TRUST4AI.Security

Security testing of AI systems

The main goal of TRUST4AI.Security is to support model engineers (or data scientists) to assessing and ensuring their models are robust against known adversarial machine learning (AML) attacks. Therefore, the component supports the improvement of security of AI systems, and particularly of ML models, by providing mechanisms for the identification of potential AML attacks against the AI under study as well as potential mitigations that can be adopted to minimise their damage.

AI4CYBER

Component

In a context where the number of works studying adversarial machine learning threats is growing exponentially, TRUST4AI.Security is a means to automate the search for all possible attack techniques and protections relevant to the AI under study. Therefore, it significantly facilitates the task of risks assessment in AI-based systems in regard to the risks to the AI robustness and security.

Model engineers or data scientists can use the TRUST4AI.Security component to conduct thorough threat assessments of their AI models. The TRUST4AI.Security component focuses on threats to the software models rather than threats to the hardware infrastructure and platforms supporting the AI.

The TRUST4AI.Security frontend is a web application developed using Flask and JavaScript offering comprehensive access to the entire threat assessment process, from initial threat analysis to the results of adversarial testing.

The TRUST4AI.Security backend is implemented in Python and is responsible for executing comprehensive and dynamic AI model threat assessments. It operates through two main phases: (i) the threat analysis phase, during which a detailed examination of potential threats to the AI model is performed, and (ii) the adversarial testing phase, during which it security testing using adversarial examples to evaluate the robustness of the AI model is conducted. It simulates AML attack scenarios to uncover weaknesses and validate the effectiveness of implemented defences. The TRUST4AI.Security includes an AI Threats & Mitigations Knowledge Base (KB), representing a comprehensive knowledge corpus structuring state-of-the-art AML tactics, techniques and mitigations from the literature and industry sources such as MITRE ATLAS. Finally, the solution uses a general-purpose open-source LLM, Llama3.1-70B, to support the analysis of testing results and generate comprehensive and easy-to-understand reports.

YouTube Video Link

The solution is still a prototype and the future tool will have commercial license.

tecnalia
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Erkuden Rios (Erkuden.rios@tecnalia.com)
Fundación Tecnalia Research & Innovation.

www.tecnalia.com
https://es.linkedin.com/company/tecnalia-research-innovation
https://x.com/tecnalia

- E. Iturbe, E. Rios, A. Rego, and N. Toledo, “Artificial Intelligence for next generation cybersecurity: The AI4CYBER framework”, in Proceedings of the 18th International Conference on Availability, Reliability and Security, 2023, pp. 1–8.
- E. Iturbe, E. Rios and N. Toledo, "Towards trustworthy Artificial Intelligence: Security risk assessment methodology for Artificial Intelligence systems," 2023 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), Naples, Italy, 2023, pp. 291-297, doi: 10.1109/CloudCom59040.2023.00054.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450. Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.