Trustworthy AI Seminars

9th February 2023
15:00 - 18:00
Conference Room 322
DIBRIS

Maura Pintor
PostDoc @ UniCa

Where ML Security Is Broken and How to Fix It

Abstract. To understand the sensitivity under attacks and to develop defense mechanisms, machine-learning model designers craft worst-case adversarial perturbations with gradient-descent optimization algorithms against the model under evaluation. However, many of the proposed defenses have been shown to provide a false sense of robustness due to failures of the attacks, rather than actual improvements in the machine-learning models' robustness, as highlighted by more rigorous evaluations. Although guidelines and best practices have been suggested to improve current adversarial robustness evaluations, the lack of automatic testing and debugging tools makes it difficult to apply these recommendations in a systematic and automated manner. To this end, the analysis of failures in the optimization of adversarial attacks is the only valid strategy to avoid repeating mistakes of the past.

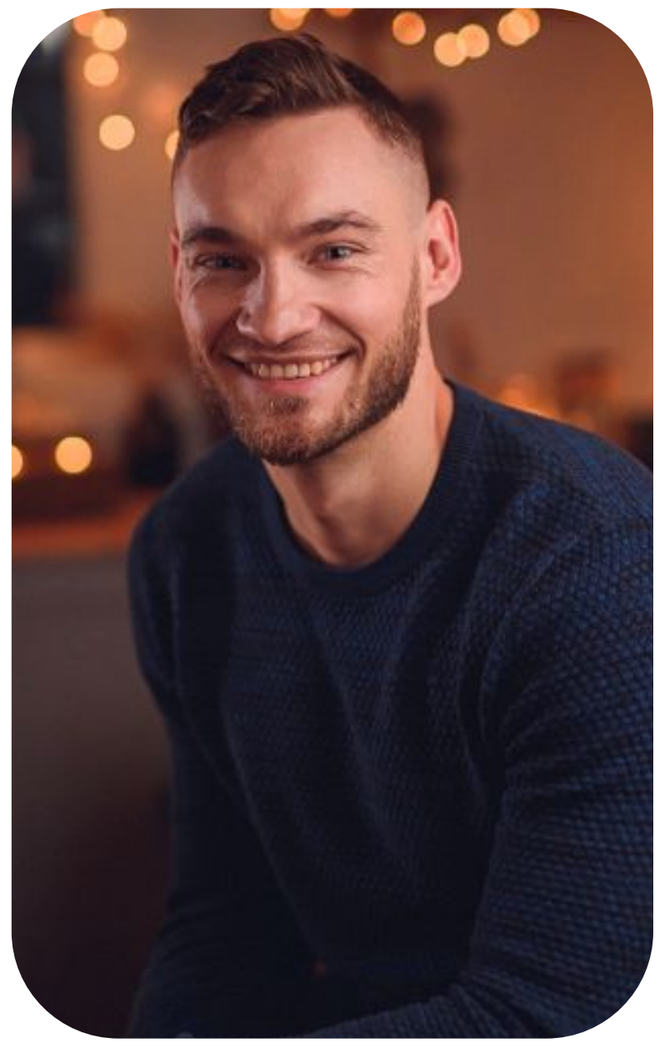
Biography. Maura Pintor is a Postdoctoral Researcher at the PRA Lab, in the University of Cagliari, and a Collaborator at the company Pluribus One. She received her PhD in Electronic and Computer Engineering from the University of Cagliari in 2022. Her thesis work provides a framework for optimizing and debugging adversarial attacks. She was a visiting student at Eberhard Karls Universitaet Tuebingen, Germany in 2020 and at the Software Competence Center Hagenberg (Austria), in 2021. She has collaborated with Pluribus One in the EU H2020 projects ALOHA and AssureMOSS.

GPT-like Pre-training for Behavioral Malware Detection

Abstract. This seminar will discuss challenges in modern information security, with areas of intrusion detection where conventional methodologies fail. We will uncover how advances in Artificial Intelligence (AI), specifically Natural Language Understanding (NLU), yielding successes of models like BERT or GPT, may help address information security needs. It will be examined how modern AI models can process technical languages like JSON on a specific example of dynamic malware analysis with Windows kernel emulation. Furthermore, extrapolation to a broader spectrum of infrastructure telemetry will be proposed, unveiling the potential of NLU techniques for cyber-security and machine languages in general.

Biography. Dmitrijs Trizna is a Ph.D. student at the University of Genoa. He has two Master's Degrees, one in Data Science received from the University of Helsinki (Finland), and one in Network Security received from the Riga Technical University (Latvia), with also 10 years of experience in commercial information security (both offensive and defensive). His research is published at scientific and industrial venues like CAMLIS, ACM CCS AISec, BlackHat and DefCon AIVillage. In addition, he holds multiple professional information security certifications like OSCP, SANS (GREM, GDAT), etc. Lastly, he held membership in NATO CCDCOE cybersecurity events.

Dmitrijs Trizna
PhD @ UniGe



Università di Genova | **DIBRIS** DIPARTIMENTO DI INFORMATICA, BIOINGEGNERIA, ROBOTICA E INGEGNERIA DEI SISTEMI

Questions about the event?
Contact the organizer:
Luca Demetrio
Assistant Professor @ UniGe
luca.demetrio@unige.it