TOWARDS A ROBUST AI REGULATORY FRAMEWORK: TECHNICAL AND LEGAL ASPECTS

Sînică ALBOAIE[1], Marco CUOMO[2,] Lenuța ALBOAIE[1]

e-mail: sinica.alboaie@axiologic.net

Abstract

Artificial Intelligence (AI) technologies are increasingly integrated into various aspects of society, raising concerns about potential risks and ethical implications. This paper presents a scoring proposal for a regulatory authority to consider when assessing AI technologies, emphasizing adopting a comprehensive AI security framework and general security methods. This proposal will facilitate the evaluation of AI systems based on security and ethical standards, promoting responsible development and deployment of AI technologies.

Keywords: AI security framework, Regulatory compliance, Ethical AI development

In the rapidly evolving world of artificial intelligence (AI), ensuring AI systems' ethical and safe development and deployment has become paramount. The article delves into the multifaceted dimensions of AI risk assessment and management. By meticulously examining AI risk classes and system protection methods, we aim to provide a comprehensive understanding of the challenges and opportunities associated with AI regulation. Furthermore, we propose an innovative scoring method for AI systems, considering risk levels and protection strategies, to aid in developing a more robust and effective regulatory framework. This approach will promote responsible AI use and foster transparency and trust between stakeholders, ultimately contributing to this transformative technology's safe and ethical advancement.

MATERIAL AND METHOD

From a legal and regulatory perspective (Cath C., 2018; Buiten M., 2019), ensuring the safety and security of AI-based systems requires an in-depth investigation of the challenges and risks associated with these systems. This investigation must consider both technical and legal measures that can be employed to ensure the responsible deployment of AI-based systems.

Technical measures, such as secure software development, access control, data encryption, and intrusion detection systems, can provide a foundation for the security and privacy of AI-based systems. However, legal and regulatory measures are also essential, including developing guidelines and standards, legal liability frameworks, and regulatory oversight.

Recognising the complexity and diversity of risks that AI-based systems can pose is important. These risks include technical challenges, such as algorithmic bias (Koene A., 2017), adversarial attacks, and legal and social challenges, such as privacy and ethical considerations. Addressing these risks requires collaboration between technical and legal experts, policymakers, and stakeholders.

Legal and regulatory efforts must also consider the potential implications of AI-based systems on society, including the impact on employment, privacy, and civil rights. Developing ethical and legal frameworks that balance innovation with responsible deployment is essential to ensure that AI-based systems benefit society while minimising potential harm (Leslie, D et al., 2021). However, as we presented above, ensuring the safety and security of AI-based systems requires a comprehensive approach that considers both technical and legal measures. Investigating the challenges and risks associated with these systems is necessary to develop appropriate mitigation measures and establish a regulatory framework that fosters innovation while protecting society. Therefore, a collaboration between technical and legal experts, policymakers, and stakeholders is essential to achieve this goal.

Adopting a comprehensive approach incorporating various methods and techniques to enhance the safety of AI-based systems is crucial. One such approach is to employ secure development and design principles, encompassing

[1] AXIOLOGIC SAAS, Iași, Romania
[2] PharmaLedger Association, Basel, Elvetia

methodologies such as Secure Software Development Lifecycle (SDLC), threat modelling, secure coding practices, and various application security testing techniques. By integrating these principles into the development process, developers can effectively identify and mitigate potential vulnerabilities and risks associated with AI-based systems, thus ensuring a robust and secure architecture.

Access control and authentication mechanisms play a vital role in safeguarding AI-based systems. Implementing strong authentication techniques, such as Multi-Factor Authentication (MFA) (Sinha A. et al., 2019; Ometov A. et al., 2018) and Role-Based Access Control (RBAC) (Kuhn D. R. et al., 2010) in conjunction with the Principle of Least Privilege (POLP) (McCarthy M., 2022), can significantly reduce the likelihood of unauthorised access to sensitive data and system resources. Furthermore, incorporating federated identity management and Single Sign-On (SSO) (Nongbri I. et al., 2017) systems can streamline the authentication process while maintaining a high-security level.

Data security and privacy (Oseni A. et al., 2021) are essential aspects of AI-based system safety. Employing robust data protection methods, such as data encryption, secure key management, and Privacy-Enhancing Technologies -PETs, can help ensure that sensitive information remains confidential and secure. Advanced cryptographic techniques like homomorphic encryption and Secure Multi-Party Computation (SMPC)(Damgård I. et al., 2012) can facilitate secure data processing and analysis without compromising privacy.

Network and infrastructure security measures are equally important in ensuring the safety of AI-based systems. Implementing defence-in-depth strategies, alongside Intrusion Detection and Prevention Systems (IDPS) and Security Information and Event Management (SIEM) tools, can significantly improve the overall security posture. Furthermore, adopting zero-trust architecture and network segmentation can help limit potential attack surfaces. At the same time, employing technologies such as blockchain and quantum-resistant cryptography can enhance the resilience of AI-based systems against advanced threats.

Lastly, leveraging AI-driven security and monitoring solutions can greatly contribute to the safety of AI-based systems.

Organisations can proactively identify and mitigate potential threats by incorporating machine learning algorithms for anomaly detection, threat intelligence, and incident response. Additionally, AI-driven Security Orchestration, Automation, and Response (SOAR) platforms can streamline security operations, allowing for rapid and efficient responses to potential risks. Through the implementation of these five classes of methods, the safety of AI-based systems can be significantly improved, ensuring their responsible and secure deployment in various domains.

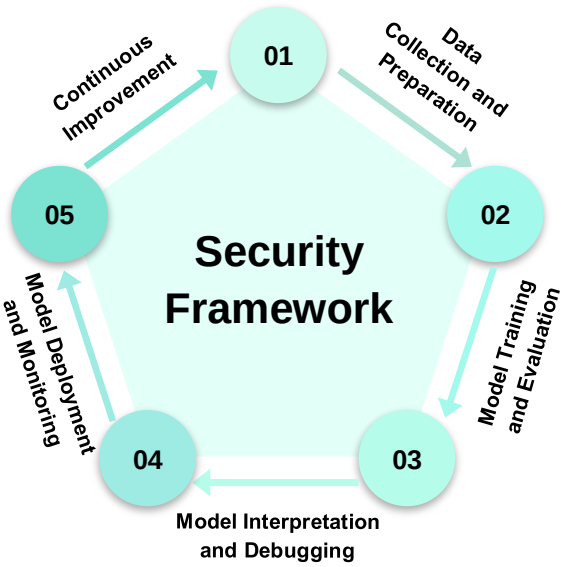

Figure 1 **AI Security Framework**

In this article, we propose a comprehensive technical framework encompassing software tools that can be employed throughout the entire lifecycle of an AI model, from training to maintenance and continuous improvement. This all-inclusive approach addresses the challenges and risks associated with AI systems, such as data security, privacy, fairness, robustness, and explainability. The framework is organised into five interconnected components that reflect critical stages in developing, deploying, and managing AI models. These components, briefly outlined below, form the foundation of our proposed framework:

The “Data Collection and Preparation” stage emphasises the importance of data security and privacy tools in collecting and preparing high-quality data while preserving user privacy. Integrating these tools ensures that AI systems are built upon secure and reliable datasets, setting the stage for addressing fairness and bias concerns.

The “Model Training and Evaluation” stage requires fairness and bias mitigation tools to ensure equitable outcomes and prevent discriminatory behaviour in AI models. By incorporating these tools into the training and evaluation process, we can create robust and generalisable AI models that perform consistently across diverse datasets and use cases.

The “Model Interpretation and Debugging” stage focuses on model robustness and generalisation tools, which enable developers to understand, interpret, and debug AI models effectively. By enhancing the transparency and interpretability of AI systems, we can facilitate better collaboration between humans and AI, ensuring that AI models align with human values and intentions.

The "Model Deployment and Monitoring" stage ensures that AI models are deployed in real-world scenarios, and then explainability and interpretability tools become essential for ensuring their safety and reliability. By integrating these tools into the deployment and monitoring process, we can maintain trust and accountability in AI systems while minimising risks and vulnerabilities.

The final stage, when productive systems are "Continuous Improvement" and continuous maintenance, underscores the iterative nature of AI development and the importance of system safety and reliability tools in fostering continuous improvement. By incorporating these tools into the maintenance and enhancement process, we can ensure that AI models remain up-to-date, secure, and effective in addressing evolving challenges and requirements.

By proposing this comprehensive technical framework, we aim to provide researchers, practitioners, and organisations with a holistic approach to AI development, deployment, and management. This framework addresses the challenges and risks associated with AI systems and promotes the responsible, ethical, and secure use of AI in various applications and domains.

This article aims to introduce a potential new AI platform at Technology Readiness Level - TRL1 (Mankins J. C., 2009), focusing on the conceptualisation stage. Therefore, we will briefly describe the envisioned tools within each node of Figure 1 as part of our comprehensive technical framework for AI systems.

At the data security and privacy stage, we envision tools that protect sensitive data from unauthorised access and ensure user privacy. Key concepts include robust encryption, access control mechanisms, privacy-preserving learning techniques (e.g., federated learning, differential privacy), and secure data storage solutions.

In the context of fairness and bias mitigation, we propose tools designed to identify and address algorithmic bias in AI models. These tools will allow preventive detection and mitigation of biases in training data and promote fairness during model development and evaluation.

For model robustness and generalisation, the envisioned tools aim to improve AI model performance, resilience against adversarial attacks, and adaptability to new tasks or domains. This category encompasses adversarial training, optimisation of transfer learning, and robust model selection and evaluation.

Regarding explainability and interpretability, we aim to develop tools that enhance the transparency and comprehensibility of AI systems. These tools will enable stakeholders to understand AI model decisions and behaviour better, fostering effective collaboration between humans and AI and ensuring alignment with human values and intentions.

Lastly, at the system safety and reliability stage, we propose tools that focus on ensuring AI systems' consistent and secure functioning. These tools will facilitate model monitoring, maintenance, incident response, and vulnerability detection and patching to promote overall AI system safety and resilience against potential threats.

By outlining these initial concepts, we hope to lay the foundation for a comprehensive technical framework that addresses AI systems' various challenges and risks throughout their entire lifecycle, from training to maintenance and continuous improvement.

RESULTS AND DISCUSSIONS

The adoption of a comprehensive AI security framework and general security methods, such as Secure Software Development Lifecycle, access control and authentication mechanisms, privacy-enhancing technologies, network and infrastructure security measures, and AI-driven security and monitoring solutions, can help ensure that AI systems are developed and deployed responsibly.

This paper proposes a solution-scoring methodology for regulatory authorities to assess AI technologies based on their adherence to the discussed tools and security methods. This scoring proposal aims to provide regulators with a quantitative and qualitative evaluation mechanism to assess and monitor AI systems' security and ethical compliance, thereby promoting the responsible and safe use of AI technologies.

The proposed methodology for scoring an AI-based solution according to the presence or absence of generic security methods throughout its lifecycle can be summarised as follows: it starts with the identification of the essential security methods relevant to the AI system, then the evaluation occurs: first the assessment of the presence or absence of these methods during data collection and preparation, followed by the evaluation of the application of security methods during preparation and model evaluation. Further, the examination of the incorporation of security measures in the interpretation and debugging of the model and the verification of the implementation of security methods during the implementation and monitoring of the model occurs. After confirmation of implementing security methods in continuous improvement and maintenance, each method is assigned a binary score (1 for presence, 0 for absence) at each stage. Then a weighted average score is calculated based on the importance of each method, and the scores from all stages are summed to obtain an overall security score. At the final, compare the overall score with a predefined

threshold to determine compliance with security requirements.

This methodology can be considered adequate and a good start for potential regulators due to several factors. Firstly, its simplicity makes it easy to understand and implement. Based on binary outcomes, the scoring system allows for a straightforward evaluation of the presence or absence of security methods during each stage of the AI system's lifecycle. This ensures that even non-experts can comprehend and apply the methodology effectively. Secondly, the method's comprehensiveness ensures that the entire lifecycle of an AI model, from training to maintenance and continuous improvement, is considered. By assessing the application of security methods across all stages, the methodology provides a holistic view of an AI system's security posture, helping stakeholders identify potential gaps and vulnerabilities. Thirdly, weighted scores enable a more nuanced evaluation of the AI system's security. By assigning different weights to various security methods based on their importance, the methodology can better reflect the system's adherence to security requirements. This allows for a more accurate representation of the AI system's overall security compliance. Finally, the methodology's flexibility allows it to be adapted to different AI systems and security requirements. The methodology can be tailored to suit various AI applications' unique needs and challenges by adjusting the specific security methods evaluated and their associated weights. This ensures the scoring system remains relevant and effective in a rapidly evolving field.

The proposed methodology offers a simple, comprehensive, and flexible approach to scoring an AI-based solution based on implementing generic security methods throughout its lifecycle. Its ability to provide a holistic and nuanced evaluation of an AI system's security posture makes it an adequate and valuable tool for assessing and ensuring compliance with security requirements.

CONCLUSIONS

This solution scoring proposal provides a comprehensive methodology for regulatory authorities to evaluate AI technologies based on their adoption of a security framework and general security methods. By implementing this scoring system, regulators can ensure that AI systems are developed and deployed responsibly, adhering to security and ethical standards. Adopting this scoring proposal will increase trust in AI technologies and promote the responsible development and use of AI systems across various domains.

In the future, the proposed solution scoring methodology can be further developed and refined to address the evolving landscape of AI technologies and regulatory requirements. One area of focus is to expand the scoring criteria to cover emerging AI-related risks and ethical considerations, particularly those related to attacks specific to language models, such as prompt injection. This will ensure the scoring proposal remains current and relevant to the latest AI research and application developments. Developing specific guidelines and best practices for each criterion in the scoring proposal is another important aspect to consider. By providing AI developers and stakeholders with a clear roadmap for achieving compliance with the security framework and general security methods, we can facilitate better implementation of tools and solutions designed to prevent attacks like prompt injection in language models. Implementing a standardised reporting template for regulatory authorities to communicate the results of their scoring assessments is also crucial. This promotes transparency and consistency in the evaluation process and helps stakeholders understand the rationale behind the scoring results.

Integrating automated scoring tools and AI-driven assessment techniques into the evaluation process can streamline the scoring methodology and provide regulators with real-time insights into AI system performance and compliance. This could include developing tools that assess AI systems' resilience against prompt injection and other language model-specific attacks. A multi-faceted approach can be employed to promote the proposed framework to future AI technology regulators. Engaging with stakeholders, such as AI developers, policymakers, industry experts, and academics, is crucial for refining and validating the framework. This collaborative approach can help establish the framework as a robust and widely accepted tool for AI regulation.

Another important strategy is disseminating the framework and its benefits through academic journals, conferences, and other platforms. Targeting the AI research community and policymakers increases awareness and foster a broader discussion on the need for comprehensive AI regulation.

Furthermore, partnering with regulatory authorities to implement pilot programs that apply the framework to real-world AI systems provides valuable insights into the framework's effectiveness. It also allows for adjustments based on practical experiences, ensuring that the framework remains relevant and effective in addressing the unique challenges of AI regulation.

Establishing a collaborative platform for regulators, AI developers, and stakeholders to share best practices, experiences, and lessons learned in adopting the security framework and general security methods is also an important research idea for future research projects. This fosters a community-driven approach to responsible AI development and deployment, particularly emphasising addressing and mitigating risks associated with language model-specific attacks.

By continuing to refine and adapt the solution-scoring proposal in these ways, regulatory authorities can maintain a proactive stance in ensuring the responsible development, deployment, and management of AI technologies, particularly concerning the unique challenges posed by language models and associated security risks.

Funding a substantial European research project is necessary to advance the proposed technologies' maturity. This project will be a large-scale, interdisciplinary endeavour that unites experts from diverse fields such as computer science, AI, cybersecurity, law, ethics, and policy. The project's main goals are to determine the essential security and ethical prerequisites for AI systems and technologies, considering the distinct demands of various domains and AI applications.

The project will also focus on developing comprehensive specifications for the proposed tools, encompassing their functionalities, interfaces, and integration with existing AI development and deployment processes. By creating open-source reference implementations for these tools, the project will foster collaboration and encourage AI developers and practitioners to adopt the tools.

To guarantee the efficacy of the security framework and the tools linked to it, the project will undergo thorough validation with existing AI systems and applications. This will encompass case studies and pilot programs spanning various AI domains and sectors, such as healthcare, finance, transportation, and manufacturing. The project will also assess the economic, social, and environmental impacts, including the benefits of improved AI security and ethical compliance on society and the broader AI ecosystem. Regarding policy and standardisation, the project will engage with policymakers and international standardisation bodies to promote the adoption of the framework as a regulatory standard for AI technologies, both within the European Union and globally. Lastly, the project will emphasise dissemination and outreach by effectively communicating its findings and results to various stakeholders, including AI researchers, developers, policymakers, industry partners, and the general public. This will contribute to a more secure and ethically compliant AI landscape.

Finally, a key objective is working towards adopting the framework as an international standard for AI regulation. Engaging with international organisations and standardisation bodies can demonstrate the framework's value in addressing global AI security and ethical concerns. This will contribute to establishing a robust regulatory ecosystem that promotes trust, accountability, and transparency in the rapidly evolving world of AI.

ACKNOWLEGMENTS

This research is co-financed by the European Fund for Regional Development through the Competitiveness Operational Program 2014 – 2020, project “Establishment and implementation of partnerships for the transfer of knowledge between the Iasi Research Institute for Agriculture and Environment and the agricultural business environment”, acronym “AGRIECOTEC”, SMIS code 119611.

REFERENCES

Buiten M., 2019. - *Towards Intelligent Regulation of Artificial Intelligence. European Journal of Risk Regulation*, 10(1), 41-59, online available at: https://doi.org/10.1017/err.2019.8.

Cath C., 2018 - *Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A*: Mathematical, Physical and Engineering Sciences, 376(2133), online available at https://doi.org/10.1098/rsta.2018.0080.

Damgård V. Pastro, N. Smart and S. Zakarias, 2012 "Multiparty computation from somewhat homomorphic encryption," Crypto 2012, vol. Springer LNCS 7417, pp. 643-662, 2012.

Koene A., 2017 - *Algorithmic bias: addressing growing concerns* [leading edge]. IEEE Technology and Society Magazine, 36(2), 31-32. available online at https://doi.org/10.1109/MTS.2017.2697080.

Kuhn D. R., Coyne, E. J., & Weil, T. R. (2010). *Adding attributes to role-based access control.* Computer, 43(6), 79-81, online available at: http://dx.doi.org/10.1109/MC.2010.155

Leslie D., Burr, C., Cowls, J., Katell, M., Briggs, M., 2021 - *Artificial intelligence, human rights, democracy, and the rule of law: A Primer*, Chapters 3-5, 12-23, available online at https://arxiv.org/abs/2104.04147.

Mankins J., 2009 - *Technology readiness assessments: A retrospective, 3.1. TRL 1 Basic principles observed and reported,* available online at https://www.sciencedirect.com/science/article/abs/pii/S0094576509002008.

McCarthy M., 2020 - *Principle of Least Privilege (PoLP): What Is It, Why Is It Important, & How to Use It, available at: https://www.strongdm.com/blog/principle-of-least-privilege*

Nongbri I., Hadem, P., Chettri, S. 2018 - *A survey on single sign-on. Int. J. Creative Res. Thoughts,*

6(2), 595-602, online available at: http://dx.doi.org/10.5281/zenodo.5763157

Ometov A., Bezzateev, S., Mäkitalo, N., Andreev, S., Mikkonen, T., Koucheryavy, Y., 2018 - *Multi-factor authentication: A survey.* Cryptography, 2(1), 1, online available at: https://doi.org/10.3390/cryptography2010001.

Oseni A., Moustafa, N., Janicke, H., Liu, P., Tari, Z., Vasilakos A., 2021 - *Security and privacy for artificial intelligence: Opportunities and challenges.* arXiv preprint arXiv:2102.04661., available at https://doi.org/10.48550/arXiv.2102.04661.

Sinha A., Shrivastava G., Kumar, P., 2019 - *A pattern-based multi-factor authentication system. Scalable Computing: Practice and Experience*, *20*(1), 101-112, available at: https://doi.org/10.12694/scpe.v20i1.1460.