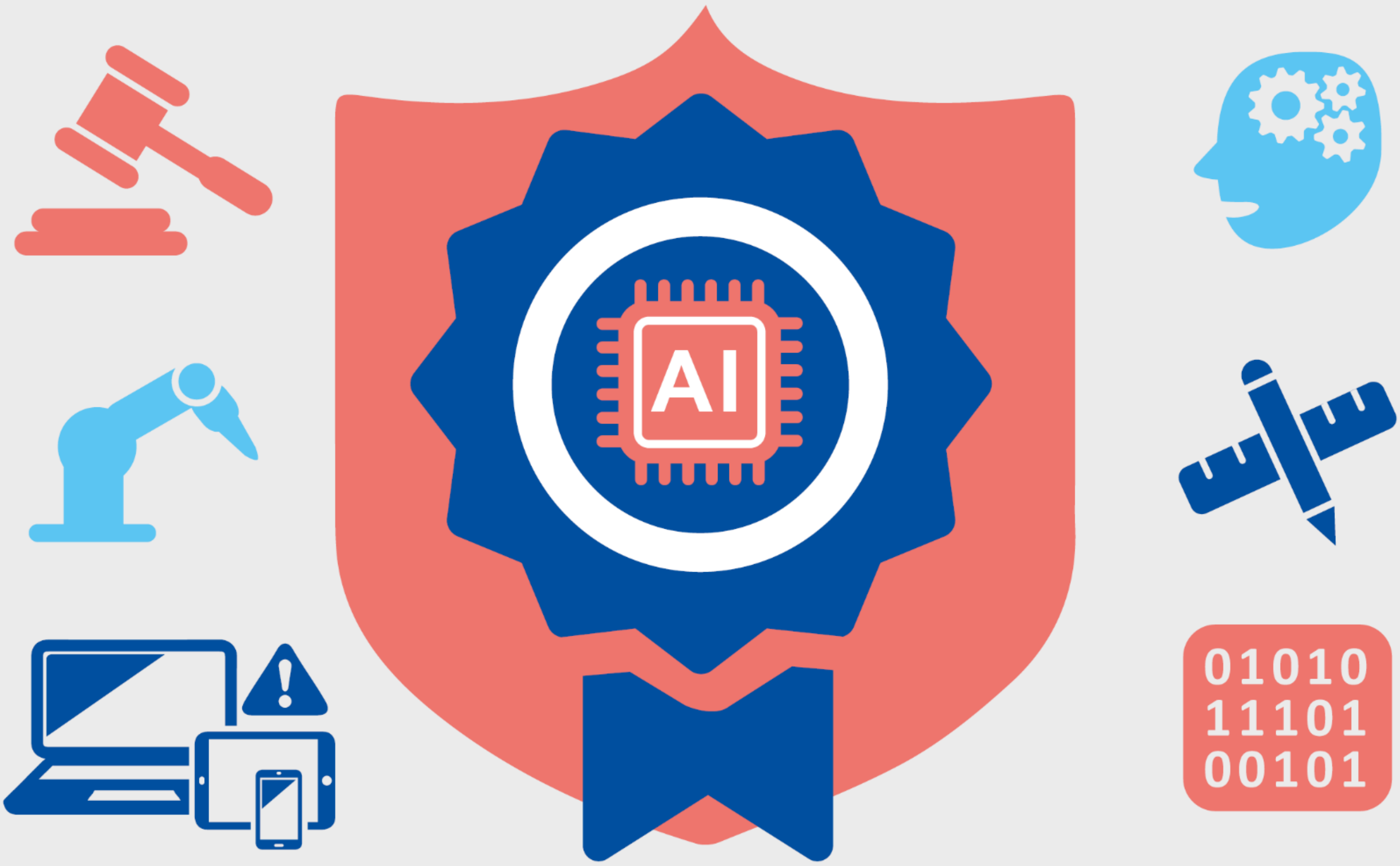
enisa

EUROPEAN UNION AGENCY
FOR CYBERSECURITY



CYBERSECURITY OF AI AND STANDARDISATION

MARCH 2023

ABBREVIATIONS

Abbreviation	Definition
AI	Artificial Intelligence
CEN-CENELEC	European Committee for Standardisation – European Committee for Electrotechnical Standardisation
CIA	Confidentiality, Integrity and Availability
EN	European Standard
ESO	European Standardisation Organisation
ETSI	European Telecommunications Standards Institute
GR	Group Report
ICT	Information And Communications Technology
ISG	Industry Specification Group
ISO	International Organization for Standardization
IT	Information Technology
JTC	Joint Technical Committee
ML	Machine Learning
NIST	National Institute of Standards and Technology
R&D	Research And Development
SAI	Security of Artificial Intelligence
SC	Subcommittee
SDO	Standards-Developing Organisation
TR	Technical Report
TS	Technical Specifications
WI	Work Item

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity 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.

CONTACT

For contacting the authors please use team@enisa.europa.eu
For media enquiries about this paper, please use press@enisa.europa.eu.

AUTHORS

P. Bezombes, S. Brunessaux, S. Cadzow

EDITOR(S)

ENISA:

E. Magonara

S. Gorniak
P. Magnabosco
E. Tsekmezoglou

ACKNOWLEDGEMENTS

We would like to thank the Joint Research Centre and the European Commission for their active contribution and comments during the drafting stage. Also, we would like to thank the ENISA Ad Hoc Expert Group on Artificial Intelligence (AI) cybersecurity for the valuable feed-back and comments in validating this report.

LEGAL NOTICE

This publication represents the views and interpretations of ENISA, unless stated otherwise. It does not endorse a regulatory obligation of ENISA or of ENISA bodies pursuant to the Regulation (EU) No 2019/881.

ENISA has the right to alter, update or remove the publication or any of its contents. It is intended for information purposes only and it must be accessible free of charge. All references to it or its use as a whole or partially must contain ENISA as its source.

Third-party sources are quoted as appropriate. ENISA is not responsible or liable for the content of the external sources including external websites referenced in this publication.

Neither ENISA nor any person acting on its behalf is responsible for the use that might be made of the information contained in this publication.

ENISA maintains its intellectual property rights in relation to this publication.

COPYRIGHT NOTICE

ISBN 978-92-9204-616-3, DOI 10.2824/277479, TP-03-23-011-EN-C

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The overall objective of the present document is to provide an overview of standards (existing, being drafted, under consideration and planned) related to the cybersecurity of artificial intelligence (AI), assess their coverage and identify gaps in standardisation. It does so by considering the specificities of AI, and in particular machine learning, and by adopting a broad view of cybersecurity, encompassing both the 'traditional' confidentiality–integrity–availability paradigm and the broader concept of AI trustworthiness. Finally, the report examines how standardisation can support the implementation of the cybersecurity aspects embedded in the proposed EU regulation laying down harmonised rules on artificial intelligence (COM(2021) 206 final) (draft AI Act).

The report describes the standardisation landscape covering AI, by depicting the activities of the main Standards-Developing Organisations (SDOs) that seem to be guided by concern about insufficient knowledge of the application of existing techniques to counter threats and vulnerabilities arising from AI. This results in the ongoing development of ad hoc reports and guidance, and of ad hoc standards.

The report argues that existing general purpose technical and organisational standards (such as ISO-IEC 27001 and ISO-IEC 9001) can contribute to mitigating some of the risks faced by AI with the help of specific guidance on how they can be applied in an AI context. This consideration stems from the fact that, in essence, AI is software and therefore software security measures can be transposed to the AI domain.

The report also specifies that this approach is not exhaustive and that it has some limitations. For example, while the report focuses on software aspects, the notion of AI can include both technical and organisational elements beyond software, such as hardware or infrastructure. Other examples include the fact that determining appropriate security measures relies on a system-specific analysis, and the fact that some aspects of cybersecurity are still the subject of research and development, and therefore might be not mature enough to be exhaustively standardised. In addition, existing standards seem not to address specific aspects such as the traceability and lineage of both data and AI components, or metrics on, for example, robustness.

The report also looks beyond the mere protection of assets, as cybersecurity can be considered as instrumental to the correct implementation of trustworthiness features of AI and – conversely –the correct implementation of trustworthiness features is key to ensuring cybersecurity. In this context, it is noted that there is a risk that trustworthiness is handled separately within AI-specific and cybersecurity-specific standardisation initiatives. One example of an area where this might happen is conformity assessment.

Last but not least, the report complements the observations above by extending the analysis to the draft AI Act. Firstly, the report stresses the importance of the inclusion of cybersecurity aspects in the risk assessment of high-risk systems in order to determine the cybersecurity risks that are specific to the intended use of each system. Secondly, the report highlights the lack of standards covering the competences and tools of the actors performing conformity assessments. Thirdly, it notes that the governance systems drawn up by the draft AI Act and the

Cybersecurity Act (CSA)[1] should work in harmony to avoid duplication of efforts at national level.

Finally, the report concludes that some standardisation gaps might become apparent only as the AI technologies advance and with further study of how standardisation can support cybersecurity.

[1] 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).

1. INTRODUCTION

1.1 DOCUMENT PURPOSE AND OBJECTIVES

The overall objective of the present document is to provide an overview of standards (existing, being drafted, under consideration and planned) related to the cybersecurity of artificial intelligence (AI), assess their coverage and identify gaps in standardisation. The report is intended to contribute to the activities preparatory to the implementation of the proposed EU regulation laying down harmonised rules on artificial intelligence (COM(2021) 206 final) (the draft AI Act) on aspects relevant to cybersecurity.

1.2 TARGET AUDIENCE AND PREREQUISITES

The target audience of this report includes a number of different stakeholders that are concerned by the cybersecurity of AI and standardisation.

The primary addressees of this report are standards-developing organisations (SDOs) and public sector / government bodies dealing with the regulation of AI technologies.

The ambition of the report is to be a useful tool that can inform a broader set of stakeholders of the role of standards in helping to address cybersecurity issues, in particular:

- academia and the research community;
- the AI technical community, AI cybersecurity experts and AI experts (designers, developers, machine learning (ML) experts, data scientists, etc.) with an interest in developing secure solutions and in integrating security and privacy by design in their solutions;
- businesses (including small and medium-sized enterprises) that make use of AI solutions and/or are engaged in cybersecurity, including operators of essential services.

The reader is expected to have a degree of familiarity with software development and with the confidentiality, integrity and availability (CIA) security model, and with the techniques of both vulnerability analysis and risk analysis.

1.3 STRUCTURE OF THE STUDY

The report is structured as follows:

- **definition of the perimeter of the analysis** (Chapter 2): introduction to the concepts of AI and cybersecurity of AI;
- **inventory of standardisation activities relevant to the cybersecurity of AI** (Chapter 3): overview of standardisation activities (both AI-specific and non-AI specific) supporting the cybersecurity of AI;
- **analysis of coverage** (Chapter 4): analysis of the coverage of the most relevant standards identified in Chapter 3 with respect to the CIA security model and to trustworthiness characteristics supporting cybersecurity;
- **wrap-up and conclusions** (Chapter 5): building on the previous sections, recommendations on actions to ensure standardisation support to the cybersecurity of AI, and on preparation for the implementation of the draft AI Act.

2. SCOPE OF THE REPORT: DEFINITION OF AI AND CYBERSECURITY OF AI

2.1 ARTIFICIAL INTELLIGENCE

Understanding AI and its scope seems to be the very first step towards defining cybersecurity of AI. Still, a clear definition and scope of AI have proven to be elusive. The concept of AI is evolving and the debate over what it is, and what it is not, is still largely unresolved – partly due to the influence of marketing behind the term ‘AI’. Even at the scientific level, the exact scope of AI remains very controversial. In this context, numerous forums have adopted/proposed definitions of AI.[2]

Box 1: Example – Definition of AI, as included in the draft AI Act

In its draft version, the AI Act proposes a definition in Article 3(1):
‘artificial intelligence system’ (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with. The techniques and approaches referred to in Annex I are:

- Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
- logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
- statistical approaches, Bayesian estimation, search and optimisation methods

In line with previous ENISA work, which considers it the driving force in terms of AI technologies, the report mainly focuses on ML. This choice is further supported by the fact that there seem to be a general consensus on the fact that ML techniques are predominant in current AI applications. Last but not least, it is considered that the specificities of ML result in vulnerabilities that affect the cybersecurity of AI in a distinctive manner. It is to be noted that the report considers AI from a life cycle perspective[3]. Considerations concerning ML only have been flagged.

[2] For example, the United Nations Educational, Scientific and Cultural Organization (UNESCO) in the ‘First draft of the recommendation on the ethics of artificial intelligence’, and the European Commission’s High-Level Expert Group on Artificial Intelligence.

[3] See the life cycle approach portrayed in the ENISA report *Securing Machine Learning Algorithms* (https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms).

Box 2: Specificities of machine learning – examples from a supervised learning model[4]

ML systems cannot achieve 100 % in both precision and recall. Depending on the situation, ML needs to trade off precision for recall and vice versa. It means that AI systems will, once in a while, make wrong predictions. This is all the more important because it is still difficult to understand when the AI system will fail, but it will eventually.

This is one of the reasons for the need for explainability of AI systems. In essence, algorithms are deemed to be explainable if the decisions they make can be understood by a human (e.g., a developer or an auditor) and then explained to an end user (ENISA, *Securing Machine Learning Algorithms*).

A major specific characteristic of ML is that it relies on the use of large amounts of data to develop ML models. Manually controlling the quality of the data can then become impossible. Specific traceability or data quality procedures need to be put in place to ensure that, to the greatest extent possible, the data being used do not contain biases (e.g. forgetting to include faces of people with specific traits), have not been deliberately poisoned (e.g. adding data to modify the outcome of the model) and have not been deliberately or unintentionally mislabelled (e.g. a picture of a dog labelled as a wolf).

2.2 CYBERSECURITY OF AI

AI and cybersecurity have been widely addressed by the literature both separately and in combination. The ENISA report Securing Machine Learning Algorithms[5] describes the multidimensional relationship between AI and cybersecurity, and identifies three dimensions:

- cybersecurity of AI: lack of robustness and the vulnerabilities of AI models and algorithms,
- AI to support cybersecurity: AI used as a tool/means to create advanced cybersecurity (e.g., by developing more effective security controls) and to facilitate the efforts of law enforcement and other public authorities to better respond to cybercrime,
- malicious use of AI: malicious/adversarial use of AI to create more sophisticated types of attacks.

The current report focuses on the first of these dimensions, namely the cybersecurity of AI. Still, there are different interpretations of the cybersecurity of AI that could be envisaged:

- a narrow and traditional scope, intended as protection against attacks on the confidentiality, integrity and availability of assets (AI components, and associated data and processes) across the life cycle of an AI system,
- a broad and extended scope, supporting and complementing the narrow scope with trustworthiness features such as data quality, oversight, robustness, accuracy, explainability, transparency and traceability.

The report adopts a narrow interpretation of cybersecurity, but it also includes considerations about the cybersecurity of AI from a broader and extended perspective. The reason is that links between cybersecurity and trustworthiness are complex and cannot be ignored: the requirements of trustworthiness complement and sometimes overlap with those of AI cybersecurity in ensuring proper functioning. As an example, oversight is necessary not only for the general monitoring of an AI system in a complex environment, but also to detect abnormal behaviours due to cyberattacks. In the same way, a data quality process (including data traceability) is an added value alongside pure data protection from cyberattack. Hence,

[4] Besides the ones mentioned in the box, the 'False Negative Rate" and the 'False Positive Rate" and the 'F measure" are examples of other relevant metrics.
[5] https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

trustworthiness features such as robustness, oversight, accuracy, traceability, explainability and transparency inherently support and complement cybersecurity.

3. STANDARDISATION IN SUPPORT OF CYBERSECURITY OF AI

3.1 RELEVANT ACTIVITIES BY THE MAIN STANDARDS-DEVELOPING ORGANISATIONS

It is recognised that many SDOs are looking at AI and preparing guides and standardisation deliverables to address AI. The rationale for much of this work is that whenever something new (in this instance AI) is developed there is a broad requirement to identify if existing provisions apply to the new domain and how. Such studies may help to understand the nature of the new and to determine if the new is sufficiently divergent from what has gone before to justify, or require, the development and application of new techniques. They could also give detailed guidance on the application of existing techniques to the new, or define additional techniques to fill the gaps.

Still, in the scope of this report, the focus is mainly on standards that can be harmonised. This limits the scope of analysis to those of the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), the European Committee for Standardization (CEN) and European Committee for Electrotechnical Standardization (CENELEC), and the European Telecommunications Standards Institute (ETSI). CEN and CENELEC may transpose standards from ISO and IEC, respectively, to EU standards under the auspices of, respectively, the Vienna and Frankfurt agreements.

3.1.1 CEN-CENELEC

CEN-CENELEC addresses AI and Cybersecurity mainly within two joint technical committees (JTCs).

- JTC 13 ‘Cybersecurity and data protection’ has as its primary objective to transpose relevant international standards (especially from ISO/IEC JTC 1 subcommittee (SC) 27) as European standards (ENs) in the information technology (IT) domain. It also develops ‘homegrown’ ENs, where gaps exist, in support of EU directives and regulations.
- JTC 21 ‘Artificial intelligence’ is responsible for the development and adoption of standards for AI and related data (especially from ISO/IEC JTC 1 SC 42), and providing guidance to other technical committees concerned with AI.

JTC 13 addresses what is described as the narrow scope of cybersecurity (see Section 2.2). The committee has identified a list of standards from ISO-IEC that are of interest for AI cybersecurity and might be adopted/adapted by CEN-CENELEC based on their technical cooperation agreement. The most prominent identified standards belong to the ISO 27000 series on information security management systems, which may be complemented by the ISO 15408 series for the development, evaluation and/or procurement of IT products with security functionality, as well as sector-specific guidance, e.g. ISO/IEC 27019:2017 *Information technology – Security techniques – Information security controls for the energy utility industry* (see the annex A.1, for the full list of relevant ISO 27000 series standards that have been identified by CEN-CENELEC).

In addition, the following guidance and use case documents are drafts under development (some at a very early stage) and explore AI more specifically. It is premature to evaluate the impacts of these standards.

- ISO/IEC AWI 27090, *Cybersecurity – Artificial intelligence – Guidance for addressing security threats and failures in artificial intelligence systems*: The document aims to provide information to organisations to help them better understand the consequences of security threats to AI systems, throughout their life cycles, and describes how to detect and mitigate such threats. The document is at the preparatory stage.
- ISO/IEC CD TR 27563, *Cybersecurity – Artificial Intelligence – Impact of security and privacy in artificial intelligence use cases*: The document is at the committee stage.

By design, JTC 21 is addressing the extended scope of cybersecurity (see Section 4.2), which includes trustworthiness characteristics, data quality, AI governance, AI management systems, etc. Given this, a first list of ISO-IEC/SC 42 standards has been identified as having direct applicability to the draft AI Act and is being considered for adoption/adaption by JTC 21:

- ISO/IEC 22989:2022, *Artificial intelligence concepts and terminology* (published),
- ISO/IEC 23053:2022, *Framework for artificial intelligence (AI) systems using machine learning (ML)* (published),
- ISO/IEC DIS 42001, *AI management system* (under development),
- ISO/IEC 23894, *Guidance on AI risk management* (publication pending),
- ISO/IEC TS 4213, *Assessment of machine learning classification performance* (published),
- ISO/IEC FDIS 24029-2, *Methodology for the use of formal methods* (under development),
- ISO/IEC CD 5259 series: *Data quality for analytics and ML* (under development).

In addition, JTC 21 has identified two gaps and has launched accordingly two ad hoc groups with the ambition of preparing new work item proposals (NWIPs) supporting the draft AI Act. The potential future standards are:

- AI systems risk catalogue and risk management,
- AI trustworthiness characterisation (e.g., robustness, accuracy, safety, explainability, transparency and traceability).

Finally, it has been determined that ISO-IEC 42001 on AI management systems and ISO-IEC 27001 on cybersecurity management systems may be complemented by ISO 9001 on quality management systems in order to have proper coverage of AI and data quality management.

3.1.2 ETSI

ETSI has set up a dedicated Operational Co-ordination Group on Artificial Intelligence, which coordinates the standardisation activities related to AI that are handled in the technical bodies, committees and industry specification groups (ISGs) of ETSI. In addition, ETSI has a specific group on the security of AI (SAI) that has been active since 2019 in developing reports that give a more detailed understanding of the problems that AI brings to systems. In addition, a large number of ETSI’s technical bodies have been addressing the role of AI in different areas, e.g., zero touch network and service management (ISG ZSM), health TC eHEALTH) and transport (TC ITS).

ISG SAI is a pre-standardisation group identifying paths to protect systems from AI, and AI from attack. This group is working on a technical level, addressing specific characteristics of AI. It has published a number of reports and is continuing to develop reports to promote a wider understanding and to give a set of requirements for more detailed normative standards if such are proven to be required.

The following are published group reports (GRs) from ISG SAI that apply to understanding and developing protections to and from AI:

- ETSI GR SAI-001: *AI Threat Ontology*,
- ETSI GR SAI-002: *Data Supply Chain Security*,
- ETSI GR SAI-004: *Problem Statement*,
- ETSI GR SAI-005*: Mitigation Strategy Report*,
- ETSI GR SAI-006: *The Role of Hardware in Security of AI*.

The following work items of ISG SAI are in development/pending publication at the time of writing:

- ETSI DGR SAI-007: *Explicability and Transparency of AI Processing* (pending publication),
- ETSI DGR SAI-008: *Privacy Aspects of AI/ML Systems* (final draft),
- ETSI DGR SAI-009: *Artificial Intelligence Computing Platform Security Framework* (pending publication),
- ETSI DGR SAI-010: *Traceability of AI Models* (under development – early draft),
- ETSI DGR/SAI-0011: *Automated manipulation of multimedia identity representations* (early draft),
- ETSI DGR/SAI-003: *Security testing of AI* (stable draft),
- ETSI DGR/SAI-0012: *Collaborative AI* (early draft).

In addition to the work already published and being developed, the group maintains a ‘roadmap’ that identifies the longer-term planning of work and how various stakeholders interact.

In addition, as a direct consequence of the draft AI Act and the Cybersecurity Act, the following potential future WIs are being discussed: AI readiness and transition, testing, and certification.

The work in ETSI ISG SAI is within the wider context of ETSI’s work in AI, which includes contributions from the other ETSI bodies, including its cybersecurity technical committee (TC Cyber). Among other projects, the committee is specifically extending TS 102 165-1, *Methods and protocols; Part 1: Method and pro forma for threat, vulnerability, risk analysis (TVRA)*.

3.1.3 ISO-IEC

ISO-IEC carries out its work on AI in JTC 1 SC 42. The list in the annex A.2 presents the standards published or under development with their publication target dates (unless already mentioned in the previous sections).

3.1.4 Others

Almost all horizontal and sectorial standardisation organisations have launched AI-related standardisation activities with very little consistency among them. The report Landscape of AI standards AI standardisation landscape published by StandICT[6] identifies more than 250 documents, and it is most likely that a lot are missing. The International Telecommunication Union (ITU), the Institute of Electrical and Electronics Engineers (IEEE) and SAE International are some of the organisations that are very active on AI. In the process of building the standardisation landscape, it has been observed that it is almost impossible to have access to the content of the documents, especially if they are in their development phase, and it is therefore impossible to assess their relevance and maturity beyond their titles.

[6] https://www.standict.eu/

One of the most interesting identified projects, though, is SAE AIR AS6983, which is dedicated to AI/ML in aeronautics and is very similar in scope to the ambition of the JTC 21 project on AI trustworthiness characterisation. Its publication is expected in 2023.

It is also recognised that major software vendors prepare their own standards and guidance on the use of their AI functional capabilities, and in many cases (e.g. where software is distributed by an app store) will require detailed review and quality controls before being made available on the market. This is in addition to the statutory obligations of the developer. Finally, the US National Institute of Standards and Technology (NIST) is also active in the area of AI and has released its *AI Risk Management Framework* (AI RMF 1.0) in January 2023[7].

[7] https://www.nist.gov/itl/ai-risk-management-framework

4. ANALYSIS OF COVERAGE

This section provides an analysis of the coverage of the most relevant standards identified in the previous chapters with respect to the CIA security model and to trustworthiness characteristics supporting cybersecurity.

4.1 STANDARDISATION IN SUPPORT OF CYBERSECURITY OF AI – NARROW SENSE

As explained in Section 2.2, in its essence the cybersecurity of AI in a narrow sense is understood as concerning the CIA of assets (AI components, and associated data and processes) throughout the life cycle of an AI system. Table 1 shows, for each of these security goals, examples of relevant attacks on AI systems.

Table 1[8]: Application of CIA paradigm in the context of AI[9]

Security goal	Contextualisation in AI (selected examples of AI-specific attacks)
Confidentiality	**Model and data stealing attacks:** **Oracle:** A type of attack in which the attacker explores a model by providing a series of carefully crafted inputs and observing outputs. These attacks can be precursor steps to more harmful types, for example evasion or poisoning. It is as if the attacker made the model talk to then better compromise it or to obtain information about it (e.g. model extraction) or its training data (e.g. membership inference attacks and inversion attacks). **Model disclosure:** This threat refers to a leak of the internals (i.e. parameter values) of the ML model. This model leakage could occur because of human error or a third party with too low a security level.
Integrity	**Evasion:** A type of attack in which the attacker works on the ML algorithm’s inputs to find small perturbations leading to large modification of its outputs (e.g. decision errors). It is as if the attacker created an ‘optical illusion for the algorithm. Such modified inputs are often called adversarial examples. **Poisoning:** A type of attack in which the attacker alters data or models to modify the ML algorithm’s behaviour in a chosen direction (e.g. to sabotage its results or to insert a back door). It is as if the attacker conditioned the algorithm according to its motivation.
Availability	**Denial of service:** ML algorithms usually consider input data in a defined format to make their predictions. Thus, a denial of service could be caused by input data whose format is inappropriate. However, it may also happen that a malicious user of the model constructs an input data (a sponge example) specifically designed to increase the computation time of the model and thus potentially cause a denial of service.

If we consider AI systems as software and we consider their whole life cycles, general-purpose standards, i.e. those that are not specific to AI and that address technical and organisational aspects, can contribute to mitigating many of the risks faced by AI. The following ones have been identified as particularly relevant:

- ISO/IEC 27001, *Information security management*, and ISO/IEC 27002, *Information security controls*: relevant to all security objectives,
- ISO/IEC 9001, *Quality management system*: especially relevant to integrity (e.g. in particular for data quality management to protect against poisoning) and availability.

[8] Based on the White Paper ‘Towards auditable AI systems’ of Germany’s Federal Office for Information Security (https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf?__blob=publicationFile&v=6) and on the ENISA report Securing Machine Learning Algorithms (https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms).

[9] There are also cybersecurity attacks that are not specific to AI, but could affect CIA even more severely. ETSI GR/SAI-004, Problem Statement, and ETSI GR/SAI-006, *The Role of Hardware in Security of AI*, can be referred to for more detailed descriptions of traditional cyberattacks on hardware and software.

This raises two questions:

- firstly, the extent to which general-purpose standards should be adapted to the specific AI context for a given threat,
- secondly, whether existing standards are sufficient to address the cybersecurity of AI or they need to be complemented.

Concerning the first question, it is suggested that general-purpose standards either apply or can be applied if guidance is provided. To simplify, although AI has some specificities, it is in its essence software; therefore, what is applicable to software can be applied to AI. Still, SDOs are actively addressing AI specificities, and many existing general-purpose standards are in the process of being supplemented to better address AI. This means that, at a general level, existing gaps concern clarification of AI terms and concepts, and the application of existing standards to an AI context, and in particular the following.

- **Shared definition of AI terminology and associated trustworthiness concepts**: Many standards attempt to define AI (e.g. ISO/IEC 22989:2022, *Artificial intelligence concepts and terminology*; ISO/IEC 23053:2022, *Framework for artificial intelligence (AI) systems using machine learning (ML)*; ETSI ISG GR SAI-001, *AI threat ontology*; NIST*, AI risk management framework*. However, in order to apply standards consistently, it is important that SDOs have a common understanding of what AI is (and what it is not), what the trustworthiness characteristics are and, therefore, where and to what related standards apply (and where they do not).
- **Guidance on how standards related to the cybersecurity of software should be applied to AI**: For example, data poisoning does not concern AI only, and good practices exist to cope with this type of threat, in particular related to quality assurance in software. However, quality assurance standards would refer to data manipulation (as opposed to data poisoning): a measure against data manipulation would not mention in its description that it also mitigates those forms of data manipulation that particularly affect AI systems. Possible guidance to be developed could explain that data poisoning is a form of data manipulation and, as such, can be addressed, at least to some extent, by standards related to data manipulation. This guidance could take the form of specific documents or could be embedded in updates of existing standards.

Concerning the second question, it is clear from the activity of the SDOs that there is concern about insufficient knowledge of the application of existing techniques to counter threats and vulnerabilities arising from AI. The concern is legitimate and, while it can be addressed with ad hoc guidance/updates, it is argued that this approach might not be exhaustive and has some limitations, as outlined below.

- **The notion of AI can include both technical and organisational elements not limited to software, such as hardware or infrastructure, which also need specific guidance.** For example, ISO/IEC/IEEE 42010 edition 2, Architecture description vocabulary, considers the cybersecurity of an entity of interest that integrates AI capabilities, including for example hardware, software, organisations and processes. In addition, new changes in AI system and application scenarios should be taken into consideration when closing the gap between general systems and AI ones.
- **The application of best practices for quality assurance in software might be hindered by the opacity of some AI models.**
- **Compliance with ISO 9001 and ISO/IEC 27001 is at organisation level, not at system level. Determining appropriate security measures relies on a system-specific analysis.** The identification of standardised methods supporting the CIA security objectives is often complex and application or domain specific, as in large part the attacks to be mitigated depend on the application or domain. Although there are general attacks on many cyber systems, and some very specific attacks that can be directed at many different systems, they

often rely upon a small set of vulnerabilities that can be exploited that are specific to a domain or an application. In this sense, ETSI TS 102 165-1, *Methods and protocols; Part 1: Method and pro forma for threat, vulnerability, risk analysis (TVRA)*[10], and ISO/IEC 15408-1, *Evaluation criteria for IT security*, can be used to perform specific risk assessments.

- **The support that standards can provide to secure AI is limited by the maturity of technological development, which should therefore be encouraged and monitored.** In other words, in some areas existing standards cannot be adapted or new standards cannot be fully defined yet, as related technologies are still being developed and not yet quite mature enough to be standardised. In some cases, first standards can be drafted (e.g. ISO/IEC TR 24029-1:2021 on the robustness of deep neural networks) but will probably need to be regularly updated and adapted as research and development (R&D) progresses. For example, from the perspective of ML research, much of the work on adversarial examples, evasion attacks, measuring and certifying adversarial robustness, addressing specificities of data poisoning for ML models, etc. is still quite active R&D. Another challenge related to R&D on AI and standardisation is benchmarking: research results are often not comparable, resulting in a situation where it is not always clear what works under what conditions.

Box 3: Example of technological gap: continuous learning[11]

Continuous learning is the ability of an AI component to evolve during its operational life through the use of in-operation data for retraining the AI component. This function is often perceived as the key ability of AI.

Model poisoning is easy to do during continuous learning / in-operation learning. For example, during continuous learning, it is very challenging to check the quality of the data in real time. When it comes to high-risk AI components, the use of continuous learning would imply continuous validation of the data used for the training of the AI component (continuous data quality assessment), continuous monitoring of the AI component, continuous risk assessment, continuous validation and continuous certification if needed. While the issues with continuous learning have been described in ISO/IEC 22989, *Information technology – Artificial intelligence – Artificial intelligence concepts and terminology*, and the activities described above are conceptually feasible, their execution is still the object of R&D.

- **The traceability and lineage of both data and AI components are not fully addressed.** The traceability of processes is addressed by several standards related to quality. In that regard, ISO 9001 is the cornerstone of quality management. However, the traceability of data and AI components throughout their life cycles remains an issue that cuts across most threats and remains largely unaddressed. Indeed, both data and AI components may have very complex life cycles, with data coming from many sources and being transformed and augmented, and, while AI components may reuse third parties' components or even open-source components, all of those are obviously a source of increased risks. This aspect implies that technologies, techniques and procedures related to traceability need to be put in place to ensure the quality of AI systems, for instance that data being used do not contain biases (e.g. forgetting to include faces of people with specific traits), have not been deliberately poisoned (e.g. adding data to modify the outcome of the model) and have not been deliberately or unintentionally mislabelled (e.g. a picture of a dog labelled as a wolf).
- **The inherent features of ML are not fully reflected in existing standards.** As introduced in Section 2.1, ML cannot, by design, be expected to be 100 % accurate. While this can also be true for (for example) ruled-based systems designed by humans, ML has a larger input space (making exhaustive testing difficult), black box properties and high sensitivity, meaning that small changes in inputs can lead to large changes in outputs. Therefore, it is even more

[10] Currently under revision to include AI as well.

[11] It is to be noted though that the concept of continuous learning is subject to different interpretations. It is not always clear how it differs from updating the system from time to time, i.e. what frequency of re-training would justify the label 'continuous learning".

important to understand, on the one hand, how the risk of failure can be mitigated and, on the other, if/when a failure is caused by a malicious actor. The most obvious aspects to be considered in existing/new standards can be summarised as follows.

- AI/ML components may be associated with hardware or other software components in order to mitigate the risk of functional failure, therefore changing the cybersecurity risks associated with the resulting set-up[12].
- Reliable metrics can help a potential user detect a failure. For example, with precision and recall metrics for AI systems relying on supervised classification, if users know the precision/recall thresholds of an AI system they should be able to detect anomalies when measuring values outside those thresholds, which may indicate a cybersecurity incident. While this would be a general check (more efficient for attacks on a massive scale than for specific attacks), the accurate definition of reliable metrics is a prerequisite to define more advanced measurements.
- Testing procedures during the development process can lead to certain levels of accuracy/precision.

It is to be noted that the subject of metrics for AI systems and of testing procedures is addressed by standardisation deliverables such as ISO/IEC DIS 5338-AI system life cycle processes (under development); ISO/IEC AWI TS 12791-*Treatment of unwanted bias in classification and regression machine learning tasks* (under development); ETSI TR 103 305-x, *Critical security controls for effective cyber defence*; and ETSI GR SAI-006, *The role of hardware in security of AI*[13]. However, the coverage of the AI systems trustworthiness metrics that are needed is incomplete, which is one reason for the CEN-CENELEC initiative on the 'AI trustworthiness characterisation' project.

4.2 STANDARDISATION IN SUPPORT OF THE CYBERSECURITY OF AI – TRUSTWORTHINESS

As explained in Section 2.2, cybersecurity can be understood as going beyond the mere protection of assets and be considered fundamental to the correct implementation of trustworthiness features of AI, and – conversely – the correct implementation of trustworthiness features is key to ensuring cybersecurity.

Table 3 exemplifies this relation in the context of the draft AI Act. It shows the role of cybersecurity within a set of requirements outlined by the act that can be considered as referring to the trustworthiness of an AI ecosystem. In fact, some of them (e.g. quality management, risk management) contribute to building an AI ecosystem of trust indirectly, but have been included because they are considered equally important and they are requirements of the draft AI Act[14].

[12] For example, a self-driving car could be automatically deactivated if the supervising system detected abnormal conditions that could signal a cybersecurity attack.

[13] Other examples include ISO/IEC 23894, *Information technology – Artificial intelligence – Guidance on risk management*; ISO/IEC DIS 42001, *Information technology – Artificial intelligence – Management system*; and ISO/IEC DIS 24029-2, *Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods*.

[14] The European Commission's High-Level Expert Group on Artificial Intelligence has identified seven characteristics of trustworthiness: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability.

Table 3[15]**:** Role of cybersecurity within a set of requirements outlined by the draft AI Act

Draft AI Act Requirement	Description	Relevance of cybersecurity
Data and data governance	High-risk AI systems which make use of techniques involving the training of models with data shall be developed on the basis of training, validation, and testing datasets that meet a set of quality criteria	The requirements here address data quality, which is key to secure data feeds, processing and outputs. Data quality can be reinforced by the use of tools that verify the source of data and the integrity of data (i.e. to prove that data have not been manipulated between source and sink), and by limiting access to data.
Record-keeping	High-risk AI systems shall be designed and developed with capabilities enabling the automatic recording of events ('logs') while the high-risk AI systems is operating. Those logging capabilities shall conform to recognised standards or common specifications.	All of the major security management control standards (e.g. ISO 27000 and ETSI TR 103 305) address the importance of event logging and having the staff to analyse the logs. These logs probably contain sensitive data, and appropriate standard cybersecurity measures, i.e. CIA, need to be deployed.
Transparency and provision of information to users	High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and of the provider set out in Chapter 3 of [COM(2021) 206 final].	As noted above, documentation in itself is not a security requirement. However, as a security control, technical documentation is a key element in system transparency and in (high-level) explainability.
Human oversight	High-risk AI systems shall be designed and developed in such a way, including with appropriate human–machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.	This form of control is identified in ISO27001 and in ETSI TS 103 305-1. ([16]) Where human oversight is required, it should form an integral part of the design of the system, and performance and other constraints should be added to the role of oversight. This may include the performance of mandatory actions and checks, and rules for escalation of an event assessment.
Risk management system	An assessment through internal checks for 'stand-alone' high-risk AI systems would require a full, effective and properly documented ex ante compliance with all requirements of the regulation and compliance with robust quality and risk management systems and post-market monitoring. A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems.	ISO/IEC 31000 is a framework for risk analysis and the management of risk analysis systems. At a more detailed level, tools for vulnerability analysis (e.g. ETSI TS 102 165-1) may apply, as well as runtime analysis tools. Many development environments will perform both static and dynamic tests on software that allow risks in the codebase to be identified. The suite of measures should operate in concert.
Quality management system	Providers of high-risk AI systems shall put a quality management system in place that ensures compliance with this Regulation. The provider should establish a sound quality management system, ensure the accomplishment of the required conformity assessment procedure, draw up the relevant documentation and establish a robust post-market monitoring system.	ISO 9001 is the overarching standard for the implementation of a quality management system in development environments, which should include security management aspects.
Conformity assessment	AI systems that create a high risk to the health and safety or fundamental rights of natural persons: in line with a risk-based approach, these high-risk AI systems are permitted on the European market subject to compliance with	This is necessary for the evaluation of all requirements, including cybersecurity.

[15] Source: adapted from Nativi, S. and De Nigris, S., AI Standardisation Landscape: State of play and link to the EC proposal for an AI regulatory framework (https://publications.jrc.ec.europa.eu/repository/handle/JRC125952).

	certain mandatory requirements and an ex-ante conformity assessment.	
Robustness	AI systems should be resilient against risks connected to the limitations of the system (e.g. errors, faults, inconsistencies, unexpected situations) as well as against malicious actions that may compromise the security of the AI system and result in harmful or otherwise undesirable behaviour.	Cybersecurity is one of the key aspects – albeit not the only one – of robustness. It concerns the protection of the AI system against attacks as well as the capacity to recover from such attacks.

The general-purpose technical and organisational standards outlined in Section 3.1 cover these trustworthiness aspects to some extent. The SDOs are actively tackling the matter and are developing AI-specific standards in support of trustworthiness. In particular, ISO/IEC SC 42 is developing most of those aspects in multiple standards, and CEN-CENELEC JTC 21 is working towards adopting/adapting those standards (see annex A.3). This is normal and, to some extent, inevitable at first. Still, in a regulatory context, one could expect a unified comprehensive, coherent and synthetic approach to trustworthiness while avoiding the multiplication – and to some extent duplication – of efforts. Furthermore, it would be inefficient and even counterproductive to have multiple sets of standards for the same characteristics (robustness, explainability, etc.), some coming from the cybersecurity domain and some coming from the AI domain, with risks of discrepancy. The result is that a unified approach to trustworthiness characteristics is highly recommended. In particular, in order to bring coherency and comprehensiveness, it is necessary to clarify who is doing what, in order to avoid needless and confusing duplication, and a certain level of coordination and liaison is vital.

Box 4: Example – Cybersecurity conformity assessment

When it comes to AI systems, conformity assessment will be performed against all requirements outlined in the draft AI Act, trustworthiness, including its cybersecurity aspects, being among them. Existing standards on trustworthiness lack conformity assessment methods, sometimes including technical requirements and metrics. While there are a lot of activities in ISO/IEC SC 42 regarding trustworthiness characteristics, there are also a lot of gaps and very few developed requirements and metrics. Therefore, there is the risk that conformity assessment methods will be addressed by different standards depending on the characteristic being evaluated. Since some characteristics overlap each other, while others might be contradictory (e.g. there might be a trade-off between transparency and cybersecurity), a global and coherent approach is needed.

Box 5: Example – Adversarial attacks

For example, identified adversarial attack threats could be used in both the ML algorithm and the testing and validation process. In that specific case, the threats could have been identified by the AI system's monitoring/oversight process and the testing process. It is likely that some technical requirements/adjustments coming from the cybersecurity threat assessment should find their place in the AI standards repository relating both to oversight and to testing.

4.3 CYBERSECURITY AND STANDARDISATION IN THE CONTEXT OF THE DRAFT AI ACT

The draft AI Act refers explicitly to the cybersecurity of high-risk AI systems. High-risk AI systems are limited to AI systems intended to be used as safety components of products that are subject to third party ex ante conformity assessment, and stand-alone AI systems mainly with fundamental rights implications (e.g. for migration, asylum and border control management) and for the management and operation of critical infrastructure. More precisely, the draft AI Act builds upon a risk-based approach to identify whether an AI system is high risk on the basis of the system's intended use and implications for health, safety and fundamental rights.

It is important to note that this approach differs from the cybersecurity risk-based approach, which sees a cybersecurity risk as a function of its adverse impact and its likelihood of occurrence. Based on the draft AI Act, cybersecurity is a requirement that applies, and therefore is assessed, only once a system is identified as high risk.

These high-risk systems are subject to a number of requirements, cybersecurity being one of them, as in Article 15, ‘Accuracy, robustness and cybersecurity’. The cybersecurity requirements outlined are legal and remain at a high level. Still, explicit reference is made to some technical aspects:

High-risk AI systems shall be resilient as regards attempts by unauthorised third parties to **alter their use or performance** by exploiting the system vulnerabilities.

[…]

The technical solutions to address AI specific vulnerabilities shall include, where appropriate, **measures to prevent and control for attacks trying to manipulate the training dataset** (‘data poisoning’), **inputs designed to cause the model to make a mistake** (‘adversarial examples’), or **model flaws**.

The draft AI Act also lays down, in Article 13, ‘Transparency and provision of information to users’, that high-risk AI systems are to be accompanied by instructions for use, specifying, among other things, the ‘**the level of accuracy, robustness and cybersecurity** referred to in Article 15 against which the high-risk AI system has been **tested** and **validated** and which can be expected, and any **known and foreseeable circumstances that may have an impact** on that expected level of accuracy, robustness and cybersecurity’.

In addition, the draft AI Act refers to cybersecurity in its recitals. In particular, recital 51 mentions that, ‘To ensure a level of cybersecurity appropriate to the risks, suitable measures should therefore be taken by the providers of high-risk AI systems, also taking into account as appropriate the underlying ICT infrastructure’.

Finally, the draft AI Act tackles cybersecurity through a number of other requirements, as exemplified in Table 2. The annexes (A.3 and A.4) contain an overview of activities of European standardisation organisations (ESOs) with respect to the requirements of the AI Act. Building on those, as well as on the previous sections, the following considerations have been outlined concerning the implementation of the draft AI Act from a cybersecurity perspective.

- **Given the applicability of AI in a wide range of domains, the identification of cybersecurity risks and the determination of appropriate security requirements should rely on a system-specific analysis and, where needed, on sectorial standards.** Sectorial standards should build coherently and efficiently on horizontal ones. In turn, the assessment of compliance to security requirements can be based on AI-specific horizontal standards[17] and on vertical/sector-specific standards as well.
- **It is important to develop the guidance necessary to back up existing technical and organisational standards that can support the cybersecurity of AI systems, while monitoring R&D advancements.** Some aspects of cybersecurity can be addressed now by developing specific guidance, while others are still under R&D. For the purposes of the AI Act, the technological gaps described and ongoing R&D processes affect some aspects of the cybersecurity requirements outlined in Article 15 (adversarial examples and data poisoning) and therefore might constitute standardisation gaps with respect to the draft AI Act, depending on how conformity assessment will be organised.

[17] For example, ISO/IEC JTC 1/SC 42 is working on an AI risk management standard (ISO 23894, Information technology– Artificial intelligence – Guidance on risk management) to be complemented by a specific JTC 21 standard on ‘AI risk catalogue and AI risk management’.

- As explained in Section 3.1, SDOs are actively working on the standardisation of trustworthiness characteristics; however, it is unclear whether those standards will be adopted in time for the adoption of the draft AI Act. Therefore, **it is recommended to monitor related developments closely.**

The draft AI Act also depicts a governance system upon which the conformity assessment of AI systems relies. Besides the specific recommendations on conformity assessment outlined above, the following are noted.

- **Ensure that the actors performing conformity assessment on AI systems have standardised tools and competences, including on cybersecurity.** In certain cases, conformity assessment may be performed by notified bodies. AI trustworthiness will therefore rely partly on the competences of those bodies. If those bodies do not have the proper competences, they could make bad assessments and even bias the market. To date there are no standards that adequately cover cybersecurity and describing the competences of organisations for auditing, certification and testing of AI systems (and AI management systems) and their evaluators. This is crucial, as it is most likely that some AI algorithms will attack AI systems while other AI algorithms will protect them. The new AI threats (threats using AI) will probably be more and more efficient at exploiting existing vulnerabilities, while AI algorithms (cybersecurity using AI) could, for example, monitor the behaviour of an AI system to protect it. To sum up, there are standardisation gaps on competences for validation, testing, auditing, certification' of AI systems and on 'competences for auditing and certification of AI management systems (Although a project on this last point is being prepared by ISO/IEC SC 42, it is unclear to what extent it will be sufficient.)
- **Ensure regulatory coherence between the draft AI Act and legislation on cybersecurity.** In particular, Article 42 of the draft AI Act sets out a presumption of conformity with cybersecurity requirements for high-risk AI systems that have been certified or for which a statement of conformity has been issued under a cybersecurity scheme pursuant to Regulation (EU) 2019/881 (the Cybersecurity Act)[18]. While no official request for a EU cybersecurity certification scheme for AI has been issued yet, it is important that, if developed, such a scheme would take due consideration of the draft AI Act – and vice versa. For example, the Cybersecurity Act sets out three levels of assurance (basic, substantial, high), which are commensurate with the level of the risk associated with the intended use of an ICT product/service/ process. These levels provide the rigour and depth of the evaluation of the ICT product/service/process and refer to technical specifications, standards and procedures, including those to mitigate or prevent incidents. It remains to be defined whether and how these assurance levels can apply in the context of the draft AI Act.
- Another regulatory development that might affect the draft AI Act is the proposal COM(2022) 454 for a regulation on horizontal cybersecurity requirements for products with digital elements (the Cyber Resilience Act)[19]. The proposal was presented in September 2022.

[18] Regulation (EU) 2019/881 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).

[19] https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act

5. CONCLUSIONS

This section sums up the report and recommends actions to ensure standardisation support to the cybersecurity of AI, and to the implementation of the draft AI Act.

5.1 WRAP-UP

The study suggests that general-purpose standards for information security and quality management (in particular ISO/IEC 27001, ISO/IEC 27002 and ISO/IEC 9001) can partially mitigate the cybersecurity risks related to the confidentiality, integrity and availability of AI systems. This conclusion relies on the assumption that AI is in its essence software, and therefore what is applicable to software can be applied to AI, if adequate guidance is provided.

This approach can suffice at a general level but needs to be complemented by a system-specific analysis (e.g. relying on ISO/IEC 15408-1:2009), as the identification of standardised methods supporting the CIA security objectives is often domain specific. It is a matter of debate to what extent the assessment of compliance with the resulting security requirements can be based on AI-specific horizontal standards and to what extent it can be based on vertical/sector-specific standards.

Still, some standardisation gaps have been identified:

- the traceability of processes is addressed by several standards, but the traceability of the data and AI components throughout their life cycles remains an issue that cuts across most threats and remains largely unaddressed in practice, despite being covered well in various standards or drafts (e.g. ISO/IEC DIS 42001 on AI management systems [20] and the ISO/IEC CD 5259 series on data quality for analytics and ML[21]);
- the inherent features of ML are not fully reflected in existing standards, especially in terms of metrics and testing procedures;
- in some areas, existing standards cannot be adapted or new standards cannot be fully defined yet, as related technologies are still being developed and not yet quite mature enough to be standardised.

Going beyond the mere CIA paradigm and considering the broader trustworthiness perspective, the main takeaway is that, since cybersecurity cuts across a number of trustworthiness requirements (e.g. data governance, transparency), it is important that standardisation activities around these requirements treat cybersecurity in a coherent manner.

Concerning the implementation of the draft AI Act, besides the considerations above, the following gaps have been identified:

- to date there are no standards that adequately cover cybersecurity and describe the competences of organisations for auditing, certification and testing of AI systems (and AI management systems) and their evaluators;
- the abovementioned gap on areas that are the subject of R&D is relevant to the implementation of the draft AI Act, in particular with respect to data poisoning and adversarial examples.

[20] ISO/IEC DIS 42001, *Information technology — Artificial intelligence — Management system* (under development)
[21] The series is under development (https://www.iso.org/ics/35.020/x/)

5.2 RECOMMENDATIONS

5.2.1 Recommendations to all organisations

The ESOs have made a commitment to standardisation in support of cybersecure AI, as is evidenced by ETSI's ISG SAI and by CEN's JTC 21. These actions are all positive and are to be encouraged and reinforced.

While it is recognised that the ESOs have different operational models and different membership profiles, it is also recognised that the ESOs operate cooperatively in many fields, and this is, again, to be encouraged. Competitive effort to develop standards is to some extent inevitable and, while that is recognised, the ESOs are strongly discouraged from negative competition. One area where harmonisation is seen as essential is in the adoption of a common AI-related terminology and set of concepts not only across SDOs but also with other stakeholders. The present report does not suggest which SDO/ESO should initiate this activity but it is strongly suggested that, without a common set of cross-domain terminology and concepts, the first risk to cybersecurity would be not understanding each other[22].

Recommendation 1: Use a standardised and harmonised AI terminology for cybersecurity, including trustworthiness characteristics and a taxonomy of different types of attacks specific to AI systems.

5.2.2 Recommendations to standards-developing organisations

The following recommendations are to standardisation organisations.

Recommendation 2: Develop specific/technical guidance on how existing standards related to the cybersecurity of software should be applied to AI. These should also include defences at different levels (before the AI system itself, e.g. infrastructure), for which the application of generic standards might be straightforward in many cases. At the same time, it is recommended to monitor and encourage areas where standardisation is limited by technological development, e.g. testing and validation for systems relying on continuous learning and mitigation of some AI-specific attacks.

Recommendation 3: The inherent features of ML should be reflected in standards. The most obvious aspects to be considered relate to risk mitigation by associating hardware/software components with AI; reliable metrics; and testing procedures. The traceability and lineage of both data and AI components should also be reflected.

Recommendation 4: Ensure that liaisons are established between cybersecurity technical committees and AI technical committees so that AI standards on trustworthiness characteristics (oversight, robustness, accuracy, explainability, transparency, etc.) and data quality include potential cybersecurity concerns.

5.2.3 Recommendations in preparation for the implementation of the draft AI Act

The following recommendations are suggested to prepare for the implementation of the draft AI Act, and should be understood as complementary to the recommendations above.

Recommendation 5: Given the applicability of AI in a wide range of domains, the identification of cybersecurity risks and the determination of appropriate security requirements should rely on

[22] Two horizontal terminology-related standards (ISO/IEC 22989 and ISO/IEC 23053) have been published recently (June and July 2022). JTC 21 will base all its work on ISO/IEC terminology.

a system-specific analysis and, where needed, sector-specific standards. Sectorial standards should build coherently and efficiently on horizontal ones.

Recommendation 6: Encourage R&D in areas where standardisation is limited by technological development, on one hand by providing funding for the advancements in specific technologies (e.g. related to countermeasures against adversarial attacks) and on the other by raising awareness of the importance of integrating standardisation aspects in research activities. In addition, it is suggested to promote benchmarking by means of a systematic approach to guide R&D efforts, which are still characterised by a proliferation of specialised approaches that work under specialised conditions.

Recommendation 7: Support the development of standards for the tools and competences of the actors performing conformity assessment.

Recommendation 8: Ensure coherence between the draft AI Act and other legislative initiatives on cybersecurity, notably Regulation (EU) 2019/881 (the Cybersecurity Act) and the proposal COM(2022) 454 for a regulation on horizontal cybersecurity requirements for products with digital elements (the Cyber Resilience Act).

5.3 FINAL OBSERVATIONS

While the report gives an overview of the state of play of standardisation in support of AI, it is likely that additional standardisation gaps and needs may become apparent only as the AI technologies advance and with further study of how standardisation can support cybersecurity. Concerning the implementation of the AI Act, the importance of some gaps may vary depending on how the conformity assessment will be conceived. Last but not least, changes in the legislative landscape, with particular reference to the proposal for a Cyber Resilience Act, are expected to affect standardisation needs.

A ANNEX:

A.1 SELECTION OF ISO 27000 SERIES STANDARDS RELEVANT TO THE CYBERSECURITY OF AI

Name	Document Reference
Information technology – Security techniques – Information security incident management – Part 1: Principles of incident management	ISO/IEC 27035-1:2016
Information technology – Security techniques – Information security incident management – Part 2: Guidelines to plan and prepare for incident response	ISO/IEC 27035-2:2017
Information technology – Information security incident management – Part 3: Guidelines for ICT incident response operations	ISO/IEC 27035-3:2020
Information technology – Security techniques – Guidelines for cybersecurity	ISO/IEC 27032:2012
Information technology – Security techniques – Guidelines for information and communication technology readiness for business continuity	ISO/IEC 27031:2011
Information technology – Security techniques – Mapping the revised editions of ISO/IEC 27001 and ISO/IEC 27002	ISO/IEC TR 27023:2015
Information technology – Guidance on information security management system processes	ISO/IEC TS 27022:2021
Information technology – Security techniques – Competence requirements for information security management systems professionals – Amendment 1: Addition of ISO/IEC 27001:2013 clauses or subclauses to competence requirements	ISO/IEC 27021:2017/AMD 1:2021
Information technology – Security techniques – Competence requirements for information security management systems professionals	ISO/IEC 27021:2017
Information technology – Security techniques – Code of practice for information security controls based on ISO/IEC 27002 for cloud services	ISO/IEC 27017:2015
Information technology – Security techniques – Information security management – Organizational economics	ISO/IEC TR 27016:2014
Information security, cybersecurity and privacy protection – Governance of information security	ISO/IEC 27014:2020
Information security, cybersecurity and privacy protection – Guidance on the integrated implementation of ISO/IEC 27001 and ISO/IEC 20000-1	ISO/IEC 27013:2021
Information technology – Security techniques – Code of practice for Information security controls based on ISO/IEC 27002 for telecommunications organizations – Technical Corrigendum 1	ISO/IEC 27011:2016/Cor 1:2018
Information technology – Security techniques – Information security management for inter-sector and inter-organizational communications	ISO/IEC 27010:2015
Information technology – Security techniques – Guidelines for the assessment of information security controls	ISO/IEC TS 27008:2019

Information security, cybersecurity and privacy protection – Guidelines for information security management systems auditing	ISO/IEC 27007:2020
Information technology – Security techniques – Requirements for bodies providing audit and certification of information security management systems – Amendment 1	ISO/IEC 27006:2015/AMD 1:2020
Information technology – Security techniques – Requirements for bodies providing audit and certification of information security management systems	ISO/IEC 27006:2015
Requirements for bodies providing audit and certification of information security management systems – Part 2: Privacy information management systems	ISO/IEC TS 27006-2:2021
Information technology – Security techniques – Information security risk management	ISO/IEC 27005:2018
Information technology – Security techniques – Information security management – Monitoring, measurement, analysis and evaluation	ISO/IEC 27004:2016
Information technology – Security techniques – Information security management systems – Guidance	ISO/IEC 27003:2017
Information security, cybersecurity and privacy protection – Information security controls	ISO/IEC 27002:2022
Information technology – Security techniques – Information security management systems – Requirements – Technical Corrigendum 2	ISO/IEC 27001:2013/Cor 2:2015
Information technology – Security techniques – Information security management systems – Requirements – Technical Corrigendum 1	ISO/IEC 27001:2013/Cor 1:2014
Information technology – Security techniques – Information security management systems – Requirements	ISO/IEC 27001:2013

A.2 RELEVANT ISO/IEC STANDARDS PUBLISHED OR PLANNED / UNDER DEVELOPMENT

Name	Document Reference	Expected publication date (at the time of writing)
Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview	ISO/IEC TR 24029-1:2021	Published
Assessment of machine learning classification performance	ISO/IEC TS 4213	Published
Big data – Overview and vocabulary	ISO/IEC 20546:2019	Published
Information technology — Artificial intelligence — Overview of ethical and societal concerns	ISO/IEC TR 24368:2022	Published
Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence	ISO/IEC TR 24028:2020	Published
Information technology — Artificial intelligence — Process management framework for big data analytics	ISO/IEC 24668:2022	Published
Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making	ISO/IEC TR 24027:2021	Published
Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems	ISO/IEC TR 24372:2021	Published
Information technology — Big data reference architecture — Part 1: Framework and application process	ISO/IEC TR 20547-1:2020	Published
Information technology — Big data reference architecture — Part 2: Use cases and derived requirements	ISO/IEC TR 20547-2:2018	Published
Information technology — Big data reference architecture — Part 3: Reference architecture	ISO/IEC 20547-3:2020	Published
Information technology — Big data reference architecture — Part 4: Security and privacy	ISO/IEC 20547-4:2020	Published
Information technology — Big data reference architecture — Part 5: Standards roadmap	ISO/IEC TR 20547-5:2018	Published
Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations	ISO/IEC 38507:2022	Published
Safety of machinery – Relationship with ISO 12100 – Part 5: Implications of embedded artificial intelligence machine learning	ISO/TR 22100-5:2021	Published
Artificial Intelligence (AI) Use cases	ISO/IEC TR 24030:2021	Published (to be revised, new version expected in May 2023)
Artificial intelligence — Functional safety and AI systems	ISO/IEC CD TR 5469	Apr-23
Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems	ISO/IEC DIS 25059	May-23
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples	ISO/IEC CD 5259-1	Jul-23
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines	ISO/IEC CD 5259-3	Jul-23

Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework	ISO/IEC CD 5259-4	Jul-23
Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems	ISO/IEC AWI TS 8200	Jul-23
Information technology — Artificial intelligence — AI system life cycle processes	ISO/IEC DIS 5338	Aug-23
Information technology — Artificial intelligence — Guidance for AI applications	ISO/IEC DIS 5339	Aug-23
Information technology — Artificial intelligence — Overview of machine learning computing devices	ISO/IEC AWI TR 17903	Nov-23
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures	ISO/IEC CD 5259-2	Jan-24
Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems	ISO/IEC AWI TS 6254	Feb-24
Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks	ISO/IEC AWI TS 12791	Feb-24
Software and systems engineering — Software testing — Part 11: Testing of AI system	ISO/IEC AWI TS 29119-11[1]	Feb-24
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance	ISO/IEC AWI 5259-5	Feb-25
Information technology — Artificial intelligence — Transparency taxonomy of AI systems	ISO/IEC AWI 12792	Feb-25
Quality evaluation guidelines for AI systems	ISO/IEC AWI TS 5471	Under consideration
Information technology — Artificial intelligence — Reference architecture of knowledge engineering	ISO/IEC DIS 5392	Under development

A.3 CEN-CENELEC JOINT TECHNICAL COMMITTEE 21 AND DRAFT AI ACT REQUIREMENTS

Name	TR, TS, EN	Action: adopt/adapt (ISO/IEC) – develop (ESOs)	Target date	Risk management system	Data and data governance	Record keeping	Transparency and provision of information to users	Human oversight	Accuracy	Robustness	Cybersecurity	Quality management system	Conformity assessment
ISO/IEC 22989:2022 **Artificial intelligence concepts and terminology**	IS	Adopt	July 2022	x	x	x	x	x	x	x	x	x	x
ISO/IEC 23053:2022 **Framework for artificial intelligence (AI) systems using machine learning (ML)**	IS	Adopt	July 2022	x	x	x	x	x	x	x	x	x	x
ISO/IEC CD 5259-1 **Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples**	IS	Adopt	December 2023	x	x	x	x	x	x	x	x	x	x
ISO/IEC 9001:2015 **Quality management systems – Requirements**	IS		2015								x	x	
ISO/IEC 42001 **Artificial intelligence – Management system**	IS	Adopt	December 2023									x	x
ISO/IEC 27001:2022 **Information security management systems – Requirements**	IS		2022								x	x	
ISO/IEC 23894 **Guidance on risk management**	IS	Adopt	December 2023	x									

CEN-CENELEC **Risk catalogue and risk management**	EN	Develop	Q1 2025	x									x
ISO/IEC TR 24029-1 **Assessment of the robustness of neural networks – Part 1: Overview**	TR	Adopt	December 2022							x			
ISO/IEC 24029-2 **Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods**	IS	Adopt	December 2023							x			
CEN-CENELEC **AI trustworthiness characterisation**	EN	Develop	Q1 2025			?	x	x	x	x	x		x
ISO/IEC CD 5259-2 **Data quality for analytics and machine learning (ML) – Part 2: Data quality measures**	IS	Adopt	Q2 2024		x								
ISO/IEC CD 5259-3 **Data quality for analytics and machine learning (ML) – Part 3: Data quality management requirements and guidelines**	IS	Adopt	Q3 2023		x								
ISO/IEC CD 5259-4 **Data quality for analytics and machine learning (ML) – Part 4: Data quality process framework**	IS	Adopt	Q4 2024		x								

A.4 ETSI ACTIVITIES AND DRAFT AI ACT REQUIREMENTS

Name	TR, TS, EN	Status ([a])	Target date	Risk management system	Data and data governance	Record keeping	Transparency and provision of information to users	Human oversight	Accuracy	Robustness	Cybersecurity	Quality management system	Conformity assessment
DTR/MTS-103910 - MTS AI Testing Test Methodology and Test Specification for AI-enabled Systems	TR	Early draft	July 2024	x									x
DTR/MTS-1191168 **MTS AI Testing AI-enabled Testing in Standardisation**	TR	WI Adopted	TBC	x									x
TR 103 911 **MTS AI testing AI-enabled testing in standardisation**	TR	Under development (late)	—	x									
EN 303 645 **Cyber security for consumer internet of things: Baseline requirements**	EN	Published	—	x									
TR 103 304 **Personally identifiable information (PII) protection in mobile and cloud services**	TR	Published	—		x								
TR 103 305 **Critical security controls for effective cyber defence**	TR	Published	—	x	x	x			x	x	x	x	
TR 103 370 **Practical introductory guide to technical standards for privacy**	TR	Published	—		x								

TR 103 404 **Network technologies (NTECH); Autonomic network engineering for the self-managing future internet (AFI); Autonomicity and self-management in the backhaul and core network parts of the 3GPP architecture**	TR	Published	—		x	x							
TR 103 473 **Evolution of management towards autonomic future internet (AFI); Autonomicity and self-management in the Broadband Forum (BBF) architectures**	TR	Published	—		x	x							
TR 103 626 **Autonomic network engineering for the self-managing future internet (AFI); An instantiation and implementation of the generic autonomic network architecture (GANA) model onto heterogeneous wireless access technologies using cognitive algorithms**	TR	Published	—		x	x							
TR 103 627 **Core network and interoperability testing (INT/WG AFI) autonomicity and self-management in IMS architecture**	TR	Published	—		x	x							
TR 103 629 **Evolution of management towards autonomic future internet (AFI); Confidence in autonomic functions; Guidelines for design and testability**	TR	Early draft	—				x	x					
TR 103 747 **Core network and interoperability testing (INT/WG AFI); Federated GANA knowledge planes (KPs) for multi-domain autonomic management & control (AMC) of slices in the NGMN® 5G end-to-end architecture framework**	TR	Published	—		x	x							
TR 103 748 **Core network and interoperability testing (INT); Artificial intelligence (AI) in test systems and testing of AI models; Use and benefits of AI technologies in testing**	TR	Published	—	x	x							x	x
TR 103 749 **INT artificial intelligence (AI) in test systems and testing AI models; Testing of AI with definition of quality metrics**	TR	Start of work	May 2023	x	x							x	x
TR 103 821 **Autonomic network engineering for the self-managing future internet (AFI); Artificial intelligence (AI) in test systems and testing AI models**	TR	Start of work	—	x									
TR 103 857 **Autonomic management and control (AMC) intelligence for self-managed fixed & mobile integrated networks (AFI); Generic framework for E2E federated GANA knowledge planes for AI-powered closed-loop self-adaptive security management & control, across multiple 5G network slices, segments, services and administrative domains**	TR	Stable draft	March 2023		x	x	x			x			

TS 102 165-1 **Methods and protocols; Part 1: Method and pro forma for threat, vulnerability, risk analysis (TVRA)**	TS	Published	—	x	x								
TS 103 195-2 **Autonomic network engineering for the self-managing future internet (AFI); Generic autonomic network architecture; Part 2: An architectural reference model for autonomic networking, cognitive networking and self-management**	TS	Published	—	x	x	x				x		x	x
TS 103 485 **Mechanisms for privacy assurance and verification**	TS	Published	—		x								
TS 103 701 **Cyber security for consumer internet of things: Conformance assessment of baseline requirements**	TS	Published	—	x									

TP-03-23-011-EN-C

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity 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 978-92-9204-613-2
doi: 10.2824/308470