Compliance Made Practical: Translating the EU AI Act into Implementable Security Actions

1st Niklas Bunzel
Fraunhofer SIT / ATHENE / TU Darmstadt
Darmstadt, Germany
niklas.bunzel@sit.fraunhofer.de

2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE) | 979-8-3315-1466-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/RAIE66699.2025.00016

***Abstract*—The EU AI Act, along with emerging regulations in other countries, mandates that AI systems meet security requirements to prevent risks associated with AI misuse and vulnerabilities. However, for practitioners, defining and achieving a *secure* AI system is complex and context-dependent, posing challenges in understanding what actions they need to take and when they are sufficient. ISO/IEC TR 24028/29 and ENISA Securing Machine Learning Algorithms offer a comprehensive framework for AI security, aligning with the EU AI Act's requirements by addressing risks, threats, and mitigation strategies. However, for practical implementation, these reports lack hands-on guidance. Industry resources like the OWASP AI Exchange and OWASP LLM Top 10 fill this gap by providing accessible, actionable insights for securing AI systems effectively. This paper addresses the question of responsibility in AI risk mitigation, especially for companies utilizing pretrained or off-the-shelf models. We want to clarify how companies can practically comply with the upcoming ISO 27090 and ensure compliance with the EU AI Act through actionable security strategies tailored to this prevalent use case.**

***Index Terms*—Artificial Intelligence, Regularization of AI, Trustworthy AI, Standardization of AI**

I. Introduction

Artificial Intelligence (AI) continues to advance rapidly, becoming more capable and versatile, with an ever-expanding range of applications. From healthcare to autonomous vehicles, AI is now integral to safety- and security-critical products. However, as AI adoption grows, it has historically operated in a regulatory vacuum, with no standardized laws or frameworks ensuring its safe and secure deployment. This changed with the introduction of the EU AI Act [1], the first comprehensive legislation mandating safety and security in AI products. The Act establishes obligations for developers and users, addressing risks and reinforcing the importance of trustworthy AI. However, the EU AI Act is not the only initiative in this area; similar initiatives are emerging around the world. In the United States, executive orders and federal directives aim to regulate AI applications, while Canada's Artificial Intelligence and Data Act and Brazil's AI Regulatory Framework reflect a worldwide shift towards structured AI governance. Despite this global momentum, these regulations vary widely, creating challenges for multinational organizations striving to comply with disparate requirements. A unified international standard is needed. The ISO 27090 [2], set to debut in 2025, promises a comprehensive framework for ensuring trustworthy AI. Unlike existing technical reports such as ISO/IEC TR 24028/24029 [3], [4] or ENISA's Secure Machine Learning Algorithms [5], we hope it provides a holistic view of threats and actionable guidance for mitigating them. Currently, practitioners rely on best practices and guidelines, including frameworks like the OWASP AI Exchange, which offers hands-on advice covering the entire AI threat landscape [6]. It emphasizes controls against specific attack vectors, risk assessment methodologies, and recommendations for technical implementations. However, these guidelines often lack direct mappings to regulatory requirements or detailed ownership responsibilities for specific security controls. In this paper, we address these gaps by:

- Mapping the EU AI Act's requirements to technical reports and best practices.
- Identifying the roles responsible for implementing various security controls.
- Exploring implementation strategies for companies integrating pretrained AI models.

We emphasize that the threats relevant to an AI system depend on the specific use case and architecture. Through risk assessment, organizations can evaluate the likelihood and impact of threats, prioritize risks, and select appropriate mitigations. For example, adversarial attacks and data leakage can have far-reaching strategic, operational, compliance, and reputational impacts.

II. Background

A. Laws and Regulation

- **EU: AI Act** The EU AI Act is a comprehensive regulatory framework governing AI within the EU, classifying AI systems by risk: unacceptable (prohibited), high (regulated), limited (transparency required), and minimal (unregulated). It aims to ensure AI is used ethically, transparently, and without compromising fundamental rights by imposing stringent requirements on high-risk systems. These include a Risk Management System (Article 9) for health, safety, and fundamental rights risks, Data Governance (Article 10) for dataset integrity, Technical Documentation (Article 11) for compliance, and Robustness and Accuracy (Article 15) to protect against errors, adversarial attacks and model

979-8-3315-1466-2/25/$31.00 ©2025 IEEE
DOI 10.1109/RAIE66699.2025.00016

vulnerabilities. This act represents a significant step toward responsible AI governance in the EU [1].

- **USA: Accountability Act, Disclosure Act, Executive Order** The White House Executive Order (October 2023) emphasizes the safe, secure, and trustworthy development of AI through policies addressing safety, security, responsible innovation, equity, privacy, and civil rights [7]. The Federal AI Disclosure Act mandates transparency in AI usage [8]. The Federal Algorithmic Accountability Act proposes regular impact assessments for AI systems, prioritizing fairness, privacy, and anti-discrimination [9].

B. Standard and Regulation Bodies

- **OWASP** The Open Worldwide Application Security Project (OWASP) is a global non-profit organization dedicated to improving the security of software and web applications by providing open-source tools, resources, and community-driven standards.
- **ISO** The International Organization for Standardization (ISO) is an independent, non-governmental international organization that develops and publishes globally recognized standards to ensure quality, safety, efficiency, and interoperability across various industries.
- **ENISA** The European Union Agency for Cybersecurity (ENISA) is an EU body that strengthens Europe's cyber resilience by providing guidance, policy support, and expertise in cybersecurity to member states, businesses, and citizens.

C. Standards and Reports

- **ISO/IEC 27090** ISO/IEC 27090 is a developing standard that provides guidance for addressing security threats and vulnerabilities within AI systems. The standard aims to help organizations understand and mitigate risks throughout the life cycle of AI systems, including threats like data poisoning, evasion attacks, membership inference, model inversion and model theft. It emphasizes the need for precise terminology and a focus on active, deliberate attacks, while also considering the broader implications of AI security [2].
- **ISO/IEC TR 24028** ISO/IEC TR 24028 provides an overview of trustworthiness in AI, focusing on enhancing trust in AI systems through principles like transparency, explainability, and controllability. It identifies threats and vulnerabilities specific to AI [3].
- **ISO/IEC TR 24029** ISO/IEC TR 24029 provides an overview of methods to assess the robustness of neural networks, focusing on their ability to maintain performance under diverse conditions such as input perturbations or domain changes. It categorizes assessment techniques into statistical methods, formal methods and empirical methods [4].
- **ENISA Securing Machine Learning Algorithms** The ENISA report on Securing Machine Learning Algorithms offers a detailed taxonomy of machine learning algorithms, outlining their core functionalities and critical stages. It identifies key threats such as data poisoning, adversarial attacks and data exfiltration. The report proposes actionable security measures, drawn from relevant literature to protect against these threats [5].
- **ISO/IEC 23894** ISO/IEC 23894 offers guidance on managing AI-specific risks. It supports integrating risk management into AI-related activities and functions within the organization. Additionally, it outlines processes to effectively implement and embed AI risk management across operations [10].

D. Guidelines and Best Practices

The OWASP Top 10's focus on awareness and education to prevent common vulnerabilities and security risks. They are consensus-driven and curated by a community of security experts around the world, aiming to provide practical security guidance and raise awareness about the critical security risks to software and applications.

- **OWASP AI Exchange** The OWASP AI Exchange [11] focuses on assessing and mitigating risks in AI systems, mapping out a comprehensive AI threat landscape to address vulnerabilities like adversarial attacks, data poisoning, and prompt injection. It emphasizes developing and sharing best practices for AI risk assessment, enabling organizations to identify weaknesses and prioritize defenses. The AI Exchange actively collaborates with industry and academia to refine its methodologies and ensures alignment with global standards and regulatory frameworks, supporting the secure deployment of trustworthy AI solutions.
- **OWASP Machine Learning Security Top 10** The OWASP Top 10 for Machine Learning (ML) outlines key security risks for ML systems [12]. These include input manipulation attacks, data poisoning, model inversion attacks, membership inference, AI supply chain attacks, transfer learning attacks and model poisoning.
- **OWASP Top 10 for Large Language Model Applications** The OWASP Top 10 for Large Language Model Applications outlines critical security risks in deploying and managing LLMs [13]. Key threats include prompt injection, insecure output handling, training data poisoning, model denial of service, supply chain vulnerabilities, sensitive information disclosure, insecure plugin design, model theft and over-reliance on LLMs.

III. Mapping EU AI Act, ISO 24028 and OWASP Guidelines

The EU AI Act enforces robustness and accuracy requirements on AI systems. From a security perspective these requirements are robustness against adversarial attacks, mitigating data and model poisoning, evasion attacks and confidentiality attacks (e.g. model inversion) [1]. The ISO/IEC TR 24028 identifies mostly the same security threats to AI systems,

EU AI Act	ISO/IEC TR 24028	OWASP AI Exchange
Accuracy, Robustness and Cybersecurity		
Manipulate training data	8.2.2 Data Poisoning	3.1.1 Data Poisoning
Manipulate pre-trained components	-	3.1.2. Development-Environment Model Poisoning 3.1.3 Supply-Chain Model Poisoning
Inputs designed to cause a mistake	8.2.3 Adversarial Attacks	2.1. Evasion
Confidentiality Attacks	8.2.4 Model Stealing	2.4. Model theft through use 4.3. Direct runtime model theft 2.3.2. Model inversion and Membership inference

TABLE I: Mapping of EU AI Act security requirements to ISO/IEC TR 24028 and OWASP AI Exchange. The EU AI Act is broad and often vague, while ISO/IEC TR 24028 provides clearer but narrower guidance, leaving gaps for full compliance with the AI Act. The OWASP AI Exchange offers a holistic framework, addressing threats and controls necessary to align with both ISO/IEC TR 24028 and the EU AI Act.

including data poisoning, evasion attacks, model stealing and hardware threats. However, Table I illustrates that existing technical reports like ISO/IEC TR 24028 do not holistically address the EU AI Act's requirements. For instance, the manipulation of pre-trained components, as required by the AI Act, is not addressed. Additionally, while the technical report covers model stealing under confidentiality-related attacks, it does not address other threats such as model inversion or membership inference. This gap emphasizes the need for a unified framework combining regulatory obligations with technical safeguards to ensure trustworthy, secure AI systems. The OWASP AI Exchange serves as such a holistic framework, but a formal standard that is equally comprehensive is still required. We hope that the upcoming ISO 27090 will fulfill this critical need. The OWASP AI Exchange provides insight and actionable controls against the threats. Data poisoning can be mitigated by e.g. data quality control such as detection of poisoned samples by integrity checks, statistical deviation or pattern recognition. Model poisoning can be mitigated by model inspection [14]. Evasion attacks can be countered with statistical detectors [15], [16]. Model theft can be prevented by limiting the query rate, and model inversion can be prevented by excluding confidence labels in the output. Risks such as data leakage, denial-of-service & denial-of-wallet attacks, and prompt injection are not explicitly mentioned in the EU AI Act, but are covered in the AI Exchange.

IV. Who is responsible?

In a typical scenario involving AI systems, three key entities play distinct roles:

- Model Provider: Responsible for designing the model architecture and training the model.
- Model Hoster: Operates the model on their infrastructure.
- Model Integrator: Embeds the model into their system to deliver a product or service to end-users.

Each entity has specific responsibilities for complying with the security aspects of laws and standards. These responsibilities depend on contractual agreements and the attack surface associated with their role. In this context, we will consider the EU AI Act as a law and ISO/IEC TR 24028 as an example of a technical report in the absence of a formal standard.

- Model Provider:
  - Ensures the robustness of the model against data poisoning attacks by curating secure datasets (AI Act: Manipulate training data; ISO 24028: Data Poisoning).
  - May be accountable for mitigating evasion attacks by employing techniques such as adversarial training, defensive distillation, or using architectures resilient to adversarial inputs (AI Act: Inputs designed to cause a mistake; ISO 24028: Adversarial Attacks).
- Model Hoster:
  - Typically responsible for classical cybersecurity measures, including protection against bot attacks like Denial-of-Service (DoS) attacks.
  - May share responsibility for safeguarding against model poisoning, if the model is obtained through his service (AI Act: Manipulate pre-trained components).
  - May be responsible for mitigating model stealing, often through query rate limitations and access controls (AI Act: Confidentiality Attacks; ISO 24028: Model Stealing).
- Model Integrator:
  - Addresses model poisoning by implementing model verification methods (AI Act:Manipulate training data,Manipulate pre-trained components; ISO 24028: Data Poisoning).
  - Ensures the robustness against evasion attacks by implementing input sanitization methods such as adversarial detectors, out of distribution detectors or input distortion. However, input distortion can reduce benign accuracy if the model is not trained with such distortions (AI Act: Inputs designed to cause a mistake; ISO 24028: Adversarial Attacks).
  - Protects against prompt injection attacks using detectors, sanitization and system prompts (AI Act: Inputs designed to cause a mistake, Confidentiality Attacks).
  - Is responsible for mitigating model stealing, model inversion, membership inference and attribute inference using detectors and query rate restrictions

(AI Act: Confidentiality Attacks; ISO 24028: Model Stealing).
  - Manages Denial-of-Wallet (e.g., sponge attacks) and DoS attacks through detection systems.
  - Handles sensitive input and output data securely to prevent leakage and ensure compliance with privacy standards.

Each stakeholder's responsibilities must be clearly delineated in agreements to address security and robustness comprehensively across the AI life cycle.

V. On the Implementation of AI Security Controls

To comply with ISO/IEC TR 24028/29, the ENISA Guidelines for Securing Machine Learning Algorithms, the upcoming ISO/IEC 27090, and the EU AI Act, organizations must implement defenses tailored to identified risks and threats. These defenses should align with the organization's risk assessment, role (e.g., model provider, host, or integrator), contractual obligations, and the specific use case scenarios. Since most companies operate as model integrators utilizing pre-trained models, this discussion focuses on their responsibilities. For companies fine-tuning models, the responsibilities of model providers would also apply.

a) Mitigations for Availability: As a model integrator, leveraging the protections provided by model hosters is critical to addressing threats such as bot activity, Denial-of-Service (DoS) and Denial-of-Wallet attacks. These are of particular concern given that bot-generated traffic accounts for approximately 47% of Internet activity [17]. Documenting these protections helps meet EU AI Act requirements. In addition, measuring inference costs, such as time or energy consumption, and implementing cut-off thresholds can prevent abuse [18]. This approach potentially eliminates the need for complex sponge attack detectors[1] while maintaining operational efficiency.

b) Mitigations for Integrity: Detecting model poisoning can be achieved through techniques like model inspection [14], which allow integrators to identify compromised models. For evasion attacks, mitigation strategies depend on the input type—whether images, video, or audio—and the attacker's level of access. Direct input access necessitates specific defenses, while indirect manipulations, such as through cameras, require different approaches. While various methods have been proposed [15], [16], [21], identifying effective detection thresholds remains an open research challenge. These thresholds should be tailored to the application's risk assessment to ensure robust security. To defend against prompt injection attacks, integrators can implement input sanitization and filtering mechanisms to detect and block malicious instructions. Prompt injection attacks not only compromise the integrity of an AI system by manipulating inputs to produce unintended outputs, but can also target confidentiality by extracting sensitive or private information from the system. While input validation is more challenging for natural language than for structured inputs like SQL, these measures remain critical.

c) Mitigations for Confidentiality: Query management plays a critical role in mitigating attacks such as model stealing and model inversion. Setting query rate limits prevents attackers from exploiting excessive queries, while restricting outputs to class labels, rather than confidence scores, effectively reduces the risk of membership inference attacks [19]. However, these measures may be insufficient against advanced label-only attacks [20], requiring further refinements.

By implementing these targeted mitigations, model integrators can address a wide range of AI security threats while aligning with regulatory and standards-based requirements. Tailoring defenses to organizational roles and risk assessments ensures both compliance and the protection of deployed AI systems.

VI. Conclusion & Future Work

Artificial Intelligence (AI) is increasingly integrated into products across industries, prompting the development of regulatory frameworks worldwide to ensure AI systems are safe and secure. The EU AI Act serves as a landmark example of such regulation. However, aligning with these regulations poses challenges, as comprehensive standards have yet to be finalized. For instance, ISO 27090 is anticipated in 2025. In this paper, we analyze the Ensia Securing Machine Learning Algorithms report alongside ISO/IEC technical reports 24028 and 24029, mapping their content to the requirements of the EU AI Act. Our findings reveal that existing technical reports fail to comprehensively address the threat landscape and are insufficient for full compliance with the Act. We delve into various threats, defense mechanisms, and the assignment of responsibility for mitigating specific risks. To illustrate practical implementation, we adopt the perspective of a model integrator, demonstrating how to address threats systematically. Our analysis shows that sophisticated detectors are necessary in only a few scenarios, such as evasion attacks. However, these detectors may negatively impact benign accuracy, emphasizing the need for careful calibration. Thresholds for such detectors remain a subject of ongoing research and should be aligned with risk assessments specific to the intended application.

Acknowledgment

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] European Commission. (2024) Artificial Intelligence Act. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

[1] In [18], the authors suggest adversarial attack detectors to detect sponge attacks.

[2] ISO/IEC. (2024) ISO/IEC 27090 - Cybersecurity - Artificial Intelligence - Guidance for addressing security threats to artificial intelligence systems. [Online]. Available: https://www.iso27001security.com/html/27090.html

[3] ISO/IEC. (2020) ISO/IEC TR 24028:2020 - Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence. [Online]. Available: https://www.iso.org/standard/77608.html

[4] ISO/IEC. (2021) ISO/IEC TR 24029-1:2021 - Artificial Intelligence (AI) - Assessment of the robustness of neural networks. [Online]. Available: https://www.iso.org/standard/77609.html

[5] ENISA. (2021) Securing machine learning algorithms. [Online]. Available: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

[6] N. Bunzel and N. Göller, "Bridging the gap: The role of owasp ai exchange in ai standardization," in *INFORMATIK 2024*. Gesellschaft für Informatik eV, 2024, pp. 263–273.

[7] The White House. (2023) Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/3

[8] 118th Congress. (2023) AI disclosure act of 2023. [Online]. Available: https://www.congress.gov/bill/118th-congress/house-bill/3831/text

[9] 117th Congress. (2022) Algorithmic accountability act of 2022. [Online]. Available: https://www.congress.gov/bill/117th-congress/house-bill/6580/text

[10] ISO/IEC. (2023) ISO/IEC 23894:2023 - Information technology - Artificial intelligence - Guidance on risk management. [Online]. Available: https://www.iso.org/standard/77304.html

[11] A. Travers, A. Tron, A. Qarry, A. Seyerlein-Klug, A. Glynn, B. Karimi, D. S. Cox, F. Tang, J. Sotiropoulos, M. Lihter, N. Bunzel, R. van der Veer, R. Sanz, S. Dunn, S. Oesch, S. Gupta, S. Francolla, W. Wei, Y. Kanellopoulos, and Z. Braiterman. (2024) Owasp ai exchange. [Online]. Available: https://owaspai.org/

[12] S. Bhure, S. Singh, R. van der Veer, M. S. Nishanth, R. M, H. Blankenship, RiccardoBiosas, A. Kenchappagol, M. Kowalczyk, and A. Nugroho. (2023) Owasp machine learning security top 10 (2023 edition) - draft release v0.3. [Online]. Available: https://mltop10.info/

[13] A. Dawson, A. Smith, A. King, B. Simonoff, D. Rowe, E. G. Junior, E. Neelou, G. Klondike, I. Golan, J. Ross, J. Sotiropoulos, K. Huang, L. Derczynski, M. S, M. Finch, R. Sood, R. Harang, S. Wilson, T. Seeparsan, and W. Chilcutt. (2024) Owasp top 10 for llm applications. [Online]. Available: https://llmtop10.com/

[14] A. Rawat, K. Levacher, and M. Sinn, "The devil is in the GAN: backdoor attacks and defenses in deep generative models," in *Computer Security - ESORICS 2022 - 27th European Symposium on Research in Computer Security, Copenhagen, Denmark, September 26-30, 2022, Proceedings, Part III*, ser. Lecture Notes in Computer Science, V. Atluri, R. D. Pietro, C. D. Jensen, and W. Meng, Eds., vol. 13556. Springer, 2022, pp. 776–783.

[15] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[16] N. Bunzel and D. Böringer, "Multi-class detection for off the shelf transfer-based black box attacks," in *Proceedings of the 2023 Secure and Trustworthy Deep Learning Systems Workshop*, ser. SecTL '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3591197.3591305

[17] Imperva Inc. (2023) Bad bot report. [Online]. Available: https://www.imperva.com/resources/reports/2023-Imperva-Bad-Bot-Report.pdf

[18] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson, "Sponge examples: Energy-latency attacks on neural networks," in *2021 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2021, pp. 212–231.

[19] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[20] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *International conference on machine learning*. PMLR, 2021, pp. 1964–1974.

[21] N. Bunzel, A. Siwakoti, and G. Klause, "Adversarial patch detection and mitigation by detecting high entropy regions," in *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 2023, pp. 124–128.