A Framework for Proactive Cyber-Resilience: Non-Intrusive Modeling for Autonomous Defense

Vincenzo Sammartino
National PhD Program in Artificial Intelligence and Society
Università di Pisa - Dipartimento di Informatica
Pisa, Italy
vincenzo.sammartino@phd.unipi.it

***Abstract*—Modern cyber-defense paradigms exhibit a critical epistemological deficit, operating in a reactive posture that is fundamentally outpaced by the operational tempo of automated threats. This research addresses this challenge by proposing a novel, end-to-end framework for proactive cyber resilience. The core problem addressed is the systemic scarcity of timely, contextually relevant data required to model intrusions and, critically, to validate the effectiveness of countermeasures in-silico before live deployment. An architecture is introduced that forges a closed-loop system for autonomous defense evaluation. By integrating a non-intrusively generated Digital Twin of a target infrastructure with an Attacker Twin, the system runs large-scale Monte Carlo simulations to perform a stochastic exploration of the adversarial state-space. The result is a high-fidelity, probabilistic model of the intrusion landscape, termed the Intrusion Graph. The central thesis is that this graph serves not merely as a static analytical instrument, but as the engine for a proactive resilience pipeline. This pipeline transforms cybersecurity from a reactive forensic discipline into a predictive, data-driven science capable of automated hypothesis testing, optimal countermeasure selection, including deception technology, and ultimately autonomous remediation.**

***Index Terms*—Proactive Security, Cyber Resilience, Autonomous Defense, Digital Twin, Distributed Simulation, Monte Carlo Methods, Deception Technology, Computational Epistemology.**

I. Introduction: The Imperative for a Proactive Epistemology

The contemporary landscape of cyber conflict is characterized by a profound asymmetry in operational tempo: adversaries leverage automation to execute attack campaigns at machine speed, while defenders are largely constrained to human-speed reaction and forensic analysis. This implies that the prevailing cybersecurity paradigm, rooted in a *reactive* posture, is structurally inadequate. It forces defenders into a perpetual cycle of catch-up, creating an untenable strategic disadvantage. This research pursues a fundamental shift towards a *proactive* security epistemology, one that seeks to anticipate, model, and neutralize threats before they manifest in a live environment. The central thesis of this work is that by leveraging large-scale distributed simulations on a high-fidelity digital replica of an infrastructure, it is possible to forge an autonomous defense loop (Fig. 1). The resulting approach is designed not merely to respond to threat attacks, but to actively hunt for its own weaknesses and automate its own defense, thereby transforming cybersecurity from a forensic, a posteriori art into a predictive, data-driven science.

II. The Reactive Quagmire and the Data Crisis

The urgency for a proactive framework stems directly from the systemic failures of the reactive model. Currently, security operations are often analogous to digital archaeology: sifting through the remnants of a breach to reconstruct the event sequence. This post-mortem approach is untenable against modern threats. Governmental bodies and security experts now operate under the assumption that a catastrophic cyber-attack on Critical National Infrastructure (CNI) is a matter of "when, not if" [1]. This reality demands a transition to genuine cyber resilience — a system's capacity to *proactively* anticipate, withstand, and adapt to adverse events [2], [3].

This transition is impeded by a dual crisis of data. First, a profound *scarcity of relevant data* exists because detailed logs of sophisticated intrusions are highly sensitive and seldom shared, creating a critical knowledge vacuum [1]. Second, available data is subject to rapid value decay due to *conceptual drift* in the adversarial feature space. The relentless pace of vulnerability discovery and the continuous evolution of attacker Tactics, Techniques, and Procedures (TTPs) mean that historical datasets quickly become obsolete [4], [5]. An AI model trained on data from last year is likely unprepared for the threats of today. This forces organizations into a perpetually reactive loop: waiting for an attack, analyzing the damage, and then applying a patch, always remaining one step behind the adversary.

III. A Non-Intrusive Framework for Infrastructure Modeling

A credible proactive strategy must be founded upon a high-fidelity, empirically-grounded model of the system it aims to protect. Its accuracy is fundamental to the quality of the results and the decisions that the results support.

A. The Security Twin: A Dynamic Digital Replica

The cornerstone of the proposed framework is the **Security Twin** (G_{Inf}), a dynamic and enriched digital replica of the target ICT/OT infrastructure [6]. A key innovation is its automated and, critically, *non-intrusive* construction. To this end, a prototype platform, NotLine, has been developed to build the

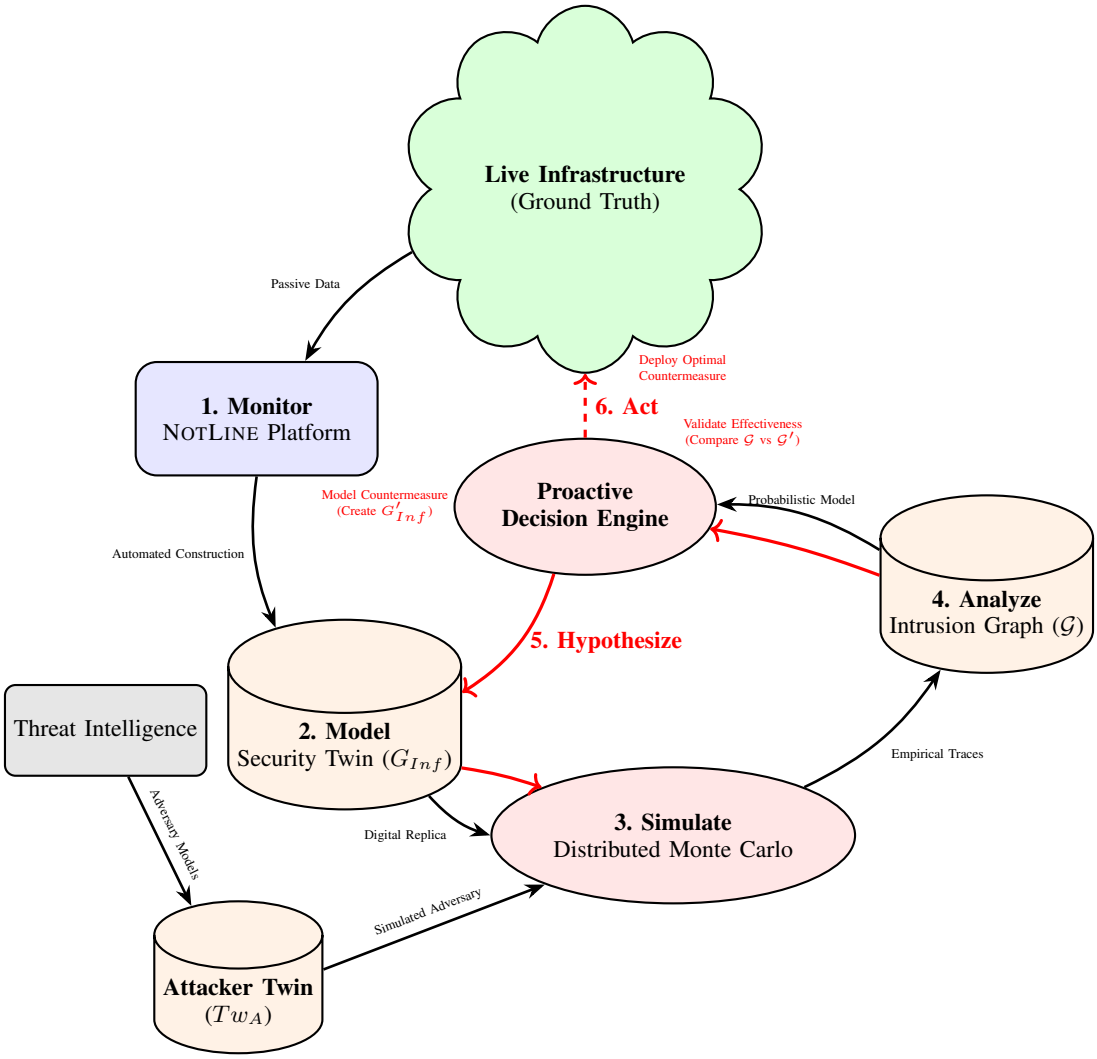

Fig. 1. The proposed **Proactive Resilience Loop**. This architecture operationalizes a continuous, data-driven cycle for automated defense, moving beyond a linear analysis pipeline. The integration of the **Attacker Twin**, fed by threat intelligence, is crucial for realistic simulations. The system is designed to *proactively* anticipate threats, validate defenses through in-silico experimentation, and automate responses, thereby breaking the cycle of reactive cybersecurity.

Security Twin by passively monitoring network traffic metadata from a wide array of protocols (e.g., ARP, DHCP, mDNS) [7]. This passive strategy is essential for live operational environments, particularly sensitive Industrial Control Systems (ICS), as it obviates the risks of destabilization or alert fatigue associated with active scanning [2]. Experiments validate this approach, showing that the node discovery process in a live network closely fits a hypo-exponential decay model [8], a pattern also observed in complex network phenomena such as worm propagation [9], which confirms the robustness of the mapping methodology.

B. The Attacker Twin: Formalizing Adversarial Rationality

To simulate realistic adversarial behaviour, the Security Twin is paired with the **Attacker Twin** (Tw_A), a formal, configurable abstraction of an attacker [2]. This is not a generic vulnerability scanner, but a sophisticated model of a goal-oriented adversary. It is defined by strategic objectives (e.g., data exfiltration, ransomware deployment) and a tactical repertoire of TTPs, often derived from frameworks like MITRE ATT&CK. This allows for the *proactive* simulation of specific, relevant threat actors and the analysis of how the infrastructure would perform against the adversaries it is most likely to face. The Attacker Twin's behavior is governed by a formal grammar of actions, each with defined preconditions, post-conditions, and success probabilities, enabling the modeling of diverse attacker rationalities, from risk-averse to opportunistic.

IV. Stochastic Sampling via Monte Carlo Simulation

With the digital battlefield constructed, a *Monte Carlo method* is employed for systematic, stochastic sampling—that is, exploration—of the state space of possible intrusions. By running thousands of independent, parallel simulations, a statistically significant sample of attack narratives is gathered. This method is inherently parallelizable and allows for the

proactive discovery of not only the most likely attack vectors but also rare, low-probability "black swan" events with potentially catastrophic consequences.

This approach overcomes the well-known limitations of classic deterministic attack graphs, which suffer from state-space explosion [10] and fail to capture the probabilistic nature of real-world attacks. The raw output—a collection of thousands of individual attack traces—is then merged into a single, unified data structure: the **Intrusion Graph ($\mathcal{G}$)** [4]. This directed acyclic multigraph constitutes a rich, probabilistic map of the adversarial landscape. Its nodes represent unique states of compromise ($n\langle P, I\rangle$), an expressive model capturing nuanced privilege escalation far more accurately than the binary "compromised/not-compromised" states of traditional Bayesian Attack Graphs (BAGs) [11], [12]. The edges are annotated with the actions that cause state transitions and, crucially, their empirical frequencies, which serve as data-driven probabilities. The entire generation process is ruled by strict logical consistency checks, ensuring every simulated path is a feasible attack sequence and guaranteeing the absence of false positives in the final model [13].

V. Self-Healing, Data-Driven Cyber-Resilience

The main innovation of this research lies in leveraging the Intrusion Graph not as a static analytical artifact, but as the dynamic core of the **Proactive Resilience Loop** (Fig. 1). This loop operationalizes the scientific method for cybersecurity, creating a self-healing system that autonomously identifies its own weaknesses and validates potential remediations.

Baseline Posture Analysis. The cycle starts with the generation of a baseline Intrusion Graph, $\mathcal{G}_{base}$, representing a detailed snapshot of the infrastructure's current security posture. The **Proactive Decision Engine** applies advanced analytical techniques to this graph. Using random walk-based centrality measures, which are ideally suited to this stochastic model [14], the engine automatically identifies critical chokepoints—nodes or edges that are pivotal for the most frequent, rapid, or damaging attack paths [6].

Automated Hypothesis Generation and In-Silico Experimentation. Based on this analysis, the engine formulates a testable hypothesis, such as: "Deploying countermeasure X will result in a Y% reduction in risk." It then models this countermeasure by creating the resulting updated Security Twin, G'_{Inf}. This "what-if" scenario is then tested by re-running the entire simulation on the new Security Twin, generating a new Intrusion Graph, $\mathcal{G}_{new}$. This constitutes a safe, automated experiment conducted entirely within the digital domain.

Quantitative Validation and Policy Optimization. The framework's final step is to act as a computational arbiter. It implements a differential analysis, quantitatively comparing $\mathcal{G}_{base}$ and $\mathcal{G}_{new}$ to produce a concrete, evidence-based "resilience impact score." This score is a user-defined vector of metrics—such as the change in Mean Time To Compromise (MTTC), the reduction in successful attack path probability, and the elimination of chokepoints—that provides an unambiguous measure of the countermeasure's effectiveness. This allows the proposed framework to *proactively* rank multiple potential countermeasures and recommend an optimal defensive policy. The final, dashed arrow in the loop represents the ultimate goal: a system with sufficient trust to automatically deploy the validated countermeasure, thus closing the loop from detection to remediation without human intervention.

VI. Proactive Defense Against AI Agents

An interesting use case for the framework is addressing the emerging threat of AI agents, i.e., AI-powered autonomous malware. Such agents can dynamically adapt their TTPs, discover novel attack paths, and operate at a speed and scale that renders human-in-the-loop defense obsolete. The framework defines a methodology to proactively model and manage these threats before they are encountered in the wild.

The process involves treating the adversary not as a static entity, but as a learning agent:

- **Modeling the Adversary as an RL Agent:** The Attacker Twin is configured not with a fixed set of TTPs, but as a Reinforcement Learning (RL) agent. This AI-powered twin is given a goal (e.g., compromise a specific data server) and learns an optimal attack policy by interacting with the Security Twin simulation environment.
- **Simulation as Adversarial Training:** The Monte Carlo simulation phase becomes a large-scale training process for the adversarial RL agent. Thousands of parallel simulations serve as training episodes, allowing the agent to explore the Security Twin and learn the most effective sequences of actions to achieve its objective.
- **Analyzing the Learned Policy:** The resulting Intrusion Graph is no longer just a map of possible static paths; it is a representation of the *learned optimal policy* of the adversarial AI. The most frequent paths in the graph represent the strategies the agent found most rewarding. This reveals the "cognitive chokepoints" and biases of the offensive AI.
- **Selecting AI-Disrupting Countermeasures:** The Proactive Decision Engine can now formulate hypotheses specifically designed to disrupt the AI's learning process or exploit its learned policy. For example:
  - *Reward Hacking:* "By placing a high-interaction honeypot that mimics a critical system, a 'reward trap' can be created. The AI agent will likely waste significant computational resources attempting to exploit this decoy, increasing its learning time and raising its probability of detection."
  - *Increasing State-Space Complexity:* "Patching a specific vulnerability identified as a key pivot in the AI's learned policy will force the agent to explore a much larger, less efficient part of the state space, thereby degrading its performance and making its behavior more noisy and detectable."
- **Validation via Adversarial Retraining:** The proposed countermeasure is modeled in a new Security Twin, G'_{Inf}, and the entire adversarial training process is repeated. By comparing the new Intrusion Graph with the

baseline, the countermeasure's effectiveness at disrupting the AI agent can be quantitatively measured. A successful countermeasure would result in a higher Mean Time To Compromise (MTTC) for the learning agent, or a policy that is demonstrably less efficient.

This application shows a paradigm shift from countering known exploits to countering the *learning process* of an intelligent adversary. It allows defenders to proactively "red team" an offensive AI in a safe, simulated environment, developing and validating defenses against next-generation threats.

VII. Future Work and Open Challenges

To date, this research has established the theoretical foundations of the framework by defining the security twin, the attacker twin, and the intrusion graph. It has also led to the development of the NotLine prototype. Immediate future work involves scaling the simulation engine and implementing the full automated countermeasure evaluation loop, including deception strategies and AI-driven anomaly detection. Key research challenges remain to be addressed:

- **Scalability and Model Abstraction:** A central challenge is defining the theoretical limits of model abstraction required to maintain high-fidelity simulations while achieving the near-real-time performance necessary for a meaningful "what-if" analysis loop, especially in the context of zero-day threats.
- **Fidelity and Simulation-to-Reality Transference:** A formal methodology for validating the "simulation-to-reality" link must be established. Investigating techniques from domain adaptation or transfer learning could be leveraged to bridge the gap between a digital twin and the complexities of the physical system it represents.
- **The Closed-Loop Safety Challenge:** In a CNI context, where an incorrect automated action could have kinetic consequences, ensuring the operational safety of a fully autonomous defense system is paramount. Promising avenues for future research include the application of formal verification methods or safe reinforcement learning paradigms to build a system that is not only proactive but also provably safe.

References

[1] Joint Committee on the National Security Strategy, "The impact of ransomware on critical national infrastructure," UK Parliament, Tech. Rep. HL Paper 20, HC 201, Jan. 2024, Accessed on July 11, 2025. [Online]. Available: https://committees.parliament.uk/publications/48694/documents/255330/default/.

[2] F. Baiardi, V. Sammartino, and S. Ruggieri, "A security twin to defeat intrusions in cyber physical systems," in *ESREL SRA-E 2025*, 2025.

[3] F. Baiardi, S. Ruggieri, and V. Sammartino, "Anticipating disasters through a security twin," in *Proceedings of the International Workshop on Dynamics of Disasters (DOD 2024)*, Workshop held at the ARES 2024 Conference., Vienna, Austria, 2024.

[4] F. Baiardi, S. Ruggieri, and V. Sammartino, "AI-enabled Cybersecurity using Synthetic Data," in *Digital Twins Ecosystems and Applications, PerCom 2025*, 2025.

[5] F. Baiardi, S. Ruggieri, and V. Sammartino, "Dati storici sulle intrusioni: utili o fuorvianti?" *ICT Security Magazine*, 2024. [Online]. Available: https://www.ictsecuritymagazine.com/pubblicazioni/security-twins-e-il-futuro-della-previsione-di-intrusioni-cyber/.

[6] F. Baiardi and V. Sammartino, "From digital twins to ai agents: A synthetic data paradigm for next-generation cybersecurity," in *Artificial Intelligence in Cybersecurity: Unlocking the Power of Large Language Models*, M. O'Brien and I. Alsmadi, Eds., Boca Raton, FL: CRC Press, Taylor & Francis Group, 2026.

[7] F. Baiardi, V. Sammartino, and S. Ruggieri, "Graph-based cyber defence using a security twin," in *29th International Symposium on Distributed Simulation and Real Time Applications (DS-RT 2025)*, 2025.

[8] G. P. Yanev, "Exponential and hypoexponential distributions: Some characterizations," *Mathematics*, vol. 8, no. 12, p. 2207, 2020.

[9] T. Wang and C. Xia, "H2p: A novel model to study the propagation of modern hybrid worm in hierarchical networks," in *Algorithms and Architectures for Parallel Processing*, Cham: Springer International Publishing, 2020, pp. 251–269, ISBN: 978-3-030-60248-2.

[10] K.-C. Huang, S.-Y. Moon, H. Lee, G.-J. Ahn, and H. Kim, "Harmer: A hierarchical attack representation model for attack scenario projection," *IEEE Access*, vol. 8, pp. 125 667–125 681, 2020.

[11] L. Muñoz-González and E. C. Lupu, "Bayesian attack graphs for security risk assessment," in *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, IEEE, 2017, pp. 224–229.

[12] A. Kazeminajafabadi and M. Imani, "Optimal monitoring and attack detection of networks modeled by bayesian attack graphs," *Cybersecurity*, vol. 6, no. 1, p. 22, 2023.

[13] F. Baiardi and V. Sammartino, "A quantitative framework for the validation of twin-based cyber defense," in *Proceedings of the 37th European Modeling & Simulation Symposium (EMSS 2025)*, I. 2. Editors, Ed., ser. 22nd International Multidisciplinary Modeling & Simulation Multiconference (I3M 2025), 2025.

[14] S. Gómez, "Centrality in networks: Finding the most important nodes," in *Business and Consumer Analytics: New Ideas*, P. Moscato and N. J. de Vries, Eds., Springer Nature Switzerland AG, 2019, pp. 401–433, ISBN: 978-3-030-06222-4. DOI: 10.1007/978-3-030-06222-4_8.