Robustness in the EU Artificial Intelligence Act

1. Introduction

The European Union (EU) recently introduced the Artificial Intelligence Act (AIA), with the goal to establish 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, such as fairness, explainability, and robustness (Rec. 27), and imposes some of them as legally binding requirements for certain types of AI systems. While it has been recognized for being one of the first legally binding regulatory frameworks for AI technologies (Chee & Hummel, 2024), it has been criticized for the incoherent use of terms (Laux et al., 2024b; Bomhard & Siglmüller, 2024), which make it difficult to implement it in practice.

Prior work has examined the AIA and its preceding documents to clarify some of these terms, such as 'explainability' (Bordt et al., 2022; Vitali, 2022; Pavlidis, 2024) and 'fairness' (Deck et al., 2024), but little attention has been given to the term 'robustness'. The AIA mandates 'robustness' in Art. 15 for specific AI systems, yet lacks clarity on its definition, leaving uncertainty about providers' legal obligations. Robustness can take different meanings depending on the context (Freiesleben & Grote, 2023). In general, a model's robustness refers to the ability of a model to maintain its performance, when its assumptions are violated (Diakonikolas et al., 2021; Zhang et al., 2023). This is a common problem when deploying ML models in the real world. Research in this field can be divided into studying non-adversarial (unintended) violations of assumptions (e.g., distribution shift), and adversarial perturbations of training data, input data or the model (*adversarial attacks*) (Tocchetti et al., 2022). In light of the variety of ML research in this field, a clear definition of the term 'robustness' in the AIA is key to translating its legal meaning into the computer science domain. This definition would enhance legal certainty and enable providers to design and develop their AI systems according to the requirements of the AIA. Compliance with the AIA will be presumed if providers adhere to upcoming harmonized standards (Art. 40 AIA) that specify legal requirements such as 'robustness'. The European Commission has requested the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation to develop these standards (Laux et al., 2024a).

In this paper, we discuss the use of the term in the AIA and make several suggestions to guide the development of harmonized standards. Specifically, we argue in this paper that the term 'robustness' as used in Art. 15(1) and (4) AIA should be understood as *non-adversarial robustness* in the ML literature, whereas the term 'cybersecurity' used in Art. 15 AIA should be understood as *adversarial robustness*.

2. Background on AIA

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 specifications (Klimas & Vaiciukaite, 2008).

In this paper, we examine ML approaches that learn from input data to generate outputs aimed at achieving specific objectives, which following Rec. (12) AIA, are an AI system as defined in Art. 3 AIA. The AIA distinguishes between general-purpose AI models and other types of AI models. The latter ones can be understood as "fixed-purpose systems" that are created with a specific objective and can only accomplish tasks they are trained to perform (e.g., translation, classification), while general-purpose AI can perform tasks that they were not originally trained for (e.g., foundation models) (Gutierrez et al., 2023). In this paper, we focus AI systems for fixed purposes, in the following 'AI systems'.

The AIA regulates AI systems differently based on their risk level (Sioli, 2021; Bomhard & Siglmüller, 2024): Those posing unacceptable risks, such as social scoring, are prohibited; high-risk systems, such as those used in medical devices, are allowed but must comply with certain requirements and undergo pre-assessment; and other systems are allowed and, depending on the risk, only require specific transparency and information obligations. In the following, we focus on high-risk AI systems. Those are the only ones that are required to fulfill 'robustness' requirements following Art. 15 AIA.

3. Legal Challenges

Art. 15(1) AIA states that high-risk AI systems must achieve an "appropriate level of accuracy, robustness and cybersecurity" and must function consistently in this respect throughout their life cycle. Art. 15(4) further specifies 'robustness', according to which high-risk AI systems must be "as resilient as possible" to "errors, faults or inconsistencies that may occur within the system or the environment in which the system operates". Even though this article describes the requirement of robustness for high-risk AI systems, it does provide a clear definition of the term. This poses three signif-

icant challenges for the interpretation of this of this article: i) the term 'robustness' remains ambiguous as it is used incoherently throughout the AIA; ii) it lacks a direct relation to the robustness terminology in the ML literature; iii) Art. 15 AIA itself offers shortcomings and creates legal uncertainty.

Incoherent Terminology. The term 'robustness' is used incoherently throughout the AIA. Art. 15(1) and (4) AIA refer to 'robustness', whereas the corresponding Rec. (27) and (75) AIA both speak of 'technical robustness'. It is unclear whether these terms are equivalent or to what extent they differ. One could argue that 'technical robustness' is used synonymous to 'robustness'. Or one could argue that they refer to different concepts: Either the term 'robustness' is being restricted in its scope to technical aspects by the recitals, or it additionally includes some form of non-technical robustness. The latter, however, remains undefined, but could refer to taking broader social implications into account (Floridi, 2021). The term 'robust' is also used in parts of the AIA that do not concern high-risk AI systems (Rec. (8) and (81)). The discussion of these is outside the scope of this paper.

Aligning Legal and ML Terminology. Aligning legal and ML terminology makes the requirements of the AIA operable and enable their technical implementation. However, we find that 'robustness' in Art. 15 AIA as a legal term does not fully overlap with the ML conception of *robustness*. First, technical advances in the ML literature often refer to ML models, whereas the AIA regulates high-risk AI systems as a whole. Second, the legal framework uses the terms 'robustness' and 'cybersecurity' and fails to reflect the subtlety terminology used in the ML domain.

First, we observe that the term 'robustness' in the ML literature is often referred to as a property of an AI model. By contrast, Art. 15 and Rec. (75) refer specifically to AI systems. A policy report by the European Commission explicitly confirms this holistic approach (Junklewitz et al., 2023). Therefore, the robustness requirements extend beyond the technical performance of the AI model itself.

Second, it is unclear whether adversarial robustness falls under requirements demanding robustness, as the name suggests, or those for cybersecurity. Art. 15(4) AIA and its corresponding Rec. (75) refer to errors, faults or inconsistencies that may occur inadvertently as the system interacts with its real-world environment. As stated above, we distinguish here between ML research on non-adversarial robustness, and adversarial robustness. *Non-adversarial robustness* refers to the ability of an AI model to maintain their performance under data distribution shifts or noisy data (Yao et al., 2022; Zhang et al., 2021). We propose that the legal term 'robustness' covers the ML concepts of *non-adversarial robustness*, as they both address non-adversarial causes for differences between the training and deployment environments.

Conversely, the cybersecurity requirements in Art. 15(5) AIA and Rec. (76) address deliberate attempts "to alter the use, outputs, or performance" of an AI system. *Adversarial robustness* in the ML domain refers to the ability of AI models to be resistant to adversarial input samples, which are intentionally perturbed inputs designed to modify the predictions of an ML model (Biggio et al., 2013; Goodfellow et al., 2015). Research that suggests methods to defend an ML model against adversarial examples sometimes also refers to assuring the *security* of an ML model (Madry et al., 2018; Athalye et al., 2018) or its *safety* (Wang et al., 2018). We propose that the legal term 'cybersecurity' covers the ML concepts of *adversarial robustness*, as they both refer to deliberate adversarial causes for data perturbations.

Important to mention is that Rec. (27) AIA also demands AI systems to be resilient "against attempts to alter the use or performance of the AI system", i.e., asking for adversarial robustness. However, following the sentence structure, the recital maps these aspects of adversarial robustness not to 'cybersecurity' as in Art. 15 AIA, but to the term 'safety'.

Here, we argue here that the term 'robustness' in Art. 15(1) and (4) AIA covers the ML concept of non-adversarial robustness, whereas the term 'cybersecurity' in Art. 15(1) and (5) AIA cover *adversarial robustness*. While there are, as mentioned above, indications from the ML literature that 'safety' could also cover *adversarial robustness*, we defer a more thorough investigation of the relationship between the terms 'cybersecurity' and 'safety' to future research.

Required Level of Robustness. The AIA creates ambiguities 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 high-risk AI systems shall be "as resilient as possible" to "errors, faults, or inconsistencies", suggesting a narrower and stricter requirement. This discrepancy leads to legal uncertainty about whether AI systems need only meet an appropriate robustness standard or strive for the highest possible level. Art. 15(5) AIA mandates an cybersecurity level that is " appropriate to the relevant circumstances and risks". Note that complex ML models can usually not be fully adversarial robust, especially if the adversary adapts to possible defense mechanisms (Xie et al., 2023; Kumar et al., 2023). An appropriate level of cybersecurity should therefore rather be understood as a requirement for a sufficient defense. We defer a thorough investigation of these points to future work.

Outlook. We call for future standardization efforts to define and differentiate the legal terms: i) 'technical robustness' and 'robustness'; ii) 'robustness' and 'cybersecurity'; iii) 'cybersecurity' and 'safety'. In future work we will extend our analysis to general-purpose AI (GPAI) systems.

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

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.

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.

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/.

Deck, L., Müller, J.-L., Braun, C., Zipperling, D., and Kühl, N. Implications of the ai act for non-discrimination law and algorithmic fairness. *arXiv preprint arXiv:2403.20089*, 2024.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Robustness meets algorithms. *Communications of the ACM*, 64(5):107–115, 2021.

Floridi, L. Establishing the rules for building trustworthy ai. *Ethics, Governance, and Policies in Artificial Intelligence*, pp. 41–45, 2021.

Freiesleben, T. and Grote, T. Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202 (4):109, 2023.

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

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.

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. Scientific analysis or review KJ-NA-31-643-EN-N (online), Luxembourg (Luxembourg), 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.

Laux, J., Wachter, S., and Mittelstadt, B. Three pathways for standardisation and ethical disclosure by default under the european union artificial intelligence act. *Computer Law & Security Review*, 53:105957, 2024a.

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, 2024b.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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.

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.

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. *arXiv preprint arXiv:2210.08906*, 2022.

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.

Wang, S., Pei, K., Whitehouse, J., Yang, J., and Jana, S. Efficient formal safety analysis of neural networks. *Advances in neural information processing systems*, 31, 2018.

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.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, Y., Zeng, D., Luo, J., Xu, Z., and King, I. A survey of trustworthy federated learning with perspectives on security, robustness and privacy. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 1167–1176, 2023.