Robustness and Cybersecurity in the EU Artificial Intelligence Act

Henrik Nolte*
University of Tübingen
Tübingen, Germany
henrik.nolte@uni-tuebingen.de

Miriam Rateike*
Saarland University
Saarbrücken, Germany
rateike@cs.uni-saarland.de

Michèle Finck
University of Tübingen
Tübingen, Germany

Abstract

The EU Artificial Intelligence Act (AIA) establishes legal principles for certain types of AI systems. While prior work has sought to clarify some of these principles, little attention has been paid to robustness and cybersecurity. This paper aims to fill this gap. We identify legal challenges and shortcomings in provisions related to robustness and cybersecurity for high-risk AI systems (Art. 15 AIA) and general-purpose AI models (Art. 55 AIA). We show that robustness and cybersecurity demand resilience against performance disruptions. Furthermore, we assess potential challenges in implementing these provisions in light of recent advancements in the machine learning (ML) literature. Our analysis informs efforts to develop harmonized standards as well as benchmarks and measurement methodologies under Art. 15(2) AIA, and seeks to bridge the gap between legal terminology and ML research to better align research and implementation efforts.

1 Introduction

The European Union (EU) recently adopted the Artificial Intelligence Act (AIA)[2] which creates a legal framework for the development, deployment, and use of "human-centered and trustworthy artificial intelligence (AI)" (Art. 1 AIA). The AIA outlines desirable "ethical principles" of AI systems (Rec. (27)) and, i.a., imposes some of these as legally binding requirements for high-risk AI systems (HRAIS), e.g, AI systems intended to be used to take university admission decisions, or to evaluate individuals' creditworthiness, and for general-purpose AI models (GPAIMs), e.g., multimodal large language models. While the AIA is recognized as being one of the first legally binding regulatory frameworks for AI [1], it has faced criticism for its imprecise and incoherent terminology [2, 3], which will complicate its practical implementation. Previous work has examined the AIA and its legislative history to clarify terms like explainability [4–6] and fairness [7]. So far, little attention has been paid to robustness and cybersecurity. Only *AI systems* classified as high-risk (HRAIS) must meet the robustness and cybersecurity requirements set out in Art. 15 AIA. This paper thus focuses on requirements for HRAIS. To provide a clearer understanding, we compare these requirements with requirements for specific *AI models*, namely for GPAIMs with systemic risk, in Art. 55 AIA.

Technical solutions to ensure the robustness and cybersecurity of AI systems are often developed within the ML domain. Therefore, it is essential to inform ML research about the legal requirements

*Equal contribution.
[2]EU Regulation 2024/1689, 12.7.2025.

2nd Workshop on Regulatable ML at NeurIPS 2024.

to ensure compliance with the AIA. However, the vagueness of requirements for cybersecurity and robustness under the AIA makes it challenging to inform ML practitioners about the specific legal requirements to further the development of solutions that can ensure compliance with the AIA. A common understanding between technical and legal domains can be facilitated through technical standards. While the AIA sets out general rules, technical standards specify these rules in detail. Standards are technical specifications designed to provide voluntary technical or quality specifications for current or future products, processes or services.[3] They prescribe technical requirements, including characteristics such as quality or performance levels, terminology, and test methods.[4] Standards have long been integral to EU product legislation under the New Legislative Framework, upon which the AIA is built [8]. Once approved by the EU Commission, technical standards become harmonized technical standards, which grants a presumption of conformity to products or processes that adhere to them. Consequently, compliance with these standards is deemed to fulfill the requirements of the AIA, thereby incentivizing providers to adopt them (Art. 40 AIA). [5]

In this paper, we make the following contributions:

- We analyze and explain the legal requirements related to robustness and cybersecurity in the AIA, identify related shortcomings, and offer possible solutions for some of these shortcomings.
- We evaluate these findings in relation to their practical implementability. This aims to inform the standardization process as well as the benchmark and measurement methodologies referred to in Art. 15(2) AIA.
- We connect the legal requirements for robustness and cybersecurity to ML terminology, aiming to inform ML research and ensure that technical solutions are conducive to legal compliance.

This paper is structured as follows: Section 2 provides a short background on robustness and cybersecurity in the ML literature. Section 3 provides an introduction to the AIA and Art. 15 AIA. Section 4 analyzes the requirements outlined in Art. 15 AIA for HRAIS, addressing both general challenges pertinent to robustness and cybersecurity, as well as specific issues related to each requirement. Section 5 examines the requirements in Art. 55 AIA relevant to GPAIMs with systemic risk. Section 6 concludes with a summary and recommendations for future research.

2 An ML Perspective on Robustness and Cybersecurity

ML research on robustness focuses on mitigating undesired changes in model outputs when deploying models in real world scenarios [9]. This issue is explored across different applications such as computer vision [10–12] and natural language processing [13, 14]. Unintended changes in model outputs can occur due to adversarial or non-adversarial factors affecting the ML model, its input (test) data, or its training data [15, 16]. Perturbations of input (test) data often present a difficult challenge (see Figure 1). While a model's output may be as expected when using "safe" test data from the original population distribution, unintended changes can occur when perturbed examples are provided as input to the ML model.

Non-adversarial (or *natural*) *robustness* often addresses changes in ML model outputs due to *distribution shifts* in input data [16, 17]. These changes occur when the distribution from which the test data is sampled differs from that of the training data [10, 11]. For instance, alterations in data collection methods, such as upgrading to a new X-ray machine, can modify the format or presentation of images [18, 19]. Importantly, distribution shifts can also result from *feedback loops*, where the ML model's outputs influence the data distribution, creating a cycle from the model's output back to its input [20, 21]. Such an effect can be found, for example, in movie recommendation systems, where user's preferences change over time in response to the ML system's suggestions, thereby influencing future recommendations [22]. Other forms of research on non-adversarial robustness investigate the robustness of ML models to noise, which frequently occurs in real-world data sets [23, 24].

[3] Art. 1, 2(1) EU Regulation 1025/2012, OJ L 316, 14.11.2012.

[4] Art. 2(4)(a) and (c) ibid.

[5] The development of harmonized technical standards for the AIA has been initiated by the EU Commission and is expected to be completed within the next years.

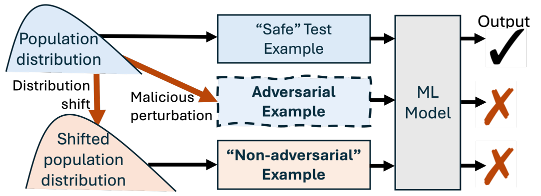

Figure 1: Robustness problems. Outputs may be as expected (✓) with "safe" test data from the original distribution; unintended changes (✗) can occur with adversarial or non-adversarial (shifted) inputs.

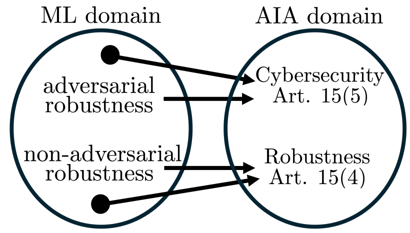

Figure 2: Technical solutions to cybersecurity can be found, i.a., in ML research on *adversarial robustness*; technical solutions to robustness can be found, i.a., in the ML research on *non-adversarial robustness*.

Adversarial robustness refers to the study and mitigation of model evasion attacks using adversarial examples. These are data samples typically drawn from the original population distribution and then modified by an adversary—often in ways that are difficult or even impossible to detect through human oversight—with the intent of altering a model's output [25]. This phenomenon can occur across various models and data types. For instance, in the image domain, small pixel perturbations in input images can lead to significant changes in a model's output [25]. In a broader sense, adversarial robustness also encompasses the study and mitigation of other forms of adversarial attacks that attempt to extract the model or reconstruct or perturb the training data set [26, 27].

From a technical standpoint, adversarial robustness is one aspect of *cybersecurity*. Research on cybersecurity focuses on developing defenses that protect computer systems from attacks compromising their confidentiality, integrity, or availability [28]. This encompasses aspects like data storage, information access and modification, and secure data transmission over networks [29]. Unlike robustness, cybersecurity is not a stand-alone concept in ML, but is discussed more broadly as both a tool for ensuring cybersecurity and a potential source of cybersecurity risks.[6] ML algorithms can be employed to detect and mitigate cybersecurity threats [29], but can also introduce specific vulnerabilities that adversaries may exploit, such as data poisoning or adversarial attacks [32, 33].

3 Background on the AIA and Art. 15 AIA

AIA. The AIA creates harmonized rules for certain AIA systems in order to incentivize the use of such systems in the EU market and prevent regulatory fragmentation between member states (Rec. (1)). Art. 3(1) AIA defines an AI system as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs [...] that can influence physical or virtual environments". These AI systems are regulated differently based on their perceived risk level [3, 34]: Those posing unacceptable risks, such as social scoring, are prohibited or subject to qualified prohibitions; HRAIS, such as those used in medical devices, are allowed but must comply with certain requirements and undergo pre-assessment; other AI systems are subject only to specific transparency and information obligations. Among these categories, only HRAIS must fulfill the robustness and cybersecurity requirements under Art. 15 AIA. According to Art. 16(a) AIA, providers of HRAIS must ensure compliance with these requirements. An AI system is considered a HRAIS if it is either a safety component of a product or a product itself regulated under specific legislation, such as medical devices, machinery, or toys (Art. 6(1) AIA, Annex I), or if it poses a significant risk of harm to the health, safety, or fundamental rights of individuals in specific areas, such as education, employment, or law enforcement (Art. 6(2) and (3) AIA, Annex III). In addition to AI systems, the AIA establishes a separate regime of legal requirements in chapter V of the AIA for a very specific type of AI models, namely GPAIM (e.g., multimodal large language models, see also Section 5).

Art. 15 AIA. Art. 15(1) AIA requires that HRAIS "shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle". The provision outlines specific product-related requirements for AI systems to ensure they are trustworthy. Art. 15(4) AIA mandates that HRAIS exhibit resilience "regarding errors, faults or inconsistencies that may occur within the

[6] E.g., NeurIPS'18 Workshop on Security in ML [30], ICML'22 Workshop on ML for Cybersecurity [31].

system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems". Additionally, Art. 15(5) AIA requires HRAIS to be "resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities". Art. 15(2) AIA requires the EU Commission, together with other relevant stakeholders, to encourage the development of benchmarks and measurement methods for assessing accuracy, robustness, and other performance metrics.

4 Requirements for High-Risk AI Systems

In this section, we provide an analysis of the overarching challenges of implementing Art. 15 AIA (Section 4.1), followed by a discussion regarding the robustness requirement in Art. 15(4) AIA (Section 4.2) and the cybersecurity requirement in Art. 15(5) AIA (Section 4.3).

4.1 General Challenges of Art. 15 AIA

We identify four legal challenges related to Art. 15 AIA that may arise in its practical implementation. First, there is no clear delineation of the legal terms of robustness and cybersecurity and its counterparts in ML literature. Second, while the AIA mandates compliance for entire AI systems, the ML literature primarily focuses on models, which may pose practical challenges for implementation. Third, while accuracy is specified as a requirement in Art. 15 AIA, the provision does not clarify its role in measuring robustness and cybersecurity. Fourth, the terms 'lifecycle' and 'consistent' performance are not defined, leaving ambiguity about how such performance can be practically ensured.

Robustness and Cybersecurity. The robustness requirement in Art. 15(4) AIA addresses "errors, faults, or inconsistencies" that may inadvertently occur as the system interacts with its real-world environment. The cybersecurity requirement in Art. 15(5) AIA targets deliberate attempts "to alter the use, outputs, or performance" of an AI system "by malicious third parties exploiting the system's vulnerabilities". While robustness is a new term in EU legislation and not explicitly defined in the AIA, the term cybersecurity has already been defined in the EU Cybersecurity Act (CSA)[7]. Art. 2(1) CSA defines cybersecurity as "the activities necessary to protect network and information systems, the users of such systems, and other persons affected by cyberthreats". Art. 42(2) AIA considers HRAIS with CSA certification or conformity declarations as compliant with cybersecurity requirements in Art. 15 AIA.[8] This suggests that the CSA definition of cybersecurity applies to the AIA. Both robustness and cybersecurity requirements aim to ensure that HRAIS perform consistently and are resilient against any factors that might compromise this performance. They, however, address different threats to consistent performance: robustness requires protection against unintentional causes, while cybersecurity protects against intentional actions.

We explore how these legal terms could be understood within the ML domain proposing a simple model as an explanatory heuristic (see Figure 2). In ML, robustness refers to maintaining consistent model performance in real-world scenarios [9]. ML research distinguishes between different types of robustness. *Non-adversarial robustness* in ML refers to a model's ability to maintain performance despite data shifts or noise [16, 17, 23, 24]. This aligns with the legal term robustness in the AIA. *Adversarial robustness* in ML refers to the model's resistance to intentional perturbations aimed at altering predictions [25]. This aspect aligns more with the legal concept of cybersecurity. The cybersecurity requirement in Art. 15(5) AIA aims to ensure AI systems' integrity, confidentiality, and availability, protecting them from threats like unauthorized access, adversarial manipulation, data modification, Denial-of-Service attacks, and theft of sensitive information (e.g., model weights). However, other scenarios within the ML domain may also fall under the relevant legal terms. For example, language model jailbreaks exploit AI vulnerabilities to bypass safety constraints [35, 36]. This aligns more closely with the notion of cybersecurity in protecting against misuse of AI systems.

System vs. Model. The AIA regulates AI systems, but not AI models, with the only exception being GPAIMs. ML research, in contrast, often focuses on developing technical solutions for *ML models*. This raises the question of whether solely relying on technical solutions for *ML models* is

[7]Regulation (EU) 2019/881, OJ L 151, 7.6.2019.

[8]Note that this holds only true in so far as the cybersecurity certificate or statement of conformity or parts thereof cover those requirements in Art. 15 AIA.

enough to ensure the compliance of a HRAIS with Art. 15 AIA—or whether additional measures are needed. Rec. (97) specifies that an AI model is an essential component of an AI system.[9] Additional components can include, i.a., user interfaces, sensors, databases, network communication components, or pre- and post-processing mechanisms for model in- and outputs (Rec. (97), [37]). All these individual components should contribute to the overall robustness of the AI system, particularly in scenarios where some components may fail. This is illustrated by Art. 15(4)(ii) AIA, which states that robustness may be ensured through technical redundancy solutions, including "back-up or contingency plans". Furthermore, Art. 15(5)(iii) AIA stipulates that the cybersecurity of AI systems shall be achieved through technical solutions that, "where appropriate", target training data, pre-trained components, the AI model or its inputs. Thus, Art. 15 AIA should not be understood as requiring a single, unified assessment of the requirements. Instead, it must be interpreted as mandating that each component, including one or more ML models, be assessed individually. The assessment of the AI system's overall performance is then derived from an aggregation of the individual performance results [38]. This requires an interdisciplinary approach that draws on expertise from fields such as ML, engineering, and human-computer interaction. To establish a common understanding, it can prove beneficial to formally describe the evaluation process of an entire AI system, including potential challenges, such as interdependencies of technical solutions.

Role of Accuracy. Art. 15(1) AIA mandates that HRAIS shall "achieve an appropriate level of accuracy". This is important because trade-offs between different desiderata can exist, such as between robustness and accuracy (see Appendix F). While accuracy is not defined in the AIA, Annex IV No. 3 AIA states that accuracy is an indicator of the capabilities and performance limits of an AI system. Accordingly, accuracy should be measured in at least two ways: i) separately for "specific persons or groups of persons on which the system is intended to be used"[10], and ii) the overall expected accuracy for the "intended purpose" of the AI system. In ML, the metric *accuracy* typically describes the overall proportion of correct predictions out of the total number of predictions made [41]. However, the term can also describe the objective of "good performance" of an AI system and, depending on its specific purpose, can also be evaluated using different metrics, such as utility [42] and f1-score [43]. Art. 15(3) AIA explicitly references 'accuracy and the relevant accuracy metrics', indicating that accuracy is understood as an objective that can be measured with various metrics, leaving the choice of the relevant metric to the provider.[11]

In ML, robustness is often measured using an *accuracy* metric. Typically, this involves comparing the *accuracy* (or error rates) evaluated on an unperturbed dataset from the original distribution with the accuracy on a perturbed test set (e.g., sampled from the shifted distribution or containing adversarial samples) [11, 44, 45]. The smaller the difference between these two *accuracy* results, the better the *robustness*. The choice of the *accuracy* metric thus has an impact on the measurement of robustness. As a result, the ML model may appear more robust under some accuracy metrics than others.[12] Technical standards should provide guidelines on how AI system providers should choose an appropriate 'accuracy' measure, especially when it is used to assess robustness in subsequent steps.

Consistent Performance Throughout the Lifecycle. AI systems must perform "consistently" in terms of accuracy, robustness, and cybersecurity "throughout their lifecycle" (Art. 15(1) AIA). However, the requirement of 'consistent' performance remains undefined and there is no specification on how it should be measured. For a brief discussion of the term term 'lifecycle', see Appendix C. In the ML literature, a model's variability in performance over time is often measured using the variance of a metric such as accuracy or robustness [49–51]. The variance of a metric over a time interval indicates its deviation from its mean within this interval. For instance, high variance in robustness indicates significant fluctuations in robustness levels between two points in time, whereas low variance indicates similar levels of robustness over time. A low variance could therefore be understood as a consistent performance.[13] In practice, performance can vary due to factors, such as

[9]Although Rec. (97) specifically refers to GPAIMs, the wording suggests that the statement about the relationship between AI systems and AI models is of a general nature.

[10]This links to fairness ML literature on diverging error rates for different sensitive groups [39, 40].

[11]The selection of the metric should consider various factors, including the specific purposes of the ML model, dataset-specific circumstances (e.g., imbalanced data) and the particular model type (e.g., classification, regression). Technical standards should clarify how AI systems' accuracy is defined and measured.

[12]The selection of favorable metrics has been studied in fair ML under the term fairness hacking [46–48].

[13]Some also consider consistency as a metric itself, rather than as a property of a (robustness) metric [52].

random initializations of weights or input data sampling. These types of variations are unavoidable. Defining level of variance considered ‘consistent’ is challenging as it is dependent on the context. Technical standards should clarify how to measure a consistent performance with respect to accuracy, robustness, and cybersecurity, and provide guidance on determining the required level of consistency.

4.2 Robustness Art. 15(4) AIA

We now turn to challenges in Art. 15(4) AIA. Art. 15(4)(i) AIA states that “technical and organisational measures shall be taken” to ensure that AI systems are “as resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment”. Art. 15(4)(ii) AIA specifies that robustness can be achieved through technical redundancy solutions, and Art. 15(4)(iii) AIA requires addressing feedback loops in online learning with possibly biased outputs.

Inconsistent Terminology. The term robustness is used inconsistently throughout the AIA. Art. 15(1) and (4) AIA refer to robustness, whereas the corresponding Rec. (27) and Rec. (75) both mention technical robustness. One could argue that technical robustness is synonymous with robustness. The term ‘technical robustness’ in Rec. (27) may be a remnant of the legislative process that built on the 2019 Ethics Guidelines for Trustworthy AI [53] developed by the AI IHEG, which introduced the principle of ‘technical robustness and safety’. These guidelines are explicitly referenced by Rec. (27). Nevertheless, it remains unclear why Rec. (75) also refers to ‘technical robustness’. It could be that the wording in Rec. (75) is borrowed from Rec. (27). Alternatively, the terms could refer to different concepts: Either the term robustness limited to technical aspects, or it additionally includes some form of non-technical robustness. The latter could refer to organizational measures that must be implemented to ensure robustness, as mandated in Art. 15(4)(i) AIA. Technical standards should clarify the definition of robustness and delineate the aspects it encompasses.

Required Level of Robustness. Art. 15(1) AIA mandates that AI systems must achieve an “appropriate level” of robustness. Art. 15(4) AIA, however, demands that AI systems shall be “as resilient as possible” to “errors, faults, or inconsistencies”, suggesting a stricter requirement. This discrepancy initially appears ambiguous, as it is unclear whether HRAIS must simply meet an appropriate standard of robustness or strive for the highest possible level. However, the “appropriate” level stated in Art. 15(1) AIA can be understood as a general principle, which is further specified by Art. 15(4) AIA. Therefore, appropriate with respect to robustness is to be understood as ‘as resilient as possible’.

When determining the appropriate level of robustness of a specific HRAIS, the intended purpose of the system and the generally acknowledged state of the art (SOTA) on AI and AI-related technologies must be taken into account (Art. 8(1) AIA). Art. 9(4) AIA acknowledges that one of the objectives of the required risk management is to achieve an “appropriate balance in the implementation of measures to fulfil” requirements. Art. 9(5) AIA further acknowledges the permissibility of a residual risk, meaning that the measures adopted under the risk management system are not expected to eliminate all existing risks, but rather to maintain these residual risks at an ’acceptable’ level. The risk management system is a continuous iterative process (Art. 9(1) AIA). This means that the appropriate level of robustness of HRAIS must be regularly determined and updated, taking into account its purpose and the SOTA while balancing it with other requirements.

Feedback Loops. Art. 15(4)(iii) AIA states that AI systems must be explicitly developed in such a way that they “duly address” feedback loops and “eliminate or reduce” the risks associated with them. According to Rec. (67), feedback loops occur when the output of an AI system influences its input in future operations, an under understanding that aligns with the concept as found in the ML literature. Feedback loops are a well-studied problem manifesting in various forms [54], with the most common issues being a distribution shift [22] or a selection bias [49, 55]. Importantly, in this context, the risk of ’biased outputs’ in feedback loops (Art. 15(4)(iii) AIA) is often studied in the literature on fairness in ML rather than in the literature on *robustness* in ML, which traditionally constitute different research fields and communities [56]. [14]

An important aspect of Art. 15(4)(iii) AIA is that it applies specifically to AI systems that learn online. Online learning ML models iteratively learn from a sequence of data and continuously update their

[14] For example, whether there is a trade-off between *robustness* and fairness, or if both pursue similar goals, remains an active discussion in the ML community [56–58].

parameters over time [59]. This adaptiveness is reflected in Art. 3(1) AIA as a factual characteristic of an AI system. The problem with feedback loops in online learning is that newly collected training data can become biased, e.g., due to selection bias, which occurs when the data collected is not representative of the overall population [60, 61]. This can distort model predictions and reinforce existing biases, ultimately impacting the model's accuracy and fairness [49–51]. Offline models, in contrast, are trained on a fixed dataset all at once [59]. Offline models can also carry risks when feedback loops are present: The outputs of an ML model can induce a distribution shift through their interaction with the environment [20, 21, 62]. Since an offline ML model is not updated, distribution shifts can influence their performance over time and possibly lead to fairness concerns [62]. Although Art. 15(4) AIA does not explicitly address feedback loops in offline systems, HRAIS are not exempt from addressing them. Since they can impact the model's accuracy, feedback loops in offline systems may still need to be addressed to comply with Art. 15(1) AIA.

4.3 Cybersecurity Art. 15(5) AIA

We now turn to legal challenges specific to Art. 15(5) AIA. Art. 15(5)(i) AIA states that AI systems shall be resilient against attempts to "alter their use, outputs, or performance by exploiting system vulnerabilities". Art. 15(5)(ii) AIA specifies that technical solutions aiming to ensure resilience against such malicious attempts "shall be appropriate to the relevant circumstances and the risks". Finally, Art. 15(5)(iii) AIA mandates specific measures "to prevent, detect, respond to, and control for attacks" exploiting AI-specific vulnerabilities. This section examines the key aspects of compliance with Art. 15(5) AIA. However, a mentioned above, providers have an additional pathway for demonstrating compliance with its cybersecurity requirements, namely a certification under the CSA [63].

Required Level of Cybersecurity. Art. 15(5)(ii) AIA mandates that technical solutions must be "appropriate to the relevant circumstances and the risks", but this needs further clarification. Specifically, it is unclear when technical solutions are appropriate to the relevant circumstances and the risks. The AIA specifically addresses only three kinds of risks: health, safety, and fundamental rights (Rec. (1)). Risks associated with these aspects can be identified and managed through a risk management system that must be put into place as stipulated by Art. 9 AIA. Relevant circumstances are any known and foreseeable circumstances that may have an impact on cybersecurity.[15]

Mandating a cybersecurity level that is 'appropriate to the relevant circumstances' acknowledges that complex ML models generally cannot be expected to be fully resistant to all types of adversarial attacks. This has two major reasons: First, it is impossible to anticipate all types of possible attacks. This is acknowledged by Art. 9(5) AIA which states that measures adopted under the risk management system are not expected to remove all existing risks. Second, complete protection against a specific attack cannot be guaranteed, especially as adversaries continuously adapt their strategies to overcome possible defense mechanisms [64, 65]. The appropriateness of a certain performance level must consider the intended purpose of the system and the generally acknowledged SOTA (see Art. 8(1) AIA). The measures to ensure cybersecurity adopted are not expected to eliminate all existing risks, but the overall residual risk must be acceptable (see Art. 9(1) and (4) AIA). Thus, when determining the appropriateness of technical solutions, all applicable requirements of the AIA must be balanced, while also mitigating risks to health, safety, and fundamental rights.

AI-specific Vulnerabilities. Art. 15(5) AIA differentiates between 'system vulnerabilities' (Art. 15(5)(i) AIA) and 'AI-specific vulnerabilities' (Art. 15(5)(iii) AIA). As the term vulnerability is not defined, we provide a working definition. The United States' Common Vulnerabilities and Exposures (CVE) system defines vulnerability as "[a]n instance of one or more weaknesses [...] that can be exploited, causing a negative impact to confidentiality, integrity, or availability" [66]. Art. 15(5)(iii) AIA provides a non-exhaustive list of components of an AI system that expose AI-specific vulnerabilities, such as training data, pre-trained components used in training, inputs, or the AI model. However, there might be additional components of the AI system that may also harbor AI-specific vulnerabilities. The question is how to identify them. We suggest performing a hypothetical test. AI models play a central role in an AI system. If a vulnerability would be eliminated by replacing the AI model with a non-AI model, it should be deemed 'AI-specific'. To define a non-AI model, we return to the definition of an AI system under the AIA. It has been argued that the central characteristic of an AI system is its ability to infer from input to output [67]. This inference ability is typically

[15] See Art. 13(3)(b)(ii) AIA and Appendix E.

performed by one or more AI models within an AI system. Therefore, non-AI models are all models lacking inference capability, such as rule-based decision-making systems that rely on predefined rules and logic defined by human experts.[16] Since AI-specific vulnerabilities relate to specific components of an AI system, we suggest viewing them as a subset of system vulnerabilities. To enhance clarity, technical standards should define both terms and mandate a process for identifying them.

Technical Solutions. Art. 15(5)(iii) AIA provides a non-exhaustive list of attacks and AI-specific vulnerabilities that must be addressed through technical solutions:[17] data poisoning, model poisoning, adversarial examples, model evasion, and confidentiality attacks, which are well-established in the ML literature. These attacks aim to induce model failures [36]: *Data poisoning* attacks manipulate training data [70], *model poisoning* attacks manipulate the trained ML model [71], and *model evasion* attacks manipulate test samples [72]. *Confidentiality attacks*, typically explored in the field of privacy in ML, refer to attempts to extract information about the training data or the model itself [73].

In addition to these attacks, Art. 15(5)(iii) AIA lists 'model flaws' as an AI-specific vulnerability. This is a vague legal term and lacks an established counterpart in the ML literature. In software contexts, the word *flaw* often refers to so-called *bugs*, which are typically the result of human errors in the coding process [38, 74]. However, the term model flaw follows the list of attacks outlined above, which are instead designed to exploit the default properties of a properly functioning ML model, and are not directly the results of errors in the coding process. Thus, it is unclear what model flaw refers to in this context, and whether technical solutions are only expected to address traditional bugs or coding errors, or whether they should address other ways of exploiting AI-specific vulnerabilities. Given that the term is situated within the cybersecurity requirements for AI system outlined in Art. 15(5) AIA, we argue that the term model flaws should be interpreted as flaws that enable the exploitation of AI-specific vulnerabilities. Technical standards should define model flaws more clearly and provide guidelines for technical solutions to address these model flaws. This should take into account the arms race between attacker and defender in the realm of adversarial robustness, where both parties are continuously adapting their strategies to outmaneuver the other [75]. This makes it infeasible to anticipate and counter all potential attacks that target AI-specific vulnerabilities.

5 Requirements for General-Purpose AI Models With Systemic Risk

In the previous section, we examined HRAIS requirements. To further elucidate them, we study GPAIMs with systemic risk, highlighting similarities and differences. The AIA establishes legal requirements for GPAIMs, such as multimodal large language models [76, 77], which can perform tasks beyond their original training objective [78]. GPAIM can be standalone or embedded in an HRAIS, with the latter requiring compliance with both GPAIM and HRAIS requirements. The AIA distinguishes between GPAIM with systemic risks and those without. Art. 3(65) AIA defines 'systemic risk' as the risk that is specific to the high-impact capabilities of GPAIMs that have a "significant impact" on the market, public health, safety, security, fundamental rights, or society.[18]. GPAIMs without systemic risks are exempt from robustness and cybersecurity obligations (Art. 53 AIA ff.).

Cybersecurity Requirements. Art. 55(1)(d) AIA mandates "an adequate level of cybersecurity protection" for GPAIMs with systemic risk. Rec. (115) further details this cybersecurity requirement. It mandates cybersecurity protection against "malicious use or attacks" and lists specific adversarial threats, such as "accidental model leakage, unauthorised releases, circumvention of safety measures", "cyberattacks", or "model theft". Notably, several of these threats have direct counterparts in the ML literature on *adversarial robustness* and *privacy* for large generative models, such as the circumvention of safety measures (jailbreaking) or model theft [82–84]. Although Art. 55(1)(d) AIA does not define the term 'cyberattacks', we infer that it includes the attacks exploiting AI-specific vulnerabilities

[16] Note that non-AI rule-based systems use human-defined rules, while rule-based ML models infer rules from data [68, 69], qualifying as AI models. In a different context, the AI IHEG ethics guidelines [53] suggest fallback plans where AI systems switch from a statistical (ML) approach to a rule-based or human-in-the-loop approach.

[17] Note that while organizational measures are explicitly mandated for the robustness of HRAIS in Art. 15(4)(i) AIA, they are omitted in the cybersecurity requirement (see Appendix D).

[18] A systemic risk is presumed when the cumulative computation during training exceeds 10^{25} Floating-Point Operations Per Second (FLOPS). GPAIMs with fewer FLOPS may still be classified as posing a systemic risk under Art. 51(1) AIA. There is an ongoing debate over this threshold and if a model's complexity truly reflects its risk level [67, 79–81]. Thresholds and criteria can be modified by the EU Commission, see Appendix B

mentioned in Art. 15 AIA (see Section 4.3). These attacks are studied in the field of *adversarial robustness* and can also affect GPAIMs [85–89]—even though specific ML techniques may be necessary to address GPAIM-specific challenges. This relation underscores that the concepts and problems explored under *adversarial robustness* are reflected in the term 'cybersecurity' as used in Art. 55(1)(d) AIA. To ensure the cybersecurity of a GPAIMs with systemic risk, providers must conduct and document internal and/or external adversarial testing of the model, such as red teaming[19]

While the AIA mandates robustness requirements for HRAIS, we observe that it does not impose an explicit equivalent legal requirement for GPAIMs, regardless of whether they present a systemic risk or not. Specifically, neither Art. 55 AIA nor Rec. (155) address unintentional causes for deviations from consistent performance. In Section 4.2, we stated that *non-adversarial robustness* is reflected in the term robustness in Art. 15 AIA. Consequently, GPAIMs, which are not required to fulfill any robustness requirement, are not mandated to be resilient against performance issues, such as data distribution shifts or noisy data. The AIA itself does not provide an explanation for the omission of a robustness requirement. It may stem from the complexity of political negotiations regarding the AIA, particularly regarding GPAIMs, which were not addressed in the initial draft of the regulation but gathered widespread media attention during the legislative procedure. However, evidence from ML research suggests that *non-adversarial robustness* is also relevant for GPAIMs [92, 93].

Required Level of Cybersecurity. Art. 55(1)(d) AIA mandates an 'adequate' level of cybersecurity protection for GPAIMs with systemic risk. This requirement contrasts with the 'appropriate' level of cybersecurity mandated for HRAIS under Art. 15(1) AIA. The use of these two different terms raises questions about whether they are equivalent or to what extent they differ. On the one hand, 'adequate' and 'appropriate' could imply different levels of cybersecurity. The Cambridge Dictionary defines the term 'adequate' as "enough or satisfactory for a particular purpose" [94] and 'appropriate' as "suitable or right for a particular situation or occasion" [95]. Accordingly, something is 'adequate' if it exceeds a minimum threshold that is good enough, while something is 'appropriate' if it meets a specific (right) level above that minimum. GPAI models can perform a wide variety of tasks in different contexts and thus be prone to a variety of different cybersecurity risks, making it difficult to identify and mitigate their specific cybersecurity risks. For this reason, it may be reasonable to only mandate an 'adequate', i.e., minimum level of cybersecurity. HRAIS, independently of whether they contain an GPAIM as a component, can be thought of as operating in a more specific contexts, potentially allowing an easier and more precise assessment of cybersecurity risks and thus a more stringent appropriate level of cybersecurity protection. On the other hand, 'adequate' and 'appropriate' could refer to the same level of cybersecurity. Rec. (115) states that "adequate technical and established solutions" must be "appropriate to the relevant circumstances and the risks". The simultaneous use of both terms in a single sentence, intended to guide the interpretation of Art. 55(1)(d) AIA, suggests that they might be intended as synonymous. This is corroborated by the observation that many official language versions of the AIA use a single term for both "adequate" and "appropriate" in Art. 15(1) AIA and Art. 55(1)(d) AIA.[20] To resolve this ambiguity, technical standards should clarify the required level of cybersecurity for GPAIMs with systemic risk.

6 Summary and Outlook

We identified legal challenges and potential limitations in implementing robustness and cybersecurity requirements for HRAIS under Art. 15(4) and (5) AIA. We also examined GPAIMs with systemic risk, which face cybersecurity but not robustness requirements, and identified additional legal challenges. Our analysis highlights the need for clearer specifications of these provisions through harmonized standards or the benchmark and measurement methodologies foreseen for robustness and cybersecurity in Art. 15(2) AIA. These would help define technical requirements and establish evaluation criteria for AI systems. Future research should focus on non-adversarial robustness for large generative AI, and explore legal intersections with frameworks like the Medical Device Regulation [96, 97]. Additionally, the focus on models in ML research versus entire AI systems in the AIA underscores the need for interdisciplinary work. Within the ML domain, future work should explore the impact of accuracy metrics on robustness, potential 'robustness hacking', and methods to measure and ensuring long-term performance consistency in the presence of feedback loops.

[19] In this context, red teaming refers to stress testing AI models by simulating adversarial attacks [90], unlike traditional cybersecurity red teaming, which focuses on assessing entire systems or networks [91].

[20] Such as FR "approprié", ES "adecuado", GER "angemessen", IT "adeguato".

References

[1] Foo Yun Chee and Tassilo Hummel. Europe sets benchmark for rest of the world with landmark ai laws. *Reuters*, May 2024. URL `https://www.reuters.com/world/europe/eu-countries-back-landmark-artificial-intelligence-rules-2024-05-21/`.

[2] Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32, 2024.

[3] David Bomhard and Jonas Siglmüller. Ai act – das trilogergebnis. *RDi*, 45, 2024.

[4] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 891–905, 2022.

[5] Fabio Vitali. A survey on methods and metrics for the assessment of explainability under the proposed ai act. In *Legal Knowledge and Information Systems: JURIX 2021: The Thirty-Fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021. Vol. 346*, page 235. IOS Press, 2022.

[6] Georgios Pavlidis. Unlocking the black box: analysing the eu artificial intelligence act's framework for explainability in ai. *Law, Innovation and Technology*, pages 1–16, 2024.

[7] Luca Deck, Jan-Laurin Müller, Conradin Braun, Domenique Zipperling, and Niklas Kühl. Implications of the ai act for non-discrimination law and algorithmic fairness. *3rd European Workshop on Algorithmic Fairness*, 2024.

[8] Lukasz Gorywoda. The new european legislative framework for the marketing of goods. *Columbia Journal of European Law*, 16:161, 2009.

[9] Leo Schwinn, Leon Bungert, An Nguyen, René Raab, Falk Pulsmeyer, Doina Precup, Björn Eskofier, and Dario Zanca. Improving robustness against real-world and worst-case distribution shifts through decision region quantification. In *International Conference on Machine Learning*, pages 19434–19449. PMLR, 2022.

[10] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639*, 2021.

[11] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[12] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 321–331, 2020.

[13] Emanuele La Malfa and Marta Kwiatkowska. The king is naked: On the notion of robustness for natural language processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11047–11057, 2022.

[14] Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. Robustness and adversarial examples in natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 22–26, 2021.

[15] Lei Cheng, Xiaowen Huang, Jitao Sang, and Jian Yu. Towards robust recommendation: A review and an adversarial robustness evaluation library. *arXiv preprint arXiv:2404.17844*, 2024.

[16] Andrea Tocchetti, Lorenzo Corti, Agathe Balayn, Mireia Yurrita, Philip Lippmann, Marco Brambilla, and Jie Yang. Ai robustness: a human-centered perspective on technological challenges and opportunities. *ACM Computing Surveys*, 2024.

[17] Gorana Gojić, Vladimir Vincan, Ognjen Kundačina, Dragiša Mišković, and Dinu Dragan. Non-adversarial robustness of deep learning methods for computer vision. In *2023 10th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN)*, pages 1–9. IEEE, 2023.

[18] Ben Glocker, Robert Robinson, Daniel C Castro, Qi Dou, and Ender Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*, 2019.

[19] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

[20] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

[21] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

[22] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[23] José A Sáez, Julián Luengo, and Francisco Herrera. Evaluating the classifier behavior with noisy data considering performance and robustness: The equalized loss of accuracy measure. *Neurocomputing*, 176:26–35, 2016.

[24] Amanda Olmin and Fredrik Lindsten. Robustness and reliability when training with noisy labels. In *International Conference on Artificial Intelligence and Statistics*, pages 922–942. PMLR, 2022.

[25] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[26] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*, 2018.

[27] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[28] Dipankar Dasgupta, Zahid Akhtar, and Sajib Sen. Machine learning in cybersecurity: a comprehensive survey. *The Journal of Defense Modeling and Simulation*, 19(1):57–106, 2022.

[29] Iqbal H Sarker, Md Hasan Furhad, and Raza Nowrozy. Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. *SN Computer Science*, 2(3):173, 2021.

[30] Nicolas Papernot, Florian Tramer, Kamalika Chaudhuri, Matt Fredrikson, and Jacob Steinhardt. Neurips 2018 workshop on security in machine learning, 2018. URL `https://secml2018.github.io/`. Accessed: 2024-10-22.

[31] Andy Applebaum, William Arbaugh, Jack W. Davidson, Joseph Edappully, John Emanuello, Howie Huang, Andrew Golczynski, Nicole Nichols, Tejas Patel, Ahmad Ridley, and Vance Wong. Icml 2024 workshop on machine learning for cybersecurity, 2022. URL `https://sites.google.com/view/icml-ml4cyber/home`. Accessed: 2024-10-22.

[32] Maryam Roshanaei, Mahir R Khan, and Natalie N Sylvester. Navigating ai cybersecurity: Evolving landscape and challenges. *Journal of Intelligent Learning Systems and Applications*, 16(3):155–174, 2024.

[33] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021.

[34] Lucilla Sioli. A european strategy for artificial intelligence. Presentation at the CEPS webinar - European approach to the regulation of artificial intelligence, April 23 2021. URL `https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf`.

[35] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

[36] Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology, 2024.

[37] Henrik Junklewitz, Ronan Hamon, Antoine-Alexandre André, Tatjana Evas, Josep Soler Garrido, and Jose Ignacio Sanchez Martin. Cybersecurity of artificial intelligence in the ai act. Technical Report KJ-NA-31-643-EN-N (online), Joint Research Center, Luxembourg (Luxembourg), 2023.

[38] Ram Shankar Siva Kumar and Hyrum Anderson. *Not with a Bug, But with a Sticker: Attacks on Machine Learning Systems and What To Do About Them*. John Wiley & Sons, 2023, 2023.

[39] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and its Application*, 8(1):141–163, 2021.

[40] Alexandra Chouldechova and Max G'Sell. Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046*, 2017.

[41] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[42] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[43] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*, pages 1015–1021. Springer, 2006.

[44] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[45] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *6th International Conference on Learning Representations*, 2015.

[46] Kristof Meding and Thilo Hagendorff. Fairness hacking: The malicious practice of shrouding unfairness in algorithms. *Philosophy & Technology*, 37(1):4, 2024.

[47] Jan Simson, Florian Pfisterer, and Christoph Kern. One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1305–1320, 2024.

[48] Emily Black, Talia Gillis, and Zara Yasmine Hall. D-hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 602–615, 2024.

[49] Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 277–287. PMLR, 2020.

[50] Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Steven Z Wu. Equal opportunity in online classification with partial feedback. *Advances in Neural Information Processing Systems*, 32, 2019.

[51] Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P Gummadi, and Isabel Valera. Don't throw it away! the utility of unlabeled data in fair decision making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1421–1433, 2022.

[52] Alexander Wei and Fred Zhang. Optimal robustness-consistency trade-offs for learning-augmented online algorithms. *Advances in Neural Information Processing Systems*, 33:8042–8053, 2020.

[53] AI IHEG. High-level expert group on artificial intelligence. *Ethics Guidelines for Trustworthy AI*, 2019.

[54] Nicolò Pagan, Joachim Baumann, Ezzat Elokda, Giulia De Pasquale, Saverio Bolognani, and Anikó Hannák. A classification of feedback loops and their relation to biases in automated decision-making systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–14, 2023.

[55] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[56] Jae-Gil Lee, Yuji Roh, Hwanjun Song, and Steven Euijong Whang. Machine learning robustness, fairness, and their convergence. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4046–4047, 2021.

[57] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pages 11492–11501. PMLR, 2021.

[58] Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3320–3331, 2021.

[59] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[60] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st International Conference on Machine learning*, page 114, 2004.

[61] Anqi Liu and Brian Ziebart. Robust classification under sample selection bias. *Advances in Neural Information Processing Systems*, 27, 2014.

[62] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[63] Federica Casarosa. Cybersecurity certification of artificial intelligence: a missed opportunity to coordinate between the artificial intelligence act and the cybersecurity act. *International Cybersecurity Law Review*, 3(1):115–130, 2022.

[64] Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496, 2023.

[65] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.

[66] CVE. Common vulnerabilities and exposures glossary. `https://www.cve.org/ResourcesSupport/Glossary`, 2024. Accessed: 2024-07-03.

[67] Philipp Hacker. Comments on the final trilogue version of the ai act. *Available at SSRN 4757603*, 2024.

[68] Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. Do machine learning models learn statistical rules inferred from data? In *International Conference on Machine Learning*, pages 25677–25693. PMLR, 2023.

[69] Sholom M Weiss and Nitin Indurkhya. Rule-based machine learning methods for functional prediction. *Journal of Artificial Intelligence Research*, 3:383–403, 1995.

[70] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pages 9389–9398. PMLR, 2021.

[71] Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2545–2555, 2022.

[72] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 387–402. Springer, 2013.

[73] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.

[74] Helen Nissenbaum. Accountability in a computerized society. *Science and Engineering Ethics*, 2:25–42, 1996.

[75] Lingwei Chen, Yanfang Ye, and Thirimachos Bourlai. Adversarial machine learning in malware detection: Arms race between evasion attack and defense. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pages 99–106. IEEE, 2017.

[76] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023.

[77] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[78] Carlos I Gutierrez, Anthony Aguirre, Risto Uuk, Claire C Boine, and Matija Franklin. A proposal for a definition of general purpose artificial intelligence systems. *Digital Society*, 2 (3):36, 2023.

[79] Claudio Novelli, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. Generative ai in eu law: liability, privacy, intellectual property, and cybersecurity. *arXiv preprint arXiv:2401.07348*, 2024.

[80] Ceyhun Necati Pehlivan. The eu artificial intelligence (ai) act: An introduction. *Global Privacy Law Review*, 5(1), 2024.

[81] Cornelia Kutterer. Regulating foundation models in the ai act: From “high” to “systemic” risk. *AI-Regulation Papers 24-01-1*, 2023.

[82] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.

[83] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[84] Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, and Tom H Luan. A survey on chatgpt: Ai-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, 2023.

[85] Yao Qiang, Xiangyu Zhou, Saleh Zare Zade, Mohammad Amin Roshani, Douglas Zytko, and Dongxiao Zhu. Learning to poison large language models during instruction tuning. *arXiv preprint arXiv:2402.13459*, 2024.

[86] Avisha Das, Amara Tariq, Felipe Batalini, Boddhisattwa Dhara, and Imon Banerjee. Exposing vulnerabilities in clinical llms through data poisoning attacks: Case study in breast cancer. *medRxiv*, 2024.

[87] Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6065–6086, 2024.

[88] Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Gunnemann. Soft prompt threats: Attacking safety alignment and unlearning in open-source llms through the embedding space. *arXiv preprint arXiv:2402.09063*, 2024.

[89] João Vitorino, Eva Maia, and Isabel Praça. Adversarial evasion attack efficiency against large language models. *arXiv preprint arXiv:2406.08050*, 2024.

[90] Michael Feffer, Anusha Sinha, Wesley H Deng, Zachary C Lipton, and Hoda Heidari. Red-teaming for generative ai: Silver bullet or security theater? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 421–437, 2024.

[91] Fabian M Teichmann and Sonia R Boticiu. An overview of the benefits, challenges, and legal aspects of penetration testing and red teaming. *International Cybersecurity Law Review*, 4(4): 387–397, 2023.

[92] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

[93] Pin-Yu Chen, Sijia Liu, Sayak Paul, and Hugging Face. Foundational robustness of foundation models. In *Annual Conference on Neural Information Processing Systems*, 2022.

[94] Cambridge University Press. Adequate, 2024. URL `https://dictionary.cambridge.org/dictionary/english/adequate`. Accessed: 2024-07-15.

[95] Cambridge University Press. Appropriate, 2024. URL `https://dictionary.cambridge.org/dictionary/english/appropriate`. Accessed: 2024-07-15.

[96] Elisabetta Biasin, Burcu Yasar, and Erik Kamenjasevic. New cybersecurity requirements for medical devices in the eu: the forthcoming european health data space, data act, and artificial intelligence act. *Law, Tech. & Hum.*, 5:43, 2023.

[97] Henrik Nolte and Zeynep Schreitmüller. Cybersicherheit ki-basierter medizinprodukte im lichte der mdr und ki-vo. *Zeitschrift für das gesamte Medizinprodukterecht*, 1:20, 2024.

[98] Tadas Klimas and Jurate Vaiciukaite. The law of recitals in european community legislation'(2008). *ILSA Journal of International & Comparative Law*, 15:61, 2008.

[99] J Scott Marcus. Promoting product longevity. how can the eu product safety and compliance framework help promote product durability and tackle planned obsolescence, foster the production of more sustainable products, and achieve more transparent supply chains for consumers? *Policy Department for Economic, Scientific and Quality of Life Policies Directorate-General for Internal Policies*, 2020.

[100] Shinsuke Murakami, Masahiro Oguchi, Tomohiro Tasaki, Ichiro Daigo, and Seiji Hashimoto. Lifespan of commodities, part i: The creation of a database and its review. *Journal of Industrial Ecology*, 14(4):598–612, 2010.

[101] Ronan Hamon, Henrik Junklewitz, Josep Soler Garrido, and Ignacio Sanchez. Three challenges to secure ai systems in the context of ai regulations. *IEEE Access*, 2024.

[102] Cambridge University Press. Relevant, 2024. URL `https://dictionary.cambridge.org/dictionary/english/relevant`. Accessed: 2024-07-17.

[103] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[104] Jingyang Li and Guoqiang Li. The triangular trade-off between robustness, accuracy and fairness in deep neural networks: A survey. *ACM Computing Surveys*, 2024.

[105] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

[106] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7909–7919, 2020.

[107] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*, 2022.

[108] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, pages 17258–17277. PMLR, 2022.

Appendix

A Legal Terminology

The AIA is formally structured into recitals (Rec.), articles (Art.), and annexes. Recitals are legally non-binding and outline the rationale behind the articles, articles delineate specific binding obligations, and the annexes provide additional details and specifications to support the articles [98].

B Notions of High-Risk AI Systems and General Purpose Models provided in AIA

AI System. The AIA defines an AI system as "a machine-based[21] system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Art. 3(1) AIA). Thereby, Rec. (12), suggests that notion of 'AI system' "should be clearly defined and should be closely aligned with the work of international organisations working on AI to ensure legal certainty, facilitate international convergence and wide acceptance, while providing the flexibility to accommodate the rapid technological developments in this field". Importantly, "the definition should be based on key characteristics of AI systems that distinguish it from simpler traditional software systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations" (Rec. (12)). Thereby a "key characteristic of AI systems is their capability to infer", which refers to "the process of obtaining the outputs, such as predictions, content, recommendations, or decisions, which can influence physical and virtual environments, and to a capability of AI systems to derive models or algorithms, or both, from inputs or data" (Rec. (12)). Examples for techniques that enable inference "include machine learning approaches that learn from data how to achieve certain objectives, and logic- and knowledge-based approaches that infer from encoded knowledge or symbolic representation of the task to be solved" (Rec. (12)).

High-risk AI Systems (HRAIS). The AIA classifies AI systems into different risk groups. Risk refers thereby to "the combination of the probability of an occurrence of harm and the severity of that harm" (Art. 3(2) AIA). An AI system is considered high-risk, if it is "intended to be used as a safety component of a product, or the AI system is itself a product" and "is required to undergo a third-party conformity assessment, with a view to the placing on the market or the putting into service of that product" covered by the Union harmonisation legislation listed in Annex I (Art. 6(1) AIA). Annex I AIA provides a list of 20 EU harmonisation legislation. For example, the Machinery Directive (Directive 2006/42/EC and amending Directive 95/16/EC), the directive on the safety of toys (Directive 2009/48/EC), or the directive concerning agricultural and forestry vehicles (Regulation (EU) No 167/2013). This means that for example, an AI system that is used in a product or as a product itself considered as a toy falling under Directive 2009/48/EC, would be considered as a high-risk AI system.

An AI systems is also considered high-risk, if it operates in types or use-cases enlisted in Annex III (Art. 6(2) AIA) and poses a significant risk of harm to the health, safety or fundamental rights of natural person following (Art. 6(3) AIA). Annex III enlists eight types or use-cases of AI systems: biometric applications (e.g., remote biometric identification systems excluding biometric verification, biometric categorisation, and emotion recognition); critical infrastructure (e.g., "critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity"); education and vocational training; employment, workers management and access to self-employment; essential private and public services; law enforcement; migration, asylum and border control management; and administration of justice and democratic processes. Thereby, the Commission shall "provide [...] a comprehensive list of practical examples of use cases of AI systems that are high-risk and not high-risk" (Art. 6(5) AIA). Rec. (52) suggests that "it is appropriate to classify them [i.e., AI systems] as high-risk if, in light of their intended purpose, they pose a high risk of harm to the health and safety or the fundamental rights of persons, taking into account both the severity of the possible harm and its probability of occurrence and they are used in a number of specifically pre-defined areas". Thereby

[21] Thereby, machine-based "refers to the fact that AI systems run on machines" (Rec. (12)).

the methodology and citeria for the identification of high-risk should be able to be adopted in order to account for "the rapid pace of technological development, as well as the potential changes in the use of AI systems" (Rec. (52)). For example, an AI system used for remote biometric identification is considered high-risk, unless it does not pose a significant risk. Remote identification refers to comparing biometric data of that individual to stored biometric data of individuals in a reference database (Rec. (15)), where a remote biometric identification system should be understood as a "AI system intended for the identification of natural persons without their active involvement, typically at a distance" (Rec. (17)).

General-purpose AI Models (GPAIM). A general-purpose AI model (GPAIM) refers to an "AI model [...] that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications" (Art. 3(63) AIA). The AIA, however, explicitly excludes from its regulation AI models that may fall under this definition but are "used for research, development or prototyping activities before they are placed on the market" (Art. 3(63) AIA). Rec. (97) suggest that the term general-purpose AI model "should be clearly defined and set apart from the notion of AI systems to enable legal certainty" taking into account the "key functional characteristics of a general-purpose AI model, in particular the generality and the capability to competently perform a wide range of distinct tasks". Thereby, GPAIMs "may be placed on the market in various ways, including through libraries, application programming interfaces (APIs), as direct download, or as physical copy", and "may be further modified or fine-tuned into new models" (Rec. (97)). Examples for GPAIMs, are large generative AI models that "allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks" (Rec. (99)). This is the case for many multimodal large language models, such as GPT-4 Omni (GPT-4o) [22], or Gemini [23].

GPAIM with Systemic Risk. The systemic risk of a GPAIM is understood as "a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain" (Art. 3(65) AIA). Rec. (110) provides as a non-exhaustive list of examples for systemic risks: "any actual or reasonably foreseeable negative effects in relation to major accidents, disruptions of critical sectors and serious consequences to public health and safety; any actual or reasonably foreseeable negative effects on democratic processes, public and economic security; the dissemination of illegal, false, or discriminatory content". Thereby "[s]ystemic risks should be understood to increase with model capabilities and model reach, can arise along the entire lifecycle of the model, and are influenced by conditions of misuse, model reliability, model fairness and model security, the level of autonomy of the model, its access to tools, novel or combined modalities, release and distribution strategies, the potential to remove guardrails and other factors" (Rec. (110)).

According to Art. 51(1)(a) AIA, a GPAIM is considered posing a systemic risk, if "it has high impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks". High-impact capabilities are understood as "capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models" (Art. 3(64) AIA). Thereby high impact capabilities of a GPAIM are to be assumed, "when the cumulative amount of computation used for its training measured in floating point operations is greater than 10^{25}" (Art. 51(2) AIA). Rec. (111) states that "according to the state of the art at the time of entry into force of this Regulation, the cumulative amount of computation used for the training of the general-purpose AI model measured in floating point operations is one of the relevant approximations for model capabilities". According to Art. 51(1)(b) AIA a GPAIM is also considered to pose a systemic risk if—"based on a decision of the Commission"—it is considered having capabilities or an impact equivalent to those set out in Art. 51(1)(a) AIA according to the criteria set out in Annex XIII. The seven criteria enlisted in Annex XIII include the number of model parameters, the quality or size of the data set, the amount of computation used for training the model, the input and output modalities of the model, the benchmarks and evaluations of capabilities of the model, whether it has a high impact on the internal market due to its reach, and the number of registered end-users.

[22] `https://openai.com/gpt-4o-contributions/`
[23] `https://gemini.google.com`

Importantly, according to Art. 51(3) AIA, the EU Commission shall be able to amend the thresholds in Art. 51(1) and (2) AIA and add "benchmarks and indicators in light of evolving technological developments, such as algorithmic improvements or increased hardware efficiency, when necessary, for these thresholds to reflect the state of the art" (see also Rec. (111)).

C Lifecycle in Art. 15(1) AIA

The exact timeframe during which consistent performance must be ensured following Art. 15(1) AIA is unclear. Particularly, the term 'lifecycle' is not defined, which leaves open whether it differs from the term 'lifetime' used in Art. 12(1) AIA and Rec. (71). It is crucial to clarify the exact timeframe during which consistent performance must be maintained. As mentioned above, the term 'lifecycle' is not defined and its distinction from the term 'lifetime' in Art. 12(1) AIA and Rec. (71) remains ambiguous. While 'lifecycle' and 'lifetime' could initially be interpreted as synonyms [99], 'lifetime' might refer specifically to the active operational period of the AI system [100], whereas 'lifecycle' could encompass a broader view of all phases from product design and development to decommissioning [101]. If this broader interpretation of 'lifecycle' is intended, it raises questions about how accuracy, robustness, and cybersecurity should be ensured beyond the operational phase (e.g., during development), and why this would be necessary when there are no immediate risks to health, safety, and fundamental rights. One explanation for using the term 'lifecycle' would be that the EU legislator intended to emphasize that the requirements of Art. 15 AIA should not only be assessed when the system is ready for deployment but also during the design process. Accordingly, technical standards should define both terms.

D Organizational Measures for Cybersecurity

Numerous EU regulations related to cybersecurity (see e.g., Art. 32 General Data Protection Regulation[24], Art. 21 NIS 2 Directive[25]) explicitly mandate both technical and organizational measures to ensure cybersecurity. In the AIA, organizational measures are not explicitly mandated for cybersecurity (Art. 15(5) AIA). Rather, Art. 15(5) AIA AIA only focuses on technical solutions for providers of HRAIS. The omission of organizational measures in Art. 15(5) AIA has been criticized in the literature accompanying the legislative process of the AIA [96]. It is unclear whether providers are still implicitly required to implement organizational measures for cybersecurity (in accordance with other EU regulations), or if it implies that such measures are not mandatory. Interestingly, organizational measures are explicitly mandated for the robustness of HRAIS in Art. 15(4) AIA. However, this ambiguity in the provisions regarding requirements for *providers* of HRAIS (Art. 15 AIA) is mitigated by the fact that *deployers* of HRAIS are required to implement both organizational and technical measures to ensure the proper use of the system in accordance with the instructions for use (Art. 26(1) AIA). These instructions include the cybersecurity measures put in place.

E Relevant Circumstance Art. 15(5)(ii) AIA

The term 'relevant circumstance' is not defined in the AIA and therefore requires interpretation. On the one hand, one could argue that the term only refers to circumstances that are "important" for a "particular purpose" or context [102]. On the other hand, the meaning of the term can also result from a comparison with other provisions of the AIA such as Art. 13(3)(b)(ii) AIA, which suggests a different understanding. The provision demands that the instructions for the use of AI systems shall contain "any known and foreseeable circumstances" that may have an impact on cybersecurity. This speaks for a broader understanding of relevance, which only excludes unknown and unforeseeable circumstances. Given this ambiguity, standards should elaborate on how to determine relevant circumstances.

F Robustness-Accuracy Trade-Off

The ML literature has found that robustness and accuracy of ML models can in some scenarios be empirically and theoretically mutually inhibiting [103–107], and this relationship remains an active

[24]EU Regulation 2016/679, OJ L 119, 4.5.2016.
[25]EU Directive 2022/2555, OJ L 333/80.

area of research [108]. As stated in Section 4.1, an AI system may also incorporate other technical components (beyond models) that must be robust and may impact the overall system's accuracy. The AIA acknowledges these trade-offs but does not offer specific guidelines on how to achieve this balance. It requires providers of HRAIS to ensure an appropriate level of both accuracy and robustness, and to find "an appropriate balance in implementing the measures to fulfil" the AIA requirements (Art. 9(4) AIA). Additionally, the technical documentation must include "decisions about any possible trade-off made regarding the technical solutions adopted to comply with the requirements" of the AIA (Annex IV Nr. 2 lit. b AIA). Standards may offer guidance on the processes and metrics to use but will not prescribe a specific balance, such as a 40-60 ratio. As a result, providers of HRAIS will ultimately need to navigate these complex trade-offs themselves, adjusting individual model parameters to find an appropriate balance and justify and document their decisions.