

AI-Enhanced Cybersecurity in Critical Infrastructures: A TRITON Framework and Review

Valeri Mladenov^{1,*}, Gergana Vacheva¹, Sophia Karagiorgou², Mariza Konidi², Evangelos Kafantaris², Ioannis Pastellas², Theodora Anastasiou², Angie Carillo³, Angeliki Vlachostergiou³, Alberto Mozo⁴, Esteban Martínez-Hoces⁴, Thomas Boufikos⁵, Panagiotis Radoglou-Grammatikis⁵, Luca Demetrio⁶

Technical University Sofia¹, UBITECH² EXUS³, Universidad Politécnica de Madrid⁴, K3Y⁵, University of Genoa⁶

valerim@tu-sofia.bg; gergana_vacheva@tu-sofia.bg; mkonidi@ubitech.eu; a.carrillo@exus.ai; a.vlachostergiou@exus.ai; a.mozo@upm.es; esteban.martinez.hoces@upm.es; tboufikos@k3y.bg; pradoglou@k3y.bg

Abstract: Artificial intelligence (AI) is reshaping the field of penetration testing by enabling faster, more adaptive simulations of cyber threats. This paper explores how AI can be ethically integrated into penetration testing processes, focusing on the European Defence Fund-backed TRITON project. TRITON proposes a comprehensive AI-driven framework for testing the security of military and critical infrastructure systems, combining technologies like machine learning, generative models, and reinforcement learning. The TRITON project reviews recent advancements in AI-supported pentesting and discusses how these tools can improve vulnerability detection, attack simulations, and threat modeling. Alongside the technical discussion, which is examined in the TRITON project, ethical concerns—including transparency, human oversight, and dual-use risks—are of the essence and must be addressed to ensure responsible use. Comparisons with other EU cybersecurity initiatives, such as AI4CYBER and CyberSecDome, highlight TRITON's unique contributions and focus areas. Ultimately, it argues that AI-enhanced penetration testing can significantly strengthen cybersecurity defenses when implemented with appropriate safeguards and ethical oversight.

KEYWORDS: ARTIFICIAL INTELLIGENCE, PENETRATION TESTING, CYBERSECURITY, TRITON PROJECT

1. Introduction

The integration of Artificial Intelligence (AI) into cybersecurity has opened new possibilities for both defensive and offensive strategies. In particular, penetration testing—traditionally a manual and time-consuming process—has seen significant transformation through automation and AI-based tools. These technologies can simulate realistic attack scenarios, identify vulnerabilities more efficiently, and adapt to evolving threat landscapes [1-3].

However, the growing reliance on AI in this context brings with it several important considerations. Ethical questions surrounding transparency, accountability, and unintended consequences are becoming more pressing, especially when AI systems are used to probe sensitive or critical infrastructure. As AI takes on a larger role in security testing, it is vital to ensure its application aligns with ethical standards and human oversight.

This paper explores the ethical implications of AI-driven penetration testing, focusing on the European Defence Fund's TRITON project. TRITON is designed to develop advanced cybersecurity solutions for military and defense-related systems, incorporating cutting-edge AI methodologies. The project's goals include enhancing the effectiveness of penetration testing, improving situational awareness, and ensuring that AI tools operate within a framework of trust and responsibility.

Our objectives in this review are threefold:

- To examine the technological advancements driving AI-supported penetration testing – including the use of machine learning, generative models, and reinforcement learning in simulating cyberattacks and detecting vulnerabilities.
- To analyze the ethical challenges associated with deploying AI in cybersecurity, particularly in contexts involving national security, where the risks of misuse or unintended consequences are heightened.
- To position the TRITON project within the broader ecosystem of European cybersecurity initiatives, highlighting its distinct approach and the measures it takes to ensure ethical compliance.

To fully understand TRITON's contribution to AI-driven cybersecurity, it's helpful to compare it with other EU-funded initiatives operating in the same domain. While many of these projects aim to advance the state of cybersecurity using artificial intelligence, each has its own focus and intended application area.

For instance, CyberSecDome is a project that leverages AI for real-time monitoring and protection in cyber-physical systems, with an emphasis on cloud infrastructure. It uses AI techniques such as anomaly detection, predictive analytics, and autonomous threat response. Unlike TRITON, which targets secure environments in the defense and critical infrastructure sectors, CyberSecDome is more focused on enhancing situational awareness in cloud and hybrid networks.

Another initiative, AI4CYBER, works on improving cyber situational awareness and decision support through AI-based technologies. AI4CYBER focuses heavily on data fusion, threat intelligence, and human-machine collaboration. While both TRITON and AI4CYBER share the goal of trustworthy AI integration, TRITON distinguishes itself through its specialized penetration testing tools and ethical framework aligned with EU defense priorities.

What sets TRITON apart is its dedicated emphasis on AI-based automated penetration testing within regulated, high-risk environments. Unlike broader cybersecurity frameworks, TRITON is tailored specifically for secure-by-design systems, incorporating generative models, reinforcement learning, and ethical.

Moreover, TRITON integrates deeply with the European Defence Fund's strategic objectives, reflecting a dual commitment to technological innovation and ethical oversight. While other projects may emphasize real-time threat mitigation or network visibility, TRITON addresses the full lifecycle of penetration testing—starting from vulnerability identification to attack simulation and risk analysis—within a tightly governed framework.

In this way, TRITON complements the broader ecosystem of EU cybersecurity initiatives while offering a distinct, focused contribution: ethically guided, AI-driven pentesting designed for use in mission-critical infrastructures.

The following document is structured as follows: Section 2 addresses the State of the art of Penetration Testing and Automation, Section 3 addresses the TRITON Project Overview, Section 4 addresses the TRITON Focus Areas for AI-Driven Penetration Testing, Section 5 addresses the Applications of AI in Penetration Testing: State of the Art, Section 6 addresses the Challenges and Ethical Considerations in AI-Driven Pentesting, and the final Section 7 presents the Conclusion and Future Directions.

2. State of the art of Penetration Testing and Automation

Penetration testing, often referred to as ethical hacking, is a vital component of modern cybersecurity strategies. It involves simulating cyberattacks to assess the strength of an organization's defenses and identify potential vulnerabilities before malicious actors can exploit them. Traditionally, penetration testing requires skilled security professionals to manually probe systems, a process that can be time-intensive and limited in scope.

The demand for faster and more scalable testing methods has led to the increased use of automation. Tools like Nessus¹, Metasploit², and Wireshark³ have become standard in many security operations. These tools offer features such as automated vulnerability scanning, exploit development, and network analysis, making it easier to perform repetitive tasks and analyze large systems more efficiently. However, they still require human expertise to interpret results, fine-tune scans, and design attack scenarios.

Automation in penetration testing has significantly advanced with the integration of Artificial Intelligence (AI) and Machine Learning (ML). These technologies empower the creation of intelligent systems capable of learning from historical penetration data, adapting to novel threat environments, and autonomously generating sophisticated attack strategies. In particular, AI encompasses a broad spectrum of approaches, including Generative Adversarial Networks (GANs), which have been explored for generating realistic payloads or simulating adversarial traffic, and Reinforcement Learning (RL), which enables agents to iteratively learn optimal exploitation strategies by interacting with dynamic environments. These techniques complement traditional rule-based or signature-driven methods by enhancing adaptability and decision-making in complex, uncertain scenarios. AI-driven tools are especially valuable for behavioral analysis, anomaly detection in large-scale networks, and real-time vulnerability assessment, offering improved scalability, efficiency, and precision over classical automated testing frameworks. Despite these advances, full automation remains a challenge. Current AI systems can assist penetration testers—as seen with tools like PentestGPT, which provides interactive guidance to professionals—but they are not yet capable of replacing human intuition, contextual understanding, or ethical reasoning. Even though emerging research into frameworks like PentestAgent aims for a more fully autonomous approach, concerns remain about the reliability, transparency, and control of AI systems, especially in high-stakes environments such as government or critical infrastructure networks.

Several EU-backed projects, just like TRITON is addressing this gap by combining automation with human oversight to ensure safe and responsible AI deployment. These initiatives aim to balance the efficiency gains enabled by machine learning with the essential requirements of accountability, transparency, and stakeholder trust. In this context, the TRITON project stands out for its emphasis on ethically guided AI integration in penetration testing scenarios, particularly in defense-related domains. A key principle underpinning TRITON's approach is the adoption of human-in-the-loop (HITL) mechanisms, where human experts systematically review and validate AI-driven actions before execution—especially in high-risk environments. Additionally, the concept of human-on-the-loop is considered, whereby human operators monitor AI behavior and intervene when anomalies or unintended consequences arise. These human-centered control structures are increasingly recognized as best practices in the design of trustworthy and auditable AI systems. [4-6].

3. TRITON Project Overview

The TRITON project is a European Defence Fund initiative aimed at strengthening cybersecurity capabilities across Europe's defense landscape. Its central mission is to develop and implement advanced tools for AI-driven penetration testing that are not only effective but also ethically responsible and compliant with European regulations. TRITON stands for "TRusTworTHY AI for the auTomedated identiFicatiON of cybersecurity threats", reflecting its focus on building trustworthy and intelligent systems.

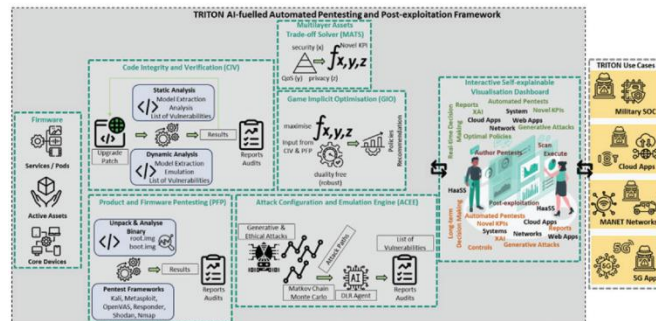


Fig.1. TRITON Conceptual Architecture[7]

Unlike traditional penetration testing tools, TRITON is designed to operate in sensitive environments such as military, defense, and other critical infrastructure systems. This raises the stakes considerably, requiring the tools to meet strict standards for safety, reliability, and ethical accountability.

At its core, TRITON seeks to automate the process of discovering and analyzing cybersecurity vulnerabilities using a combination of AI methods, including:

- Machine Learning (ML): For pattern recognition and behavioral analysis of networks and systems.
- Reinforcement Learning (RL): To train models that adapt and evolve based on simulated attack outcomes.
- Generative AI: To create synthetic attack and regular traffic scenarios and test how well systems respond to novel threats.

A distinguishing feature of TRITON is its emphasis on building "human-in-the-loop" systems, ensuring that AI tools are guided by expert judgment and ethical oversight. This approach helps mitigate the risks of over-automation and ensures that the results of AI-driven testing are interpretable, actionable, and aligned with human values.

The project also includes a modular platform architecture, allowing different tools, engines and AI modules to be integrated and customized depending on the operational context. This flexibility supports a wide range of testing environments, from enterprise IT systems to mission-critical defense infrastructure.

TRITON is part of a broader movement in the EU to establish "Trustworthy AI"—systems that are secure, fair, transparent, and aligned with fundamental rights. The project's outcomes are expected to contribute significantly to both the technological and ethical standards[10] for future AI applications in cybersecurity.

4. TRITON Focus Areas for AI-Driven Penetration Testing

The TRITON project introduces a structured, multi-phase areas for integrating AI into penetration testing, specifically used for high-security environments such as military and defense systems. This areas are designed to enhance the effectiveness of traditional testing approaches while addressing the ethical and operational challenges posed by autonomous systems.

¹ <https://www.tenable.com/products/nessus>

² <https://www.metasploit.com/>

³ <https://www.wireshark.org/>

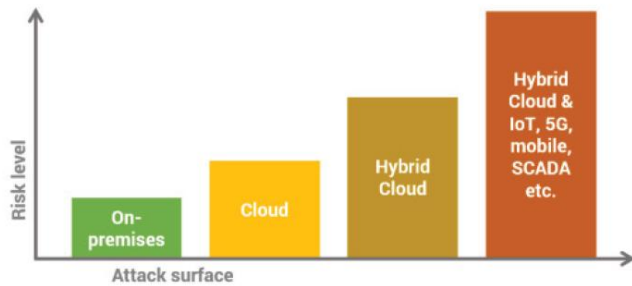


Fig.2. Cyberspace Trends in Networks, Systems, IoT-Cloud, and 5G infrastructures[7]

4.1 Automated Discovery of Vulnerabilities

At the initial stage, TRITON focuses on automatically identifying vulnerabilities within a system. This is achieved through advanced AI techniques like machine learning and pattern recognition. Unlike traditional scanners that rely on known vulnerabilities, TRITON's approach enables the system to detect novel or context-specific threats by analyzing traffic behavior, system configurations, and historical security data.

The platform combines signature-based detection with anomaly-based learning to improve accuracy. It adapts over time, using feedback from previous scans to fine-tune its models and reduce false positives. This capability is especially important in complex or evolving infrastructures, where static tools often fall short.

4.2 Simulation of Realistic Cyberattacks

Once vulnerabilities are identified, TRITON employs generative AI models to simulate realistic attack scenarios. Specifically, the system utilizes Large Language Models (LLMs) as a traffic generator, capable of creating distinct traffic types—such as realistic DDoS attacks—on demand. This allows the framework to synthesize network flows entirely from scratch without needing prior traffic samples, resulting in highly authentic connections. These simulations allow security teams to test system resilience under conditions that mirror actual cyber threats, including zero-day exploits and sophisticated multi-vector attacks.

The AI models are trained on a wide variety of attack techniques and behaviors, enabling them to generate dynamic and unpredictable test cases. This helps uncover weaknesses that might not be detected through conventional means. Importantly, these simulations are conducted in controlled environments to avoid unintended consequences or real-world disruptions.

4.3 Adaptive Learning and Reinforcement

A key innovation in TRITON's purposes is the use of reinforcement learning. Here, AI agents learn to navigate systems and develop attack strategies through trial and error, guided by reward signals based on their success in exploiting vulnerabilities. This allows for more intelligent and adaptive penetration tests that can mimic human decision-making.

However, reinforcement learning also introduces risks—such as overly aggressive or unsafe behaviors—if not carefully monitored. TRITON addresses this by integrating constraints and human oversight into the learning loop. Ethical safeguards ensure that AI agents operate within clearly defined limits, maintaining control and safety.

4.4 Reporting and Human Oversight

Throughout the process, TRITON emphasizes transparency and human involvement. Results are presented in a clear, interpretable format, allowing cybersecurity professionals to assess findings and

take corrective actions. Analysts remain in control of the decision-making process, using AI as a tool rather than a replacement.

This “human-in-the-loop” approach ensures accountability and prevents the misuse of autonomous systems. By keeping experts actively engaged, TRITON reinforces trust in AI-driven penetration testing while maintaining high ethical standards.

5. Applications of AI in Penetration Testing: State of the Art

AI has begun to reshape the landscape of penetration testing by introducing automation, adaptability, and advanced data analysis. While traditional pentesting depends heavily on manual processes, AI allows for a more scalable and dynamic approach to identifying and exploiting system vulnerabilities. As cyber threats become more complex, AI tools provide a way to respond quickly and intelligently.

One of the most significant developments in this area is the use of machine learning (ML) to analyze large volumes of network traffic, system logs, and behavioral data. By learning patterns of normal and abnormal activity, ML models can detect subtle anomalies that might indicate a security weakness or an active threat. This predictive capability improves the accuracy and efficiency of vulnerability assessments.

Reinforcement learning (RL) is another AI technique gaining traction in the pentesting field. It enables systems to learn through trial-and-error interactions, gradually improving their ability to simulate attacker behavior. RL agents can navigate complex environments, test different exploitation paths, and adapt strategies based on real-time feedback. This makes them particularly useful for discovering multi-step vulnerabilities or weaknesses in layered security defenses.

Generative AI models, such as Generative Adversarial Networks (GANs), offer a different kind of capability. These models can create realistic synthetic data, such as attack payloads or exploit scripts, which can be used to test system resilience under diverse scenarios. Generative models can also simulate variations of known attacks, helping organizations prepare for future threats that may not yet be documented.

Despite these advances, current AI systems still rely on human oversight. While they excel at processing data and running automated tasks, they often lack contextual understanding or the ability to make nuanced ethical decisions. For example, an AI system may identify a vulnerability but be unable to judge whether exploiting it in a particular context is safe or appropriate. Human expertise remains essential for interpreting results, validating findings, and making informed decisions.

A number of research initiatives [7-8] and tools have begun to integrate these AI techniques into operational workflows. Some focus on automating routine scanning and reconnaissance tasks, while others explore more advanced applications such as attack path generation, behavioral modeling, or proactive threat hunting. Collectively, these developments represent a shift toward more intelligent and autonomous penetration testing environments.

In sum, the state of the art in AI-driven penetration testing reflects a balance between automation and oversight. While AI offers significant gains in efficiency and coverage, the most effective systems still require collaboration between algorithms and skilled professionals. The TRITON project builds on these foundations, combining leading-edge AI methods with a strong commitment to ethical and secure deployment.

As AI continues to advance and integrate into cybersecurity operations, its role in penetration testing brings both powerful capabilities and serious ethical concerns. While AI can enhance efficiency and simulate complex attack behaviors, its deployment in sensitive domains—especially in military and critical infrastructure—demands careful scrutiny.

One of the central challenges is ensuring transparency. AI models, particularly those based on deep learning, often function as “black boxes,” making it difficult to understand how they arrive at certain conclusions or decisions. In the context of penetration testing, this lack of interpretability can pose risks—such as triggering unintended system disruptions or overlooking contextual nuances that a human tester might catch.

Accountability is another major concern. When an AI system autonomously performs penetration testing, it raises questions about who is responsible for any negative outcomes. Unlike human testers, AI tools may act unpredictably or make decisions that are difficult to trace back to specific design choices or user inputs. Assigning responsibility becomes even more complex in collaborative environments involving multiple stakeholders, such as defense contractors and government agencies.

There is also the issue of bias and fairness in AI algorithms. If training data used for AI-based pentesting does not represent a wide range of systems or threat models, the resulting tools may be less effective—or even harmful—when applied in unfamiliar contexts. For instance, a tool trained primarily on enterprise networks may fail to identify vulnerabilities in industrial control systems or embedded military platforms.

The dual-use nature of AI tools further complicates ethical deployment. Technologies developed for defensive penetration testing can potentially be misused for offensive purposes. Ensuring these tools are used responsibly—and not repurposed for unauthorized attacks—requires strict access control, legal oversight, and clear boundaries for acceptable use.

The TRITON project addresses these concerns by aligning its development with the European Union’s Ethics Guidelines for Trustworthy AI. These guidelines emphasize principles such as human oversight, robustness, privacy protection, and societal well-being. Within TRITON, ethical safeguards are embedded into both the design of AI tools and their operational use, ensuring that automation does not come at the expense of accountability or safety.

By maintaining a strong focus on “human-in-the-loop” governance, TRITON ensures that AI-enhanced testing remains interpretable and controllable. Human experts are kept involved in critical stages of analysis, decision-making, and validation. This balanced approach promotes both innovation and responsibility, allowing AI to enhance cybersecurity without undermining the ethical standards that are crucial in high-risk environments.

8. Conclusion and Future Directions

Artificial intelligence is rapidly transforming the field of penetration testing, offering new opportunities for automating threat discovery, simulating complex attack scenarios, and improving security analysis. The TRITON project demonstrates how AI can be thoughtfully integrated into cybersecurity practices, especially in environments where security and accountability are critical, such as defense and critical infrastructure.

In this manuscript, we have presented some conceptual architecture (Figure 1) behind TRITON, highlighting how it blends cutting-edge AI techniques—such as reinforcement learning and generative models—with a strong commitment to ethical design. Unlike many generic automation tools, TRITON is purpose-built to ensure human oversight, transparency, and alignment with EU principles for trustworthy AI.

We have also placed TRITON in the context of similar European initiatives, showing that while there is a growing ecosystem of AI-enabled cybersecurity projects, TRITON stands out for its specialized focus on automated penetration testing within ethically sensitive environments.

Looking ahead, the continued development of AI in cybersecurity must overcome several practical and strategic challenges. These include ensuring model interpretability, managing dual-use risks, preventing algorithmic bias, and maintaining meaningful human oversight. Within TRITON, these challenges are addressed through actionable mechanisms such as the use of AI Tool Registers, predefined Rules of Engagement, and ethics checkpoints throughout the testing pipeline.

Future directions are likely to involve the integration of explainable AI (XAI) techniques into AI-assisted pentesting tools, enabling human analysts to understand and audit the reasoning behind attack simulations and detection decisions. Additionally, more granular access controls, real-time risk scoring, and adaptive response strategies—driven by reinforcement learning—may be introduced to better contain potential misuse. Projects like TRITON also highlight the importance of developing sector-specific regulatory templates that align AI capabilities with GDPR, NIS2, and AI Act requirements.

Ultimately, while AI is not expected to replace human security professionals, it will play a transformative role in scaling threat detection, simulating novel attack paths, and augmenting human-led investigations. With properly enforced safeguards and governance models, AI-enhanced penetration testing can become an essential layer in defending critical infrastructure against advanced persistent threats.

Acknowledgements:

This work has received funding from the European Defence Fund 2022 Programme (EDF-2023-RA-SI) under Grant Agreement No. 101168103. 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 granting authority can be held responsible for them.

9. References

- Al-Sinani, H. S., & Mitchell, C. J. (2025). *PenTest++: Elevating Ethical Hacking with AI and Automation*. arXiv preprint arXiv:2502.09484. <https://arxiv.org/abs/2502.09484>
- Ariadna Moreno, A., Sánchez, J., & Luna, J. (2025). *Analysis of Autonomous Penetration Testing Through Reinforcement Learning and Recommender Systems*. *Sensors*, 25(1), 211. <https://doi.org/10.3390/s25010211>
- European Commission. (2022). *Trustworthy Artificial Intelligence for Cybersecurity Reinforcement and System Resilience (AI4CYBER)*. CORDIS Project ID 101070450. <https://cordis.europa.eu/project/id/101070450>
- TRITON Consortium. (2025). *Generative Automation of Security Penetration Tests – Project Website*. <https://triton-edf.eu>
- UBITECH. (2024, December 4). *UBITECH Hosts Kick-off of TRITON EDF Action Project to Revolutionize Automated Security Penetration Testing*. <https://ubitech.eu>
- Jabir, R., Le, J., Nguyen, C., Öner, U., & Ashour, M. (2024). *Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors*. *AI*, 6(8), 174. <https://doi.org/10.3390/ai6080174>
- TRITON Project Consortium. *TRITON Project Partners Overview*. Available online: <https://triton-edf.eu/triton-consortium/>
- Ghanem, A., Ali, H., & Wang, Y. (2023). *Hierarchical Reinforcement Learning for AI-Driven Cybersecurity Attack Path Discovery*. *Computers & Security*, 130, 102868. <https://doi.org/10.1016/j.cose.2023.102868>
- Fernandes, D. A., Rodrigues, J. J., & Kumar, N. (2024). *Using Reinforcement Learning to Simulate Advanced Persistent Threats*. *Journal of Network and Computer Applications*, 228, 103487. <https://doi.org/10.1016/j.jnca.2023.103487>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). *Generative*

- Adversarial Nets*. Advances in Neural Information Processing Systems, 27, 2672–2680.
10. Doshi-Velez, F., & Kim, B. (2017). *Towards a Rigorous Science of Interpretable Machine Learning*. arXiv preprint arXiv:1702.08608. <https://arxiv.org/abs/1702.08608>
11. ENISA. (2021). *Artificial Intelligence Threat Landscape*. European Union Agency for Cybersecurity. <https://www.enisa.europa.eu/publications/artificial-intelligence-threat-landscape>
12. High-Level Expert Group on AI. (2019). *Ethics Guidelines for Trustworthy AI*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
13. Shrestha, R., Mahmood, A., & Hu, J. (2021). *Review of Deep Learning Algorithms and Architectures for Cybersecurity Applications*. IEEE Access, 9, 29684–29713. <https://doi.org/10.1109/ACCESS.2021.3058794>
14. Sommer, R., & Paxson, V. (2010). *Outside the Closed World: On Using Machine Learning for Network Intrusion Detection*. IEEE Symposium on Security and Privacy, 305–316. <https://doi.org/10.1109/SP.2010.25>
15. CyberSecDome Consortium. (2025). *CyberSecDome Project Overview and News Release*. EIT Digital. <https://28digital.eu/news/cybersecdome-eit-cybersecurity-platform>
16. Deng, G., Liu, Y., Mayoral-Vilches, V., Liu, P., Li, Y., Xu, Y., Zhang, T., Liu, Y., Pinzger, M., & Rass, S. (2024). *PentestGPT: An LLM-empowered Automatic Penetration Testing Tool*. <https://arxiv.org/abs/2308.06782>
17. Shen, X., Wang, L., Li, Z., Chen, Y., Zhao, W., Sun, D., Wang, J., & Ruan, W. (2025). *PentestAgent: Incorporating LLM Agents to Automated Penetration Testing*. <https://arxiv.org/abs/2411.05185>