry 4.0 Cybersecurity: Challenges and recommendations, 2019, https://www.enisa.europa.eu/publications/industry-4-0-cybersecurity-challenges-and-recommendations.
[79] ENERGY EXPERT CYBER SECURITY PLATFORM, Cyber Security in the Energy Sector, February 2017, https://energy.ec.europa.eu/system/files/2017-03/eecsp_report_final_0.pdf
[80] MIT Technology Review Insights, Transforming the Energy Industry with AI, 2021, https://assets.siemens-energy.com/siemens/assets/api/uuid:4b6f1e50-6639-4cb9-8a5d-85ac8e29c807/siemensreport-v10.pdf?ste_sid=88ed48911b29356753651e2fd4237fae.
[81] Macwan, R., King, R., Artificial Intelligence for Energy Systems Cybersecurity, National Renewable Energy Laboratory, NREL/PR-5R00-81098, 2021, https://www.nrel.gov/docs/fy22osti/81098.pdf.
[82] ENISA, EU Cybersecurity Market Analysis – IoT in distribution grids, 2022, https://www.enisa.europa.eu/publications/eu-cybersecurity-market-analysis-iot-in-distribution-grid.
[83] Wired, A Ransomware Attack Has Struck a Major US Hospital Chain, 2020, https://www.wired.com/story/universal-health-services-ransomware-attack/.

from the body area network to the caregiver device. Hence, patient privacy is breached.

- **Applications and software.** Cybercriminals can exploit vulnerabilities in web applications and related software for connected devices. For example, web applications can be targeted to steal user credentials or push malware. There is an urgent need to provide a solution where manufacturers can easily identify, estimate, mitigate and audit by design all cybersecurity risks of connected devices (hardware, software and integrated medical frameworks consisting of various modular components), in order to ensure their security and resilience and progress towards a resilient and trustworthy EU healthcare ecosystem.

Regulators around the globe have increasingly pursued medical device cybersecurity as a policy objective over the past few years. In the EU, the first piece of guidance on cybersecurity on medical devices (MDCG-2019-16)[84] was issued in July 2020 by the EU's Medical Devices Coordination Group.

The EU has included the health sector among its critical information infrastructures and is developing cybersecurity legislation and directives that impose cybersecurity and privacy RM (e.g. GDPR, NIS), supply chain security (e.g. NIS 2), secure authentication and access of healthcare e-services (e.g. eIDAS) and cybersecurity certification (e.g. CSA, AI liability directive, European Chips Act).

The following best practices can provide guidance to AI stakeholders in the healthcare sector.

- Definitions/Characteristics of Artificial Intelligence in Health Care (ANSI/CTA-2089.1)[85]
- *Whitepaper for the ITU/WHO focus group on artificial intelligence for health*[86],
- ENISA report *Smart Hospitals – Security and resilience for smart health service and infrastructures*[87]
- ENISA report *Deploying Pseudonymisation Techniques – The case in the health sector*[88].

Automotive

New generations of cars are making use of advances in the field of AI. Autonomous vehicles are systems that rely on autonomous driving capabilities using AI on a perception–planning–control pipeline. Designing an is a challenging problem that requires tackling a wide range of environmental conditions (lightning, weather, etc.) and multiple complex tasks. These include road following, obstacle avoidance, abiding with traffic laws, smooth driving style, manoeuvre coordination with other elements of the ecosystem (e.g. vehicles, scooters, bikes, pedestrians) and control of the commands of the vehicle.

The joint ENISA/JRC report *Cybersecurity challenges in the uptake of artificial intelligence in autonomous driving*[89] analyses cybersecurity vulnerabilities related to AI, identifies related challenges and provides recommendations for securing autonomous vehicles. Five hypothetical scenarios are presented to illustrate the exploitation of AI vulnerabilities in an automotive context, using both classical cybersecurity and AI-specific vulnerabilities:

- adversarial perturbations against image processing models for street sign recognition and lane detection;
- man-in-the-middle attacks on the planning module;
- data poisoning attacks on stop sign detection;
- attacks related to large-scale deployment of rogue firmware after hacking backend servers of original equipment manufacturers;

[84] MDCG 2019-16 Guidance on Cybersecurity for medical devices, 2019, https://ec.europa.eu/docsroom/documents/41863/attachments/1/translations/en/renditions/native.
[85] ANSI, *ANSI/CTA-2089.1-2020 – Definitions/characteristics of artificial intelligence in health care*, 2020, https://webstore.ansi.org/Standards/ANSI/ANSICTA20892020.
[86] Wiegand, T., Lee, N., Pujari, S., Singh, M., Xu, S., Kuglitsch, M., Lecoultre, M., Riviere-Cinnamond, A., Weicken, E., Wenzel, M., Werneck Leite, A., Campos, S. and Quast, B., *Whitepaper for the ITU/WHO focus group on artificial intelligence for health*, Focus Group on Artificial Intelligence for Health, ITU and WHO, 2023, https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FG-AI4H_Whitepaper.pdf.
[87] ENISA, *Smart Hospitals – Security and resilience for smart health service and infrastructures*, 2016, https://www.enisa.europa.eu/publications/cyber-security-and-resilience-for-smart-hospitals.
[88] ENISA, *Deploying Pseudonymisation Techniques – The case of the health sector*, 2022, https://www.enisa.europa.eu/publications/deploying-pseudonymisation-techniques.
[89] Dede, G., Hamon, R., Junklewitz, H., Naydenov, R., Malatras, A. and Sanchez, I., *Cybersecurity challenges in the uptake of artificial intelligence in autonomous driving*, ENISA and Joint Research Centre, Publications Office of the European Union, Luxembourg, 2021, https://www.enisa.europa.eu/publications/enisa-jrc-cybersecurity-challenges-in-the-uptake-of-artificial-intelligence-in-autonomous-driving/.

- attacks related to sensor/communication jamming and global navigation satellite system spoofing.

In the 2019 report *ENISA Good Practices for Security of Smart Cars*[90], security measures against AI vulnerabilities, such as being tricked by adversarial attacks and data falsification/manipulation, were already identified.
The International Telecommunication Union (ITU) Focus Group on AI for autonomous and assisted driving supports standardisation activities for services and applications enabled by AI systems[91]. The group focuses on the behavioural evaluation of AI responsible for the dynamic driving task in accordance with the 1949 and 1968 Convention on Road Traffic of the UNECE Global Forum for Road Safety. In 2021, the group also published the report *FGAI4AD-02 – Automated driving safety data protocol – Ethical and legal considerations of continual monitoring*[92].

Telecommunications

While modern networks are becoming more sophisticated, the telecommunications industry can benefit from data recovered from networks, mobile applications, customer insight, profile, technology, billing data and services through the integration of AI and help the industry in self-optimising networks, security and predictive measures. AI use cases related to telecom include the following.

- **Network optimisation.** Networks are managed by AI systems and ML algorithms that predict and detect network abnormalities. AI is also used to optimise and configure various networks, so that it is easy for end users to leverage the advantage of stable network performance.
- **Virtual assistants and chatbots.** The telecommunications industry is leveraging the power of AI to implement chatbots and virtual assistants, which can deliver round-the-clock support and assistance to customers without any waiting time.
- **Predictive maintenance.** AI-enabled predictive analytics is helping the telecom sector to maintain high levels of service and products to customers.
- **Security and fraud detection.** ML algorithms are used to detect and prevent fraudulent activities. AI-driven alerts can notify customers and telecom operators in real time.

The ITU Focus Group on Machine Learning for Future Networks including 5G has published a technical specification on unified architecture for ML in 5G and future networks[93]. The presented logical architecture establishes a common vocabulary and nomenclature for ML functions and their interfaces to allow standardisation and interoperability for ML functions in 5G and future networks.

The Dutch Radiocommunications Agency published the report *Managing AI use in telecom infrastructures – Advice to the supervisory body on establishing risk-based AI supervision*[94], which addresses the current and future risks of applying AI in the telecom sector, along with their supervision and ways to mitigate them.

New challenges

Horizontal threats and cybersecurity challenges exist in every economic sector (automotive, energy, health, etc.), independently of how AI is being used. Fragmented recommendations, best practices, solutions and tools for horizontal issues become stumbling blocks for guiding sectoral stakeholders. Collaboration among sectoral stakeholders and information sharing and analysis centres (ISACs) is recommended to best address horizontal challenges. Sector-specific issues and mitigation measures need to be listed and published to serve as 'lessons learned' for other sectors.

[90] https://www.enisa.europa.eu/publications/smart-cars.
[91] https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/default.aspx.
[92] ITU, *Automated driving safety data protocol – Ethical and legal considerations of continual monitoring*, Focus Group on AI for autonomous and assisted driving (FG-AI4AD), Technical Report, 2021, https://www.itu.int/pub/T-FG-AI4AD-2021-02.
[93] ITU, *Unified architecture for machine learning in 5G and future networks*, Focus group on Machine Learning for Future Networks including 5G (FG-ML5G), Technical Specification, 2019, https://www.itu.int/dms_pub/itu-t/opb/fg/T-FG-ML5G-2019-PDF-E.pdf.
[94] van der Vorst, T., Jelicic, N., van Rees, J., Bekkers, R., Brennenraedts, R. and Bakhyshov, R., *Managing AI use in telecom infrastructures – Advice to the supervisory body on establishing risk-based AI supervision*, Dialogic innovatie & interactie, Utrecht, 2020, https://www.dialogic.nl/en/projects/managing-ai-use-in-telecom-infrastructures/.

3. SURVEY ANALYSIS

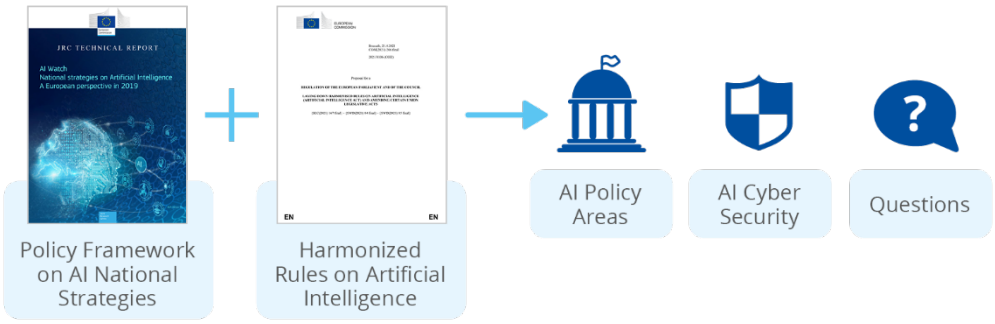
The proposal for the AI Act[95] regulation contains various requirements for providers and operators of AI systems, including cybersecurity requirements for high-risk AI systems. The objective of the survey was to collect information from the EU NCAs about existing national cybersecurity requirements for AI, to determine how compliance with these requirements is monitored and enforced nationally and to evaluate whether certain practices/requirements established under the AI Act have already been implemented at the MS level.

In this section of the report, the methodology used to develop the survey is described and the results of the survey are analysed. The complete questionnaire is available in Annex I.

3.1. METHODOLOGY

The questionnaire is based on the cybersecurity concepts described in the FAICP framework, on the main principles (related to cybersecurity) of the proposed AI Act and the Coordinated Plan on AI[96] and on the harmonised rules related to cybersecurity as reported in the explanatory memorandum.

Figure 11: Methodology to structure the questionnaire



Based on this, we identified and structured the survey according to five policy areas identified in the AI national strategies.

- **Human capital** (education, training and lifelong learning / labour market / intelligence and skills demand). This area targets all policies to foster the educational development of people. The focus of the survey is to identify whether and how AI security is being considered in all forms of education and existing awareness initiatives, and about hands-on skills and practical capabilities about AI.
- **From the lab to the market** (R & D/innovation/testing). This encompasses policy initiatives to encourage research and innovation in AI towards business growth in the private sector and increased efficiency of public services. The focus here is to understand national funding and research activities on the security of AI, along with sandboxes, cyber-ranges, simulation, testing environments and methodologies and tools.
- **Networking** (collaboration/dissemination and uptake). This includes policy initiatives related to AI collaborations across private and/or public sectors and directed to increasing the (inter)national attractiveness of the country. It also includes policies related to the dissemination and uptake of AI.

[95] See footnote [1].
[96] See footnote [4].

- **Infrastructure** (digital and telecom/data). This area highlights policies for the development of ethical guidelines, legislative reforms and (international) standardisation.
- **Regulation** (NLF legislation and trustworthy frameworks / AI standards / compliance with the GDPR). The focus here is on policies for the development of ethical guidelines, legislative reforms and (international) standardisation.

In Annex I, Table 2 provides a detailed overview of the policy areas under consideration, the associated types of public and private initiatives and their relation to cybersecurity in AI.

3.2. SURVEY ANALYSIS

The survey was distributed to NCAs that deal with cybersecurity and/or AI. We received 10 responses to the survey, which are analysed below. The survey contained 30 questions, of which 14 were mandatory, organised within the five policy areas mentioned above. **The mandatory questions are marked with (M).**

Human capital

Cybersecurity plays a crucial role in ensuring that AI systems are resilient against attempts to alter their use, behaviour, performance or compromise their security properties by malicious third parties exploiting the systems' vulnerabilities. Raising practical skills and capabilities in handling emerging AI cyber threats and challenges is important in the future development of AI systems.

(1) Have you built / do you plan to build synergies with educational authorities/institutions to increase AI cybersecurity capabilities at all levels of education? If yes, please elaborate. (M)

At this level, most of the MS mention formal and informal collaboration between different entities, such as universities, computer societies, centres for cybersecurity and legal authorities, on the promotion of cybersecurity capabilities. However, most of them do not yet have AI security topics ready.
Concerning AI-specific security, one of the MS mentioned an AI observatory that is part of the national strategy for AI and that will bring together all levels of education on AI, including security. Another MS reported on studies conducted in collaboration with universities that focus on analyses of legal aspects of AI and its cybersecurity implications, in particular the usage of AI to carry out cyberattacks.

(2) Do you offer awareness raising campaigns about the secure development and use of AI solutions? If yes, please elaborate. (M)

The MS recognise AI as an emerging and disruptive technology. They encourage public discussion on AI security and support promoting safe, trustworthy and democratic AI which respects the rights and welfare of humans and EU principles. No specific campaigns have been conducted, but two MS mentioned the regular publication of studies and white papers related to this topic. Another MS has already published guidance on security of AI, including a label for digitally responsible businesses to cover cybersecurity, privacy and trustworthy AI.

(3) Do you provide guidance and best practices on how to improve AI security? If yes, please elaborate. (M)

Four MS mentioned related measures, such as: (i) the collection of information on successful examples and best practices of using AI both in the private and public sectors, as well as information on the impact of AI activities on the fundamental rights of natural persons; and (ii) the publication of rules for governments on how to maintain and develop emerging and disruptive technologies without national security disruption. More thorough initiatives were also reported: two MS provide guidance on the security and RM measures during the AI life cycle. One of these reports was also disseminated as a webinar that reached over 500 people in this MS. One MS supports organisations by providing a self-assessment tool for AI security.

(4) Do you consider AI cybersecurity in syllabus of courses dedicated to AI or to cybersecurity? If yes, please elaborate.

(5) Do you offer practical trainings to the AI stakeholders and can elaborate to which stakeholders? If yes, please elaborate.

For these two questions, none of the 10 MS reported on ongoing specific modules on AI cybersecurity, although these concerns are expected to be addressed in AI courses. Regarding the practical trainings, self-learning material was mentioned by one MS.

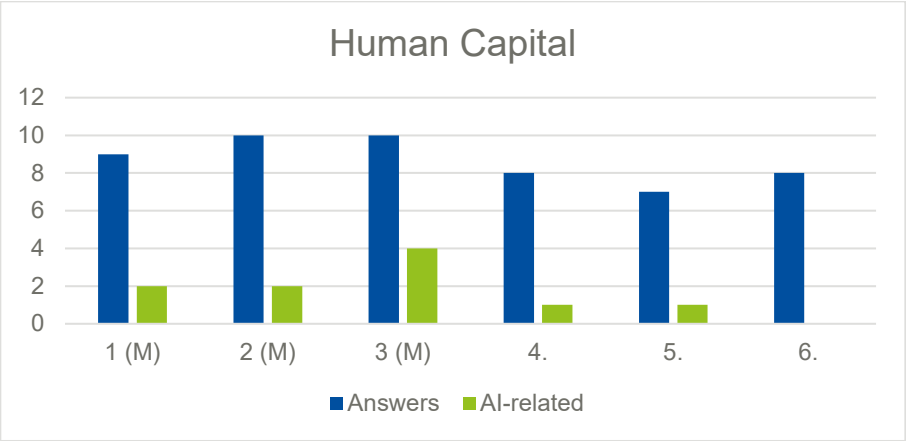
(6) Do you organise national, regional and cross-border cybersecurity exercises enabling the upskilling of the AI stakeholders? If yes, please elaborate.

Despite existing cross-border cybersecurity exercises, nothing specific to AI security was reported in response to this question.

Conclusions

MS recognise the need to consider the security of AI at all levels of education and a good engagement with the educational community on cybersecurity topics. However, they do not yet provide specific courses or training dedicated to AI security. Regarding awareness campaigns and guidance, it is worth mentioning that one MS mentioned a label for digitally responsible businesses to cover both cybersecurity, privacy, and trustworthy AI, while another mentioned a report on self-assessment guidance for AI/ML risk over the life cycle. Figure 12 illustrates this tendency.

Figure 12: Overview of AI-related 'Human capital' answers



From the lab to the market

According to the coordinated plan on AI, supporting AI research and innovation related to threats and attacks on AI and offering solutions for testing promising AI solutions is critical to ensure cybersecurity obligations in the uptake of developments from the lab to the market.

(7) What type of support (funding/scholarships/collaboration opportunities) do you offer to increase the cybersecurity capabilities of newly innovative solutions that rely on AI? (M)

Two MS mentioned projects where AI is used as an enabler of cybersecurity and one MS mentioned involvement on research initiatives about the secure use of AI. Two other MS reported on the opening of calls regarding AI.

(8) Have you informed national AI stakeholders on cybersecurity requirements set by the NCAs for their AI products and how do you do it? (M)

Most of the MS are not yet applying any information regarding cybersecurity requirements for AI products and expect the AI Act to provide guidance on how to do it. One of the MS mentioned some workshops and labs with this aim, while another mentioned the existence of a direct contact with national and international stakeholders to inform them about new requirements.

(9) How do you monitor if such requirements have been met? (M)

There were no answers relating to monitoring whether external AI solutions meet security requirements, however one MS mentioned that this is done by external companies.

(10) What are the means that you use to support SMEs/MEs to secure their AI products?

Aligned with the answers from question 3, under the topic of human labour, two MS provide guidance about AI security awareness, although used voluntarily, and the self-assessment mode regarding risks through the entire AI life cycle. Another MS related that, besides studies and white papers, there is also direct communication to discuss new trends and developments in the area. However, this MS does not provide any kind of classical consulting or assistance during development.

(11) Do you have/promote testing environments, like sandboxes/cyber ranges/simulation platforms to test and evaluate AI vulnerabilities before market? How?

Three MS mentioned the usage of sandboxes, cyber ranges and test platforms, where companies can test their solutions, however these solutions are of a general ICT nature and not AI-specific.

(12) Do you have specific measurements / key performance indicators (KPIs)/metrics that the AI stakeholders are required to use?

Nothing specific to AI was mentioned, with some MS mentioning the lack of experience on security issues related to AI, and others mentioning the expectation about the AI Act to provide guidance in this area.

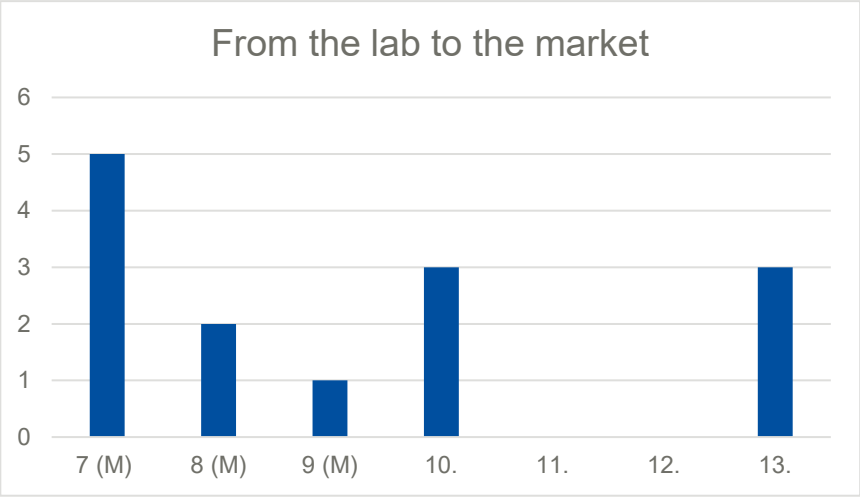
(13) How do you inform the national stakeholders about the relevant legal instruments and standards available (e.g. regulatory sandboxes)?

Three MS reported having that information available and sharing it through channels such as white papers, websites and social media.

Conclusions

As we can see in Figure 13, few MS mentioned having mechanisms concerning R & D, innovation and testing dedicated to AI security. We also found some expectations the MS that the AI Act will bring some light to this topic.

Figure 13: Overview of AI-related 'From the lab to the market' answers



Networking

According to the AI Act (but also to NIS and the NIS 2), AI providers will be obliged, among other aspects, to inform NCAs about serious incidents or breaches as soon as they become aware of them, along with any recalls or withdrawals of AI systems from the market. NCAs will then collect all the necessary information and investigate the incidents/malfunctions.

(14) Have you developed/plan to develop national incident management or handling procedures considering AI?

Three MS mentioned that AI-related incidents must be reported following regular cybersecurity incident reporting procedures.

(15) Are there national initiatives with focus on collaboration about threat intelligence (AI threats, vulnerabilities and security controls) to the users/community?

MS recognise AI threats as very significant and highlight the potential use of AI by criminals. Despite not having explicitly and dedicated initiatives related to security of AI, three MS mentioned that AI threats are shared through existing mechanisms, provided by the national units on threat intelligence.

(16) Is there a collaboration with the national CSIRT/CERTs, ISACS for the efficient handling of AI-related incidents?

No MS reported on specific mechanisms or entities to handle AI-specific incidents, but three MS (aligned with the answers given on the three previous questions) clarified that regular cybersecurity procedures and mechanisms should be used and that information will then be shared with existing ISACs when appropriate. One of the MS mentioned that at present, no AI-related incidents were registered.

(17) Do you/promote/inform about new initiatives on AI security and vulnerabilities sharing? Like a catalogue of pointers to initiatives (e.g. NIST AI framework)?

One MS mentioned that it is aware of the NIST AI framework, but uses its own guidelines and a set of rules on how to maintain and develop emerging and disruptive technologies, including AI, without national security disruption.

(18) Have you developed appropriate collaboration with the national AI stakeholders for information sharing?

One MS reported to be in direct contact with the most important stakeholders, while another MS mentioned that the collaboration is just starting and is happening on an informal basis through the Competent Authorities on AI (CA@AI) working group, which was established in 2021.

High-risk AI systems should perform consistently throughout their life cycle and meet an appropriate level of cybersecurity in accordance with the generally acknowledged state of the art. The level of accuracy and accuracy metrics should be communicated to the users (rule 49 in the explanatory memorandum).

(19) Have you defined/developed/use specific cyber measurements/metrics at national level that AI stakeholders are required to use? If yes, please elaborate.

Nothing specific to AI was reported.

(20) How do you monitor/audit the level of the cybersecurity of the AI systems throughout their life cycle? Please elaborate.

One of the MS reported the establishment of a national committee for AI ethics and reliability, and another MS mentioned that all public sector organisations and all medium and large-sized enterprises that operate AI systems are obliged to maintain a registry with information about their AI systems (AI systems register), containing the measures taken by the organisation or enterprise to ensure the safe usage and operation of its AI systems.

(21) Do you impose dynamic risk assessment to be conducted by the AI stakeholders? Please elaborate.

Nothing related to AI was mentioned.

(22) What kind of sanctions have you set up for non-compliance with integrity of data and models? Please elaborate.

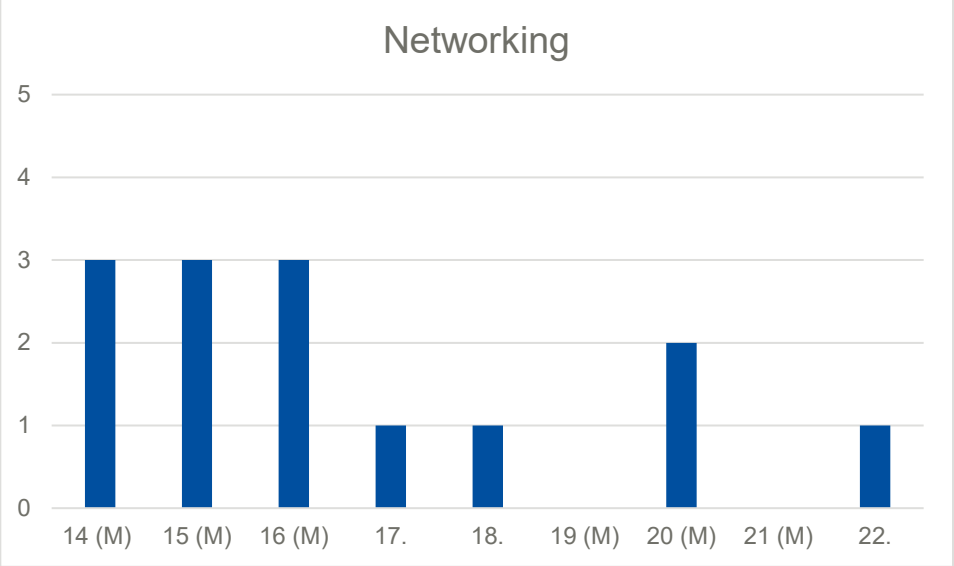
One MS mentioned the creation of a legislative framework for ethical and credible AI, focusing also on cybersecurity requirements.

Conclusions

Figure 14 illustrates the number of AI-related networking questions, including two questions that did not receive any answers: '*(19) Have you defined/developed/used specific cyber measurements/metrics at the national level that AI stakeholders are required to use?'* and '*(21) Do you impose dynamic risk assessments to be conducted by the AI stakeholders?'*.

Incident handling, collaboration and threat intelligence regarding AI security are essentially expected to follow the same mechanisms as cybersecurity in general. Some MS mention progressing towards AI-specific frameworks and the strengthening of European collaboration under the Competent Authorities on AI working group, which already discusses ways to implement the AI Act. One of the MS reported on the establishment of a national committee for AI ethics and reliability, while another MS mentioned that all public sector organisations and all medium and large-sized enterprises that operate AI systems are obliged to maintain a register containing the measures taken to ensure the safe usage and operation of their AI systems.

Figure 14: Overview of AI-related 'Networking' answers



Infrastructure

In accordance with the AI Act, to ensure a level of cybersecurity appropriate to the risks, suitable measures would have to be taken by the providers of high-risk AI systems, also considering as appropriate the underlying ICT infrastructure (rule 51 in the explanatory memorandum).

(23) How do you monitor/audit the appropriateness of the controls undertaken by the AI stakeholders (developers, integrators, providers of critical infrastructure e.g. telecom operators) to adequately secure the underlying ICT infrastructure? Please elaborate. (M)

Nothing was reported about specific measures for providers of services or products based on AI technologies.

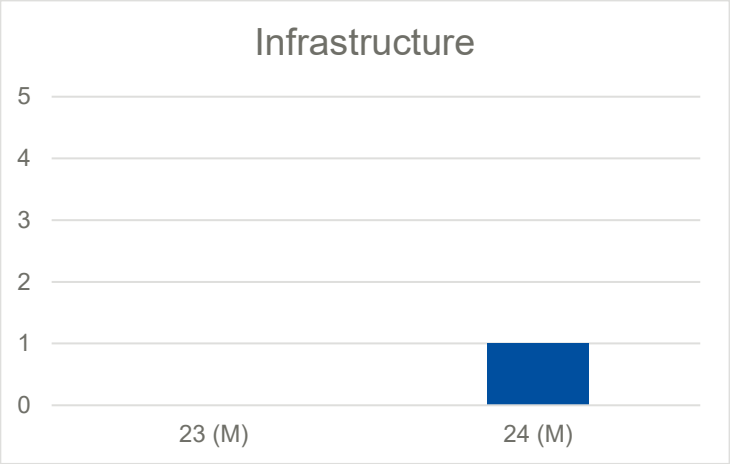
(24) Have you specified/defined measurements and KPIs which the AI stakeholders can use to assess the appropriateness of the controls undertaken? If yes, please elaborate. (M)

One MS reported the regular publishing of criteria catalogues to assess cybersecurity, also for AI, in a cloud environment.

Conclusions

Figure 15 illustrates that just one MS referred to criteria catalogues for AI, in the scope of the question about the specification/definition of measurements and KPIs to assess security controls for AI.

Figure 15: Overview of AI-related 'Infrastructure' answers



Regulation

One of the AI cybersecurity challenges is that a breach of integrity (e.g. poor data quality or biased input data sets) can lead to automated decision-making systems that wrongly classify individuals and exclude them from certain services or deprive them from their rights. The AI Act aims to minimise the risk of algorithmic discrimination.

(25) How do you monitor the integrity and quality of data sets used for the development of AI systems? Please elaborate.

One MS reported on a registry of AI systems, where information on the measures taken to ensure their safe operation is kept. Furthermore, they mentioned that all public sector organisations that acquire AI systems must perform algorithmic impact assessments and data protection impact assessments before the first use of the systems.

(26) Have national auditors as well as certification and accreditation bodies been established for assessing the security of the AI systems? If yes, please elaborate.

Nothing specific to AI was reported.

(27) How do you evaluate security of the AI systems (e.g. via conformity assessment, certification, standards compliance, risk assessment)? Please elaborate.

Nothing specific to AI was reported.

(28) What are the obligations you have imposed for testing, risk management, documentation and human oversight throughout the AI systems' life cycle to ensure continuous data and training model integrity? Please elaborate.

One MS stated that the same criteria as proposed in the AI Act are applied.

According to the proposed AI regulation, the requirements of a high-risk AI system related to products covered by the NLF legislation (e.g. machinery, medical devices, toys) need to be assessed.

(29) Do you have a process where you are notified about the high-risk AI systems used in various NLF-regulated products? If yes, please elaborate.

Nothing specific to AI was reported.

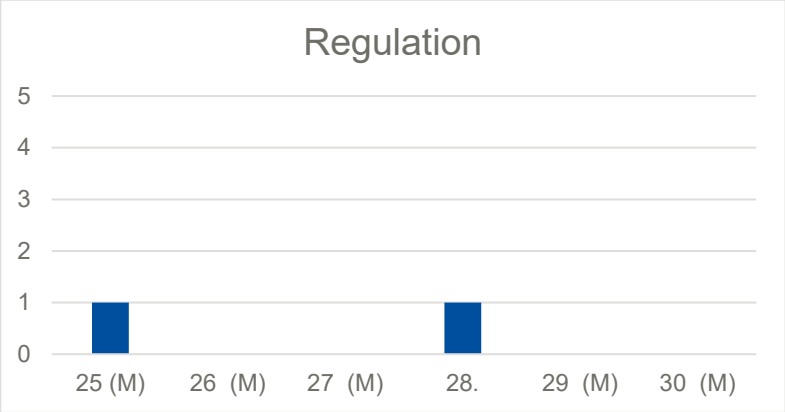
(30) Do you have rules in relation to NLF products that may be relevant to cybersecurity? If yes, please elaborate.

Nothing specific to AI was reported.

Conclusions

In this section the number of answers was not significant. For question (25) 'How do you monitor the integrity and quality of data sets used for the development of AI systems?', only one MS reported about a registry of AI systems, with algorithm impact assessments and data protection impact assessments. For question (28) 'What are the obligations you have imposed for testing, risk management, documentation and human oversight throughout the AI systems' life cycle to ensure continuous data and training models' integrity?', another single MS reported using the same criteria as proposed by the AI Act. Figure 16 illustrates these results.

Figure 16: Overview of AI-related 'Regulation' answers



3.3. SURVEY CONCLUSIONS

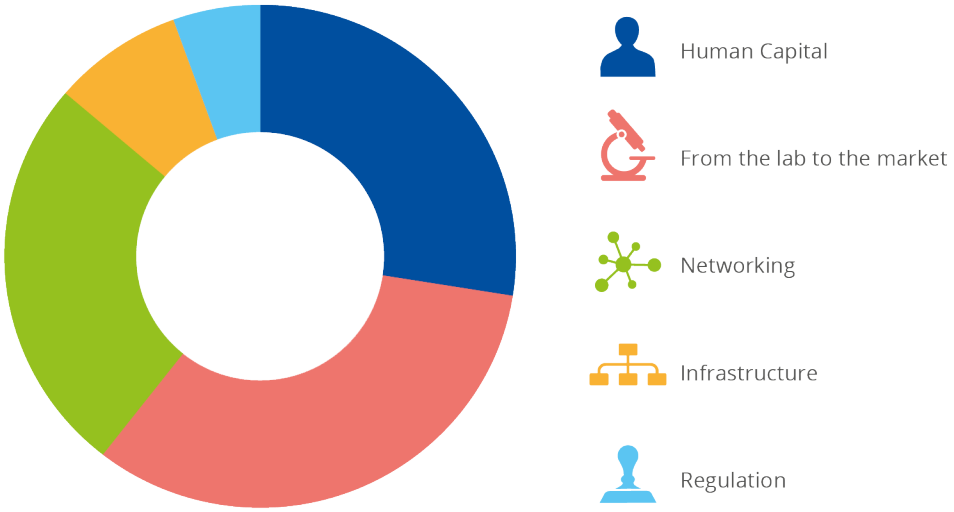
After analysing the answers in each of the policy areas, we conclude that the MS are aware of the new challenges and risks brought by the generalised usage of AI in society and in all kinds of critical infrastructures. Some countries have already started to disseminate procedures related to AI assessment, although the general expectation is that the AI Act will help clarify the way forward.

The number of answers regarding effective measures and mechanisms dedicated specifically to AI shows that, up to now, MS essentially expect to follow the same mechanisms as for other cybersecurity threats or incidents. However, two MS already have some guidance and self-assessment tools specific to AI security.

In Figure 17, we can see that 'Human capital', 'From the lab to the market' and 'Networking' are the policy areas where most insights were given.

- **Human capital.** Collaboration with universities is well recognised, as is the need for security of AI at all levels of education, although security of AI topics is so far handled essentially at AI courses.
- **From the lab to the market.** Regarding awareness campaigns and guidance, it is worth mentioning that one MS mentioned a label for digitally responsible businesses to cover both cybersecurity, privacy and trustworthy AI, and another mentioned a report with self-assessment guidance for AI/ML risk over the life cycle.
- **Networking.** One MS is establishing a national committee for AI ethics and reliability, while another MS maintains an AI systems registry, containing the measures taken by enterprises to ensure the safe usage and operation of their AI systems.

Figure 17: Distribution of answers related to each policy area



It is also worth mentioning the collaboration that already exists among MS under the initiative of the Competent Authorities on AI working group, where topics such as the supervision of AI and the foreseen roles of competent authorities are being discussed and analysed.

4. CONCLUSIONS AND THE WAY FORWARD

The report provides a **framework (FAICP) consisting of three layers (basic cybersecurity relevant to AI, AI-specific cybersecurity and sector-specific cybersecurity for AI)** that categorises the various identified best practices and standards in a way that can be used by NCAs and AI stakeholders to address the cybersecurity challenges of their AI systems. It also adopts the view that AI systems are hosted by an ICT infrastructure and, as such, the stakeholders need to first conduct their basic cybersecurity practices (Layer I). Then they need to pay attention to additional cybersecurity challenges that the AI systems reveal due to their dynamic and socio-technical nature and complement their efforts with additional cybersecurity practices (Layer II). Finally, the use of AI systems in various economic sectors require further cybersecurity practices to be applied (Layer III). For each layer we identified **open issues and research activities that still need to be conducted and resolved.** Below we present our recommendations for various stakeholders.

Cybersecurity and AI experts, including those who represent standardisation organisations.

- **Integrity of data sources and data.** The trustworthiness of AI algorithms relies on the integrity of the data and the data sources that generate this data, therefore we need to dynamically and continuously assess them before using them. Best practices on how to assess all types of data sources (e.g. surveillance cameras, biometric systems, smart traffic lights) are needed.
- **Continuous monitoring of the data life cycle security.** All processes in data management need to be assessed, from data collection to labelling to cleaning to using and storing. Poisoning of data can take place at any stage of the process. Methodologies and dynamic tools need to be developed.
- **Longitudinal risk assessment.** AI systems continue to learn and consequently evolve after their deployment, meaning that vulnerabilities can be exploited at various stages of their life cycle and thus risk evaluation cannot be static. Traditional methodologies and tools are not efficient. New approaches to cover dynamic threat assessment and RM are needed to cover the entire AI life cycle, which address not only technical but also societal threats (e.g. bias, discrimination, lack of explicability, interpretability, explainability, transparency, accountability).

Multidisciplinary experts.

- **Collaboration and interdisciplinarity.** Multi-perceptive approaches are needed for the development of trustworthy AI with clear design principles that meet societal and human requirements and specificities. Collaboration of experts representing various disciplines (sociologists, psychologists, data scientists, computer scientists and cybersecurity engineers) is needed to be able to design, implement, operate, measure and audit human-centric AI systems.

The Commission, other EU institutions and MS need to collaborate in support of the following.

- **Global framework for AI ethics.** The AI Act is based on the EU ethical principles for AI. However, these are not universal and not globally accepted. Globally accepted ethical frameworks are needed. Only then can we develop universal acceptable measures and scales for the security and trustworthiness of AI.
- **From policy requirements to design principles to technical specifications.** Ethical measurements, KPIs and AI design best practices need to be developed and disseminated to guide AI designers and developers to improve AI security.
- **Enhance skills and capabilities.** Favourable conditions and funding opportunities

need to be created to support collaboration of data scientists and cybersecurity experts in order to develop knowledge needed to advance the security and resilience of the AI systems as well as the AI-attacks management. ENISA's *User Manual – European cybersecurity skills framework (ECSF)*[97] can be used for this purpose.

- **Periodic distribution of the survey** to capture the NCAs' monitoring achievements of national AI stakeholders is recommended to accelerate good cybersecurity practices and identify open issues.

The growing use of AI means that its security will become a key challenge for the future. According to ENISA Top 10 Emerging Cybersecurity Threats for 2030[98], misuse of AI will become a significant threat. State-sponsored operatives or cyber criminals who attack blockchain technology and issue deep fakes might not just exist in fiction. On its own, AI is not going to solve today's or tomorrow's complex societal, business or security challenges. However, AI's ability to identify patterns and adaptively learn in real time as events warrant can accelerate detection, containment and response. It can also help reduce the heavy load on analysts working in security operations centres (SOCs) and enable them to be more proactive. These workers will likely remain in high demand, but AI will change their roles.

Finally, as the elements of AI- and ML-driven security threats begin to emerge, AI can help security teams prepare for the eventual development of AI-driven cybercrimes ([99]). In order for this transformation to take place experts need to have a good understanding of both AI's contribution to cybersecurity and cybersecurity issues in AI.

This report recommends to stakeholders to realise that AI systems are hosted in their ICT ecosystem and they need to continue protecting all the layers (physical, network, IT, data, users) of the ecosystem by following traditional good cybersecurity practices (FAICP Layer I). Additional practices are needed due to the dynamic nature of the AI (FAICP Layer II) or the security requirements of the environment that AI operate (FAICP Layer III). Research efforts are needed to further develop comprehensive complementary practices.

[97] ENISA, *User Manual – European cybersecurity skills framework (ECSF)*, 2022, https://www.enisa.europa.eu/publications/european-cybersecurity-skills-framework-ecsf.
[98] https://www.enisa.europa.eu/news/cybersecurity-threats-fast-forward-2030
[99] Aubley, C., Frank, W., Bowen, E. and Golden, D., 'Cyber AI: Real defense – Augmenting security teams with data and machine intelligence', Deloitte Insights, Deloitte, 2021, https://www2.deloitte.com/us/en/insights/focus/tech-trends/2022/future-of-cybersecurity-and-ai.html.

ANNEX I: QUESTIONNAIRE

Goal of the questionnaire: The proposal for a regulation (AI Act) provides various harmonised rules. The goal of the questionnaire is to assess the level of preparedness of authorities for monitoring and enforcement of these future requirements by the NCAs and evaluate whether certain practices/requirements established under the AI Act have been already implemented at the level of the MS.

Methodology: Our questionnaire is based on the cybersecurity concepts described in the FAICP framework, on the main principles (related to cybersecurity) of the proposed AI Act and the coordinated plan on AI and on the harmonised rules related to cybersecurity as reported in the explanatory memorandum (see Table 2).

Table 2: Summary of AI cybersecurity aspects relating with policy areas and national strategies

Policy areas		Public and private sector initiatives	Cybersecurity of AI
Human capital	Education/training / lifelong learning Labour market Intelligence/skills Demand	Enhancement of AI-related skills Education reforms Initiatives targeting teachers and educators Lifelong learning and continuing education: Upskilling and reskilling Future labour market: skills intelligence	Inclusion of AI security in all forms of education/training Awareness initiatives about security of AI Hands-on skills and practical capabilities of AI cybersecurity
From the lab to the market	R & D Innovation	Support AI research Initiatives for innovations towards business growth and increased efficiency of public services	Funding programs including AI security National research centres with AI security activities Research for advancing cybersecurity offensive and defensive practices, methodologies and tools
	Testing	Promotion of experimentation facilities to test promising AI applications: • Innovation sandboxes • Open piloting and testing environments.	Sandboxes, pilots and testing environments for AI security Cyber ranges, simulation platforms and testing environments for secure AI systems Cybersecurity exercises for enhancing AI cybersecurity defences
Networking	Collaboration	Enhancement of collaboration opportunities	Collaboration for certifying AI systems, training and secure interoperable data exchange

		Network of interdisciplinary institutions Improve international collaboration	Collaboration with CSIRTS, CERTs
	Dissemination and uptake	Monitoring the dissemination and uptake of AI	Trustworthy AI systems made in the EU
Infrastructure	Data	Enhancement of data access, usage, sharing and protection: increase availability and quality of data without violating personal rights	Develop fair, equitable and secure data sharing frameworks and interoperable data sets and models Development of secure training models
	Digital & telecom	Improvement of digital infrastructure to leverage opportunities of AI: • deployment of large-scale computing infrastructures; • development of network infrastructure (e.g. 5G standard).	Secure and certified AI systems and applications Enhancing the security of national critical infrastructures
Regulation	Ethical Legal Standardisation	Implementation of norms and ethical principles of AI Review of the legal framework for AI-based applications Enhance and define interoperable technical standards	Trustworthiness of AI legislation and frameworks Apply standards to AI security AI security compliance with the GDPR

List of questions *(mandatory questions are in bold)*

Policy Areas	Questions related to Cybersecurity of AI
Human Capital	Cybersecurity plays a crucial role in ensuring that AI systems are resilient against attempts to alter their use, behaviour, performance or compromise their security properties by malicious third parties exploiting the system's vulnerabilities. Raising practical skills and capabilities in handling emerging AI cyber threats and challenges is important in the future development of AI systems. 1. **Have you built/do you plan to build synergies with educational authorities/institutions to increase AI cybersecurity capabilities at all levels of education?** 2. **Do you offer awareness campaigns about the secure development and use of AI solutions?** 3. **Do you provide guidance and best practices on how to improve AI security?** 4. Do you consider AI cybersecurity in syllabus of courses dedicated to AI or to CS? 5. Do you offer practical trainings to the AI stakeholders and can elaborate to which stakeholders? 6. Do you organise national, regional and cross-border cybersecurity exercises enabling the upskilling of the AI stakeholders?

From the lab to the market	According to the Coordinated Plan on AI supporting AI research and innovation related to threats and attacks on AI and offering solutions for testing promising AI solutions is critical to ensure cybersecurity obligations in the uptake of developments from the lab to the market: **7. What type of support (funding/scholarships/ collaboration opportunities) do you offer to increase the cybersecurity capabilities of newly innovative solutions that rely on AI?** 8. What are the means that you use to support SMEs/MEs to secure their AI products? 9. Do you have/promote testing environments, like sandboxes/cyber ranges/simulation platforms to test and evaluate AI vulnerabilities before market? How? 10. Do you have specific measurements/KPIs/metrics that the AI stakeholders are imposed to use? **11. Have you informed national AI stakeholders on cybersecurity requirements set by the NCAs for their AI products and how do you do it?** **12. How do you monitor if such requirements have been met?** 13. How do you inform the national stakeholders about the relevant legal instruments and standards available? (e.g. regulatory sandboxes)
Networking	According to the AI Act (but also by the NIS and the NIS 2), AI providers will be obliged, among other aspects, to inform NCAs about serious incidents or a breach as soon as they become aware of them, as well as any recalls or withdrawals of AI systems from the market. NCAs will then collect all the necessary information and investigate the incidents/ malfunctions. **14. Have you developed / plan to develop national incident management or handling procedures considering AI?** **15. Are there national initiatives that focus on collaboration about threat intelligence (AI threats, vulnerabilities and security controls) to the users/community?** **16. Is there collaboration with the national CSIRTs/ CERTs and ISACs for the efficient handling of AI-related incidents?** 17. Do you promote/inform about new initiatives on AI security and vulnerabilities sharing? Like a catalogue of pointers to initiatives (e.g. NIST AI framework)? 18. Have you developed appropriate collaboration with the national AI stakeholders for information sharing? High-risk AI systems should perform consistently throughout their life cycle and meet an appropriate level of cybersecurity in accordance with the generally acknowledged state of the art. The level of accuracy and accuracy metrics should be communicated to the users (rule 49 in the explanatory memorandum). **19. Have you defined/developed/used specific cyber measurements/metrics at the national level that AI stakeholders are required to use?**

	20. How do you monitor/audit the level of the cybersecurity of the AI systems throughout their life cycle? **21. Do you impose dynamic risk assessment to be conducted by the AI stakeholders?** **22. What kind of sanctions have you set up for non-compliance with integrity of data and models?**
Infrastructure	In accordance with the AI Act, to ensure a level of cybersecurity appropriate to the risks, suitable measures would have to be taken by the providers of high-risk AI systems, also considering as appropriate the underlying ICT infrastructure (rule 51 in the explanatory memorandum). **23. How do you monitor/audit the appropriateness of the controls undertaken by the AI stakeholders (developers, integrators, critical infrastructures, e.g. telecom operators) to adequately secure the underlying ICT infrastructure?** **Have you specified/defined measurements and KPIs which the AI stakeholders can use to assess the appropriateness of the controls undertaken?**
Regulation	One of the AI cybersecurity challenges is the breach of integrity (e.g. poor data quality or biased input data sets) that can lead to automated decision-making systems that wrongly classify individuals and exclude them from certain services or deprive them from their rights (ENISA). The AI Act aims to minimise the risk of algorithmic discrimination. **25. How do you monitor the integrity and quality of data sets used for the development of AI systems?** **26. Have national auditors and certification and accreditation bodies been established for assessing the security of the AI systems?** **27. How do you evaluate security of the AI systems (e.g. via conformity assessment, certification, standards compliance, risk assessment, etc.)?** 28. What are the obligations you have imposed for testing, risk management, documentation and human oversight throughout the AI systems' life cycle to ensure continuous data and training model integrity? According to the proposed AI regulation, the requirements of a high-risk AI system related to products covered by the NLF legislation (e.g. machinery, medical devices, toys) need to be assessed. **29. Do you have a process where you are notified about the high-risk AI systems used in various NLF regulated products?** 30. **Do you have rules in relation to NLF products that may be relevant to cybersecurity?**

ANNEX II: AI-RELATED STANDARDS

A.1 AI SECURITY-RELATED STANDARDS

ETSI ISG[100]

- ETSI GR SAI 006 V1.1.1 (2022-03) The role of hardware in security of AI
- ETSI GR SAI 001 V1.1.1 (2022-01) AI Threat Ontology
- ETSI GR SAI 002 V1.1.1 (2021-08) Data Supply Chain Security
- ETSI GR SAI 005 V1.1.1 (2021-03) Mitigation Strategy Report
- ETSI GR SAI 004 V1.1.1 (2020-12) Problem Statement

ISO/IEC[101]

- ISO/IEC 24368:2022 Artificial Intelligence – overview of ethical and societal concerns
- ISO/IEC 22989:2022 Artificial Intelligence – concepts and terminology
- ISO/IEC DIS 23894 – Information technology – Artificial intelligence – Risk management – under development
- ISO/IEC CD 42001.2 – Information Technology – Artificial intelligence – Management system – under development
- ISO/IEC TR 24027:2021 – Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision-making
- ISO/IEC AWI TS 12791 – Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks – under development
- ISO/IEC TR 24028:2020 – Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence
- ISO/IEC TR 24029-1:2021 – Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview
- ISO/IEC CD 24029-2 – Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods – under development
- ISO/IEC DTR 24368 – Information technology – Artificial intelligence – Overview of ethical and societal concerns – under development
- ISO/IEC DTR 27563 – Impact of security and privacy in Artificial Intelligence – under development
- ISO/IEC AWI 12792 – Information technology – Artificial intelligence – Transparency taxonomy of AI systems – under development

A.2 DESIGN-RELATED STANDARDS

The work of IEEE addresses some of the most important characteristics of AI as a determinant for its trustworthiness, with a particular focus on explainability, but also on model distribution and management.

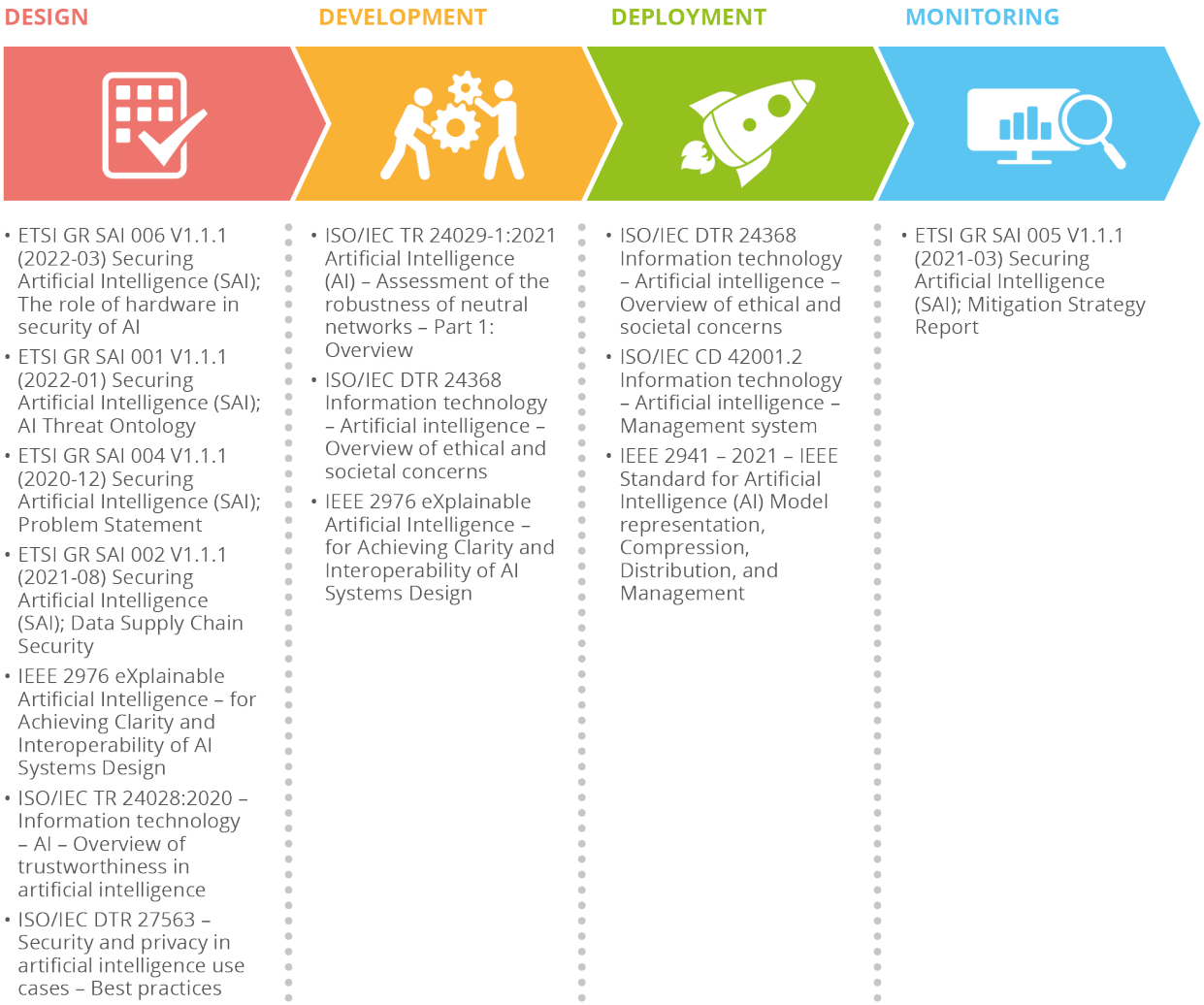
- IEEE 7000 Standards for Building Ethical Systems
- IEEE 2894 Guide for an Architectural Framework for Explainable Artificial Intelligence
- IEEE 2941 Approved Draft Standard for Artificial Intelligence (AI) Model Representation, Compression, Distribution and Management

[100] https://www.etsi.org/committee/1640-sai.
[101] https://www.iso.org.

- IEEE P2941.1 Standard for Operator Interfaces of Artificial Intelligence
- IEEE 2976 Standard for XAI – eXplainable Artificial Intelligence – for Achieving Clarity and Interoperability of AI Systems Design

In the figure below, we provide a summary of the standards analysed in previous sections and map them into the different stages of the AI life cycle.

Figure 18: AI-related standards along the AI life cycle



ANNEX III: LIST OF ABBREVIATIONS

AI	artificial intelligence
AI Act	Artificial Intelligence Act
CEPS	Centre for European Policy Studies
CSA	Cybersecurity Act
CSIRTs	computer security incident response teams
DL	deep learning
ENISA	European Union Agency for Cybersecurity
ETSI	European Telecommunications Standards Institute
EU	European Union
GDPR	general data protection regulation
FAICP	framework for AI good cybersecurity practices
ICT	information and communications technology
IEEE	Institute of Electrical and Electronics Engineers
IoT	internet of things
ISAC	information sharing and analysis centres
ISO	International Organization for Standardization
ITU	International Telecommunication Union
JRC	European Joint Research Centre
KPI	key performance indicator
MEs	micro enterprises, minor enterprises
ML	machine learning
MS	EU Member State
NCA	national competent authorities
NIS/NIS 2	EU directives on measures for a high common level of cybersecurity across the EU
NIST	National Institute for Standards and Technology
NLF	new legislative framework
OECD	Organisation for Economic Co-operation and Development
RM	risk management
SMEs	small and medium-sized enterprises

TP-09-23-025-EN-N

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the EU agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, ENISA contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found here: www.enisa.europa.eu.

ENISA
European Union Agency for Cybersecurity

Athens Office
Agamemnonos 14, Chalandri 15231, Attiki, Greece

Heraklion Office
95 Nikolaou Plastira
700 13 Vassilika Vouton, Heraklion, Greece

enisa.europa.eu

ISBN 978-92-9204-619-4
doi 10.2824/588830