PhD in Information Technology and Electrical Engineering
Università degli Studi di Napoli Federico II

PhD Student:

Cycle: XXXIX

Training and Research Activities Report

Year: First

Pellegrino Casoria

Tutor: prof. Simon Pietro Romano

Simon Pietro Romano

Date: October 31, 2024

1. Information:

- **PhD student: Pellegrino Casoria** **PhD Cycle: XXXIX**
- **DR number: DR997204**
- **Date of birth: 17/05/1988**
- **Master Science degree: Computer Engineering**
- **University: University of Naples Federico II**
- **Scholarship type: N/A**
- **Tutor: Simon Pietro Romano**
- **Co-tutor: N/A**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
SEC595 - Applied Data Science and AI/Machine Learning for Cybersecurity Professionals	Courses	36	4	03/11/2023 - 11/11/2023	SANS Institute - David Hoelzer	Y
Virtualization technologies and their applications	Courses	23	5	08/01/2024 – 26/02/2024	DIETI – Luigi De Simone	Y
Hands-on Network Intrusion Detection via Machine and Deep Learning	Courses	14	4	09/01/2024 – 26/02/2024	DIETI – Antonio Montieri	Y
Strategic Orientation for STEM Research & Writing	Courses	28	5	07/12/2023 – 23/02/2034	DIETI - Chie Shin Fraser	Y
Machine Learning & Cheshire Cat	Courses	16	2	03/04/2024 – 04/04/2024	Accenture / Cheshire Cat AI – Piero Savastano	Y
Introduction to AI and Leveraging it in Cybersecurity	Seminar	1	0.2	16/04/2024	SANS Institute - Lance Spitzner, Rob Lee	Y
SANS AI Cybersecurity Forum: Insights from the Front Lines	Seminar	3	0.6	25/04/2024	SANS - Rob Lee, David Hoelzer et al.	Y
IEEE Authorship and Open Access Symposium: Tips and Best Practices to Get Published from IEEE Editors	Seminar	1.5	0.3	07/05/2024	IEEE - Petar Popovski, Eszter Lukács, Judy Brady	Y
Regolazione in tema di	Seminar	5	1	13/05/2024	5G Academy	Y

Intelligenza Artificiale alla luce dell'AI Act					**- Elvira Raviele**	
Sustainable IT: strategies and best practices for a green engineering future	**Seminar**	**5**	**1**	**27/05/2024**	**5G Academy - Annalisa Di Leva et al.**	**Y**
Generative AI for software engineering: strategies, impacts, and practical applications	**Seminar**	**5**	**1**	**29/05/2024**	**5G Academy - Annalisa Di Leva et al.**	**Y**
GCFA – GIAC Certified Forensic Analyst	**Course**	**n.a.**	**4**	**01/07/2024 – 15/07/2024**	**SANS Institute – self-paced**	**Y**
NIS2 Directive Readiness: Compliance, Challenges and Recommendations	**Seminar**	**1**	**0.2**	**28/10/2024**	**SANS Institute – Bojan Zdrnja**	**Y**
2024 SANS Artificial Intelligence Solutions Track	**Seminar**	**3**	**0.6**	**29/10/2024**	**SANS Institute – Matt Bromiley**	**Y**
Dark Energy - Cyber Threats to Satellite Communications During Conflict	**Seminar**	**1**	**0.2**	**31/10/2024**	**SANS Institute – Justin Parker**	**Y**

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	**4**	**0**	**6**	**0**	**10**
Bimonth 2	**14**	**0**	**2**	**0**	**16**
Bimonth 3	**2**	**0.8**	**6.2**	**0**	**9**
Bimonth 4	**0**	**3.3**	**6**	**0**	**9.3**
Bimonth 5	**4**	**0**	**6**	**0**	**10**
Bimonth 6	**0**	**1**	**7**	**0**	**8**
Total	**24**	**5.1**	**33.2**	**0**	**62.3**
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

3. Research activity:

Cybersecurity threats and opportunities in the field of emerging AI Technologies

The rapid evolution of technological solutions, along with the growing frequency and complexity of cyberattacks targeting these technologies, demands ever-greater efforts to keep pace with the continuous shifting of global Cyber Threat Landscape. In this context, AI and Generative AI (GenAI) technologies presents both significant opportunities and challenges for cybersecurity, enabling implementation of new defensive capabilities, but also introducing new threats that attackers may exploit. For this reason, this year's research has focused on **analyzing AI paradigms from a security perspective**, with a

particular emphasis on **Large Language Models (LLMs) and GenAI**. On one hand, the study aims at identifying **advanced GenAI security applications** to enhance Security Operations and improve response efficacy; on the other, it analyzes **potential malicious applications** and threats introduced by these technologies to develop effective remediation strategies.

This study spans multiple application areas, which will be detailed in the following sections.

AREA 1: GenAI for Enhanced Security Operations

As the complexity and volume of cybersecurity threats continue to grow, Security Operation Centers (SOCs) require more efficient, automated solutions to stay ahead. This research area has been dedicated to **exploring how GenAI could be applied within SOCs** to improve threat detection, response times, and overall operational efficiency.

A key component of this work has been the study of **state-of-the-art for GenAI** through academic, industry knowledge and direct client conversations (in collaboration with Accenture). Significant efforts have been dedicated to analyze existing multi-agent architectures and model fine-tuning approaches – such as instruction-based prompting and Parameter Efficient Fine-Tuning (PEFT). This process has enabled standardization of skills, tasks, and knowledge for virtual LLM agents, ensuring alignment with industry frameworks like the NIST NICE Workforce Framework for Cybersecurity. Part of the study was also dedicated to **evaluating Retrieval-Augmented Generation (RAG)** approaches to reduce bias and hallucination phenomena, involving identification, evaluation, and generation of **security-relevant datasets** across different cybersecurity contexts. To support experimentation activities, it has been developed a **Multi-Agent lab environment** based on langGraph and integrated with different LLMs (including OpenAI and Llama), providing a controlled environment to explore advanced prompting techniques and to automate security workflows through chains of Virtual Agents specialized in different cybersecurity domains – implemented use cases include automated incident triage, Q&A with security virtual experts, policy and mitigation inference, enhanced security reporting.

Future research will focus on **engineering the current lab environment** to make it accessible to the broader research community, **extending security datasets** – also through synthetic data generation – and enabling **specialization of GenAI-driven virtual agents** in additional cybersecurity areas, such as incident response, vulnerability management, active deception, and cyber threat intelligence.

AREA 2: GenAI-powered Threat Landscape

This research area has focused on analyzing the growing use of AI and GenAI by malicious actors, investigating new attack strategies emerging from these technologies. On one hand, this has involved a deep **study of GenAI threat landscape** to understand how attackers are currently employing these tools; on the other hand, an experimental phase aimed at **replicating such attacks for research purposes**, with the goal of enhancing awareness and developing effective countermeasures.

The main research focus has been usage of **GenAI for advanced Social Engineering** attacks in relation to the evolving DeepFake technologies. This required an extensive analysis to identify trends, methodologies, and enabling tools for creating and detecting malicious artificial content, including video, audio, text, and code. This analysis has enabled the development of a general vision of DeepFake threats and potential solutions to mitigate user exposure. A laboratory environment has been setup to conduct **GenAI-powered phishing campaigns** specifically aimed at measuring user awareness on deepfake-related threats. This study has also generated a parallel analysis of tools and techniques to generate ransomwares using GenAI. This included a preliminary analysis of ransomware attacks –

particularly within energy sector – and explored AI-assisted malware generation approaches along with LLM jailbreaking techniques to bypass GenAI guardrails. The study identified the core components of ransomwares at code level, providing the basis for a laboratory activity integrating LangChain with OpenAI models to automatically **generate executable Python-based ransomware** files.

Future research will focus on exploring additional methods to generate different types of malicious contents in the context of offensive security, enabling extensive experimentation in real-world scenarios.

AREA 3: AI-powered robot inspection

Traditional security methods often fall short when extended to the physical world, where timely detection and hard-to-reach environments can compromise both safety and security. With reference to emerging regulatory frameworks – like NIS Directive, emphasizing robust protection for Critical Infrastructures – my research has focused on exploring **robot-based inspection paradigms integrated with AI** to provide continuous monitoring and minimize human exposure to hazardous environments.

Key aspects of the research included analysis of relevant **Threats Actors on energy sector** to design threat scenarios applicable to industrial physical security. The research has led to development of a prototype on Boston Dynamics' Spot robot, involving **integration with Computer Vision** algorithms on **manual and automated inspections** to detect unauthorized accesses and provide real-time centralized anomaly alerting. The research has also explored **integration with open-source LLMs** to provide command inference, speech-to-text conversion, and task reasoning, leveraging edge computing capabilities of industrial robots to minimize the human-machine interaction gap. Future developments involve consolidation of a **risk model for physical security** aligned with existing standards, and extended experimentation on **integrated security fleets** that incorporates diverse types of robots.

AREA 4: Trustworthy AI & LLM Security Performance

AI and LLMs introduce **new security risks** to be exploited by malicious actors – such as manipulation, data breaches, inherent biases, model poisoning and adversarial attacks – driving the need for robust security standards. This area of research aims at **consolidating trustworthy AI frameworks and approaches** that embeds security at their core, ensuring transparent, robust, ethical, and reliable systems, ultimately safeguarding users and organizations.

The activity involved an in-depth study of threats, attacks, and trends associated with AI tools, in addition to existing **frameworks and regulations for Responsible/Trustworthy AI** – such as NIST AI Risk Management Framework, MITRE ATLUS, OWASP Top 10 for LLM and EU AI Act. This study will represent an enabler for future research on a **unified AI Threat and Security framework**, aiming to become a foundation guideline for implementing security mitigations on AI-enabling platforms and to reduce attack surface throughout the entire AI lifecycle. The research has also included an analysis of potential **anonymization and tokenization approaches for LLMs**, investigating open-source Data Loss Prevention solutions, and defining a methodology to measure performance impacts of these tools on the efficacy of AI models, to be further developed in future works.

The research also focused on identifying metrics and indicators to effectively evaluate security **performances of LLMs**. The study explored various metrics, including deterministic, LLM-as-a-judge human-based approaches. The ongoing work aims at developing a **unified performance measurement model** specifically tailored to cybersecurity applications. This model will serve as a foundational tool for fine-tuning LLMs and evaluating their effectiveness across different cybersecurity use cases.

4. Research products:

- **(Prototype) Multi-Agent Cognitive Lab**: Prototype architecture based on LangGraph, RAG embedding with Qdrant vectorial DB and integration with different LLMs (including Llama and OpenAI), to automate security workflows through chains of Virtual Agents specialized in different cybersecurity domains.
- **(Prototype) The Dark Side of GenAI:** Prototype application to integrate LangChain with OpenAI models and automatically generate executable Python-based ransomware files. The application leverages LLM jailbreaking techniques to bypass guardrails for malicious purposes.
- **(Prototype) Spot the Security Threat:** Prototype of robot-based inspection integrated with Computer Vision to identify unauthorized access in controlled environments. It includes robot integration with edge LLMs to enable voice-based reasoning and interaction.
- **(Prototype) GenAI-powered Phishing:** Prototype tool to execute GenAI-powered phishing campaigns to assess level of user awareness on emerging deepfake threats.

5. Conferences and seminars attended

- (Faculty) The Dark Side of GenAI: Your first exploit with LLM, Red Hot Cyber Conference 2024, Rome, 19-20-Apr-2024.
- (Faculty) Spot the Cyber Threat: Robot-based inspection for industrial security, Red Hot Cyber Conference 2024, Rome, 19-20-Apr-2024.
- (Speaker) Phishing 2.0 - Unmasking Deep Fake Threats with GenAI, Cybertech Europe 2024, Rome, 8-9-Oct-2024.
- (Speaker) Industrial Security Reinvented: AI-powered robot inspection for anomaly detection in critical systems, Cybertech Europe 2024, Rome, 8-9-Oct-2024.
- (Speaker) Phishing 2.0 - Unmasking Deep Fake Threats with GenAI, Codemotion Conference 2024, Milan, 22-23-Oct-2024.
- (Attendee) Live Event: Briefing Pass, Black Hat USA 2024, Las Vegas, 7-8-Aug-2024.

6. Activity abroad:

N/A

7. Activity in partner companies:

Research activities have been conducted in partnership with **Accenture S.p.A**. Specifically, I am currently onboarded on Accenture “Security Cyber Resilience” team, responsible to support Clients on identification, detection, and monitoring of security threats on IT services and infrastructures. I am currently holding the position of Security Consulting Senior Manager, coordinating the “Cyber Next” group, with the objective to develop innovative propositions, enforce collaborations with innovation

stakeholders, and perform **R&D on emerging security topic**s. This includes research in all areas described in the previous sections (sec. 3).

During this first year, my activities in Accenture have focused on analyzing security needs across different markets, organizing **extended campaigns** on emerging security opportunities with potential Clients – through direct discussions as well as **participation in multiple conferences** and security events (sec. 5) – **exploring innovative solutions** to be further designed and developed. My work has involved **creation of many prototypes** and Proofs of Concept in partnership with academia, vendors, technology partners, and corporate R&D labs and research centers – including Accenture Labs in Tel Aviv and Sophia Antipolis, and Accenture Liquid Studios in Brussels.

Throughout these activities, Accenture has supported my research by providing access to its entire innovation ecosystem, named **Accenture Innovation Architecture**. This comprehensive framework has enabled support throughout the full research lifecycle – from market analysis to prototyping and large-scale engineering – providing specialized expertise in relevant security sectors, technologies and markets.

8. Tutorship

N/A