ISSN 1831-9424

JRC SCIENCE FOR POLICY REPORT

Cybersecurity of Artificial Intelligence in the AI Act

Guiding principles to address the cybersecurity requirement for high-risk AI systems

Junklewitz, H., Hamon, R., André, A., Evas, T., Soler Garrido, J., Sanchez Martin, J.

2023

Joint Research Centre

EUR 31643 EN

This publication is a Science for Policy report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The contents of this publication do not necessarily reflect the position or opinion of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contact information
Name: Ignacio Sanchez
Address: DG Joint Research Centre, Directorate T – Digital Transformation and Data, Unit T.2 – Cybersecurity and Digital Technologies, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy
Email: Ignacio.Sanchez@ec.europa.eu
Tel.: +39 033278 5998

EU Science Hub
https://joint-research-centre.ec.europa.eu

JRC134461

EUR 31643 EN

PDF ISBN 978-92-68-07045-1 ISSN 1831-9424 doi:10.2760/271009 KJ-NA-31-643-EN-N

Luxembourg: Publications Office of the European Union, 2023

How to cite this report: Junklewitz, H., Hamon, R., André, A., Evas, T., Soler Garrido, J. and Sanchez Martin, J.I., *Cybersecurity of Artificial Intelligence in the AI Act*, Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/271009, JRC134461.

Contents

Contents I

Abstract 1

Authors 2

Executive summary 3

1 Introduction 5

2 Background 6

2.1 Machine Learning 6

2.2 AI Cybersecurity 6

2.3 AI Cybersecurity in the AI Act 7

2.4 AI Cybersecurity Standardisation 7

3 Guiding principles to address the cybersecurity requirement of the AI Act 9

Guiding principle 1: The focus of the AI Act is on AI systems. 9

Guiding principle 2: Compliance with the AI Act necessarily requires a security risk assessment. 9

Guiding principle 3: Securing AI systems requires an integrated and continuous approach using proven practices and AI-specific controls. 10

Guiding principle 4: There are limits in the state of the art for securing AI models. 10

4 Conclusions 12

References 13

List of abbreviations 15

Abstract

The European Commission's proposal for the AI Act represents a significant milestone in the regulation of Artificial Intelligence (AI). This report focuses on the cybersecurity requirement for high-risk AI systems, as set out in Article 15 of the regulation. It presents a high level analysis in the context of the rapidly evolving AI landscape, and provides a set of key guiding principles to achieve compliance with the AI Act.

The proposed AI Act focuses on AI systems. The internal structure of AI systems involves a range of components. Although AI models are essential components of AI systems, they do not constitute AI systems on their own. The AI Act cybersecurity requirement applies to the AI system as a whole and not directly to its internal components.

In order to ensure compliance, a security risk assessment should be conducted taking into account the design of the system, to identify risks, and implement the necessary mitigation measures. This process requires an integrated and continuous approach using proven cybersecurity practices and procedures combined with AI-specific controls.

Although the state of the art for securing AI models has limitations, AI systems may still achieve compliance with the AI Act's cybersecurity requirement as long as their cybersecurity risks are effectively mitigated through other measures not exclusively deployed at AI model level. However, this may not always be possible, and indeed for some high-risk AI systems using emerging AI technologies, it may not be feasible to achieve compliance with the cybersecurity requirement of the AI Act unless in their design these system additionally introduce new cybersecurity controls and mitigation measures of proven effectiveness

Authors

Junklewitz, Henrik (Joint Research Centre, European Commission)

Hamon, Ronan (Joint Research Centre, European Commission)

André, Antoine-Alexandre (DG CONNECT, European Commission)

Evas, Tatjana (DG CONNECT, European Commission)

Soler Garrido, Josep (Joint Research Centre, European Commission)

Sanchez, Ignacio (Joint Research Centre, European Commission)

Executive summary

The European Commission's proposal for the AI Act represents a significant milestone in the regulation of Artificial Intelligence (AI). This report focuses on the cybersecurity requirement for high-risk AI systems, as set out in Article 15 of the Commission proposal of the AI Act. It provides a high level analysis of the practical applications of this requirement to AI systems in the context of the rapidly evolving AI landscape, and provides a set of key guiding principles to achieve compliance with the AI Act.

Background

The increased adoption of Artificial Intelligence represents a paradigm shift in software development. AI enables solving problems and carrying out tasks that otherwise would be intractable for computers. Artificial Intelligence has the potential to revolutionise many sectors and bring substantial benefits to society. However, due to the nature of their underlying technologies, AI also introduces new risks and challenges that need to be properly addressed.

AI systems are computer systems and, as such, they inherit all the cybersecurity risks already connected to traditional digital systems that operate in similar contexts. In addition, the uptake of AI technologies introduces new cybersecurity risks connected to the emergence of new classes of AI-specific vulnerabilities. AI cybersecurity is an emerging field that aims to research and address these AI specific vulnerabilities, including adversarial machine learning attacks, data poisoning or backdoors embedded in AI models.

The global AI landscape is evolving rapidly, with new AI techniques being introduced regularly. The current AI ecosystem is composed of a wide variety of AI models and techniques at various levels of maturity. Against this background, current cybersecurity practices and procedures, already well established and used to secure traditional software (and hardware based) systems, are limited in their capacity to address the wider range of cybersecurity risks of AI systems. There are ongoing research and operational efforts at global level to better understand and address these risks extending the current practices and procedures and develop new ones in order to ensure the cybersecurity of AI systems. These efforts are already resulting in the development of AI cybersecurity risk management tools and the definition of security controls that are tailored to the particularities of securing AI systems, including AI cybersecurity metrics and mitigation measures to address AI-specific vulnerabilities.

Policy context

The proposed AI Act represents a significant milestone in the regulation of Artificial Intelligence. This legislation aims to establish a horizontal framework for trustworthy AI. At its core, the proposed AI Act is designed to address risks to health, safety and fundamental rights specifically associated with AI technologies by setting legally binding requirements for high-risk AI systems. Prior to the deployment on the EU market, a provider of a high-risk AI application, will have the obligation to adopt the necessary organisational and technical measures to achieve conformity with the requirements.

In line with a well-established EU system of product safety regulation, harmonised standards are one of the main means to achieve compliance and conformity with the legislative requirements. European standards will be developed by European Standardisation Organisations following the standardisation request of the European Commission published on 22 May 2023 and will eventually become harmonised standards at the time of application of the AI Act. It is important to note that the standardisation request makes clear that the reach of AI standardisation is determined by the technological maturity of AI technologies, explicitly referring to the technological state of the art as the basis of standardisation. Therefore, in terms of technical cybersecurity measures, AI standards are expected to cover mature and generally accepted measures for cybersecurity of AI systems. This anticipates an ongoing need for standardisation work on AI cybersecurity in the coming years. However, this does only imply technological limits for securing individual models, and not a limit to the development of standards. Instead, an initial set of cybersecurity standards in response to the initial AI Act standardisation request should provide a strong basis in terms of available technical controls for achieving and measuring AI cybersecurity together with general outcome-based horizontal cybersecurity requirements for high-risk AI systems.

Guiding principles

The results of the report's analysis are summarised in these four guiding principles:

1. The focus of the AI Act is on AI systems. The structure of AI systems involves a range of internal components, some of which are AI-related, whilst others are not. Although AI models are essential components of AI systems, they do not constitute AI systems on their own. The cybersecurity requirement set out by the AI Act applies to the AI system as a whole and not directly to its internal components.
2. Compliance with the AI Act necessarily requires a security risk assessment. In order to ensure that an AI system complies with the cybersecurity requirement of the AI Act, a security risk assessment should be conducted considering the internal architecture of the AI system and the intended application context. This cybersecurity risk assessment, carried out in the context of the Risk Management System described in the Article 9 of the AI Act, aims to identify the specific risks, translate the higher level cybersecurity requirement of the regulation down to specific requirements for the components of the system and implement the necessary mitigation measures.
3. Securing AI systems requires an integrated and continuous approach using proven practices and AI-specific controls. This process should leverage current cybersecurity practices and procedures, using a combination of existing controls for software systems and AI-model specific measures. AI systems are the sum of all their components and of their interactions. A holistic approach following the security-in-depth and security-by-design principles should be adopted to ensure that AI systems achieve compliance with the cybersecurity requirement of the AI Act.
4. There are limits in the state of the art for securing AI models. A wide variety of AI technologies of different degrees of maturity coexist in the current AI landscape. Not all AI technologies might be ready for use in AI systems designed to be deployed in high-risk scenarios, unless their limitations in terms of cybersecurity are properly addressed. In some cases, particularly for emergent AI technologies, there are inherent limitations that cannot be exclusively addressed at the level of the AI model. In those cases, compliance with the cybersecurity requirement of the AI Act can only be achieved following the holistic approach described earlier.

Conclusions

The guiding principles laid out in this report are intended to help all relevant stakeholders in addressing the cybersecurity requirement for high-risk AI systems, as set out in Article 15 of the proposal of the European Commission. It is important to understand that limitations in the state of the art for securing AI models exist and that harmonised standards developed for the AI Act are expected to provide horizontal requirements to ensure the cybersecurity of AI systems, but not expected to cover approaches lacking maturity beyond the generally accepted state of the art in AI cybersecurity.

Nonetheless, these technical limitations do not necessarily impede compliance of AI systems with the cybersecurity requirement of the AI Act. High-risk AI systems can still achieve compliance if they appropriately mitigate the overall cybersecurity risks of the system through other complementary measures, following an integrated approach combining well established cybersecurity practices and procedures with AI specific measures.

However, it is important to acknowledge that this may not always be possible and that for some high-risk AI systems making use of emerging AI technologies, it may not be feasible to achieve compliance with the cybersecurity requirement of the AI Act using currently existing and mature cybersecurity techniques and measures, highlighting the need for technological advancements in parallel to standardisation in this area.

1 Introduction

The EU AI Act (European Commission 2021b) represents a significant milestone in the regulation of artificial intelligence (AI) technologies. This legislation aims to establish a horizontal framework for trustworthy AI systems that are safe, secure and compliant with European fundamental rights and values. At its core, the AI Act is designed to address risks to health, safety and fundamental rights specifically associated with AI technologies by setting legally binding requirements for high-risk AI systems. Prior to the deployment on the EU market, a provider of a high-risk AI application must adopt the necessary organisational and technical measures to achieve conformity with the requirements. Behind this approach is the intention to build public trust in AI as a transformative technology and ensure that it benefits the society as a whole. Note that at the time of publication the AI Act is still a legislative proposal and awaits final adoption. By "AI Act", this report refers to the original European Commission proposal (European Commission 2021b).

In line with a well-established EU system of product safety regulation (European Parliament and Council of the European Union 2012), harmonised standards are one of the main means to achieve compliance and conformity with the legislative requirements. Harmonised standards will be developed by European Standardisation Organisations following a standardisation request of the European Commission published in May 2023 (European Commission 2023). It is important to note that the standardisation request makes clear that the reach of AI standardisation is determined by the technological maturity of AI technologies, explicitly referring to the state of the art (SOTA) in technology[1] as the basis of standardisation.

AI is a rapidly evolving field, with novel and emergent approaches, models, and tools being introduced on an increasingly frequent basis (OECD 2023; 2022). This pace has dramatically accelerated in recent months driven by new developments and products on large-scale AI models (Bommasani et al. 2021). In the current AI scientific ecosystem, a wide variety of techniques and approaches of different levels of maturity coexist. Whilst some AI models are based on well-established techniques that have been used for decades, many techniques, in particular those driving the current innovative developments, have only been in use for a few years or even months. Additionally, research has focused primarily on improving the accuracy of models, and the shift towards the consideration of trustworthy requirements has only gained momentum over the past few years, in the light of potential negative consequences of the use of AI in the society (High Level Expert Group on Artificial Intelligence 2019). Therefore, considerations such as robustness, explainability or cybersecurity of AI models are often in earlier stages of research and development.

This report focusses on the requirement of cybersecurity for high-risk AI systems, as set out in Article 15 of the proposed AI Act. The requirements of cybersecurity, accuracy and robustness, are connected to the technical dimension of AI systems and require a deep understanding of the inner workings of AI systems, established technical practices and standards.

Even though established standards and practices in cybersecurity may apply to AI systems as they do to other software systems, AI-specific technological challenges exist and have not yet been the subject of established security practices or specific standards. However, work is increasingly being dedicated to the topic in form of reports, studies and first international standardisation work items (Tabassi et al. 2019; Berghoff et al. 2021; Malatras, Agrafiotis, and Adamczyk 2021; The MITRE Corporation 2022). Currently, for security engineering purposes such as the practical implementation of processes and techniques to secure systems, many AI-specific security approaches and tools may not be considered mature enough to be directly used for properly securing certain AI models individually (Berghoff et al. 2021). The cybersecurity of AI (or AI cybersecurity) is an emerging field that aims to fill this gap, and that strongly relies on ongoing research activities in fields such as security engineering or adversarial machine learning (Papernot et al. 2016). In fact, the main purpose of the AI Act is to address the AI-specific risk in AI technology – as discussed in detail in the AI Act impact assessment (European Commission 2021a) - and as such it can be expected that considerations are needed that go beyond established practice in software security to address its requirements.

In the report, these considerations are elaborated and guidance is provided for standardisation bodies and AI providers that seek to comply with the cybersecurity requirement of the proposed AI Act. These results are summarised in four key messages and recommendations. The report is conducted as part of an ongoing collaboration between the JRC and DG CONNECT, providing scientific and technical support to the development of the AI Act and related standardisation activities.

[1] Defined as generally accepted practices and not as the latest developments in research.

2 Background

Article 3 (1) of the proposed AI Act defines an AI system as *"(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"*, while Annex I covers both machine learning and other approaches to AI such as knowledge-based systems. In this report, the analysis is focused on machine learning, as it currently accounts for the most prominent technical approaches in AI and also poses most of the challenges for AI cybersecurity relevant for the AI Act.

2.1 Machine Learning

Machine learning itself is not homogeneous, and different approaches co-exist that have emerged since the creation of the field. Broadly, the main features of machine learning compared to traditional software can be summarised in three points: 1) reasoning and learning capabilities, 2) strong reliance on data, and 3) stochastic nature of outcomes.

Essentially, in machine learning a so-called model is trained from existing data using statistical and optimisation techniques, which then can be used for a range of reasoning, prediction, decision or generation tasks. For the purpose of this report, three main categories of machine learning approaches are distinguished, providing a rough classification into increasingly more complex AI models and approaches, usually relying on progressively larger training data sizes and more computing power:

1. Traditional machine learning, processing pre-processed features, e.g., linear regression, decision trees, Bayesian network classifiers, or support vector machines, see, for example (Bishop 2007).

2. Advanced machine learning, based on deep learning using neural networks, e.g., deep convolutional neural networks or recurrent neural networks. See, for example (Goodfellow, Bengio, and Courville 2016).

3. Large-scale deep learning systems, such as attention-based large-scale neural networks trained on very large data sets. See (Bommasani et al. 2021).

2.2 AI Cybersecurity

AI cybersecurity is an emerging field of study, collecting and combining knowledge and approaches from different fields such as AI research, adversarial machine learning and general cybersecurity. Different angles can be considered in the interaction between these fields, most prominently the application of AI to enhance and reinforce cybersecurity, the misuse of AI systems for malicious purposes, and the cybersecurity of AI systems. In the context of this report, only the latter is considered.

For the purposes of cybersecurity, AI can be viewed as a type of software and, thus, a goal of the field is to rely on already established practices where possible. However, a range of AI-specific technological challenges exist for AI cybersecurity, and neither many tested security practices nor specific standards have been introduced yet to address them (Papernot et al. 2016). These challenges are mostly connected to newly introduced computing and product lifecycle paradigms by machine learning systems and due to the fact that a growing number of new AI-specific vulnerabilities are being identified for machine learning systems, such as evasion attacks, data poisoning, model stealing, model inversion, or backdoors embedded in models (Berghoff et al. 2021; Malatras, Agrafiotis, and Adamczyk 2021). The MITRE corporation (The MITRE Corporation 2022) provides an especially useful, continuously updated and developed taxonomy and kill chain analysis about such AI-specific attacks.

Challenges for AI Cybersecurity can be roughly grouped into two categories:

1. Organisational challenges related to processes, e.g. harmonising terminologies, threat taxonomies, and definitions across fields and standards; managing AI lifecycle security and AI-specific supply chain security problems; adapting existing security controls for AI software.

2. Research and Development challenges related to techniques, e.g. assessment of attacks on machine learning models; developing AI-specific security measures and hardening models for more advanced AI methodologies; defining metrics and measures for AI cybersecurity and adversarial robustness of AI models; evaluating trade-offs with other requirements such as between accuracy and cybersecurity; developing practical AI threat modelling experience.

Cybersecurity is largely focusing on risk-based approaches, and, as with any novel technology, AI should be appreciated in the light of a trade-off between opportunities and challenges (Nai Fovino et al. 2020). It remains clear that the risks for many of the listed challenges need to be carefully assessed, as well as whether they can be addressed directly. However, cybersecurity has a long history of securing new technologies under new risks and many proven practices will be applicable in this case too, such as organising the cybersecurity of a software systems in such a way that insecure systems are encapsulated in several layers, with outer security controls like access management ('security-in-depth' principle) (Chapple, Stewart, and Gibson 2021).

2.3 AI Cybersecurity in the AI Act

AI cybersecurity is covered in the AI Act in Article 15, albeit not as an individual requirement, but together with accuracy and robustness. In addition, the cybersecurity requirement is further explained in Recital 51. Building on Article 15 and Recital 51, the standardisation request, Annex II (2.8), provides further operational details on the cybersecurity requirement. As summarised by the JRC (Soler Garrido et al. 2023), operationally, the cybersecurity requirement includes four main elements:

- High-risk AI systems should be ensured and designed to be resilient against attempts to alter their use, behaviour, and performance and to compromise their security properties by malicious third parties exploiting the AI systems' vulnerabilities.
- Organisational and technical solutions shall be implemented to address these goals.
- A cybersecurity risk assessment shall be done for high-risk AI systems.
- Technical solutions shall be appropriate to the relevant circumstances and risks.

After the AI Act becomes applicable all high-risk AI systems as defined by the legislation would have to undergo a conformity assessment and comply with the cybersecurity requirement before they can be used or put into service in the EU market.

There are two main possibilities to ensure conformity. One option is compliance with harmonised standards as laid down in Chapter 5 of the AI Act. Recital 61 and Article 40 state how harmonised standards can provide a presumption of conformity with the requirements of the legislation.

Harmonised standards are however always voluntary and the provider of an AI system can always demonstrate a conformity with the requirements of AI Act without relying on harmonised standards, which provides a second option to ensure conformity. It should be noted however that according to the European Commission 2022 standardisation strategy *"standards are at the core of the EU single market"* and *"delivered great benefits (…) creating a level playing field in the single market (…)"* and the goal should be that harmonised standards will remain the main means of showing compliance also in case of the AI Act.

2.4 AI Cybersecurity Standardisation

In compliance with regulation (EU) No 1025/2012 on European standardisation (European Parliament and Council of the European Union 2012) and with the new standardisation strategy of the European Commission (European Commission 2022a), the AI Act foresees in Recital 61 that *"standardisation should play a key role to provide technical solutions to providers to ensure compliance with this Regulation"*. The first AI standardisation request published by the European Commission in May 2023 provides a formal mandate to develop the standards required in support of the future AI regulation.

The standardisation request sets the level of expectations related to the AI Act requirements, including on cybersecurity. In addition, it explicitly acknowledges that standards should be developed considering the state of the art (SOTA) in technology. In addition to the central reference to the AI Act, for cybersecurity, the AI standardisation request also refers to the proposed Cyber Resilience Act (CRA) (European Commission 2022b). European standardisation deliverables shall take due account of the essential requirements for products with digital elements as listed in Sections 1 and 2 of Annex I in the CRA. If an AI system falls under the scope of both CRA and AI Act, and if it fulfils the essential requirements of the CRA, it should be deemed compliant with the cybersecurity requirement in Article 15 of the AI Act (see Recital 29, CRA). In this case, the main arguments of this paper can be considered equally applicable, since also the CRA requests a risk-based approach to the cybersecurity of systems.

The standardisation request is addressed to CEN-CENELEC (European Committee for Standardization and European Committee for Electrotechnical Standardization), with a requirement to consult ETSI (European Telecommunications Standards Institute) in specific areas of standardisation, such as, notably, in cybersecurity.

It is important to note that the standardisation request makes clear that the reach of AI standardisation is determined by the technological maturity of AI technologies, explicitly referring to the technological state of the art as the basis of standardisation. Therefore, in terms of technical cybersecurity measures, AI standards are expected to cover mature and generally accepted measures for cybersecurity of AI systems. This anticipates an ongoing need for standardisation work on AI cybersecurity in the coming years. However, this does only imply technological limits for securing individual models, and not a limit to the development of standards. Instead, an initial set of cybersecurity standards in response to the initial AI Act standardisation request should provide a strong basis in terms of available technical controls for achieving and measuring AI cybersecurity together with general outcome-based horizontal cybersecurity requirements for high-risk AI systems. Recently, the JRC published an analysis of the preliminary AI standardisation work plan in support of the AI Act (Soler Garrido et al. 2023) and these considerations regarding cybersecurity were summarised as follows:

- Many non-AI-specific security measures can largely be taken from the ISO 27000 series, which includes well established procedures on organisational principles, risk management and security controls. However, current existing standards are not yet adapted to be used for AI software.
- Standards on AI-specific cybersecurity are beginning to be developed on the international level, but are not yet available, most notably ISO 27090 on AI-specific mitigation and controls. Work at European level is just beginning, and may cover AI cybersecurity elements in dedicated standards on risk and trustworthiness.

3 Guiding principles to address the cybersecurity requirement of the AI Act

This section presents four guiding principles to facilitate the understanding and compliance with the cybersecurity requirement of the AI Act. These guiding principles are developed in the policy context of the AI Act and its standardisation, and are based on the analysis of the current state of play in AI cybersecurity and on established cybersecurity principles.

Cybersecurity is a field that has gained a lot of experience with the advent of new computing paradigms such as cloud computing, Internet of Things (IoT) and edge systems, and it is now evolving to address the risks brought by the uptake of AI technologies.

Guiding principle 1: The focus of the AI Act is on AI systems.

As stated in its Article 1, the AI Act lays down specific requirements for high-risk AI systems. These requirements, including the one on cybersecurity, apply to the high-risk AI system, and not directly to the AI models (or any other internal system component) contained within it.

An AI system, formally defined in Article 3(1), should be understood as a software that includes one or several AI models as key components alongside other types of internal components such as interfaces, sensors, databases, network communication components, computing units, pre-processing software, or monitoring systems.

Although AI models are essential components of AI systems, they do not constitute AI systems on their own, as they will always require other software components to be able to function and interact with users and the virtual or physical environment. The cybersecurity requirement of the AI Act does not apply directly the internal components of the systems (such as the AI models), but instead to the AI system as a whole. For example, an AI-based computer vision system does not only include the AI model, but also the sensors and software components to pre-process and manipulate inputs and outputs. Similarly, an AI chatbot may rely on large language models, but also on a traditional cloud infrastructure to be accessed through the internet and manage the hardware required to run the application.

This perspective is aligned with known approaches in cybersecurity, where the system-level view is often applied when considering the security of complex software systems with concepts such as security-in-depth, applying security controls at different layers of an ICT system (Chapple, Stewart, and Gibson 2021), and top-down threat modelling from the system to its components (Shostack 2014).

Guiding principle 2: Compliance with the AI Act necessarily requires a security risk assessment.

In order to ensure the compliance of a high-risk AI system to the AI Act, the cybersecurity requirement needs to be mapped from the system to individual components in the context of the Risk Management System described in Article 9. Practically speaking, this involves a cybersecurity risk assessment of the system and its components (Ross 2012). In particular, for the risk assessment, AI models shall be considered by taking into account their limitations and vulnerabilities in the context of their interactions with other non-AI components of the system.

This risk-based approach is crucial to determine the details of a cybersecurity solution for individual AI products. This is in line with established practice in cybersecurity, where risk assessments already play an important role, particularly in the most widely used information security standards of the ISO 27000 series (ISO/IEC JTC 1/SC 27 2022). Crucially, within the scope of the AI Act, two levels of risk assessment should be considered:

- The higher regulatory level risk assessment involving all other requirements as described in Article 9.
- A cybersecurity risk assessment to determine security measures appropriate to the specific risks of the AI system and its components, whether AI or not, as described in Recital 51 of the AI Act and Annex II (2.8) of the standardisation request.

This implies that, although the use of an AI system in a given specific context may be considered as high–risk from a regulatory standpoint (Annex III of the AI Act), the system could present limited cybersecurity risks due to the way it is designed and operates. In the regulatory risk-assessment stage, the interplay with other requirements of the AI Act should be considered. Further, the cybersecurity risk assessment will determine the type and implementation method of the appropriate specific security risk mitigation measures for the system and its components, taking into account the internal architecture of the system and its intended application context. In practice, cybersecurity measures will be balanced in complexity with other requirements for specific AI systems.

As an illustration, an AI system designed to only be run locally by trained practitioners where the inputs to the AI model cannot be directly controlled by an adversary may not need a high level of robustness against adversarial attacks.

Guiding principle 3: Securing AI systems requires an integrated and continuous approach using proven practices and AI-specific controls.

The cybersecurity of AI systems should rely on a combination of existing controls for software systems and AI-specific controls on individual models. This should be applied at different system levels throughout the lifecycle of the system following a holistic approach based on the security-in-depth and security-by-design principles. In general, when possible, AI systems should not be treated differently from other known complex software architectures. For instance, securing AI systems could take as a model already well developed cybersecurity approaches for other complex digital systems, such as for cloud or Internet of Things (IoT) systems (Google 2023).

However, there is no one-size-fits-all solution, as securing individual AI systems will depend on the outcomes of the risk assessment considering the internal design of the system and its intended application context. An approach following known security principles should be adopted to ensure that AI systems achieve compliance with the cybersecurity requirement of the AI Act during their lifecycle.

Cybersecurity risks of AI systems may be addressed with the "technical solution appropriate to the relevant circumstances and risks", as stated in Annex II (2.8) of the standardisation request of the European Commission. These risks can be addressed through known technical and organisational measures, either at the level of components or at any suitable system level.

AI-specific vulnerabilities may be addressed at AI component level using specific measures, whenever the state of the art provides suitable solutions. However, they can also be addressed at system level using complementary non-AI specific technical and organisational measures, possibly adapted to be used as measures to secure AI components, if needed. Explicitly, it may even be more suitable or "appropriate to the relevant circumstances and estimated risks" to handle an AI-specific vulnerability at a higher system level following a security-in-depth approach. Typical AI-specific cybersecurity challenges for machine learning models are listed in Sec. 2.2. Examples of AI-specific security controls at model level are adversarial training against specific types of evasion attacks and knowledge distillation. At system level, examples of non- AI-specific controls are access control measures to restrict user interaction with the input to AI models, data sanitisation methods on inputs data, and the broader set of security controls implementing the security in-depth and security by default paradigms.

It is important to note that, as it is the case with any software component, vulnerabilities of AI models need to be assessed in the context role that the model plays within the internal architecture and design of the system, and its interactions with other components.

Guiding principle 4: There are limits in the state of the art for securing AI models.

A wide variety of AI technologies of different degrees of maturity coexist in the current AI landscape, ranging from traditional machine learning models to the latest large scale deep learning architectures (see Section 2.1). Open technological questions and challenges, including those for securing AI systems, are connected especially to the more complex and newer models. This will likely lead to an increasing trend of complexity for securing AI systems with the growing more widespread adoption of such advanced AI models.

Therefore, not all AI technologies might be ready for use in AI systems designed to be deployed in high-risk scenarios, unless their limitations in terms of cybersecurity are properly addressed. Some examples for current technological challenges in securing machine learning models are listed in Sec. 2.2.

As defined in the Annex II of the standardisation request, the harmonised standards that will be developed for the AI Act are not expected to go beyond the available state of the art. This should include AI-specific security measures that have reached a certain level of maturity in practice, but consequently, standards will not specifically cover all limitations in AI technologies, particularly the emergent ones.

Instead, cybersecurity standards are expected to endorse elements of the presented integrated approach and security-in-depth system perspective need not be limited themselves by any technological limitation in securing individual AI models. Deploying AI-specific security measures beyond the state of the start, i.e. not covered in harmonised standards, remains possible, but it will be up to the provider of the AI system to ensure conformity, as the presumption of compliance does not apply outside of coverage of harmonised standards.

4 Conclusions

The global AI landscape is undertaking a rapid transformation, with new emergent technologies moving quickly from the academic field into practical applications. This enables new use-cases that were considered until recently out of reach using the available technology. Whilst AI is expected to bring many benefits to the European society, it also creates new threats that need to be properly addressed.

In this context, the AI Act is of particular relevance, becoming a crucial element to ensure that AI systems deployed in high-risk scenarios are safe to use and respect fundamental rights. This report focuses on the cybersecurity requirement for high-risk AI systems, as set out in Article 15 of the regulation. The result of the high level analysis carried out on the implementation of this requirement are the following four guiding principles summarised hereafter:

1. The focus of the AI Act is on AI systems. The structure of AI systems involves a range of internal components, some of which are AI-related, whilst others are not. Although AI models are essential components of AI systems, they do not constitute AI systems on their own. The cybersecurity requirement set out by the AI Act apply to the AI system as a whole and not directly to its internal components.
2. Compliance with the AI Act necessarily requires a security risk assessment. In order to ensure that an AI system complies with the cybersecurity requirement of the AI Act, a security risk assessment should be conducted considering the internal architecture of the AI system and the intended application context. This cybersecurity risk assessment, carried out in the context of the Risk Management System described in the Article 9 of the AI Act, aims to identify the specific security risks, translate the higher level cybersecurity requirement of the regulation down to specific requirements for the components of the system, and implement the necessary mitigation measures.
3. Securing AI systems requires an integrated and continuous approach using proven practices and AI-specific controls. This process should leverage current cybersecurity practices and procedures, using a combination of existing controls for software systems and AI-model specific measures. AI systems are the sum of all their components and of their interactions. A holistic approach following the security-in-depth and security-by-design principles should be adopted to ensure that AI systems achieve compliance with the cybersecurity requirement of the AI Act during their lifecycle.
4. There are limits in the state of the art for securing AI models. A wide variety of AI technologies of different degrees of maturity coexist in the current AI landscape. Not all AI technologies might be ready for use in AI systems designed to be deployed in high-risk scenarios, unless their limitations in terms of cybersecurity are properly addressed. In some cases, particularly for emergent AI technologies, there are inherent limitations that cannot be exclusively addressed at the level of the AI model. In those cases, compliance with the cybersecurity requirement of the AI Act can only be achieved following the holistic approach described earlier.

These guiding principles are intended to help all relevant stakeholders in addressing the cybersecurity requirement for high-risk AI systems, as set out in Article 15 of the proposal of the European Commission for the AI Act.

In the context of the rapidly evolving AI landscape, harmonised standards developed for the AI Act are not expected to cover approaches beyond the currently available state of the art in AI cybersecurity. However, these limitations in the state of the art, do not necessarily impede the compliance of AI systems that make use of emergent AI technologies with the cybersecurity requirement of the AI Act, nor should they hinder the development of standards. High-risk AI systems can still achieve compliance if they mitigate appropriately the overall cybersecurity risks of the system through other complementary measures, following an integrated approach combining well-established cybersecurity practices and procedures with AI specific measures.

It is nonetheless important to acknowledge that this may not always be possible and that for some high-risk AI systems making use of emerging AI technologies it may not be feasible to achieve compliance with the cybersecurity requirement of the AI Act using currently existing and mature cybersecurity techniques and measures, highlighting the need for technological advancements in parallel to standardisation in this area.

References

Berghoff, Christian, Battista Biggio, Elisa Brummel, Vasilios Danos, Thomas Doms, Heiko Ehrich, Barbara Hammer, et al. 2021. 'Towards Auditable AI Systems - Current Status and Future Directions'. Whitepaper. Bundesamt für Sicherheit in der Informationstechnik, Deutschland.

Bishop, Christopher M. 2007. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 1st ed. Springer.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2021. 'On the Opportunities and Risks of Foundation Models'. Preprint. arXiv. http://arxiv.org/abs/2108.07258.

Chapple, Mike, James Michael Stewart, and Darril Gibson. 2021. *(ISC)^2 CISSP Certified Information Systems Security Professional Official Study Guide*. 9th ed. Indianapolis: John Wiley and Sons.

European Commission. 2021a. 'Impact Assessment Accompanying the Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (SWD(2021) 84 Final)'. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021SC0084.

———. 2021b. 'Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (COM(2021) 206 Final)'. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

———. 2022a. 'An EU Strategy on Standardisation: Setting Global Standards in Support of a Resilient, Green and Digital EU Single Market (COM(2022) 31 Final)'. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022DC0031.

———. 2022b. 'Proposal for a Regulation on Horizontal Cybersecurity Requirements for Products with Digital Elements and Amending Regulation (EU) 2019/1020 (COM(2022) 454 Final)'. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0454.

———. 2023. 'Commission Implementing Decision on a Standardisation Request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in Support of Union Policy on Artificial Intelligence (C(2023) 3215 Final)'. https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215.

European Parliament and Council of the European Union. 2012. 'Regulation (EU) No 1025/2012 on European Standardisation'. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32012R1025.

Goodfellow, Ian J., Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA, USA: MIT Press.

Google. 2023. 'Google Secure AI Framework (SAIF)'. *Google Safety & Security* (blog). 2023. https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/.

High Level Expert Group on Artificial Intelligence. 2019. 'Ethics Guidelines for Trustworthy AI'.

ISO/IEC JTC 1/SC 27. 2022. 'ISO/IEC 27001 - Information Security Management Systems - Requirements'. https://www.iso.org/standard/27001.

Malatras, Apostolos, Ioannis Agrafiotis, and Monika Adamczyk. 2021. 'Securing Machine Learning Algorithms'. ENISA. https://doi.org/10.2824/874249.

Nai Fovino, Igor, Geraldine Barry, Stephane Chaudron, Iwen Coisel, Marion Dewar, Henrik Junklewitz, G. Kambourakis, et al. 2020. 'Cybersecurity, Our Digital Anchor'. JRC121051. European Commission - Joint Research Centre. https://doi.org/10.2760/352218.

OECD. 2022. 'Harnessing the Power of AI and Emerging Technologies'. https://www.oecd-ilibrary.org/content/paper/f94df8ec-en.

———. 2023. 'The OECD Artificial Intelligence Policy Observatory'. 2023. https://oecd.ai/en/.

Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. 2016. 'Towards the Science of Security and Privacy in Machine Learning', 1–19.

Ross, Ronald S. 2012. 'Guide for Conducting Risk Assessments'. 800-30 Rev 1. NIST. https://doi.org/NIST.SP.800-30r1.

Shostack, Adam. 2014. *Threat Modeling: Designing for Security*. 1st ed. Wiley Publishing.

Soler Garrido, Josep, Delia Fano Yela, Cecilia Panigutti, Henrik Junklewitz, Ronan Hamon, Tatjana Evas, Antoine-Alexandre André, and Salvatore Scalzo. 2023. 'Analysis of the Preliminary AI Standardisation Work Plan in Support of the AI Act', no. KJ-NA-31-518-EN-N (online). https://doi.org/10.2760/5847 (online).

Tabassi, Elham, Kevin J. Burns, Michael Hadjimichael, Andres D. Molina-Markham, and Julian T. Sexton. 2019. 'A Taxonomy and Terminology of Adversarial Machine Learning'. Draft NISTIR 8269. https://doi.org/10.6028/NIST.IR.8269-draft.

The MITRE Corporation. 2022. 'MITRE ATLAS'. https://atlas.mitre.org/.

List of abbreviations

AI	Artificial Intelligence
CEN	European Committee for Standardization
CENELEC	European Committee for Electrotechnical Standardization
CRA	Cyber Resilience Act
ETSI	European Telecommunications Standards Institute
ISO	International Organization for Standardization
IoT	Internet of Things
OECD	Organisation for Economic Co-operation and Development

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

— at the following standard number: +32 22999696,

— via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

EU Science Hub
joint-research-centre.ec.europa.eu

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

@eu_science

Publications Office
of the European Union