IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 22 January 2024, accepted 27 March 2024, date of publication 22 April 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3391021

RESEARCH ARTICLE

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, 21027 Ispra, Italy
[2]European Commission, Joint Research Centre, 41092 Seville, Spain

Corresponding author: Ignacio Sanchez (ignacio.sanchez@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], [2], [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 corpora 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 respect to fundamental rights and societal values. At the core of the proposed European Union's AI Act [7], for which a political agreement has been reached at the end of 2023 [8], lies the idea that it is necessary to take into account the purpose and the context of use of AI systems in order to fulfil the essential trustworthiness requirements in an effective and proportionate manner and establish harmonised rules for AI products. The AI Act exemplifies growing governmental efforts around the world to establish regulatory guardrails for AI [9], [10], [11], [12].

This paper focuses on the cybersecurity requirement as laid out in Article 15(4) of the AI Act, which also connects to other international policy initiatives that emphasise AI safety and security [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

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Yuan Chen.

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 [13], [14], [15].

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 AI regulation: 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 [16], 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 empowers cybersecurity by enabling a range of tasks such as the prediction and detection of threats in real-time, the anticipation of future attacks, or the automation of mitigation measures and incident response processes [17], [18].
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 [19].
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 adversaries [20], [21].
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 [22].

The focus of cybersecurity requirements of current AI regulation is on the security of AI systems (point 2). In the rest of this section, we describe this aspect in greater detail, 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 safeguard AI systems and their data from cyberthreats. 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 [23], taxonomies [24], [25] or AI threat landscape analyses [19], [26], [27].

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 [14]. The field built upon early theoretical work on learning against adversaries [28], [29] and went closer to cybersecurity applications such as spam filters [30], [31], before evolving more concretely into security principles for machine learning systems [5], [32], [33]. 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 [27].

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 regulatory rules on AI, a question that has now become central with the expected adoption of the AI Act [8]. Harmonised standards [34] 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 [35].

From a conformity assessment perspective, a clear distinction should be made between an AI model and an AI system [16]. 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 interact with users and the virtual or physical environment. Thus, ensuring the conformity of AI systems with future regulatory rules on cybersecurity does not necessarily require making single models secure. Even if limitations due to technical gaps and lack of scientific maturity of AI technology may limit compliance [35], [36], [37], alternative approaches at system level may exist.

Many activities have been initiated to standardise various technical, ethical and organisational aspects of AI systems on topics ranging from risk modelling to AI hardware security [25], [35], [36], [38], [39], [40]. The analysis of the efforts required to standardise AI cybersecurity [16], [37], [41], [42] 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 such as procedures on organisational principles, risk management and security controls, can largely be taken from the ISO/IEC 27000 series, current standards are, however, not yet adapted to be used for AI software. At the European level, this adaptation is just beginning and plans to 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 [43], natural language processing [44], robotics [45], or computer vision [46]. AI-based software typically exhibits a range of features that sets them apart from traditional software: *reasoning and learning* [47], 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* [48], 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* [49], describing the absence of explicit mechanisms and rules that limits the understanding of the functioning of the system; *unpredictability* [50], 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

Legislation needs 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 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 [34]. 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 [36] and led to the development of terminologies [19], [24], [26] 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 [51], 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) [52]. 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 [30], [31], [32], [33] and were extended to deep neural networks [53], [54], [55], [56], [57] later. They are defined in [24] as *"[...] manipulat[ing] input samples to evade (cause a misclassification) a trained classifier at test time."*, in [23] as *"[...] craft[ing] adversarial data that prevent a machine learning model from correctly identifying the contents of the data."*, and in [58] 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 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 [59], 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* [60], 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 [61]; *federated learning* [62], a distributed approach to train models on devices, and that can be compromised by exploiting the decentralised nature of the training [63], [64]; *foundation models and general purpose AI models* [65], self-supervised models trained on massive amounts of data, and that have been particularly notable for text [66], [67], [68] and text-to-image generation [69], [70], but also pose for additional security risks [71], [72] in terms of supply chain.

C. SECURITY, SAFETY, AND GENERALISABILITY

Safety and generalisability are two concepts that overlap 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, whether induced intentionally or not. As for generalisability, it is more about the capacity of machine learning systems to operate reliably on unknown data, going beyond the robustness against attacks. 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 [73] 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. It also touches upon more controversial discussions about artificial general intelligence (AGI) and the adequate measures to ensure that advanced AI systems will remain aligned with human values, a topic which is usually out of scope of AI regulation. This makes AI safety a component of cybersecurity, promoting a resilience against malicious attacks, but without encompassing all aspects 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 [53], [74] that are not necessarily relevant for cybersecurity problems [14], [31], [75]. 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 [76]. 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 [75], [77], or properly assessing the strength of defences [78].

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], [49], [79], 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 in 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 [80], [81], 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 [49], [79]. 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 [82], [83], and comes with additional development costs. In other 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 that could operate offline 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 fundamental 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 [24], [25], [84].

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 [25], [84], [86], [87] from an analysis of the following factors (see Fig. 1):

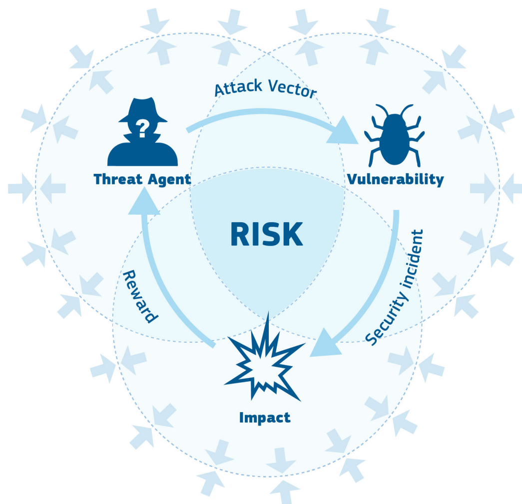


FIGURE 1. Conceptual model depicting the logical links between the different components of the cybersecurity risk (adapted from [85]).

- *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 [14].

The level of knowledge of an agent about AI systems includes different elements [13]: 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 [13], [15], [31], [37]. This remains particularly complex to evaluate for actual cybersecurity risks since any long-term and widespread deployment of AI products is still only in 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 [19], [26], [27]. 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 [19].

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 [27], including: *data poisoning* [88], [89], manipulating or injecting false data with the intention to compromise the training of machine learning models; *backdoors* [90], installing AI-specific mechanisms triggering, for specific patterns, unwanted behaviours; *evasion (or adversarial) attacks* [91], crafting inputs to alter the outcome of an AI system; *model extraction and inversion* [92], [93], retrieving the parameters of a model; *membership inference* [94], [95], retrieving data or specific features used in training and testing sets; *latency attacks* [96], 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 in AI systems [97]. 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) [98], and their disclosure [99], 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 [90] 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 [100] and a better anticipation.

C. IMPACT

The evaluation of the impact of an attack is traditionally conducted using the CIA model [84], [86], which has already been translated to analysing AI systems [13], [24], [26]:

- *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 [101] 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.

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 [35], [36], [37], [102] 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 [26]: *data* (e.g., data sets for training, testing and

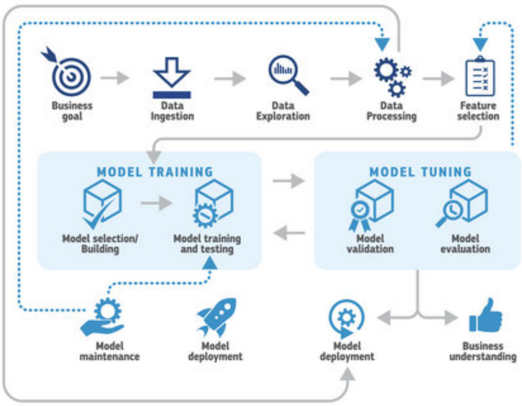


FIGURE 2. Depiction of the typical lifecycle of an AI system based on machine learning techniques [26]. The lifecycle illustrates the multiple stages involved in the development of the system, and includes feedback loops and potentially automated decisions.

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, 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 models (such as General Purpose AI or foundation models [65]); 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 [103], 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 management 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 [19], [23], [24], [26], [39]. 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 ones listed in ISO/IEC 27002, will remain important. Others, more specific to AI technology, such as data validation, input sanitation, model watermarking or crucially incorporating hardened models and defences, may need additional work and an adaptation of practices. These problems are exacerbated when considering whole AI systems made of several AI subsystems. Any set of security controls for relevant AI assets will always depend on the particular AI system at hand.

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 took place in the context of evasion attacks, with the establishment of controls for applications such as network intrusion detection systems or spam filtering [31], [32], [33]. Subsequent works expanded on these [13], [104], 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 [53], [54], by including in the training set adversarial examples, at increased computational cost and potentially decreased performance on benign data.

As detailed in [16], security controls for AI vulnerabilities may not need to be themselves based on AI, or directly fixing flaws in AI models. For example, protecting models against membership inference [105] may rely on a combination of model-based controls of differential privacy [106], 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 [107], 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 [102] and close to a certain degree to the one used in machine learning [77] for determining the performance of a trained model.

A measure of interest in the context of AI regulation is the robustness of AI systems, in particular against cyberattacks exploiting AI-specific vulnerabilities [75], [76], [77], [78]. Generally speaking, two approaches co-exist: *formal methods* [108], [109] and *statistical and empirical approaches* [52], [75]. Methods for formal verification 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 [110], [111], many of the known methods are scalable only with difficulty to complex deep neural networks [37]. 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 [112], [113]. 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 [37]. 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 [114], actively finding vulnerabilities in systems and exploit them to determine their associated risks. Its application to AI systems as part of red-teaming [115] requirements has emerged recently with the latest developments in generative AI [116]. Putting under stress controls aims at preventing unexpected behaviours, such as hallucinations, jailbreaks or bias, and addressing the risks that come from automating these new technologies at scale [67]. Its current implementation for AI systems (e.g., as presented 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 [117]. Adapting classical cybersecurity practices [118] 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 [119], [120], [121] 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 self-supervised 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 [112], or by applying penetration testing principles and finding potential flaws in the model. For the latter, prompt hacking [115] 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 [117]. While referred to as red-teaming, prompt hacking leaves out the traditional red-teaming objectives [122] related to breaking into the system, and focuses instead on the safety issues at the detriment of more security-relevant aspects.

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 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 [123] 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}.

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.

1) 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. While the possibility to conduct such evasion attacks against computer vision models is well understood and proven [91], it 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 [72] is presented below:

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

The 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 and the explanation relies on provided information to justify the relevance of the candidate to the position.

2) 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 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 [16], 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, without guarantees of efficiency [124]. 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.

3) 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 [72]. 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 [76] 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 achieved with traditional means without conducting such an elaborate adversarial attack [14]. 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 [125], [126], [127], and that current alignment techniques may prove inefficient to protect users [128].

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 EU 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 development and 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 new computing and product lifecycle paradigms, introduced by machine learning systems. This is due to the fact that not only a growing number of new AI-specific vulnerabilities are being identified, without generally accepted and established practices to address them [104], but also theoretical limits to the securing of individual models exist. 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 measuring and guaranteeing security against known threat vectors cannot at the moment 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 [34].

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 and should 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 [129]. 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 technologies 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 et al., "Artificial intelligence: A European perspective," Eur. Commission Joint Res. Centre, Sci. Policy, 2018.
[2] R. Righi, C. P. Leon, M. Cardona, J. S. Garrido, M. Papazoglou, S. Samoili, and M. V.-P. Baillet, "AI watch index 2021," Publications Office Eur. Union, Luxembourg, U.K., Tech. Rep., JRC128744, 2022, doi: 10.2760/921564.
[3] N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J. C. Niebles, V. Parli, "The AI index 2023 annual report," Inst. Human-Centered AI, Stanford Univ., Stanford, CA, USA, 2023.
[4] A. Karpathy, "Software 2.0," *Medium*, 2021. Accessed: Nov. 1, 2023. [Online]. Available: https://karpathy.medium.com/software-2-0-a64152b37c35
[5] S. Samoili, M. L. Cobo, B. Delipetrev, F. Martínez-Plumed, E. Gómez, and G. De Prato, "AI watch—Defining artificial intelligence 2.0," Publications Office Eur. Union, Luxembourg, U.K., Tech. Rep. JRC126426, 2021, doi: 10.2760/019901.
[6] High Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, Eur. Commission, 2019.
[7] European Commission, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence*, 2021.
[8] European Commission, *Commission Welcomes Political Agreement on AI Act*, 2023.
[9] The White House. (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. [Online]. Available: 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/
[10] *Hiroshima Process International Code of Conduct for Advanced AI Systems*, 2023.
[11] B. Hawes and D. W. Hall. (2023). *After the Summit: Progress in Public Policy on AI*. [Online]. Available: https://eprints.soton.ac.uk/485533/
[12] OECD Legal Instruments, *Recommendation of the Council on Artificial Intelligence*, OECD, Paris, France, 2023.
[13] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and privacy in machine learning," in *Proc. IEEE Eur. Symp. Secur. Privacy*, 2018, pp. 399–414.
[14] J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen, and G. E. Dahl, "Motivating the rules of the game for adversarial example research," 2018, *arXiv:1807.06732*.
[15] K. Grosse, L. Bieringer, T. R. Besold, B. Biggio, and K. Krombholz, "'Why do so?'—A practical perspective on machine learning security," in *Proc. Int. Conf. Machin. Learn., New Frontiers Adversarial Machine Learn.*, 2022, pp. 1–14.
[16] H. Junklewitz, R. Hamon, A.-A. André, T. Evas, J. S. Garrido, and S. Martin, "Cybersecurity of artificial intelligence in the AI act," Publications Office Eur. Union, Luxembourg, U.K., Tech. Rep. JRC134461, 2023.
[17] L. F. Sikos, *AI in Cybersecurity*, vol. 151. Berlin, Germany: Springer, 2018.
[18] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018.
[19] A. Malatras, I. Agrafiotis, and M. Adamczyk, "Securing machine learning algorithms," ENISA, 2021, doi: 10.2824/874249.
[20] J. Chen, "AI-based deterrence in the cyber domain," in *Proc. 14th Int. Conf. Cyber Warfare Security (ICCWS)*, 2019, p. 38.
[21] P. Svenmarck, L. Luotsinen, M. Nilsson, and J. Schubert, "Possibilities and challenges for artificial intelligence in military applications," in *Proc. NATO Big Data Artif. Intell. Military Decision Making Spec. Meeting*, 2018, pp. 1–16.
[22] T. C. King, N. Aggarwal, M. Taddeo, and L. Floridi, "Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions," *Sci. Eng. Ethics*, vol. 26, no. 1, pp. 89–120, Feb. 2020.
[23] The MITRE Corporation. (2022). *MITRE ATLAS*. [Online]. Available: https://atlas.mitre.org/

[24] E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. Sexton, "A taxonomy and terminology of adversarial machine learning," NIST, Gaithersburg, MD, USA, Tech. Rep. Draft NISTIR 8269, 2019, doi: 10.6028/NIST.IR.8269-draft.
[25] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST, Gaithersburg, MD, USA, Tech. Rep. NIST AI 100-1, 2023.
[26] ENISA, *Artificial Intelligence Cybersecurity Challenges—Threat Landscape for Artificial Intelligence*, Brussels, Belgium, 2020, doi: 10.2824/238222.
[27] A. Vassilev and A. Oprea, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, NIST Artif. Intell. (AI) Tech. Rep., 100-2 E2023, Jan. 2024.
[28] M. Kearns and M. Li, "Learning in the presence of malicious errors," in *Proc. 20th Annu. ACM Symp. Theory Comput.*, 1988, pp. 267–280.
[29] N. H. Bshouty, N. Eiron, and E. Kushilevitz, "PAC learning with nasty noise," *Theor. Comput. Sci.*, vol. 288, no. 2, pp. 255–275, Oct. 2002.
[30] N. Dalvi, P. Domingos, S. Sanghai, and D. Verma, "Adversarial classification," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2004, pp. 99–108.
[31] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, Dec. 2018.
[32] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proc. ACM Symp. Inf., Comput. Commun. Secur.*, Mar. 2006, pp. 16–25.
[33] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. 4th ACM Workshop Secur. Artif. Intell.*, 2011, pp. 43–58.
[34] 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*, 2023.
[35] W. Wahlster and C. Winterhalter, *German Standardization Roadmap on Artificial Intelligence*, 2nd ed. Berlin, Germany: DIN, 2023.
[36] CEN-CENELEC Focus Group Report, *Road Map on Artificial Intelligence*, CEN-CENELEC, Brussels, Belgium, 2020.
[37] C. Berghoff et al., "Towards auditable AI systems current status and future directions based on the workshop, auditing AI-systems: From basics to applications," TUV, BSI, Fraunhofer, 2021.
[38] 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," Publications Office Eur. Union, Luxembourg, U.K., Tech. Rep. JRC125952, 2021, doi: 10.2760/376602.
[39] *ETSI GR SAI 004—Securing AI Problem Statement*, ETSI, Sophia Antipolis, France, 2020.
[40] *ETSI GR/SAI-005—Mitigation Strategy Report*, ETSI, Sophia Antipolis, France, 2021.
[41] P. Bezombes, S. Brunessaux, and S. Cadzow, "Cybersecurity of AI and standardisation," ENISA, Brussels, Belgium, 2023.
[42] J. Soler Garrido, S. Tolan, I. H. Torres, D. F. Llorca, V. Charisi, E. G. Gutierrez, H. Junklewitz, R. Hamon, D. F. Yela, and C. Panigutti, "AI watch: Artificial intelligence standardisation landscape update," Publications Office Eur. Union, Luxembourg, U.K., Tech. Rep. JRC131155, 2023, doi: KJ-NA-31-343-EN-N.
[43] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.
[44] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.
[45] K. M. Lynch and F. C. Park, *Modern Robotics*. Cambridge, U.K.: Cambridge Univ. Press, 2017.
[46] R. Szeliski, *Computer Vision: Algorithms and Applications*. Berlin, Germany: Springer, 2011.
[47] S. J. Russell, P. Norvig, and E. Davis, *Artificial Intelligence: A Modern Approach*, 3rd ed. London, U.K.: Pearson, 2009.
[48] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics), 2nd ed. New York, NY, USA: Springer-Verlag, 2009.
[49] C. Panigutti, R. Hamon, I. Hupont, D. F. Llorca, D. F. Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. S. Garrido, and E. Gomez, "The role of explainable AI in the context of the AI act," in *Proc. ACM Conf. Fairness, Accountability, Transparency*. New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 1139–1150.
[50] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data Soc.*, vol. 3, no. 1, 2016, Art. no. 2053951715622512.
[51] N. Humbatova, G. Jahangirova, G. Bavota, V. Riccio, A. Stocco, and P. Tonella, "Taxonomy of real faults in deep learning systems," in *Proc. ACM/IEEE 42nd Int. Conf. Softw. Eng.* New York, NY, USA: Association for Computing Machinery, 2020, pp. 1110–1121.
[52] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "RobustBench: A standardized adversarial robustness benchmark," in *Proc. Conf. Neural Inf. Process. Syst.-Datasets Benchmarks Track*, 2021, pp. 1–29.
[53] C. Szegedy, W. Zaremba I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–10.
[54] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015, *arXiv:1412.6572*.
[55] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.
[56] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.
[57] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.
[58] G. Mcgraw, R. Bonett, H. Figueroa, and V. Shepardson, "Security engineering for machine learning," *Computer*, vol. 52, no. 8, pp. 54–57, Aug. 2019.
[59] *Explanatory Memorandum on the Updated OECD Definition of an AI System*, OECD Artificial Intelligence Papers, Paris, France, 2023.
[60] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[61] T. Everitt, V. Krakovna, L. Orseau, and S. Legg, "Reinforcement learning with a corrupted reward channel," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1–24.
[62] J. Konecny, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2017, *arXiv:1610.05492*.
[63] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to Byzantine-Robust federated learning," in *Proc. 29th USENIX Secur. Symp.*, 2020, pp. 1605–1622.
[64] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 7232–7241.
[65] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, and E. Brynjolfsson, "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.
[66] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.
[67] OpenAI, *GPT-4 System Card*, 2023.
[68] H. Touvron, T. Lavril G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, B. Roziere, N. Goyal, E. Hambro, F. Azhar, and A. Rodriguez, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[69] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[70] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
[71] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–11.
[72] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," 2023, *arXiv:2307.15043*.
[73] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mane, "Concrete problems in AI safety," 2016, *arXiv:1606.06565*.

[74] J. Gilmer, N. Ford, N. Carlini, and E. D. Cubuk, ''Adversarial examples are a natural consequence of test error in noise,'' in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 2280–2289.

[75] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, ''On evaluating adversarial robustness,'' 2019, *arXiv:1902.06705*.

[76] R. Hamon and H. Junklewitz, ''Exploring the feasibility of physical adversarial attacks: A cybersecurity study,'' in *Proc. 26th Eur. Conf. Artif. Intell.*, vol. 372, 2023, pp. 980–987.

[77] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, ''Machine learning testing: Survey, landscapes and horizons,'' *IEEE Trans. Softw. Eng.*, vol. 48, no. 1, pp. 1–36, Jan. 2022.

[78] F. Tramèr, N. Carlini, W. Brendel, and A. Madry, ''On adaptive attacks to adversarial example defenses,'' in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 1633–1645.

[79] R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, and P. De Hert, ''Bridging the gap between AI and explainability in the GDPR: Towards trustworthiness-by-design in automated decision-making,'' *IEEE Comput. Intell. Mag.*, vol. 17, no. 1, pp. 72–85, Feb. 2022.

[80] C. Kittel and M. Alberts. (2023). *Security Tradeoffs—Microsoft Azure Well-Architected Framework*. [Online]. Available: https://learn.microsoft.com/en-us/azure/well-architected/security/tradeoffs

[81] K. Wolter and P. Reinecke, ''Performance and security tradeoff,'' in *Formal Methods for Quantitative Aspects of Programming Languages* (Lecture Notes in Computer Science), A. Aldini, and M. Bernardo, Eds., Bertinoro, Italy. Berlin, Germany: Springer, 2010, pp. 135–167.

[82] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang, ''Understanding and mitigating the tradeoff between robustness and accuracy,'' in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1–26.

[83] S. Hong, V. Chandrasekaran, Y. Kaya, T. Dumitras, and N. Papernot, ''On the effectiveness of mitigating data poisoning attacks with gradient shaping,'' 2020, *arXiv:2002.11497*.

[84] R. S. Ross, ''Guide for conducting risk assessments,'' NIST, Gaithersburg, MD, USA, Tech. Rep. 800-30 Rev 1, 2012, doi: NIST.SP.800-30r1.

[85] I. N. Fovino et al., ''Cybersecurity, our digital anchor,'' Luxembourg, U.K., Tech. Rep. JRC121051, 2020, doi: 10.2760/352218.

[86] C. P. Pfleeger and S. L. Pfleeger, *Analyzing Computer Security: A Threat/Vulnerability/Countermeasure Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, 2012.

[87] A. Shostack, *Threat Modeling: Designing for Security*, 1st ed. Hoboken, NJ, USA: Wiley, 2014.

[88] B. Biggio, B. Nelson, and P. Laskov, ''Poisoning attacks against support vector machines,'' in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 1467–1474.

[89] L. Munoz-Gonzalez, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, ''Towards poisoning of deep learning algorithms with back-gradient optimization,'' in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, New York, NY, USA, 2017, pp. 27–38.

[90] A. E. Cina, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, and B. Biggio, ''Wild patterns reloaded: A survey of machine learning security against training data poisoning,'' *ACM Comput. Surv.*, vol. 55, no. 13, pp. 1–39, 2022.

[91] K. Chen, H. Zhu, L. Yan, and J. Wang, ''A survey on adversarial examples in deep learning,'' *J. Big Data*, vol. 2, no. 2, pp. 71–84, 2020.

[92] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, ''Stealing machine learning models via prediction APIs,'' in *Proc. 25th USENIX Secur. Symp.*, 2016, pp. 601–618.

[93] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, ''When machine learning meets privacy: A survey and outlook,'' *ACM Comput. Surv.*, vol. 54, no. 2, p. 1–31, 2021.

[94] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, ''Membership inference attacks against machine learning models,'' in *Proc. IEEE Symp. Secur. Privacy*, May 2017, pp. 3–18.

[95] C. Song, T. Ristenpart, and V. Shmatikov, ''Machine learning models that remember too much,'' in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, New York, NY, USA, Oct. 2017, pp. 587–601.

[96] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson, ''Sponge examples: Energy-latency attacks on neural networks,'' in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Sep. 2021, pp. 212–231.

[97] M. Musser et al., ''Adversarial machine learning and cybersecurity,'' Center Secur. Emerg. Technol., Washington, DC, USA, 2023, doi: 10.51593/2022CA003.

[98] J. M. Spring, A. Galyardt, A. D. Householder, and N. VanHoudnos, ''On managing vulnerabilities in AI/ML systems,'' in *Proc. New Secur. Paradigms Workshop*, 2021, pp. 111–126.

[99] G. C. M. Moura and J. Heidemann, ''Vulnerability disclosure considered stressful,'' *ACM SIGCOMM Comput. Commun. Rev.*, vol. 53, no. 2, pp. 2–10, Apr. 2023.

[100] R. Sommer and V. Paxson, ''Outside the closed world: On using machine learning for network intrusion detection,'' in *Proc. IEEE Symp. Secur. Privacy*, May 2010, pp. 305–316.

[101] S. McGregor, ''Preventing repeated real world AI failures by cataloging incidents: The AI incident database,'' in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 17, pp. 15458–15463.

[102] P. Ammann and J. Offutt, *Introduction to Software Testing*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[103] J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramer, ''Red-teaming the stable diffusion safety filter,'' 2022, *arXiv:2210.04610*.

[104] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, ''The limitations of deep learning in adversarial settings,'' in *Proc. IEEE Eur. Symp. Secur. Privacy*, Mar. 2016, pp. 372–387.

[105] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. De Cristofaro, M. Fritz, and Y. Zhang, ''ML-doctor: Holistic risk assessment of inference attacks against machine learning models,'' in *Proc. 31st USENIX Security Symp.*, Aug. 2022, pp. 4525–4542.

[106] Z. Ji, Z. C. Lipton, and C. Elkan, ''Differential privacy and machine learning: A survey and review,'' 2014, *arXiv:1412.7584*.

[107] K. Schroeder and H. Trinh, ''Performance measurement guide for information security (initial working draft),'' Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. SP 800-55 Rev. 2, 2022.

[108] W. Xiang, P. Musau, A. A. Wild, D. M. Lopez, N. Hamilton, X. Yang, J. Rosenfeld, and T. T. Johnson, ''Verification for machine learning, autonomy, and neural networks survey,'' 2018, *arXiv:1810.01989*.

[109] C. Urban and A. Mine, ''A review of formal methods applied to machine learning,'' 2021, *arXiv:2104.02466*.

[110] H. Salman, M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter, ''Denoised smoothing: A provable defense for pretrained classifiers,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–13.

[111] K. Hu, K. Leino, Z. Wang, and M. Fredrikson, ''A recipe for improved certifiable robustness: Capacity and data,'' 2023, *arXiv:2310.02513*.

[112] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, and B. Newman, ''Holistic evaluation of language models,'' 2023, *arXiv:2211.09110*.

[113] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, and M. Kang, ''Holistic evaluation of text-to-image models,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 1–31.

[114] F. M. Teichmann and S. R. Boticiu, ''An overview of the benefits, challenges, and legal aspects of penetration testing and red teaming,'' *Int. Cybersecurity Law Rev.*, vol. 4, no. 4, pp. 387–397, 2023.

[115] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, and A. Jones, ''Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned,'' 2022, *arXiv:2209.07858*.

[116] E. David, ''Biden releases AI executive order directing agencies to develop safety guidelines,'' *The Verge*, 2023. Accessed: Nov. 1, 2023. [Online]. Available: https://www.theverge.com/2023/10/30/23914507/biden-ai-executive-order-regulation-standards

[117] E. Groll, ''Fifty minutes to hack ChatGPT: Inside the DEF CON competition to break AI,'' *CyberScoop*, 2023. Accessed: Nov. 1, 2023. [Online]. Available: https://cyberscoop.com/def-con-ai-hacking-red-team/

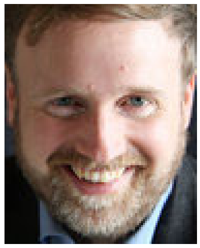
[118] OWASP Foundation. (2020). *WSTG—V4.2*. [Online]. Available: https://owasp.org/www-project-web-security-testing-guide/v42/

[119] W. Pearce and J. Lucas, ''NVIDIA AI red team: An introduction,'' *NVIDIA Tech. Blog*, 2023. Accessed: Nov. 1, 2023. [Online]. Available: https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction/

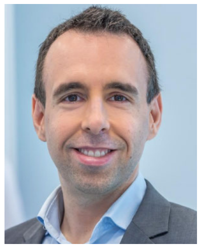
[120] R. S. S. Kumar. (2023). *Microsoft AI Red Team Building Future of Safer AI*. [Online]. Available: https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/

[121] D. Fabian and J. Crisp, "Why red teams play a central role in helping organizations secure AI systems," Google, 2023.
[122] T. Baker, "What does AI red-teaming actually mean?" *Center Secur. Emerg. Technol.*, 2023. Accessed: Nov. 1, 2023. [Online]. Available: https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/
[123] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf, "Zephyr: Direct distillation of LM alignment," 2023, *arXiv:2310.16944*.
[124] N. Jain, A. Schwarzschild, Y. Wen, G. Somepalli, J. Kirchenbauer, P.-Y. Chiang, M. Goldblum, A. Saha, J. Geiping, and T. Goldstein, "Baseline defenses for adversarial attacks against aligned language models," 2023, *arXiv:2309.00614*.
[125] N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramer, "Poisoning web-scale training datasets is practical," 2023, *arXiv:2302.10149*.
[126] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," in *Proc. 32nd USENIX Secur. Symp.*, 2023, pp. 5253–5270.
[127] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, "Jailbreaking black box large language models in twenty queries," 2023, *arXiv:2310.08419*.
[128] X. Qi, Y. Zeng, T. Xie, P. Y. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!" 2023, *arXiv:2310.03693*.
[129] R. Hamon and H. Junklewitz, "Putting adversarial machine learning to the test: Towards AI threat modelling," in *New Frontiers in Adversarial Machine Learning*. Baltimore, MD, USA: AdvML-ICML, 2022.

RONAN HAMON received the Ph.D. degree in physics from Laboratoire de Physique de l'École Normale Supérieure de Lyon (ENS de Lyon), France, in 2015. He has been a Postdoctoral Researcher with Laboratoire d'Informatique and Systémes (LIS), Marseille, France, and a Scientist with CMRE NATO, La Spezia, Italy. He is currently a Scientific Project Officer with the European Commission, Joint Research Centre in Ispra, Italy. His research interests include adversarial machine learning, graph signal processing, and matrix factorization techniques, with applications to cybersecurity, audio signal processing, and underwater acoustics.



HENRIK JUNKLEWITZ received the Diploma and Ph.D. degrees in physics from Ludwig-Maximilians-Universität München, in 2009 and 2014, respectively. He has been a Postdoctoral Researcher in astrophysics and statistical inference with Friedrich-Wilhelms-Universität Bonn. Currently, he is a Scientific Project Officer of machine learning with the European Commission, DG Joint Research Centre 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 has been an official with the European Commission, Joint Research Centre (JRC), since 2021. He currently leads the scientific team working on Digital Services Act enforcement with the European Centre for Algorithmic Transparency (ECAT). In addition, he represents the European Commission in technical standardization activities in support of the Artificial Intelligence Act. Prior to joining the European Commission, he held various technical and leadership research and development positions in the private sector and working on innovation projects in microelectronics, digital systems, and artificial intelligence across multiple industries.

IGNACIO SANCHEZ received the M.Sc. degree in computer engineering and the Ph.D. degree in cybersecurity. He is an European Commission official at the Joint Research Centre (JRC), serving in the Cybersecurity and Digital Technologies unit within the Digital Transformation and Data directorate. In his role, he leads several scientific projects focusing on Artificial Intelligence and Cybersecurity. Additionally, Ignacio is a co-leader of the Trustworthy AI portfolio, which coordinates the JRC's efforts to address the multidisciplinary scientific and policy challenges posed by the increasing adoption of Artificial Intelligence. He is also a Certified Information Systems Security Professional (CISSP) with over 20 years of expertise in the field of cybersecurity. This current research is focused on artificial intelligence and cybersecurity, exploring the interplay of these fields in the context of the related technical and EU policy frameworks.