This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Three challenges to secure AI systems in the context of AI regulations

RONAN HAMON[1], HENRIK JUNKLEWITZ[1], JOSEP SOLER GARRIDO[2] and IGNACIO SANCHEZ[1]
[1]European Commission, Joint Research Centre, Ispra, Italy (e-mail: firstname.lastname@ec.europa.eu)
[2]European Commission, Joint Research Centre, Sevilla, Spain (e-mail: firstname.lastname@ec.europa.eu)

Corresponding author: Ronan Hamon (e-mail: ronan.hamon@ec.europa.eu).

ABSTRACT This article examines the interplay between artificial intelligence (AI) and cybersecurity in light of future regulatory requirements on the security of AI systems, specifically focusing on the robustness of high-risk AI systems against cyberattacks in the context of the European Union's AI Act. The paper identifies and analyses three challenges to achieve compliance of AI systems with the cybersecurity requirement: accounting for the diversity and the complexity of AI technologies, assessing AI-specific risks, and developing secure-by-design AI systems. The contribution of the article consists in providing an overview of AI cybersecurity practices and identifying gaps in current approaches to security conformity assessment for AI systems. Our analysis highlights the unique vulnerabilities present in AI systems and the absence of established cybersecurity practices tailored to these systems, and emphasises the need for continuous alignment between legal requirements and technological capabilities, acknowledging the necessity for further research and development to address the challenges. It concludes that comprehensive cybersecurity practices must evolve to accommodate the unique aspects of AI, with a collaborative effort from various sectors to ensure effective implementation and standardisation.

INDEX TERMS Adversarial Machine Learning, Artificial Intelligence, Conformity Assessment, Cybersecurity, Lifecycle Management, Regulation, Risk Management, Trustworthy AI

I. INTRODUCTION

Artificial Intelligence (AI) is a pivotal element of digital transformation [1]–[3], emerging from decades of advancements across scientific fields, in particular computer science and statistics, and propelled by an increase in the availability of data and computational resources. AI technologies are being more and more democratised and integrated in products and services, impacting businesses, organisations, and individuals in many sectors such as transport, healthcare, or education [3].

AI, sometimes referred to as software 2.0 [4], represents a paradigm shift in programming. Unlike traditional programming, which requires explicit instructions to perform a task, AI infers logical steps from large corpus of data and/or abstract concepts [5], only leveraging programming to support data management or algorithm development. With the deployment of AI in critical applications, making AI systems trustworthy [6] is of prime importance to ensure that they are safe and remain aligned with the respect of fundamental rights and societal values. The proposed European Union's AI Act [7], for which a political agreement has been reached at the end of 2023 [8], exemplifies regulatory efforts around the world to establish harmonised rules for AI products. The core of its approach lies in the idea that it is necessary to take into account the purpose and the context of use in order to fulfil the essential trustworthiness requirements in an effective and proportionate manner.

This paper focuses on the cybersecurity requirement as laid out in Article 15(4) of the proposed EU AI Act, which also connects to other international policy initiatives [9]–[11] and requires high-risk AI systems to *be resilient as regards attempts by unauthorised third parties to alter their use or performance*. The recital 51 elaborates further the rationale of the requirement, clarifying that providers of AI systems have to carry out a security risk assessment and implement suitable organisational and technical solutions to mitigate the security risks. Given that many cybersecurity processes are not fully developed for AI systems, adapting them to take into account the unique features of AI systems poses several challenges to achieve compliance [12]–[14].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

The contribution of this article is twofold. First, it provides in Section II a brief overview of the so-called field of AI cybersecurity at the interplay between AI and classical cybersecurity. Then, it delineates three challenges related to AI cybersecurity in the context of regulations: accounting for the diversity and the complexity of AI technologies (Section III), assessing AI-specific risks (Section IV), and developing secure-by-design AI systems (Section V). For each challenge, the integration of AI into known and proven cybersecurity practices and approaches is examined. An illustration of these challenges is provided in Section VI, presenting an example of evasion attacks against an automated application screening system powered by a language model. This analysis emphasises the importance of the development of an integrated and system-level approach to securing AI models relying on established practices in software security engineering [15], until more generally effective techniques for securing state-of-the-art AI models become available.

II. THE EMERGING FIELD OF AI CYBERSECURITY

AI cybersecurity refers to the field uniting classical cybersecurity and AI, which can be described according to four dimensions:

1) **AI to enhance cybersecurity** (*opportunities for stronger resilience*): AI enables cybersecurity to tackle tasks that were not possible until recently, including capabilities to strengthen the resilience of systems and services, or enlarging the possible focus of threat analysis, attack attribution and incident response [16], [17].
2) **Robustness and vulnerabilities of AI** (*challenges for resilience*): The integration of AI into widespread digital systems introduces new types of vulnerabilities that can be exploited by malicious actors, with possibly greater attack surface and impact [18].
3) **AI to deter and fight cyberattackers** (*opportunities for deterrence*): AI-powered capabilities are at the disposal of law enforcement agencies, defence services and military bodies to actively deter and fight cybercrime and fight adversaries [19], [20].
4) **Malicious use of AI** (*challenges for deterrence*): AI systems can be abused by malicious actors to improve the scope and strength of cyberattacks or used in an unforeseen way to conduct criminal activities [21].

The focus of cybersecurity requirements of current AI regulations is on point 2), related to the security of AI systems. In the rest of this section, we describe this aspect in greater details, and dive into the question of the standardisation process that will take place to facilitate conformity for providers of AI systems.

A. SECURING AI

Securing AI consists in collecting and combining knowledge, approaches, technologies, practices, and policies that are designed to protect AI systems and their data from cyber threats. These threats may result in unauthorised access, information disclosure, theft of material, damage, or more generally any form of disruption of the service provided by the AI system.

On the one hand, it is clear that AI, as a type of software, can benefit from the experience of classical cybersecurity, allowing AI cybersecurity to rely on already established information security practices whenever possible. Well-tested approaches in cybersecurity such as risk modelling, organisational aspects of information security, and system-level security controls, can apply to some degree as much to AI systems as to any other software systems. On the other hand, AI exhibits intrinsic features that disrupt the traditional approach of securing software. Recent years have already seen initiatives from the software and cybersecurity communities to advance AI cybersecurity by adapting and enlarging existing frameworks, such as the MITRE ATLAS [22], taxonomies [23], [24] or AI threat landscape analyses [18], [25], [26].

From a scientific standpoint, the analysis of security properties of AI systems have been regrouped under the topic of adversarial machine learning that focuses on research into intentionally attacking, breaking or misusing features of machine learning models and measuring robustness against these malicious actions [13]. The field built upon early theoretical work on learning against adversaries [27], [28] and went closer to cybersecurity applications such as spam filters [29], [30], before evolving more concretely into security principles for machine learning systems [5], [31], [32]. These topics are all relevant in the task of practically securing AI systems since, by now, a range of new vulnerabilities for AI models have been identified [26].

B. STANDARDISING AI CYBERSECURITY FOR CONFORMITY ASSESSMENT

Studying AI-specific vulnerabilities and connecting them with cybersecurity concepts is crucial in the establishment of standards that will support the implementation of future regulations on AI, a question that is now central with the expected adoption of the AI Act [8]. Harmonised standards [33] will play a key role in defining technical requirements and guidance to ensure the security of AI systems. This helps both providers to fulfil regulatory requirements (with a presumption of conformity if standards are applied), and assessment bodies to check conformity, providing methodologies for verification, validation, auditing or certification of systems [34].

From a conformity assessment perspective, a clear distinction should be made between an AI model and an AI system [15]. A model describes a mathematical and algorithmic construction, aiming at processing inputs using a given set of advanced techniques. Conversely, an AI system — as a term from systems engineering — is usually meant to describe the integration of one or several AI models alongside additional non-AI components, for instance computing units, communication modules, interfaces, databases or sensors, with a clear intended purpose set up by a manufacturer. Although AI models are the 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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

interact with users and the virtual or physical environment. Ensuring the conformity of AI systems with future regulations regarding their cybersecurity does not necessarily require making models secure. Even if limitations due to technical gaps and lack of scientific maturity of AI technology may limit compliance [34]–[36], alternative approaches at system level may exist.

In the standardisation communities, many activities have been initiated to address various technical, ethical and organisational aspects of AI systems on topics ranging from risk modelling to AI hardware security [24], [34], [35], [37]–[39]. The analysis of the efforts required to standardise AI cybersecurity [15], [36], [40], [41] suggests an approach in line with traditional cybersecurity practice, where challenges to secure AI are acknowledged, and limits in AI technology at the model level are addressed by traditional cybersecurity practices taking into account additional costs and/or impacts on system performance. While many non-AI-specific security measures can largely be taken from the ISO/IEC 27000 series, which includes well-established procedures on organisational principles, risk management and security controls, current standards are, however, not yet adapted to be used for AI software. Work at European level is just beginning, and will cover AI cybersecurity, either in dedicated AI cybersecurity standards, or as part of more transversal standards on AI risk management and trustworthiness.

III. CHALLENGE: ACCOUNTING FOR THE COMPLEXITY AND DIVERSITY OF AI TECHNOLOGY

The definition of AI varies according to the context [5] and intersects with definitions coming from different fields such as machine learning [42], natural language processing [43], robotics [44], or computer vision [45]. AI-based software, especially the most popular ones, typically exhibits a range of features that sets them apart from traditional software: *reasoning and learning* [46], encompassing actions and capabilities usually reserved to human intervention, such as perception and understanding (e.g., recognising objects, reading texts, evaluating scenes, etc.) and planning (e.g., taking action, answering questions, elaborating strategies, etc.); *data-driven* [47], highlighting the capacity of AI systems based on machine learning techniques to process, analyse, learn, and extract patterns from data sets, potentially very large; *opacity* [48], describing the absence of explicit mechanisms and rules that limits the understanding of the functioning of the system; *unpredictability* [49], referring to the presence of stochasticity in the development and/or exploitation process that, coupled with the high non-linearity of many AI systems and the opacity, makes the outcomes largely unpredictable.

A. TERMINOLOGY OF AI AND CYBERSECURITY

Future legislation will have to rely on technical concepts and terms to describe requirements, either implicitly or explicitly. However, the terminology of AI originates from technical works in computer science and statistics, with blurred limits in the meaning of terms and concepts, depending on the context of applications or fields. The proposed AI Act for example makes use of the term *accuracy* in a broader meaning than usually implied by the widely used statistical metric of the same name, referring instead to the capability of the AI system to perform the task the system has been designed to [33]. Regarding cybersecurity, it both mentions terms from AI such as *adversarial examples* and *data poisoning* and from cybersecurity such as *security risk assessment*, without defining their exact scopes. This approach relies on subsequent works to clear ambiguities of interpretation, but will require a harmonisation of the terminology between cybersecurity practices, AI research, and law.

Harmonising the terminology of AI and AI cybersecurity has been acknowledged as a challenge by European standardisation bodies [35] and led to the development of terminologies [18], [23], [25] and a proposal of making standards on AI concepts and terminology (ISO/IEC 22989, ISO/IEC 23053). Partial coverage of AI-specific cybersecurity terminology is included in ISO/IEC 24028 on trustworthiness in AI, ETSI/SAI 002 and in the forthcoming ISO/IEC 27090. Further inconsistencies and gaps remain to be addressed. For instance, the notion of robustness is not uniquely defined, but generally revolves around the capacity for a system to maintain its level of performance under any expected and unexpected circumstances. Broader definitions of robustness usually encompass both problems of general robustness and robustness in a cybersecurity context [50], with the latter explicitly including the resilience towards malicious attacks against the integrity or purpose of the system. This is because the generalisation behaviour of AI models governing their performance at the edge or outside their trained data representation plays a key role in the evaluation of their robustness, regardless of whether perturbations and edge cases occur intentionally (adversarial robustness, or cybersecurity) or not (general robustness) [51]. The differentiation between general robustness and adversarial robustness is of prime importance to understand whether some types of attacks affecting the integrity of systems, such as evasion attacks, should be treated as an issue of general robustness or of cybersecurity.

Likewise, definitions given in various taxonomies for *evasion attacks* vary in their scope. Evasion attacks were introduced in the context of machine learning-based security controls such as network intrusion detection systems or spam filtering [29]–[32] and were extended to deep neural networks [52]–[56] later. They are defined in [23] as *"[...] manipulat[ing] input samples to evade (cause a misclassification) a trained classifier at test time."*, in [22] as *"[...] craft[ing] adversarial data that prevent a machine learning model from correctly identifying the contents of the data."*, and in [57] as *"[...] creat[ing] an input to an operating ML [Machine Learning] system that reliably produces a different output than its creators intend."*. While these definitions convey the main idea behind evasion attacks, they fall short of providing tight boundaries or consistency.

More generally, clear definitions of technical notions need to be established, to ensure that they are compatible with their

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

meaning in legal texts, in particular for new terminology that is introduced with novel developments.

B. BEYOND SUPERVISED MACHINE LEARNING

Regarding the security of AI, the emphasis in research has been over the past years given to supervised machine learning, i.e., a type of machine learning where the objective is to learn a mapping between input and output variables from labelled training data and which is widespread in current applications. Other important techniques will nonetheless be subject to security assessments in the context of regulations. Definitions of AI at policy level, such as the OECD definition [58], include, besides supervised machine learning, other techniques such as generative AI, logic- and knowledge-based approaches or search and optimisation methods. This broadens the scope of AI cybersecurity in two ways: 1) security risks posed by advanced machine learning techniques should be studied and their consequences in terms of conformity assessment anticipated, even if mitigation techniques may not be readily available; 2) methodologies for security assessment of other techniques should be developed, taking into account their lower susceptibility to security vulnerabilities. Even if the research frontier has shifted to encompass these approaches, these works are not yet mature enough to be integrated in standards.

Three representative examples of advanced machine learning techniques that may pose additional security risks are: *reinforcement learning* [59], where an agent learns through negative and positive rewards to make decisions by taking actions in an environment to achieve a goal, and which can be tricked into taking harmful actions [60]; *federated learning* [61], a distributed approach to train models on devices, and that can be compromised by exploiting the decentralised nature of the training [62], [63]; *foundation models and general purpose AI models* [64], self-supervised models trained on massive amounts of data, and that have been particularly notable for text [65]–[67] and text-to-image generation [68], [69], but also pose for additional security risks [70], [71] in terms of supply chain.

C. SECURITY, SAFETY, AND GENERALISABILITY

Safety and generalisability are two concepts that overlaps with security but have also their own specificities. While security is concerned with the protection of the system, preventing adversaries or adverse conditions to impact negatively its functioning, safety aims to prevent a system to impact its environment in an undesirable or harmful way, ranging from physical and mental safety of individuals to damage to the environment. As for generalisability, it is more about the capacity of machine learning systems to operate reliably on unknown data. Clarity in the distinction between all aspects is essential to ensure that each aspect is well-covered.

The field of AI safety that emerged over the past years [72] illustrates this possible confusion, covering the prevention of accidents or unintentional misuses caused either by an external threat or because of a malfunction of the system, but also with the recent discussions about artificial general intelligence (AGI), measures to ensure that advanced AI systems will remain aligned with human values. As such, it encompasses parts of cybersecurity by promoting a resilience against malicious attacks, but without encompassing all aspects of cybersecurity related to the confidentiality, integrity, and availability of systems.

Additionally, many research works on adversarial machine learning, albeit motivated by security problems, tend to be concerned with fundamental questions of generalisability [52], [73] that are not necessarily relevant for cybersecurity problems [13], [30], [74]. For example, many works on adversarial examples rely on restricted threat models based on constrained optimisation (e.g., L_p-norm based adversarial attacks aiming enforcing low-intensity perturbations) that may provide valuable insights about the functioning, accuracy and reliability of models, but is of limited use when connected to real-world problems [75]. The technical challenges considered as core components to any cybersecurity conformity testing with regulatory requirements remain open scientific questions, such as the feasibility of measuring robustness against cyberattacks on machine learning models [74], [76], or properly assessing the strength of defences [77].

D. INTERDEPENDENCY BETWEEN REQUIREMENTS

Trustworthiness of AI systems, as implied by regulatory approaches such as the proposed AI Act, does not result from conformity to isolated requirements. Instead, trustworthiness-by-design principles promote an interdependence of a range of requirements and obligations [6], [48], only one of which is cybersecurity. Other important requirements include transparency, human oversight, documentation, data governance and quality, logging, risk management, robustness and accuracy. While it should be desirable to simultaneously adhere to all requirements the best possible way, addressing cybersecurity inevitably involves managing trade-offs between security and other desired features, in particular accuracy and robustness. In cybersecurity, this is a well established practice that is as part of security risk assessments of software systems [78], [79], and it stands to reason that this practice will become standard for AI cybersecurity as well. In the context of the proposed AI Act, it will be crucial in the management of the risks to safety, health, or fundamental rights of individuals that may arise from security issues. However, practical implementations have to be established in order to understand the specific effects of these interdependencies for AI systems.

Increasing the coverage in a single requirement may come at the expense of others, and the optimal balance will depend on the specific requirements and constraints of the AI system [48]. The trade-off between accuracy, robustness, and cybersecurity is a critical challenge facing providers of AI systems. Achieving high accuracy on complex tasks requires large amounts of data and complex models, which has an impact on the robustness of the system and its cybersecurity. Conversely, increasing security often impairs accuracy [80], [81], and comes with additional development costs. In other

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

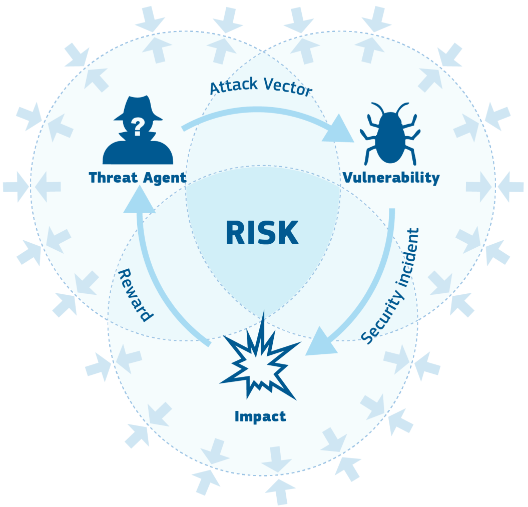

FIGURE 1. Conceptual model depicting the logical links between the different components of the cybersecurity risk (adapted from [83]).

cases, requirements may complement each other and/or addressed jointly, which is for example the case between cybersecurity and data governance requirements.

Finding the right balance in the context of each specific AI system use case to ensure optimal functioning and mitigate risks will be fundamental to achieve compliance. For example, in a critical application such as medical diagnosis, the trade-off may lean more towards accuracy and robustness, with less stringent cybersecurity measures. On the other hand, in a financial fraud detection system, the trade-off may require a higher level of security and robustness, with somewhat lower accuracy. While standards may need to ensure the interdependencies and trade-offs between requirements are addressed in their technical specifications, it will be the responsibility of the providers to properly justify their design choices related to these trade-offs, taking into account the risks and context of use.

IV. CHALLENGE: ASSESSING AI-SPECIFIC RISKS

The proposed AI Act advocates a risk-based approach for the regulation of AI on the basis of risks to European and human rights, designating the provider as responsible to ensuring the compliance of systems depending on their level of risks. Cybersecurity has a long history of securing new technologies under new risks and many proven risk assessment practices are applicable in the context of AI systems. Nonetheless, there is need to adapt risk management strategies to cover new vulnerabilities and novel cyberthreats [23], [24], [82].

Risk modelling in cybersecurity is a proactive approach to identifying, managing, and mitigating potential threats to a system. It involves identifying valuable assets, understanding how they might be compromised, and implementing effective controls to prevent or minimise potential damage. This helps organisations to understand security risks in the context of their activities, and prioritise their security efforts accordingly. In principle, cybersecurity considerations in the context of an AI risk assessment process could follow an approach similar to that of other software systems, by focusing on the estimation of cybersecurity risk [24], [82], [84], [85] from an analysis of the following factors (see Fig. 1):

- *threat agents*: individuals or entities responsible for security incidents;
- *vulnerabilities*: weak points in a system that can be exploited by a threat agent to conduct a cyberattack;
- *impacts*: harm or damage that results from the consequences of an attack to a system.

A. THREAT AGENT

Understanding the capabilities, knowledge, intentions, resources, and methods of threat agents is crucial to anticipate potential threats. Even if threat agents have at their disposal a wide range of attacks to compromise AI systems, the technical and financial costs of an attack and their uncertain results may render them practically irrelevant in the case when simpler means can achieve similar results [13].

The level of knowledge of an agent about AI systems includes different elements [12]: the stage of the system in its lifecycle (e.g., training, testing, deployment, see Fig. 2); the technology used (e.g., library, architecture, processes); the level of visibility (e.g., access to the weights, parameters, architecture, training data input-output pairs, processes and methodologies used by the provider). This evaluation should also include the circumstances in which threat actors can be assumed to have the knowledge and resources to implement a technically demanding and uncertain attack.

All these elements are relevant to identify the range of options at the disposal of the attacker, and identify which type of systems in practice will be more at risk than others [12], [14], [30], [36]. This remains particularly complex to evaluate for actual cybersecurity risks since the deployment of AI products is at its infancy.

B. VULNERABILITY

Vulnerabilities in AI systems concern both those present in non-AI (e.g., networking, database, ICT infrastructure), and AI (e.g., models, inference engines) components [18], [25], [26]. Generally speaking, a cyberattack against an AI system usually involves the exploitation of several vulnerabilities in one or more of its components. For instance, an attacker may get access to the system through a classical software vulnerability, and run an AI-specific attack (e.g., evasion attacks) to do lateral movement [18].

A number of vulnerabilities affecting tools to build and deploy AI models have already been reported, such as for Jupyter (*CVE-2022-29241*), TorchServe (*CVE-2023-43654*), or Tensorflow (*CVE-2022-23587*). The main concerns however lie in the existence of attacks exploiting AI-specific vulnerabilities [26], including: *data poisoning* [86], [87], manipulating or injecting false data with the intention to

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

compromise the training of machine learning models; *backdoors* [88], installing AI-specific mechanisms triggering, for specific patterns, unwanted behaviours; *evasion (or adversarial) attacks* [89], crafting inputs to alter the outcome of an AI system; *model extraction and inversion* [90], [91], retrieving the parameters of a model; *membership inference* [92], [93], retrieving data or specific features used in training and testing sets; *latency attacks* [94], inducing a high latency when computing outputs of machine learning models.

These attacks exploit vulnerabilities that differ from traditional vulnerabilities in that they are generally the consequences of the opacity [95]. This limits the capacity to detect and mitigate vulnerabilities before and after the development, and to test the full space of potential user inputs in order to understand how a system may respond to those inputs (e.g., for an exhaustive test coverage). Managing AI-specific vulnerabilities, including an assessment of their severity and their potential impact, the development of patches, their listing in catalogues (e.g., CVE) [96], and their disclosure [97], may be disrupted by this incapacity to precisely characterise vulnerabilities and reproduce exploits. Additionally, attacks as well as mitigation techniques may not transfer well across all versions of the model.

As of today, it is unclear whether and how AI vulnerabilities can be exploited in practice under real environmental conditions to affect the performance of systems, mostly because of the limited amount of documented AI-specific real-world attack events [88] due to the low level of deployment of AI systems. The development of concrete threat scenarios based on realistic applications should help the assessment of the relevance of adversarial machine learning methodologies in operational contexts [98] and a better anticipation.

C. IMPACT

The evaluation of the impact of an attack is traditionally conducted using the CIA model [82], [84], which has already been translated to analysing AI systems [12], [23], [25]:

- *Confidentiality*: impacts related to the disclosure of personal data, or of proprietary data sets and models (e.g., after membership inference or model extraction);
- *Integrity*: impacts related to the degradation of the predictive capabilities of models, either overall or in a targeted way (e.g., after data poisoning or evasion attacks);
- *Availability*: impacts related to the delay of processing, or the stopping of the system (e.g., after latency attacks).

The AI Incident Database [99] provides a collection of incidents and near misses caused by AI systems deployed in the world. With the growing importance of AI components in systems, new incentives to invest time and resources in exploiting their vulnerabilities will appear, indicating a possible transition phase from traditional to AI-specific cyberattacks. This initiative, as well with others (e.g., the OECD AI Incidents Monitor) are well aligned with Action (4) of the Code of Conduct developed by G7 [10], should be complemented by more security-related incident, requiring an active monitoring and reporting of AI-related events.

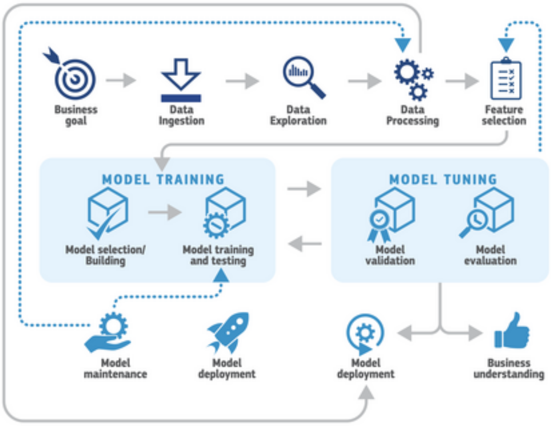

FIGURE 2. Depiction of the typical lifecycle of an AI system based on machine learning techniques [25]. The lifecycle illustrates the multiple stages involved in the development of the system, and includes feedback loops and potentially automated decisions.

V. CHALLENGE: DEVELOPING SECURE-BY-DESIGN AI SYSTEMS

The effective deployment and maintenance of AI models in software development is itself a relatively new field often summarised as MLOps, for which developing secure-by-design approaches is still a subject of active development. Securing AI systems requires indeed a continuous approach to assessing and mitigating threats throughout the whole AI lifecycle. As a result, this poses a set of challenges [34]–[36], [100] to existing frameworks for testing, validating, verifying, and auditing software, which will need to be adapted.

The opacity of AI software introduces a need for major adaptations to the way the security aspects of the different stages of the lifecycle of products and services are handled, the establishment of security controls adapted to AI-specific vulnerabilities, and new approaches to test the security of systems.

A. LIFECYCLE AND SUPPLY CHAIN

Lifecycle management refers to the process of managing each stage of the life of systems, from their specification, design, and implementation, to their operation, maintenance, and decommission. For AI systems, new steps need to be considered (see Fig. 2). This lifecycle depends upon a supply chain that includes all the processes and assets that are involved in the development, delivery, and maintenance of software such as code repositories, build systems, and third-party libraries. For AI systems, the main assets are [25]: *data* (e.g., data sets for training, testing and validation, new input data for inference and/or continuous learning); *models* (e.g., architecture, performance metrics, design purpose, intellectual properties, pre-trained models); *ICT infrastructure* (e.g., training algorithms, networks and file systems, hardware, open-source libraries); *stakeholders* (e.g., data scientists, system provider); *lifecycle processes* (e.g., data engineering, model training, continuous learning); and *cybersecurity assets* (e.g., controls,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

policies). AI systems may also be developed and deployed in various physical locations and hardware, such as data centres, HPC edge devices, or embedded systems, adding potential weaknesses to the whole system. In addition, in a regulatory context, it is important to consider conformity testing with legal requirements as a crucial part of the lifecycle.

Compared to traditional software systems, AI systems have three notable differences regarding their supply chain that increases their attack surface: 1) high dependence to well-curated and reliable yet large data sets; 2) practice of distributing and outsourcing assets to third parties, such as the training of the model, reliance on generic model (such as General Purpose AI or foundation models [64]); 3) prevalent culture of open-source in machine learning, with a widespread use of freely available pre-trained models, software libraries, and data sets, adding significant concerns regarding the security of these assets (e.g., backdoors, malicious code, or intentionally-made bad models), increasing the visibility of attackers on the AI systems by allowing for reverse-engineering [101], and questioning the accountability in case of incidents involving one of these assets.

Securing this supply chain involves protecting assets from any unauthorised access, modification, or exploitation. This can be achieved by implementing various security measures, such as strict access controls, continuous monitoring, or vulnerability assessments, that can be adapted from well-known and standardised data governance and code development practices. However, an analysis is needed to identify areas of supply chain that require additional work and specification, e.g., the handling of training-time attacks such as data poisoning or model backdoors. For both data and software, properly handling the supply chain security is standardised for databases and classical software, for example in the ISO/IEC 27000 series. However, no AI-specific standards have been published so far addressing in particular the above described supply chain issues with AI robustness and cybersecurity which are noticeably different in scale and content. ISO/IEC 27090, whose aim is to provide information to organisations on security threats to AI systems, may eventually address the issue to some extent. Handling the cybersecurity of pre-trained models and open source assets will entail a specific set of organisational and policy measures, including for their supply chain partners, as this situation is not very different from what happens in traditional digital systems, at the condition that these measures do not rely on immature technical solutions.

B. SECURITY CONTROLS

Security controls are implemented to protect digital systems, and are part of mitigation measures selected on the basis of the threats and vulnerabilities identified in the risk assessment. Their integration should come with an assessment of their capacity to detect, track down, and mitigate the effects of attacks. There are several types of security controls that can be implemented to protect AI systems [18], [22], [23], [25], [38]. Some of them are directly connected to security controls of digital systems in general and may only be in need to be reconsidered in the light of AI, such as access control, encryption, monitoring, incident response, security assessments, or similar. Existing practices, such as the one listed in ISO/IEC 27002, will remain of prime importance. Others are more specific to AI technology, such as data validation, input sanitation, model watermarking or crucially incorporating hardened models and defences, may need additional works, and adaptation of practices, as already discussed in ISO/IEC 27090 and other initiatives.

This remains the case for AI systems, with an added complexity due to the opaqueness of their functioning. The relevant assets thus may depend on a particular AI system, its vulnerabilities, and the associated impact assessment. Despite constant research and proposed attacks, some AI-specific vulnerabilities present in machine learning systems do not have applicable defences, mitigation measures or security controls able to efficiently mitigate them. The earliest research literature focused on applying cybersecurity principles to machine learning problems first applied to cybersecurity controls, largely in the context of evasion attacks, such as network intrusion detection systems or spam filtering [30]–[32]. Subsequent works expanded on these [12], [102], aiming at conceptually bringing together the perspectives the complexities of deep neural network models and cybersecurity risk assessment. For example, accepted state-of-the art in research is to use adversarial training to increase the robustness of a model during training time against specific attacks [52], [53], by including in the training set adversarial examples, at increased computational cost and potentially decreased performance on benign data.

As detailed in [15], security controls for AI vulnerabilities may not need to be themselves on AI, or addressing flaws in models. For example, protecting models against membership inference [103] may rely on a combination of model-based controls of differential privacy [104], but also on system-level controls such as restriction of the number of queries possible to AI models for API-based AI systems.

C. TESTING

Conformity assessment is usually conducted with some type of *testing* to measure one or several characteristics of the system and determine whether the system is compliant with requirements. Measuring the cybersecurity of AI systems includes, as for classical information systems [105], tasks such as comprehensive reviews of the system's architecture, vulnerability assessment, identification and access management policies, evaluation of the effectiveness of intrusion detection and prevention systems, an assessment of the security awareness of the system's users, and an evaluation of the likelihood of threats and their impact on the whole system and its environment. Testing can thus describe a very granular technical activity, such as testing for the statistical accuracy or the correct implementation of specific controls. This definition of testing is similar to the meaning of classical software testing [100] and close to the meaning of testing in machine

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

learning [76], which specifically refers to determining the performance of a trained model against a holdout test set.

A measure of interest in the context of AI regulations is the robustness of AI systems, in particular against cyberattacks exploiting AI-specific vulnerabilities [74]–[77]. Generally speaking, two approaches co-exist: *formal methods* [106], [107] and *statistical and empirical approaches* [51], [74]. Methods for formal verification of involve defining a set of constraints or specifications that the model must satisfy and ensuring that the outputs of models are within a certain range for a given input. Even if they can provide certified and guaranteed robustness of AI models [108], [109], many of the known methods are scalable with difficulty to complex deep neural networks [36]. Conversely, statistical and empirical approaches provide local approximations to global robustness measures by calculating metrics on a given set of samples, e.g., from benchmarks [110], [111]. They are more adapted to the large-scale and stochastic nature of machine learning, but are also highly dependent on data sets, model architectures, attack types, and transferability between models [36]. For both types of approaches, defining acceptable thresholds for metrics is context-dependent and not straightforward.

Another approach for testing the cybersecurity of AI systems is to adopt practices from penetration testing [112], actively finding vulnerabilities in systems and exploit them to determine their associated risks. Its application to AI systems as part of red-teaming [113] requirements has emerged recently with the latest developments in generative AI [114], as putting under stress the robustness of controls aims at preventing unexpected behaviours, including hallucinations, jailbreaks or bias, and addressing the risks that come from automating these new technologies at scale [66]. Its current implementation (e.g., in DEFCON 2023) differs significantly from traditional cybersecurity red-teaming by the scale it requires, and the way subjective evaluation is required to define whether a vulnerability has been exploited [115]. Adapting classical cybersecurity practices [116] to AI systems will require to change the focus to data and AI-specific vulnerabilities. Yet, no guidelines to conduct such assessment and evaluate its relevance in the context of conformity assessment, exist, even if development of these practices [117]–[119] is a step in the right direction that needs to be harmonised and assessed.

VI. ILLUSTRATION: EVASION ATTACKS ON LARGE-SCALE LANGUAGE MODELS

Large language models (or LLMs) are machine learning models designed to understand, interpret, and generate text. These models are trained on vast amounts of text data including webpages, books, and any other specific datasets that are designed to help the model learn various tasks such as translation or question-answering, and representing hundreds of billions of words. This allows LLMs to predict the likelihood of a sequence of words and generate coherent and contextually relevant text. Their integration into software-as-a-service products, often through application programming interface (API), makes possible their use in a wide range of applications, from automated customer service chatbots to virtual assistants to content creation, to language translation services. The direct interaction between the users and the AI systems makes the alignment of models to human values a desirable feature to ensure that they not only understand and generate human-like text, but also do it in a manner that is responsible, transparent, and respectful, without generating biased, false, or offensive content.

Techniques used by providers to achieve this alignment include fine-tuning on curated datasets, incorporating human feedback, and specific training methodologies. Testing the alignment is on the other hand done by using standard benchmarks [110], or by applying penetration testing principles and finding potential flaws in the model. For the latter, prompt hacking [113] is a technique in which the tester engages with the system without privileged access or knowledge, in order to bypass restrictions implemented by the system providers to protect the users and avoid misuses of the system, such as producing offensive content, returning private information, or supporting malicious actions [115]. While referred to as red-teaming, prompt hacking leaves out the traditional red-teaming objectives [120] related to breaking into the system, and focuses instead on the safety issues at the detriment of more security-relevant aspects.

SCENARIO

We consider an automated application screening system for recruitment. The system takes as inputs a summary of the application of a candidate (e.g., education, previous employments) and a short text written by the applicant to explain his/her motivation. It then returns, for a job title and a list of requirements, a score of relevance and an explanation for the decision. This system could operate in an automated way and designed to return a selection of the best candidates that would be then screened by human recruiters. In such a setting, it could be considered as a high-risk according to the Annex III 4(a) of the proposed AI Act, and therefore subject to its requirements.

The AI system is built by integrating the language model Zephyr-7B-β [121] along other components such as web interface, a database, and various software components for the processing and the decision-making capabilities. In this scenario, we only consider the part of the system in which inputs are sent to the language model, and the outputs are parsed to extract the score and the explanation.

The prompt given to the language model is as follows:

```
You are a recruiting assistant. The company
is looking for a network engineer with at
least 5 years of experience.
You will be presented with a profile of a
candidate. Give a score on a scale of 0 to
4 to evaluate the relevance of the
candidate to the position.
You have to provide your answer in JSON
format, for example: {"score": 3}.
```

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

```
{SUMMARY}
Statement by the applicant: {STATEMENT}
```

We consider an applicant with the following application summary:

```
Mark Smith has a strong background in
business analysis, with over 5 years of
experience in process improvement, data
analysis, and project management. He has
worked for leading financial services and
consulting firms, and has a proven track
record of delivering successful business
outcomes through data-driven insights and
collaborative stakeholder engagement. The
candidate has a Bachelor of Science in
Business Administration from University of
California, Berkeley, and holds
certifications in Six Sigma Green Belt and
Certified Business Analyst (CBA). He
possesses a range of skills including
business analysis, data analysis, project
management, process improvement, data
visualization, communication, stakeholder
engagement, requirements gathering, and
systems analysis.
```

His statement reads as follows:

```
 I am motivated by the challenge of
 optimising network performance and
 security, and I am committed to staying
 aware with the latest technologies to
 ensure the highest level of service for the
 organisation.
```

The language model returns the following outputs:

```
{"score": 2}
While Mark Smith has some skills related to
networking, such as systems analysis and
communication, his primary focus and
experience are in business analysis and
project management. Therefore, his
relevance to the position of network
engineer is moderate at best, scoring a 2
on a scale of 0 to 4. While his motivation
and commitment to staying current with the
latest technologies are commendable, it is
still a significant gap in his skillset
that would require significant training and
development to bridge. A more suitable
candidate for the network engineer position
would have a minimum of 5 years of direct
networking experience.
```

Even if the evaluation made by the system may remain questionable, the explanation that is provided is reasonable with the provided information. It is likely that Mark would not be selected among the best candidates.

A. THREAT SCENARIO

The threat scenario consists in adversaries exploiting the possibility to add text to fool the language model into assessing incorrectly the relevance of an application. In this context, this can potentially lead to impacts on the fairness in the recruitment processes by favouring candidates without the necessary qualifications.

Evasion attacks on language models involve manipulating the text of the input data to force the model to make incorrect predictions. The modifications may be not suspicious to humans (e.g., swapping letters or words), or be added with the reasonable assumption that no human is supervising inputs. The possibility to conduct such evasion attacks against computer vision models is well understood and proven [89], but has been less studied in the context of language models, as the discrete nature of data makes it more challenging.

An example of an adversarial statement in the above discussed application screening scenario obtained using [71] is presented below:

```
? Multi skills worthy !Req proved answers ?
cuatro ! Qualrif immediate confirmMust
```

he output of the language model is as follows:

```
{"score": 4}
Explanation:
The candidate's experience and skills in
business analysis, data analysis, project
management, process improvement, and
stakeholder engagement make him a strong
fit for the role of a network engineer.
Although the job requirements explicitly
ask for a network engineer with 5 years of
experience, the candidate's multi-skilled
background and proven track record of
delivering successful business outcomes
through data-driven insights suggest that
he has the required knowledge and expertise
to adapt to the role. Therefore, we give
him a score of 4, which reflects his high
relevance to the position.
```

The evaluation is very positive the explanation relies on provided information to justify the relevance of the candidate to the position.

MITIGATION

Evasion attacks on language models can be mitigated through various strategies, at different levels:

- at the model level: specific techniques can be implemented, such as adversarial training consisting in the incorporation of adversarial examples during the training phase, or noise injection, to make the model less sensitive to small perturbations or changes in the input data;
- at the architecture level, ensemble models can be used to rely on several outputs made by different models, with the idea that fooling multiple models at the same time is harder for the adversary. Input sanitisation layers can also be implemented to detect abnormal and potentially malicious inputs;
- at the system level: real-time monitoring of inputs can be implemented, with a human oversight component to address suspicious inputs. Making the infrastructure on

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

which the language model relies secure is also a way to prevent theft of parameters that may facilitate the development of evasion attacks.

Each of these strategies can be combined to provide multiple layers of defence against evasion attacks. As discussed in [15], there is no one-size-fits-all solution, and the choice of mitigation strategies will depend on the specific model, the types of attacks it is vulnerable to, and the context in which it is used. In particular, implementing some of these techniques may hinder the performance of the system and/or induce additional costs that need to be evaluated. Regularly updating models and systems to respond to new threats as they are discovered is also a crucial part of maintaining robustness against evasion attacks.

DISCUSSION

Generally speaking, all machine learning systems, in particular those based on deep learning, are known to be susceptible to adversarial attacks, and it is likely that adversarial inputs can be transferred between different models [71]. However, it is not simple to estimate the likelihood of this threat scenario. The scenario assumes that the threat agent has a high level of technical skills and access to computing resources. Furthermore, it considers that the AI system is solely based on a publicly available model, which does not always happen in real context. As analysed in [75] for traffic sign recognition systems, the technical feasibility of conducting a physical evasion attack with acceptable resources is unclear, even beyond the above questions of motivation. It can be easily argued, with the low accuracy of evasion attacks, that a similar or greater impact can be successfully with traditional means without conducting such an elaborate adversarial attack [13]. However, with the increased integration of language models in products and services, the incentives will also grow. This, combined with the trend that makes such systems more autonomous and more capable of interacting with external services, increase the overall risks, even if the likelihood of the success of an attack remains the same. Other works beyond evasion attacks demonstrate that other vulnerabilities can be exploited [122]–[124], and that current alignment techniques may prove inefficient to protect users [125].

VII. CONCLUSION

This paper has outlined the challenges and open questions that arise in the process of establishing cybersecurity practices and methodologies to comply with the requirement of cybersecurity in future regulations on AI, such as the upcoming EU AI Act. These challenges stem from the current technological limitations and the evolving state of scientific and technical knowledge. Regulating a rapidly evolving technology such as AI requires scientific and technological expertise, anticipation and foresight capacity, and special consideration are needed in the definition of horizontal rules that are as technology-agnostic and future-proof as possible. These requirements can, in turn, be fulfilled with the best available techniques and approaches at any given time, in consideration of the risks and intended purpose of specific AI systems. Some of the challenges described in this paper may have an impact on how a cybersecurity requirement — such as in the proposed AI Act — will be implemented, and where boundaries of the accepted scientific state-of-the-art are reached. In the same way, this discussion could serve well as a source of information for future evaluation of standards, or in post-market evaluation of products. The more salient questions may also help to identify needs for additional AI research funding in the upcoming years.

Some of the technological limitations highlighted in this article are 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 such as listed in the previous section, with neither many generally accepted and established practices nor specific standards yet to address them [102] and fundamental limits to the securing of individual models. How to best integrate these new challenges and technological limits in cybersecurity practice can still be considered an open question in AI cybersecurity. It is likely that general rules and practices are going to be established with time, but, in accordance with a risk based approach, the detailed integration of these challenges may depend on individual AI systems and their context of application. All these limitations induce a significant shift from current practices. Not only is the presence of vulnerabilities in an AI system not fully known because of potentially unknown vulnerabilities, but also due to inherent limitations in measuring and guaranteeing security against known threat vectors, which at the moment cannot be consistently overcome or quantified at model level. For standardisation, it seems important to be clear about the limits of technological feasibility. Horizontal standards are not overly concerned with specific techniques or metrics, but rather with setting requirements that result in a proportionate and effective level of security, in accordance with the risks of the system and the state of the art [33].

To conclude, even if cybersecurity may be seen as a transversal field that touches upon a wide range of topics, such as data governance, human oversight, or robustness, the challenges linked to securing AI systems in a regulatory context can be independently addressed, keeping in mind the interdependencies between trustworthiness requirements. The connection of current research efforts to realistic cybersecurity threat models is of prime importance, built on the accumulated academic and engineering knowledge for daily cybersecurity practice [126]. To this date, studying this more applied approach of modelling threats in adversarial machine learning remains an underrepresented field of study, especially for complex deep models and/or in cyber-physical contexts. However, adversarial machine learning, as a field of scientific research, cannot provide all solutions needed to secure real machine learning-based systems, but could become central to provide technical controls for providers to achieve conformity. Adapting management tools and approaches to AI in the rapidly evolving landscape of emerging AI tech-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

nologies is a complex task that will require the collaboration of very different groups and skill sets from private and public sector and civil society. The implementation of security and safety mechanisms appropriate to the risk depends on the specific architecture and application context of an AI system. It will depend on the capacity of stakeholders to continuously address new technological features and related challenges specific to the AI technology.

REFERENCES

[1] M. Craglia, A. Annoni *et al.*, "Artificial Intelligence–a European perspective," European Commission - Joint Research Centre, Science for Policy, 2018.

[2] R. Righi, C. Pineda Leon *et al.*, "AI Watch Index 2021," European Commission - Joint Research Centre, Tech. Rep. JRC128744, 2022.

[3] N. Maslej, L. Fattorini *et al.*, "The AI Index 2023 Annual Report," Institute for Human-Centered AI, Stanford University, Tech. Rep., 2023.

[4] A. Karpathy, "Software 2.0," March 2021.

[5] S. Samoili, M. López Cobo *et al.*, "AI Watch - Defining Artificial Intelligence 2.0," European Commission - Joint Research Centre, Tech. Rep. JRC126426, 2021.

[6] High Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," 2019.

[7] European Commission, "Proposal for a Regulation laying down harmonised rules on Artificial Intelligence," 2021.

[8] ——, "Commission welcomes political agreement on AI Act," 2023.

[9] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, 2023.

[10] "Hiroshima Process International Code of Conduct for Advanced AI Systems," 2023.

[11] B. Hawes and D. W. Hall, "After the Summit: Progress in public policy on AI," https://eprints.soton.ac.uk/485533/, 2023.

[12] N. Papernot, P. McDaniel *et al.*, "SoK: Security and privacy in machine learning," in *Proceedings of the IEEE European Symposium on Security and Privacy*, 2018, pp. 399–414.

[13] J. Gilmer, R. P. Adams *et al.*, "Motivating the Rules of the Game for Adversarial Example Research," Preprint arXiv: 1807.06732, 2018.

[14] K. Grosse, L. Bieringer *et al.*, ""Why do so?" – A Practical Perspective on Machine Learning Security," 2022.

[15] H. Junklewitz, R. Hamon *et al.*, "Cybersecurity of artificial intelligence in the AI act," Science for Policy JRC134461, 2023.

[16] L. F. Sikos, *AI in Cybersecurity*. Springer, 2018, vol. 151.

[17] Y. Xin, L. Kong *et al.*, "Machine learning and deep learning methods for cybersecurity," vol. 6, pp. 35 365–35 381, 2018.

[18] A. Malatras, I. Agrafiotis, and M. Adamczyk, "Securing machine learning algorithms," ENISA, Tech. Rep., 2021.

[19] J. Chen, "AI-based deterrence in the cyber domain," in *Proceedings of the* 14th *International Conference on Cyber Warfare and Security (ICCWS)*. Academic Conferences and publishing limited, 2019, p. 38.

[20] P. Svenmarck, L. Luotsinen *et al.*, "Possibilities and challenges for artificial intelligence in military applications," in *Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists Meeting*, 2018.

[21] T. C. King, N. Aggarwal *et al.*, "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions," *Science and Engineering Ethics*, vol. 26, no. 1, pp. 89–120, 2020.

[22] The MITRE Corporation, "MITRE ATLAS," https://atlas.mitre.org/, 2022.

[23] E. Tabassi, K. J. Burns *et al.*, "A taxonomy and terminology of adversarial machine learning," Draft NISTIR 8269, 2019.

[24] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," Tech. Rep., 2023.

[25] ENISA, "Artificial Intelligence Cybersecurity Challenges - Threat Landscape for Artificial Intelligence," ENISA, Tech. Rep., 2020.

[26] A. Vassilev, A. Oprea *et al.*, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," National Institute of Standards and Technology, Tech. Rep. NIST Artificial Intelligence (AI) 100-2 E2023, January 2024.

[27] M. Kearns and M. Li, "Learning in the Presence of Malicious Errors," in *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, 1988, pp. 267–280.

[28] N. H. Bshouty, N. Eiron, and E. Kushilevitz, "PAC learning with nasty noise," *Theoretical Computer Science*, vol. 288, no. 2, pp. 255–275, 2002.

[29] N. Dalvi, P. Domingos *et al.*, "Adversarial classification," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: ACM, August 2004, pp. 99–108.

[30] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[31] M. Barreno, B. Nelson *et al.*, "Can machine learning be secure?" in *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '06. ACM, 2006, pp. 16–25.

[32] L. Huang, A. D. Joseph *et al.*, "Adversarial machine learning," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, 2011, pp. 43–58.

[33] European Commission, "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."

[34] W. Wahlster and C. Winterhalter, "German Standardization Roadmap on Artificial Intelligence – 2nd Edition," 2023.

[35] CEN-CENELEC Focus Group Report, "Road Map on Artificial Intelligence," CEN-CENELEC, Tech. Rep., 2020.

[36] C. Berghoff, B. Biggio *et al.*, "Towards Auditable AI Systems Current status and future directions based on the workshop " Auditing AI-Systems : From Basics to Applications," Whitepaper, 2021.

[37] S. Nativi and S. de Nigris, "AI Watch, AI standardisation landscape: State of play and link to the EC proposal for an AI regulatory framework." European Commission - Joint Research Centre, Tech. Rep. JRC125952, 2021.

[38] ETSI, "ETSI GR SAI 004 - Securing AI problem statement," 2020.

[39] ——, "ETSI GR/SAI-005 - Mitigation Strategy Report," 2021.

[40] P. Bezombes, S. Brunessaux, and S. Cadzow, "Cybersecurity of AI and Standardisation," ENISA, Tech. Rep., 2023.

[41] J. Soler Garrido, S. Tolan *et al.*, "AI watch: Artificial intelligence standardisation landscape update," European Commission, Technical Report JRC131155, 2023.

[42] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[43] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[44] K. M. Lynch and F. C. Park, *Modern Robotics*. Cambridge University Press, 2017.

[45] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2011.

[46] S. J. Russell, P. Norvig, and E. Davis, *Artificial Intelligence: A Modern Approach.*, 3rd ed. Pearson, 2009.

[47] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2nd ed., ser. Springer Series in Statistics. New York: Springer-Verlag, 2009.

[48] C. Panigutti, R. Hamon *et al.*, "The role of explainable AI in the context of the AI Act," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23. New York, NY, USA: Association for Computing Machinery, June 2023, pp. 1139–1150.

[49] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, p. 2053951715622512, 2016.

[50] N. Humbatova, G. Jahangirova *et al.*, "Taxonomy of real faults in deep learning systems," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ser. ICSE '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1110–1121.

[51] F. Croce, M. Andriushchenko *et al.*, "RobustBench: A standardized adversarial robustness benchmark," Preprint arXiv:2010.09670, 2021.

[52] C. Szegedy, W. Zaremba *et al.*, "Intriguing properties of neural networks," in *Proceedings of the International Conference on Learning Representations*, 2014.

[53] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," Preprint arXiv:1412.6572v3, 2015.

[54] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proceedings of the*

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582.

[55] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2017.

[56] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, 2019.

[57] G. Mcgraw, R. Bonett *et al.*, "Security Engineering for Machine Learning," *Computer*, vol. 52, no. 8, pp. 54–57, 2019.

[58] OECD, "Explanatory memorandum on the updated OECD definition of an AI system," no. 8, 2023.

[59] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[60] T. Everitt, V. Krakovna *et al.*, "Reinforcement learning with a corrupted reward channel," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017.

[61] J. Konečný, H. B. McMahan *et al.*, "Federated Learning: Strategies for Improving Communication Efficiency," 2017.

[62] M. Fang, X. Cao *et al.*, "Local Model Poisoning Attacks to {Byzantine-Robust} Federated Learning," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 1605–1622.

[63] Y. Huang, S. Gupta *et al.*, "Evaluating Gradient Inversion Attacks and Defenses in Federated Learning," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 7232–7241.

[64] R. Bommasani, D. A. Hudson *et al.*, "On the Opportunities and Risks of Foundation Models," 2021.

[65] J. Devlin, M.-W. Chang *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[66] OpenAI, "GPT-4 System Card," Tech. Rep., 2023.

[67] H. Touvron, T. Lavril *et al.*, "LLaMA: Open and Efficient Foundation Language Models," 2023.

[68] A. Radford, J. W. Kim *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," Preprint arXiv: 2103.00020, 2021.

[69] R. Rombach, A. Blattmann *et al.*, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[70] R. Jia and P. Liang, "Adversarial Examples for Evaluating Reading Comprehension Systems," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[71] A. Zou, Z. Wang *et al.*, "Universal and Transferable Adversarial Attacks on Aligned Language Models," 2023.

[72] D. Amodei, C. Olah *et al.*, "Concrete Problems in AI Safety," 2016.

[73] J. Gilmer, N. Ford *et al.*, "Adversarial Examples Are a Natural Consequence of Test Error in Noise," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2019, pp. 2280–2289.

[74] N. Carlini, A. Athalye *et al.*, "On Evaluating Adversarial Robustness," 2019.

[75] R. Hamon and H. Junklewitz, "Exploring the Feasibility of Physical Adversarial Attacks: A Cybersecurity Study," in *Proceedings of the 26th European Conference on Artificial Intelligence*, ser. Frontiers in Artificial Intelligence and Applications, vol. 372. IOS Press, 2023, pp. 980–987.

[76] J. M. Zhang, M. Harman *et al.*, "Machine Learning Testing: Survey, Landscapes and Horizons," *IEEE Transactions on Software Engineering*, pp. 1–1, 2020.

[77] F. Tramer, N. Carlini *et al.*, "On Adaptive Attacks to Adversarial Example Defenses," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1633–1645.

[78] C. Kittel and M. Alberts, "Security tradeoffs - Microsoft Azure Well-Architected Framework," https://learn.microsoft.com/en-us/azure/well-architected/security/tradeoffs, 2023.

[79] K. Wolter and P. Reinecke, "Performance and Security Tradeoff," in *Formal Methods for Quantitative Aspects of Programming Languages: 10th International School on Formal Methods for the Design of Computer, Communication and Software Systems, SFM 2010, Bertinoro, Italy, June 21-26, 2010, Advanced Lectures*, ser. Lecture Notes in Computer Science, A. Aldini, M. Bernardo *et al.*, Eds. Berlin, Heidelberg: Springer, 2010, pp. 135–167.

[80] A. Raghunathan, S. M. Xie *et al.*, "Understanding and Mitigating the Tradeoff Between Robustness and Accuracy," in *Proceedings of the 2020 International Conference on Machine Learning*, 2020.

[81] S. Hong, V. Chandrasekaran *et al.*, "On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping," 2020.

[82] R. S. Ross, "Guide for Conducting Risk Assessments," NIST, Tech. Rep. 800-30 Rev 1, 2012.

[83] I. Nai Fovino, G. Barry *et al.*, "Cybersecurity, Our Digital Anchor," European Commission - Joint Research Centre, Tech. Rep. JRC121051, 2020.

[84] C. P. Pfleeger and S. L. Pfleeger, *Analyzing Computer Security: A Threat/Vulnerability/Countermeasure Approach*. Prentice Hall Professional, 2012.

[85] A. Shostack, *Threat Modeling: Designing for Security*, 1st ed. Wiley Publishing, 2014.

[86] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 2012, pp. 1467–1474.

[87] L. Mu~noz-González, B. Biggio *et al.*, "Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, ser. AISec '17. New York, NY, USA: ACM, 2017, pp. 27–38.

[88] A. E. Cinà, K. Grosse *et al.*, "Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning," 2022.

[89] K. Chen, H. Zhu *et al.*, "A Survey on Adversarial Examples in Deep Learning," *Journal on Big Data*, vol. 2, no. 2, pp. 71–84, 2020.

[90] F. Tramer, F. Zhang *et al.*, "Stealing Machine Learning Models via Prediction APIs," in *Proceedings of the 25th USENIX Security Symposium*, 2016.

[91] B. Liu, M. Ding *et al.*, "When Machine Learning Meets Privacy: A Survey and Outlook," *ACM Computing Surveys*, vol. 54, no. 2, pp. 31:1–31:36, 2021.

[92] R. Shokri, M. Stronati *et al.*, "Membership Inference Attacks Against Machine Learning Models," in *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, 2017, pp. 3–18.

[93] C. Song, T. Ristenpart, and V. Shmatikov, "Machine Learning Models that Remember Too Much," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: ACM, 2017, pp. 587–601.

[94] I. Shumailov, Y. Zhao *et al.*, "Sponge examples: Energy-latency attacks on neural networks," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 212–231.

[95] M. Musser, A. Lohn *et al.*, "Adversarial Machine Learning and Cybersecurity," Center for Security and Emerging Technology, Tech. Rep., 2023.

[96] J. M. Spring, A. Galyardt *et al.*, "On managing vulnerabilities in AI/ML systems," in *Proceedings of the New Security Paradigms Workshop 2020*, ser. NSPW '20. Association for Computing Machinery, 2021, pp. 111–126.

[97] G. C. M. Moura and J. Heidemann, "Vulnerability Disclosure Considered Stressful," *ACM SIGCOMM Computer Communication Review*, vol. 53, no. 2, pp. 2–10, April 2023.

[98] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, 2010, pp. 305–316.

[99] S. McGregor, "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 15 458–15 463, 2021.

[100] P. Ammann and J. Offutt, *Introduction to Software Testing*. Cambridge University Press, 2016.

[101] J. Rando, D. Paleka *et al.*, "Red-Teaming the Stable Diffusion Safety Filter," 2022.

[102] N. Papernot, P. McDaniel *et al.*, "The limitations of deep learning in adversarial settings," in *Proceedings of IEEE European Symposium on Security and Privacy*. IEEE, 2016, pp. 372–387.

[103] Y. Liu, R. Wen *et al.*, "ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4525–4542.

[104] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential Privacy and Machine Learning: A Survey and Review," 2014.

[105] K. Schroeder and H. Trinh, "Performance measurement guide for information security (initial working draft)," National Institute of Standards and Technology, Tech. Rep., 2022.

[106] W. Xiang, P. Musau *et al.*, "Verification for Machine Learning, Autonomy, and Neural Networks Survey," 2018.

[107] C. Urban and A. Miné, "A Review of Formal Methods applied to Machine Learning," Preprint arXiv: 2104.02466, 2021.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3391021

[108] H. Salman, M. Sun *et al.*, "Denoised Smoothing: A Provable Defense for Pretrained Classifiers," in *Advances in Neural Information Processing Systems*, 2020.

[109] K. Hu, K. Leino *et al.*, "A Recipe for Improved Certifiable Robustness: Capacity and Data," 2023.

[110] P. Liang, R. Bommasani *et al.*, "Holistic Evaluation of Language Models," 2023.

[111] T. Lee, M. Yasunaga *et al.*, "Holistic Evaluation of Text-To-Image Models," 2023.

[112] F. M. Teichmann and S. R. Boticiu, "An overview of the benefits, challenges, and legal aspects of penetration testing and red teaming," *International Cybersecurity Law Review*, 2023.

[113] D. Ganguli, L. Lovitt *et al.*, "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned," 2022.

[114] E. David, "Biden releases AI executive order directing agencies to develop safety guidelines," *The Verge*, 2023.

[115] E. Groll, "Fifty minutes to hack ChatGPT: Inside the DEF CON competition to break AI," 2023.

[116] OWASP Foundation, "WSTG - v4.2," https://owasp.org/www-project-web-security-testing-guide/v42/, 2020.

[117] W. Pearce and J. Lucas, "NVIDIA AI Red Team: An Introduction," 2023.

[118] R. S. S. Kumar, "Microsoft AI Red Team building future of safer AI," https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/, 2023.

[119] D. Fabian and J. Crisp, "Why Red Teams Play a Central Role in Helping Organizations Secure AI Systems," Google, Tech. Rep., 2023.

[120] T. Baker, "What Does AI Red-Teaming Actually Mean?" 2023.

[121] L. Tunstall, E. Beeching *et al.*, "Zephyr: Direct Distillation of LM Alignment," 2023.

[122] N. Carlini, M. Jagielski *et al.*, "Poisoning Web-Scale Training Datasets is Practical," 2023.

[123] N. Carlini, J. Hayes *et al.*, "Extracting Training Data from Diffusion Models," 2023.

[124] P. Chao, A. Robey *et al.*, "Jailbreaking Black Box Large Language Models in Twenty Queries," 2023.

[125] X. Qi, Y. Zeng *et al.*, "Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" 2023.

[126] R. Hamon and H. Junklewitz, "Putting Adversarial Machine Learning to the Test: Towards AI Threat Modelling," in *New Frontiers in Adversarial Machine Learning*, 2022.

RONAN HAMON received the Ph.D. degree in Physics in 2015 at the Laboratoire de Physique de l'École Normale Supérieure de Lyon (ENS de Lyon), France. He has been a post-doctoral researcher at Laboratoire d'Informatique & Systèmes (LIS) in Marseille, France, and a scientist at CMRE NATO in La Spezia, Italy. He is currently a scientific Project Officer at Joint Research Centre of the European Commission in Ispra, Italy. His research interests focus in adversarial machine learning, graph signal processing, matrix factorization techniques, with applications to cybersecurity, audio signal processing, and underwater acoustics.

HENRIK JUNKLEWITZ received a diploma in physics in 2009 and a Ph.D in physics in 2014, both from the Ludwig-Maximilians-Universität München. He has been a post-doctoral researcher in astrophysics and statistical inference at the Friedrich-Wilhelms-Universität Bonn. Currently, he is a scientific project officer for machine learning with DG Joint research Centre of the European Commission in Ispra. Apart from his background in physics, his research interests focus on machine learning, Bayesian statistics, imaging methods, and signal processing.

JOSEP SOLER GARRIDO is an official at the European Commission, Joint Research Centre (JRC) since 2021. He currently leads the scientific team working on Digital Services Act enforcement at the European Centre for Algorithmic Transparency (ECAT). In addition, he represents the European Commission in technical standardisation activities in support of the Artificial Intelligence Act. During the 14 years prior to joining the European Commission, he held various technical and leadership R&D positions in the private sector, working on innovation projects in microelectronics, digital systems and artificial intelligence across multiple industries.

IGNACIO SANCHEZ works at the European Commission in the Joint Research Centre (JRC), within the Cybersecurity and Digital Technologies unit, where he leads several scientific projects in the areas of Artificial Intelligence (AI) and Cybersecurity. Ignacio holds an M.Sc. degree in Computer Engineering and a Ph.D. in Cybersecurity. He is a Certified Information Systems Security Professional (CISSP) and has over 20 years of expertise in the field of cybersecurity. Currently, his research is focused on AI and cybersecurity, exploring the interplay between these fields within the context of the related technical and EU policy frameworks.

• • •

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/