Robustness and Cybersecurity in the EU Artificial Intelligence Act

Henrik Nolte [*1] **Miriam Rateike** [*2] **Michèle Finck** [1]

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 in provisions related to robustness and cybersecurity for high-risk AI systems (Art. 15 AIA) and general-purpose AI models (Art. 55 AIA). We demonstrate 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 identifies shortcomings in the relevant provisions, 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 in relation to the AIA.

1. Introduction

The European Union (EU) recently adopted the Artificial Intelligence Act (AIA)[1] 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, inter alia, imposes some of these as legally binding requirements for high-risk AI systems (HRAIS) and general-purpose AI models (GPAIMs). While the AIA is recognized as being one of the first legally binding regulatory frameworks for AI (Chee & Hummel, 2024), it has faced criticism for its imprecise and incoherent terminology (Laux et al., 2024; Bomhard & Siglmüller, 2024), which will complicate its practical implementation. Previous work has examined the AIA and its legislative history to clarify terms like explainability (Bordt et al., 2022; Vitali, 2022; Pavlidis, 2024) and fairness (Deck et al., 2024). So far, little attention has been paid to other relevant terms such as robustness and cybersecurity.

This paper focuses on requirements for HRAIS, which are the only types of AI systems mandated to meet robustness and cybersecurity requirements set out in Art. 15 AIA. The structure of this provision indicates that Art. 15(4) AIA specifies robustness, while Art. 15(5) AIA specifies cybersecurity. Art. 15(4) AIA mandates resilience "regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems". Art. 15(5) AIA further requires resilience "against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities". Accordingly, under the AIA, robustness and cybersecurity[2] refer to different but related concepts. Both are concerned with the resilience of AI systems, which must have the ability to withstand performance disruptions. Robustness refers to the ability of a HRAIS to remain resilient against errors from internal malfunctions or from its interactions with the environment they operate in.[3] Cybersecurity instead focuses on resilience against attacks from unauthorized third parties.

To better understand these requirements, we also contrast them to similar requirements for GPAIMs with systemic risk, which must "ensure an adequate level of cybersecurity protection" (Art. 55(1)(d) AIA). We analyze how cybersecurity requirements for GPAIMs with systemic risk differ from or align with those for HRAIS. Notably, the AIA does not contain any robustness requirements for GPAIMs. However, evidence from ML research suggests that robustness is also relevant for machine learning (ML) models that qualify as GPAIMs under the AIA (Yuan et al., 2023; Chen et al., 2022).

[*]Equal contribution [1]University of Tübingen, Tübingen, Germany [2]Saarland University, Saarbrücken, Germany. Correspondence to: Henrik Nolte <henrik.nolte@uni-tuebingen.de>, Miriam Rateike <rateike@cs.uni-saarland.de>.

Generative AI and Law (GenLaw '24) Workshop at 41st International Conference on Machine Learning, Vienna, Austria. 2024.

[1]EU Regulation 2024/1689, 12.7.2025.

[2]The term cybersecurity is defined in the earlier dated EU Cybersecurity Act (Regulation (EU) 2019/881, OJ L 151, 7.6.2019).

[3]The environment can be a real-world setting, like the physical surroundings of a robot, or a virtual one, such as a simulation (James et al., 2020; Mahmood et al., 2018).

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 to ensure compliance with the AIA. However, 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.[4] They prescribe technical requirements, including characteristics such as quality or performance levels, terminology, and test methods.[5] Standards have long been integral to EU product legislation under the New Legislative Framework, upon which the AIA is built (Gorywoda, 2009). 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). [6]

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 analyze these findings in relation to recent advancements in ML technology. This aims to inform the standardization process as well as the benchmark and measurement methodologies referred to in Art. 15(2) AIA.
- We aim to inform ML research about the legal requirements for robustness and cybersecurity to 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 the rationale behind 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.

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

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

[6] 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.

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 (Schwinn et al., 2022). This issue is explored across various applications such as computer vision (Drenkow et al., 2021; Taori et al., 2020; Dong et al., 2020) and natural language processing (La Malfa & Kwiatkowska, 2022; Chang et al., 2021). 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 (Cheng et al., 2024; Tocchetti et al., 2024). Perturbations of input (test) data often present a significant 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.

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 invisible to the human eye, with the intent of altering a model's output (Szegedy et al., 2013). For instance, minor pixel perturbations in an image can lead to significant changes in model output (Szegedy et al., 2013). 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 (Nicolae et al., 2018; Chen et al., 2017b).

Non-adversarial (or *natural*) *robustness* often adresses changes in ML model outputs due to *distribution shifts* in input data (Gojić et al., 2023; Tocchetti et al., 2024). These changes occur when the distribution from which the test data is sampled differs from that of the training data (Taori et al., 2020; Drenkow et al., 2021). For instance, alterations in data collection methods, such as upgrading to a new X-ray machine, can modify the format or presentation of X-ray images (Glocker et al., 2019; Castro et al., 2020). 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 (D'Amour et al., 2020; Zhang et al., 2020). 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 (Perdomo et al., 2020). Other forms of research on *non-adversarial robustness* in-

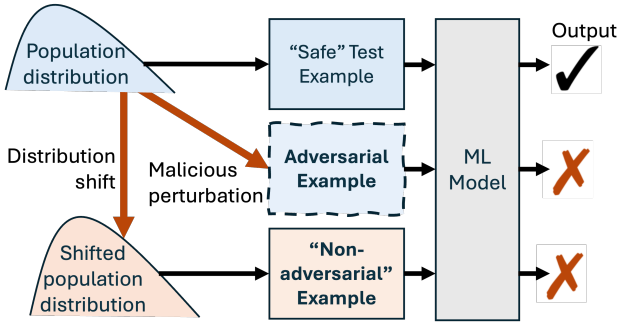

Figure 1: Examples of key robustness problems. Model outputs may be as expected (✓) with "safe" test data from the original distribution; unintended changes (✗) can occur with adversarial or non-adversarial (shifted) inputs.

vestigates the robustness of ML models to noise, which frequently occurs in real-world data sets (Sáez et al., 2016; Olmin & Lindsten, 2022).

As discussed in Section 1, this paper examines the terms robustness and cybersecurity in the AIA. 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 (Dasgupta et al., 2022). This encompasses aspects like data storage, information access and modification, and secure data transmission over networks (Sarker et al., 2021). Unlike robustness, cybersecurity is not a stand-alone concept in ML,[7] but is discussed more broadly as both a tool for ensuring cybersecurity and a potential source of cybersecurity risks. ML algorithms can be employed to detect and mitigate cybersecurity threats (Sarker et al., 2021), but can also introduce specific vulnerabilities that adversaries may exploit, such as data poisoning or adversarial attacks (Roshanaei et al., 2024; Rosenberg et al., 2021). *Adversarial robustness* specifically studies attacks that use manipulated input data to alter the performance of an ML model.

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 internal market and prevent regulatory fragmentation between member states. 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 (Sioli, 2021; Bomhard & Siglmüller, 2024): Those posing unacceptable risks, such as social scoring, are prohibited or subject to qualified prohibitions; high-risk AI systems (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.[8] According to Art. 16(a) AIA, providers of HRAIS must ensure compliance with these requirements. To support the implementation of these requirements, Art. 15(2) AIA mandates that the EU Commission "shall, in cooperation with relevant stakeholders and organisations [...], encourage, as appropriate, the development of benchmarks and measurement methodologies" to "address the technical aspects of how to measure the appropriate levels of accuracy, robustness and any other relevant performance metrics". 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.

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. As discussed in Section 1, Art. 15(4) AIA mandates that HRAIS exhibit resilience "regarding errors, faults or inconsistencies that may occur within the 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". The architecture and rationale behind the AIA, as well as its legal history (AI IHEG, 2019), suggests that one of legislator's main objectives was to foster trust in AI (see Art. 1 AIA). Consequently, Art. 15 AIA should be interpreted and implemented in light of its purpose to promote widespread societal adoption of trustworthy AI systems and enhancing the competitiveness in the EU market (see Art. 1 AIA).

[7] For example, at the top-tier ML conference Advances in Neural Information Processing Systems in 2023, 'robust' appeared in around 170 paper titles, whereas 'cybersecurity' did not appear in any. (Oh et al., 2023).

[8] The term robust is also used in parts of the AIA that do not concern HRAIS and is used in a different context (Rec. (8) and Rec. (81)).

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 begin by identifying four legal challenges related to Art. 15 AIA: First, a clear delineation of the legal terms of robustness and cybersecurity and its counterparts in ML literature is missing. Second, the ML literature predominantly focuses on ML models, while the AIA mandates compliance for the entire AI system. This discrepancy might create practical challenges in implementing the AIA. For instance, while the robustness of a HRAIS could be ensured by a robust AI model, it is unclear whether this alone suffices or if other components must also meet robustness requirements. In an autonomous vehicle, even if the AI model is robust, the robustness of the overall system could still be compromised if a camera system fails to produce images with sufficient contrast in certain lighting conditions.

Third, while accuracy is specified as a requirement in Art. 15 AIA, the provision does not clarify its role in measuring the robustness and cybersecurity of AI systems. Fourth, the terms 'lifecycle' and 'consistent' performance are not defined, leaving ambiguity as to how such a performance can be ensured in practice.

Robustness and Cybersecurity. In this section, we address two issues: First, the AIA lacks legal definitions for robustness and cybersecurity in the AIA. While robustness is a new term in EU legislation, a definition of cybersecurity can be found in the CSA. Second, implementing Art. 15 AIA requires technical solutions from the ML domain. However, concepts and terms often differ between domains. To address these issues, we: i) provide a legal interpretation of both terms; and ii) determine how these terms should be understood in the ML domain.

The robustness and cybersecurity requirements in the AIA both stem from the principle of 'technical robustness and safety' introduced in the 2019 Ethics Guidelines for Trustworthy AI (AI IHEG, 2019). Given their shared origin, we deem it crucial to explore the similarities and differences between these two requirements to gain a clearer understanding of both terms. While Art. 15(1) AIA mentions both terms, they are elaborated separately in subsequent provisions: robustness in Art. 15(4) AIA and cybersecurity in Art. 15(5) AIA. The robustness requirement in Art. 15(4) AIA and its corresponding Rec. (75) address "errors, faults, or inconsistencies" that may inadvertently occur as the system interacts with its real-world environment. In contrast, the cybersecurity requirements in Art. 15(5) AIA and Rec. (76) focus on deliberate attempts "to alter the use, outputs, or performance" of an AI system "by malicious third parties exploiting the system's vulnerabilities".

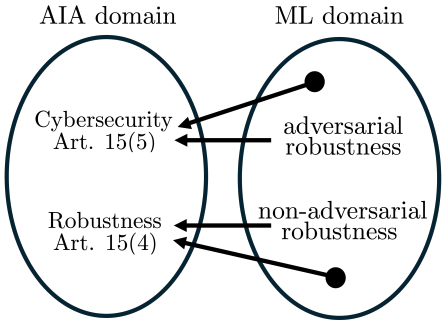

Figure 2: Technical solutions to cybersecurity (Art. 15(5) AIA) can be found, inter alia, in ML research on *adversarial robustness*, and technical solutions to robustness (Art. 15(4) AIA) can be found, inter alia, in the ML research on *non-adversarial robustness*.

Notably, cybersecurity is explicitly defined in Art. 2(1) CSA as "the activities necessary to protect network and information systems, the users of such systems, and other persons affected by cyber threats". Although the AIA does not directly reference this definition, Art. 42(2) AIA states that HRAIS with certification or a conformity declaration in accordance with the CSA are considered compliant with the cybersecurity requirements of Art. 15 AIA.[9] As both certification and conformity declarations are based on the definition in Art. 2(1) CSA and can cover the requirements of Art. 15 AIA, this suggests that the definition in Art. 2 CSA can be applied to the AIA as well.

The objective of the robustness and cybersecurity requirements is to ensure that HRAIS function properly and are resilient against any factors that might compromise consistent performance. However, they differ in the type of cause that can affect consistent performance. Robustness pertains to unintentional behavior, whereas cybersecurity focuses on malicious acts by third parties. Thus, while both aim to ensure consistent performance, the protections they mandate differ. Robustness requires protecting HRAIS against unintentional causes that could compromise performance, while cybersecurity mandates mitigating intentional causes that compromise the performance of a HRAIS.

We now explore how these legal terms could be understood within the ML domain (see Figure 2 for a visualization). By doing so, we aim to facilitate the implementation of the AIA and inform ML research. As explained above, robustness

[9] 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.

as a legal term refers to the resilience of a HRAIS against unintentional causes that might compromise its consistent performance. In the context of ML, robustness refers to mitigating undesired changes in model outputs when deploying models in real-world scenarios (see section 2). While both the legal and the ML understanding of robustness focus on maintaining consistent performance, ML research further distinguishes between *non-adversarial robustness* and *adversarial robustness*. *Non-adversarial robustness* refers to to a model's ability to maintain performance under data distribution shifts or noisy data- unwanted causes that can affect the consistent performance of a model. The legal term of robustness aligns with the ML literature's concept of *non-adversarial robustness*. However, the ML literature also addresses *adversarial robustness*, which is the ability of a model to resist intentionally perturbed inputs aimed at altering its predictions. This aspect of robustness, however, is not reflected in the legal understanding of robustness but is associated with the legal term cybersecurity. Consequently, the legal term robustness and its understanding in the ML literature only partially overlap.

Cybersecurity as mandated in Art. 15(1) AIA focuses on the protection against malicious attempts to alter the consistent performance of a HRAIS. However, as outlined above, cybersecurity is not a term frequently used in the ML domain. Instead, the ML literature refers to specific concepts and attacks on AI models that are potential sources of cybersecurity risks. Deliberate adversarial causes for data perturbations are studied under the concept of *adversarial robustness* and are also reflected in the cybersecurity requirements in Art. 15(5) AIA. For instance, the concept of model evasion, which is typically studied in the ML literature on *adversarial robustness*, is explicitly mentioned in Art. 15(5) AIA. Therefore, the legal term of cybersecurity encompasses aspects of the ML literature on *adversarial robustness*.

System vs. Model. Art. 15 AIA applies to *HRAIS*. However, technical solutions for robustness and cybersecurity in the ML domain typically focus on *ML models*. This raises the question of whether solely relying on technical solutions for *ML models* is enough to ensure the compliance of a HRAIS with Art. 15 AIA—or whether additional measures are needed. To answer this, it is first necessary to clarify the general relationship between AI models and AI systems.

The AIA regulates AI systems and not AI models, with the only exception being GPAIMs. Rec. (97) specifies that an AI model is an essential component of an AI system. 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. This relationship is also emphasized in a report by the Joint Research Centre of the EU Commission (Junklewitz et al., 2023), according to which an AI system comprises one or more AI models and additional components. These components can include, inter alia, user interfaces, sensors, databases, network communication components, or pre- and post-processing mechanisms for model in- and outputs (Rec. (97), Junklewitz et al. (2023)).

Art. 15(4)(ii) AIA states that robustness may be ensured through technical redundancy solutions, including "back-up or contingency plans". This implies that multiple individual components should contribute to the overall robustness of the AI system, particularly in scenarios where some components may fail. 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. 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 measures. For instance, if a measure designed to enhance sensor robustness alters the sensor's outputs, it may necessitate retraining the AI model.

Role of Accuracy. Having outlined the relationship between robustness and cybersecurity above, we now turn to the role of accuracy in measuring these attributes. Art. 15(1) AIA mandates that HRAIS shall "achieve an appropriate level of accuracy". First, we show that accuracy in the AIA seamlessly corresponds to *accuracy* in the ML domain. Second, we highlight that *accuracy* plays a critical role in *robustness*, as *robustness* in the ML literature is often measured using *accuracy* metrics on a robustness test dataset, and selecting certain favorable accuracy metrics may make an ML model seem more robust compared to other accuracy metrics. Third, we argue that there can be trade-offs between *robustness* and *accuracy*.

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 the

[10]This links to fairness ML literature on the possible divergence of error rates for different sensitive groups. (Mitchell et al., 2021; Chouldechova & G'Sell, 2017)

ML literature, the term *accuracy* is used both as a metric and as an objective. As a metric, *accuracy* typically describes the overall proportion of correct predictions out of the total number of predictions made (Carvalho et al., 2019). As an objective, *accuracy* describes "how well" the AI system performs given its specific purpose, and can consequently be measured with different metrics, such as utility (Corbett-Davies et al., 2017) and f1-score (Sokolova et al., 2006). 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). Given these two uses of the term, the question is how accuracy is understood in the AIA. 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. It remains up to technical standards to clarify how AI systems' accuracy is to be defined and measured.

In the ML literature, robustness is often measured using *accuracy* as a 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) (Taori et al., 2020; Hendrycks et al., 2021; Goodfellow et al., 2015). A small difference between these two *accuracy* results indicates greater (i.e., better) *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. The selection of favorable metrics has been studied in the fairness literature under the term fairness hacking (Meding & Hagendorff, 2024; Simson et al., 2024; Black et al., 2024). 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.

Lastly, without entering the debate, we note that there is an ongoing discussion in the ML literature regarding the existence and characteristics of a potential trade-off between *robustness* and *accuracy*. While some studies suggest that enhancing *robustness* can lead to a drop in *test accuracy* (Zhang et al., 2019; Rade & Moosavi-Dezfooli, 2022; Tsipras et al., 2019), other research argues that *robustness* and *accuracy* are not inherently conflicting goals and can be achieved concurrently (Yang et al., 2020; Raghunathan et al., 2020). These trade-offs are not addressed by the AIA; leading providers to make the normative choice of which objective to give preference.

Consistent Performance Throughout the Lifecycle. AI systems must perform "consistently" in terms of accuracy, robustness, and cybersecurity "throughout their lifecycle" (Art. 15(1) AIA). This presents two challenges: i) the requirement of 'consistent' performance remains ambiguous, as there is no specification on how it should be measured; ii) the exact timeframe during which consistent performance must be ensured 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).

First, the meaning of 'consistent' and its measurement remains undefined. 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 (Kilbertus et al., 2020; Bechavod et al., 2019; Rateike et al., 2022). In practice, performance can vary due to different factors, such as random initializations of weights or input data sampling. These types of variations are unavoidable. 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.[11] It has yet to be defined which maximum value of variance would be considered consistent. 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.

Second, 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 (Marcus, 2020), 'lifetime' might refer specifically to the active operational period of the AI system (Murakami et al., 2010), whereas 'lifecycle' could encompass a broader view of all phases from product design and development to decommissioning (Hamon et al., 2024). 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.

[11] Note that some also consider consistency as a metric itself (which may be in trade-off with robustness), rather than as a property of a (robustness) metric (Wei & Zhang, 2020).

4.2. Robustness Art. 15(4) AIA

We now turn to challenges specific to 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. Lastly, Art. 15(4)(iii) AIA requires addressing feedback loops in online learning with possibly biased outputs.

Incoherent 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. 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 (AI IHEG, 2019) developed by the AI IHEG, which introduced the principle of 'technical robustness and safety' (Section 4.1). 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).

On this basis one could argue that technical robustness is synonymous with robustness. 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. The AIA is ambiguous regarding the required level of robustness. Art. 15(1) AIA mandates that AI systems must achieve an "appropriate level" of robustness. On the other hand, Art. 15(4) AIA 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. Regarding robustness, the latter provision clarifies that "appropriate" means "as resilient as possible". Pursuant to Art. 8(1) AIA, the intended purpose of the system and the generally acknowledged state of the art (STOA) on AI and AI-related technologies must be taken into account when determining the appropriate level of robustness of a specific HRAIS. 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 further to be understood as 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 STOA while balancing it with other requirements.

Feedback Loops. Art. 15(4)(iii) AIA specifies robustness measures to address the risks introduced by feedback loops, stating 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 (Pagan et al., 2023), with the most common issues being a distribution shift (Perdomo et al., 2020) or a selection bias (Lum & Isaac, 2016; Kilbertus et al., 2020). Importantly, in this context the risk of "biased outputs" in feedback loops (Art. 15(4)(iii) AIA) is rather studied in the literature on fairness in ML than in the literature on *robustness* in ML, which traditionally constitute different research fields and communities (Lee et al., 2021). Whether there is a trade-off between *robustness* and fairness, or if both pursue similar goals, remains an active discussion in the ML community (Xu et al., 2021; Pruksachatkun et al., 2021; Lee et al., 2021).

An important aspect of Art. 15(4)(iii) AIA is that it applies specifically to AI systems that learn online. In the ML literature it is common to distinguish between online and offline learning. Online learning ML models iteratively learn from a sequence of data and continuously update their parameters over time (Hoi et al., 2021). This adaptiveness is reflected in Art. 3(1) AIA as a factual characteristic of an AI system. Conversely, offline models are trained on a fixed dataset all at once (Hoi et al., 2021).

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 (Zadrozny, 2004; Liu & Ziebart, 2014). This can distort model predictions and reinforce existing biases, ultimately impacting the model's accuracy and fairness (Kilbertus et al., 2020; Bechavod et al., 2019; Rateike et al., 2022). Offline systems, however, 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 (Liu et al., 2018;

D'Amour et al., 2020; Zhang et al., 2020). Since an offline ML model is not updated, such a distribution shift can influence their performance over time and possibly lead to fairness concerns (Liu et al., 2018). Although Art. 15(4) AIA) does not explicitly address feedback loops in offline systems, HRAIS are not exempt from addressing them. Since feedback loops can impact the model's consistent accuracy performance, 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. Notably, the AIA provides an additional pathway to demonstrate compliance with its cybersecurity requirements (Casarosa, 2022). Art. 42(2) AIA explicitly states that HRAIS certified under the EU Cybersecurity Act (CSA) [12] "shall be presumed to be in compliance with the cybersecurity requirements" outlined in Art. 15 AIA. Consequently, our findings and interpretations in Art. 15(5) AIA offer insights that support not only the development of harmonized standards but also for the potential creation of EU cybersecurity certification schemes for AI systems.

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: i) what constitutes a 'relevant circumstance'; and ii) 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.

First, we analyze the term 'relevant circumstance'. On the one hand, one could argue that the term only refers to circumstances that are "important" for a "particular purpose" or context (Cambridge University Press, 2024c). 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.

Second, 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, particularly highlighted in the above-mentioned arms race. 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 (Xie et al., 2023; Kumar et al., 2023). Therefore, an appropriate level of cybersecurity should therefore be understood as a requirement for a sufficient defense.

The CSA defines cybersecurity but focuses primarily on technical methodologies for testing it, rather than specifying appropriate levels of cybersecurity. The CSA certificate itself does not guarantee that the certified level of cybersecurity will be always be deemed sufficient in a risk analysis (Kipker et al., 2023). Nonetheless, as indicated above, under specific circumstances compliance with Art. 15(5)(ii) AIA is assumed for AI systems certified under the CSA under specific circumstances.

It remains to define what 'appropriate to the relevant risks' means. As outlined in Section 4.2, the appropriateness of a certain performance level must consider the intended purpose of the system and the generally acknowledged STOA (see Art. 8(1) AIA). The measures to ensure cybersecurity adopted under the risk management system are not expected to eliminate all existing risks, but the overall residual risk must be acceptable (see Art. 9(1) and (4) AIA). When determining the appropriateness of technical solutions, the known and foreseeable risks following their intended purpose (Art. 9(2)(a) AIA) and risks of reasonably foreseeable misuse (Art. 9(2)(b) AIA) must be taken into account. As demonstrated above, the risk management system is to be understood as a continuously and iteratively process. Furthermore, Art. 9(4) AIA shows that the process of determining appropriateness typically involves balancing various requirements. This means that the risk management system mandates the identification of risks to health, safety, and fundamental rights and the cybersecurity requirements in Art. 15(5) AIA are intended to address these risks.

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

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" (CVE, 2024). We focus on: i) identifying components that are susceptible to 'AI-specific' vulnerabilities; and ii) the distinction between 'system' and 'AI-specific' vulnerabilities.

First, 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 these vulnerabilities. We suggest performing a hypothetical test. Consider the central role of AI models 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 AI systems regulated under the AIA. It has been argued that the central characteristic of an AI system is its ability to infer from input to output (Hacker, 2024). This inference ability is typically 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.[13]

Second, 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 the terms 'AI-specific vulnerabilities' and 'system vulnerabilities' 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. The legal terms data poisoning, model poisoning, adversarial examples, model evasion, and confidentiality attacks are well-established in the ML literature, whereas 'model flaws' remains a vague term.

In ML research, the aforementioned attacks aim to induce model failures (Vassilev et al., 2024): *Data poisoning* attacks manipulate training data (Schwarzschild et al., 2021), *model poisoning* attacks manipulate the trained ML model (Zhang et al., 2022), and *model evasion* attacks manipulate test samples (Biggio et al., 2013). *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 (Rigaki & Garcia, 2023).

In addition to these attacks, Art. 15(5)(iii) AIA lists 'model flaws' as an AI-specific vulnerability. This term, however, 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 (Kumar & Anderson, 2023; Nissenbaum, 1996). 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 should are only expected to address traditional 'bugs' or coding errors, or whether they should address other ways of exploiting AI-specific vulnerabilities that should be addressed.

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 (Chen et al., 2017a). Consequently, it is impractical to anticipate and counter all potential attacks targeting AI-specific vulnerabilities.

Organizational Measures. Numerous EU regulations related to cybersecurity (see e.g., Art. 32 General Data Protection Regulation[14], Art. 21 NIS 2 Directive[15]) explicitly mandate both technical and organizational measures to ensure cybersecurity. While Art. 15(5)(ii) and (iii) AIA provide more details on technical solutions, they do not explicitly state that these are the sole measures required for cybersecurity. This omission raises the question of whether it is an obligation to take organizational measures to ensure cybersecurity because such measures are inherently included in the term cybersecurity, or whether their absence implies that they are not obligatory. The omission of organizational measures to fulfill cybersecurity goals has been criticized in the literature accompanying the legislative process of the AIA (Biasin et al., 2023). Interestingly, organizational measures are explicitly mandated for the robustness of HRAIS in Art. 15(4)(i) AIA.

[13] A similar idea in a different context can be found the ethics guidelines (AI IHEG, 2019), which suggest that fallback plans in case of problems can foresee AI systems switching from a statistical (i.e., ML) to a rule-based or human-in-the-loop approach.

[14] EU Regulation 2016/679, OJ L 119, 4.5.2016.

[15] EU Directive 2022/2555, OJ L 333/80.

5. Requirements for General-Purpose AI Models With Systemic Risk

The AIA establishes legal requirements for a particular category of AI models, namely so-called general-purpose AI models (GPAIM). GPAIM are AI models that can perform tasks that they were not originally trained for (Gutierrez et al., 2023), such as large language models (OpenAI, 2023; Gemini Team et al., 2023), or large text-to-image models (Ramesh et al., 2022). GPAIM can either be "provided as a standalone model" or be "embedded in an AI system" (Rec. (114)). If a GPAIM is embedded in an HRAIS, a provider needs to adhere both to the legal requirements for both GPAIM and HRAIS simultaneously. The AIA differentiates between GPAIMs with systemic risk, and those that do not present such risks. It is important to note that other types of AI models are not subject to the AIA. These are models that are created with a specific objective and can only accomplish tasks they are trained to perform (e.g., translation, classification).

In the previous section, we examined requirements for HRAIS. To further elucidate these requirements, we now focus on GPAIMs with systemic risk and highlight the similarities and differences between them, as GPAIMs without systemic risks do not need to fulfill any robustness and cybersecurity obligations (see Art. 53 AIA ff.). The term 'systemic risk' is defined in Art. 3(65) AIA 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.[16] A detailed analysis of this definition is beyond the scope of this paper; for an in-depth discussion, we refer the interested reader to the existing literature (Novelli et al., 2024; Hacker, 2024; Pehlivan, 2024; Kutterer, 2023).

In the following, we will focus on two aspects: i) Contrary to HRAIS, the AIA does not impose any robustness requirements on GPAIMs with systemic risk although it does impose cybersecurity requirements; and ii) it remains unclear whether the required level of cybersecurity for GPAIMs with systemic risk is the same as for HRAIS.

Cybersecurity Requirements. Art. 55(1)(d) AIA mandates "an adequate level of cybersecurity protection" for GPAIMs with systemic risk. Rec. (115) details the cybersecurity requirement for GPAIMs with systemic risk set out in Art. 55(1)(d) AIA. 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 (Yao et al., 2024; Li et al., 2023; Wang et al., 2023). Although Art. 55(1)(d) AIA does not define the term 'cyberattacks', we infer that it includes the attacks exploiting AI-specific vulnerabilities mentioned in Art. 15 AIA (see Section 4.3). These attacks are studied in the field of *adversarial robustness* and can also affect GPAIMs (Qiang et al., 2024; Das et al., 2024; Yan et al., 2024; Schwinn et al., 2024; Vitorino et al., 2024)—even though specific ML techniques may be necessary to address GPAIM-specific challenges. This correlation underscores that the concepts and problems explored under *adversarial robustness* are reflected in the term 'cybersecurity' as used in Art. 55(1)(d) AIA.

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 he AIA, particularly regarding GPAIMs, which were not addressed in the initial draft of the regulation but gathered widespread media attention during the legislative procedure. Nevertheless, evidence from ML research suggests that *non-adversarial robustness* is also relevant for GPAIMs (Yuan et al., 2023; Chen et al., 2022).

Required Level of Cybersecurity. Art. 55(1)(d) AIA mandates an 'adequate' level of cybersecurity protection for GPAI models with systemic risks. This requirement contrasts with the 'appropriate' level of cybersecurity mandated for HRAIS under Art. 15(1) AIA. The distinction between these two terms raises questions about their equivalence and the extent of their differences.

On the one hand, they could imply different levels of cybersecurity. The Cambridge Dictionary defines the term 'adequate' as "enough or satisfactory for a particular purpose" (Cambridge University Press, 2024a) and 'appropriate' as "suitable or right for a particular situation or occasion" (Cambridge University Press, 2024b). According to these definitions, "appropriate" mandates a higher level than "adequate". 'Adequate' mandates surpassing a minimum threshold that is "enough", whereas 'appropriate' mandates ensuring a specific "right" level that can be considered above

[16] 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.

the bare minimum. Mandating an 'adequate' level of cybersecurity aligns with the nature of stand-alone GPAI models with systemic risk. 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 a minimum level of cybersecurity. Conversely, HRAIS including those with GPAI models as components, 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, Rec. (115) introduces ambiguity by stating 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 that they may be intended to be 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 (such as FR "approprié", ES "adecuado", GER "angemessen", IT "adeguato"). To resolve this ambiguity, technical standards should clarify the required level of cybersecurity for GPAIMs with systemic risk.

6. Summary and Outlook

We have identified several legal challenges and potential limitations regarding the practical implementation of robustness and cybersecurity requirements for HRAIS as stipulated in Art. 15(4) and (5) AIA. To elucidate these requirements, we also examined GPAIMs with systemic risk, which are subject to cybersecurity requirements, but are not mandated to meet specific robustness requirements. Therefore, we examined the cybersecurity requirements for GPAIMs with systemic risk and identified additional legal challenges.

Our analysis shows that these provisions are vague and require further specifications, such as through harmonized standards or the benchmark and measuremenet methodologies foreseen by Article 15(2) in relation to robustness. This could: i) further specify what the loosely defined concepts of robustness and cybersecurity require from a technical perspective; ii) establish the required levels of robustness and cybersecurity as well as of other related concepts such as consistency; and iii) define the requirements for evaluating and assessing AI systems and their components.

Our analysis also aimed to inform ML research in the field of robustness and cybersecurity about the legislative changes introduced by the AIA. For HRAIS, we demonstrated that the concept of *non-adversarial robustness* in the ML literature is reflected in the term robustness used in Art. 15(1) and (4) AIA, while *adversarial robustness* is reflected in the term cybersecurity used in Art. 15(1) and (5) AIA and Art. 55(1)(d) AIA., and *accuracy* in the ML literature aligns with the term accuracy used in Art. 15(1) AIA. We found that in the ML domain, *robustness* is often assessed by the drop in accuracy under non-adversarial or adversarial conditions. Consequently, the concept of accuracy may play a crucial role in measuring both robustness and cybersecurity. While we have attributed the omission of robustness requirements for GPAIMs with systemic risk to the policy process, the ML literature indicates that large generative models, which may be classified as GPAIM, also face robustness challenges.

Future work should provide an in-depth overview of the ML literature on *non-adversarial robustness* for large generative AI to identify research gaps in this area. Moreover, there is still limited legal literature on the robustness and cybersecurity requirements in the AIA (Casarosa, 2024; Ludvigsen et al., 2022) and their relationship to other legal frameworks, such as the Medical Device Regulation (Biasin et al., 2023; Nolte & Schreitmüller, 2024) or the Machinery Regulation. Research should explore these intersections to ensure coherent product safety standards.

Additionally, several challenges arise within the ML domain. First, while the AIA regulates AI systems, ML research often focuses on models. This highlights the need for interdisciplinary research addressing the robustness of entire AI systems. Second, there is limited literature on how the choice of accuracy metrics can affect robustness, and research should explore potential 'robustness hacking' effects. Third, the AIA emphasizes the severe consequences of unattended feedback loops and the importance of performance consistency over time. Questions remain about measuring and achieving the long-term stability of these metrics.

To bridge the gap between ML and legal terminology and support the implementation of the AIA, we recommend that scholars and practitioners from both domains strive for better mutual understanding and collaborate to address existing ambiguities.

Acknowledgements

Thank you to Tommaso Fia, and Sebastian Bordt for helpful feedback and comments. Special thank you to Marie-Sophie Müller for a thorough review of our text. Michèle Finck and Henrik Nolte are members of the Machine Learning Cluster of Excellence, EXC number 2064/1- Project number 390727645. Miriam Rateike is grateful for the generous funding support by the 2023 Google PhD Fellowship in Machine Learning.

References

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

Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, S. Z. Equal opportunity in online classification with partial feedback. *Advances in Neural Information Processing Systems*, 32, 2019.

Biasin, E., Yasar, B., and Kamenjasevic, E. 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.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. 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*, pp. 387–402. Springer, 2013.

Black, E., Gillis, T., and Hall, Z. Y. D-hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 602–615, 2024.

Bomhard, D. and Siglmüller, J. Ai act – das trilogergebnis. *RDi*, 45, 2024.

Bordt, S., Finck, M., Raidl, E., and von Luxburg, U. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 891–905, 2022.

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

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

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

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Casarosa, F. 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.

Casarosa, F. The risk of unreliable standards: Cybersecurity and the artificial intelligence act. *Internet Policy Review*, 2024. URL https://policyreview.info/articles/news/cybersecurity-and-artificial-intelligence-act/1742. Accessed: 2024-07-15.

Castro, D. C., Walker, I., and Glocker, B. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Chang, K.-W., He, H., Jia, R., and Singh, S. Robustness and adversarial examples in natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pp. 22–26, 2021.

Chee, F. Y. and Hummel, T. 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/.

Chen, L., Ye, Y., and Bourlai, T. Adversarial machine learning in malware detection: Arms race between evasion attack and defense. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pp. 99–106. IEEE, 2017a.

Chen, P.-Y., Liu, S., Paul, S., and Face, H. Foundational robustness of foundation models. In *Annual Conference on Neural Information Processing Systems*, 2022.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017b.

Cheng, L., Huang, X., Sang, J., and Yu, J. Towards robust recommendation: A review and an adversarial robustness evaluation library. *arXiv preprint arXiv:2404.17844*, 2024.

Chouldechova, A. and G'Sell, M. Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046*, 2017.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 797–806, 2017.

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

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020.

Das, A., Tariq, A., Batalini, F., Dhara, B., and Banerjee, I. Exposing vulnerabilities in clinical llms through data poisoning attacks: Case study in breast cancer. *medRxiv*, 2024.

Dasgupta, D., Akhtar, Z., and Sen, S. Machine learning in cybersecurity: a comprehensive survey. *The Journal of Defense Modeling and Simulation*, 19(1):57–106, 2022.

Deck, L., Müller, J.-L., Braun, C., Zipperling, D., and Kühl, N. Implications of the ai act for non-discrimination law and algorithmic fairness. *3rd European Workshop on Algorithmic Fairness*, 2024.

Dong, Y., Fu, Q.-A., Yang, X., Pang, T., Su, H., Xiao, Z., and Zhu, J. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 321–331, 2020.

Drenkow, N., Sani, N., Shpitser, I., and Unberath, M. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639*, 2021.

Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Glocker, B., Robinson, R., Castro, D. C., Dou, Q., and Konukoglu, E. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*, 2019.

Gojić, G., Vincan, V., Kundačina, O., Mišković, D., and Dragan, D. Non-adversarial robustness of deep learning methods for computer vision. In *2023 10th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN)*, pp. 1–9. IEEE, 2023.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *6th International Conference on Learning Representations*, 2015.

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

Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C. C., and Franklin, M. A proposal for a definition of general purpose artificial intelligence systems. *Digital Society*, 2(3): 36, 2023.

Hacker, P. Comments on the final trilogue version of the ai act. *Available at SSRN 4757603*, 2024.

Hamon, R., Junklewitz, H., Garrido, J. S., and Sanchez, I. Three challenges to secure ai systems in the context of ai regulations. *IEEE Access*, 2024.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., 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*, pp. 8340–8349, 2021.

Hoi, S. C., Sahoo, D., Lu, J., and Zhao, P. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

James, S., Ma, Z., Arrojo, D. R., and Davison, A. J. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Junklewitz, H., Hamon, R., André, A.-A., Evas, T., Soler Garrido, J., and Sanchez Martin, J. I. Cybersecurity of artificial intelligence in the ai act. Technical Report KJ-NA-31-643-EN-N (online), Joint Research Center, Luxembourg (Luxembourg), 2023.

Kilbertus, N., Rodriguez, M. G., Schölkopf, B., Muandet, K., and Valera, I. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 277–287. PMLR, 2020.

Kipker, D.-K., Reusch, P., Ritter, S., and Beucher, K. *Recht der Informationssicherheit: BSIG, EU Cybersecurity Act, DS-GVO. Kommentar*. Beck, 2023.

Klimas, T. and Vaiciukaite, J. The law of recitals in european community legislation'(2008). *ILSA Journal of International & Comparative Law*, 15:61, 2008.

Kumar, A., Agarwal, C., Srinivas, S., Feizi, S., and Lakkaraju, H. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.

Kumar, R. and Anderson, H. *Not with a Bug, But with a Sticker: Attacks on Machine Learning Systems and What To Do About Them*. John Wiley & Sons, 2023, 2023.

Kutterer, C. Regulating foundation models in the ai act: From "high" to "systemic" risk. *AI-Regulation Papers 24-01-1*, 2023.

La Malfa, E. and Kwiatkowska, M. The king is naked: On the notion of robustness for natural language processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11047–11057, 2022.

Laux, J., Wachter, S., and Mittelstadt, B. 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.

Lee, J.-G., Roh, Y., Song, H., and Whang, S. E. Machine learning robustness, fairness, and their convergence. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 4046–4047, 2021.

Li, H., Guo, D., Fan, W., Xu, M., Huang, J., Meng, F., and Song, Y. Multi-step jailbreaking privacy attacks on chatgpt. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Liu, A. and Ziebart, B. Robust classification under sample selection bias. *Advances in Neural Information Processing Systems*, 27, 2014.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Ludvigsen, K. R., Nagaraja, S., and Daly, A. The dangers of computational law and cybersecurity; perspectives from engineering and the ai act. *arXiv preprint arXiv:2207.00295*, 2022.

Lum, K. and Isaac, W. To predict and serve? *Significance*, 13(5):14–19, 2016.

Mahmood, A. R., Korenkevych, D., Komer, B. J., and Bergstra, J. Setting up a reinforcement learning task with a real-world robot. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4635–4640. IEEE, 2018.

Marcus, J. S. 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.

Meding, K. and Hagendorff, T. Fairness hacking: The malicious practice of shrouding unfairness in algorithms. *Philosophy & Technology*, 37(1):4, 2024.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and its Application*, 8(1):141–163, 2021.

Murakami, S., Oguchi, M., Tasaki, T., Daigo, I., and Hashimoto, S. Lifespan of commodities, part i: The creation of a database and its review. *Journal of Industrial Ecology*, 14(4):598–612, 2010.

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*, 2018.

Nissenbaum, H. Accountability in a computerized society. *Science and Engineering Ethics*, 2:25–42, 1996.

Nolte, H. and Schreitmüller, Z. Cybersicherheit ki-basierter medizinprodukte im lichte der mdr und ki-vo. *Zeitschrift für das gesamte Medizinprodukterecht*, 1:20, 2024.

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., and Floridi, L. Generative ai in eu law: liability, privacy, intellectual property, and cybersecurity. *arXiv preprint arXiv:2401.07348*, 2024.

Oh, A., Naumann, T., Globerson, A., K. Saenko, K., Hardt, M., and Levine, S. (eds.). *Advances in Neural Information Processing Systems 36*, 2023.

Olmin, A. and Lindsten, F. Robustness and reliability when training with noisy labels. In *International Conference on Artificial Intelligence and Statistics*, pp. 922–942. PMLR, 2022.

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

Pagan, N., Baumann, J., Elokda, E., De Pasquale, G., Bolognani, S., and Hannák, A. 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*, pp. 1–14, 2023.

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

Pehlivan, C. N. The eu artificial intelligence (ai) act: An introduction. *Global Privacy Law Review*, 5(1), 2024.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Pruksachatkun, Y., Krishna, S., Dhamala, J., Gupta, R., and Chang, K.-W. 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*, pp. 3320–3331, 2021.

Qiang, Y., Zhou, X., Zade, S. Z., Roshani, M. A., Zytko, D., and Zhu, D. Learning to poison large language models during instruction tuning. *arXiv preprint arXiv:2402.13459*, 2024.

Rade, R. and Moosavi-Dezfooli, S.-M. Reducing excessive margin to achieve a better accuracy vs. robustness tradeoff. In *International Conference on Learning Representations*, 2022.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7909–7919. PMLR, 2020.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rateike, M., Majumdar, A., Mineeva, O., Gummadi, K. P., and Valera, I. 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*, pp. 1421–1433, 2022.

Rigaki, M. and Garcia, S. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.

Rosenberg, I., Shabtai, A., Elovici, Y., and Rokach, L. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021.

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

Sáez, J. A., Luengo, J., and Herrera, F. Evaluating the classifier behavior with noisy data considering performance and robustness: The equalized loss of accuracy measure. *Neurocomputing*, 176:26–35, 2016.

Sarker, I. H., Furhad, M. H., and Nowrozy, R. Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. *SN Computer Science*, 2(3):173, 2021.

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., and Goldstein, T. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pp. 9389–9398. PMLR, 2021.

Schwinn, L., Bungert, L., Nguyen, A., Raab, R., Pulsmeyer, F., Precup, D., Eskofier, B., and Zanca, D. Improving robustness against real-world and worst-case distribution shifts through decision region quantification. In *International Conference on Machine Learning*, pp. 19434–19449. PMLR, 2022.

Schwinn, L., Dobre, D., Xhonneux, S., Gidel, G., and Gunnemann, S. Soft prompt threats: Attacking safety alignment and unlearning in open-source llms through the embedding space. *arXiv preprint arXiv:2402.09063*, 2024.

Simson, J., Pfisterer, F., and Kern, C. 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*, pp. 1305–1320, 2024.

Sioli, L. 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.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*, pp. 1015–1021. Springer, 2006.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Tocchetti, A., Corti, L., Balayn, A., Yurrita, M., Lippmann, P., Brambilla, M., and Yang, J. Ai robustness: a human-centered perspective on technological challenges and opportunities. *ACM Computing Surveys*, 2024.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Vassilev, A., Oprea, A., Fordyce, A., and Anderson, H. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology, 2024.

Vitali, F. 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*, pp. 235. IOS Press, 2022.

Vitorino, J., Maia, E., and Praça, I. Adversarial evasion attack efficiency against large language models. *arXiv preprint arXiv:2406.08050*, 2024.

Wang, Y., Pan, Y., Yan, M., Su, Z., and Luan, T. H. A survey on chatgpt: Ai-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, 2023.

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

Xie, Y., Yi, J., Shao, J., Curl, J., Lyu, L., Chen, Q., Xie, X., and Wu, F. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496, 2023.

Xu, H., Liu, X., Li, Y., Jain, A., and Tang, J. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pp. 11492–11501. PMLR, 2021.

Yan, J., Yadav, V., Li, S., Chen, L., Tang, Z., Wang, H., Srinivasan, V., Ren, X., and Jin, H. 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)*, pp. 6065–6086, 2024.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems*, 33:8588–8601, 2020.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, pp. 100211, 2024.

Yuan, L., Chen, Y., Cui, G., Gao, H., Zou, F., Cheng, X., Ji, H., Liu, Z., and Sun, M. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

Zadrozny, B. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st International Conference on Machine learning*, pp. 114, 2004.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

Zhang, X., Tu, R., Liu, Y., Liu, M., Kjellstrom, H., Zhang, K., and Zhang, C. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

Zhang, Z., Cao, X., Jia, J., and Gong, N. Z. 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*, pp. 2545–2555, 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 (Klimas & Vaiciukaite, 2008).