Human-Aware Multi-Agent Cyber Defense for Distributed Nuclear Operational Technology Systems

Brian G. Rodiles Delgado[1], Marco A. Alanis Komiyama[1], Carmela Gonzales[2], Khadija Taki[3]

[1]Lewis Bear Jr. College of Business, University of West Florida, USA
[2]Center for Cybersecurity Research and Analysis, Capitol Technology University, USA
[3]Heinz College of Information Systems and Management, Carnegie Mellon University, USA
bgr14@students.uwf.edu, maa146@students.uwf.edu, cagonzales@captechu.edu, ktaki@andrew.cmu.edu

Abstract

Operational Technology (OT) systems used in small modular reactor deployments (SMR) are becoming increasingly interconnected, increasing exposure to cyber threats that can affect tightly coupled physical processes. These systems operate under strict real-time and reliability requirements, which limit how security mechanisms can be deployed and evaluated. We present a multi-agent system for cybersecurity analysis across distributed nuclear OT environments in SMR fleets. The system is built on a simulation-based framework and uses a supervisory agent to coordinate specialized agents for vulnerability analysis and remediation. Agent interactions are restricted through controlled interfaces, and safeguards are applied to prevent unsafe actions during analysis. The system is designed to support continuous inspection of simulated environments while maintaining separation from operational systems. Human operators remain involved in decision-making when actions affect system behavior, enabling security analysis to be performed without introducing risk to live infrastructure.

Introduction

The global energy sector is undergoing a period of transformation driven by rising electricity demand, growing diversification of energy sources, and continued investment in grid infrastructure (MarketLine 2025). Nuclear energy remains a key component of this transition due to its role as a stable, low-carbon power source that supports long-term energy security and decarbonization goals. At the same time, technological advancements and policy initiatives are accelerating the development of next-generation nuclear systems, including small modular reactors (SMRs), which are expected to reshape how nuclear energy is deployed and integrated within modern power systems (MarketLine 2025). As power systems incorporate distributed energy resources, they are becoming more interconnected and reliant on digital infrastructure, expanding potential pathways through which cyber threats can propagate across operational environments. This shift is particularly significant in nuclear systems, where digital control, monitoring, and communication systems are coupled with physical processes, requiring both high reliability and secure operation (MarketLine 2025). In this context, SMRs introduce new operational flexibility and deployment models, and create a more distributed and interconnected attack surface.

In the United States, this transition is reflected in the restructuring of electric power markets following deregulation. This shift has enabled the growth of independent power producers, whose contribution to total energy generation continues to increase, approaching levels comparable to traditional utility companies. This trend is supported by steady nuclear energy market growth, with revenues estimated at 40.3 billion dollars, increasing by 0.9 percent from 2021 to 2026 and projected to grow by 1.4 percent through 2031 (Al Bari 2026). In parallel, this industry is evolving through the development of SMRs, which enable deployment as distributed energy resources rather than large centralized facilities.

The integration of SMRs into distributed grid environments introduces cybersecurity challenges that extend beyond traditional power system design. As they are deployed as distributed energy resources, they increase exposure to cyber-attacks, strengthen the coupling between cyber and physical components, and raise the risk of disruptions across interconnected infrastructure (Haseltine and Albert 2025). Without proactive threat modeling, these vulnerabilities may lead to energy supply interruptions, equipment damage, or broader grid instability.

These challenges are further shaped by the characteristics of Operational Technology (OT) environments. Industrial control systems (ICS) and SCADA networks must support time-sensitive, high-throughput, and secure communication while maintaining strict reliability requirements (Rodiles Delgado et al. 2024). At the same time, these environments are inherently difficult to modify, limiting their ability to adapt to evolving threats or changing operational conditions. As a result, systems must not only detect threats but also coordinate responses in real time, raising questions

about how intelligent agents should operate within critical infrastructure and how they should interact with human operators. At the same time, increasing use of autonomous AI systems in cyber operations introduces the possibility that larger portions of the attack lifecycle may be automated, increasing the speed and scale of cyber-attacks in cyber-physical and OT environments (Pati 2025).

Notably, OT systems are typically organized into layered architectures that segment field devices, control systems, and enterprise services through enforced trust boundaries using firewalls, intrusion detection systems, and controlled communication pathways (Rodiles Delgado 2025c). While this structure enables defense-in-depth, it constrains dynamic response as systems become more interconnected.

Improving adaptability in these environments requires approaches that operate across distributed systems without compromising data sensitivity or performance. Network slicing and virtualization enable the creation of isolated, reconfigurable network segments aligned with specific performance and security requirements. These capabilities are present in public cloud architectures, such as Google Cloud Platform (GCP), Amazon Web Services (AWS), and more that work as virtual infrastructure managers (Rodiles Delgado et al. 2025a, 2025b). Recent advances in agent-based AI enable systems where multiple specialized agents are orchestrated to perform coordinated tasks under shared objectives.

This work takes the position that Artificial Intelligence (AI) driven systems must operate in coordination with human decision makers in environments where actions have direct physical consequences. Fully autonomous responses may improve speed but often lack the contextual awareness required in critical infrastructure systems. Human operators remain essential for oversight and decision making, yet they are constrained by scale, limited visibility, and cognitive load.

This paper proposes a framework that integrates OT-aware agents in a virtualized network environment for digital twin architectures to address these challenges. By leveraging Industrial Control Systems Security Simulation (ICSSIM) along with AWS technologies and LLM models from various providers, the proposed system ensures a production-ready and OT-requirement-compliant environment to perform attack and defense exercises on critical infrastructure, specifically an SMR fleet catered to the nuclear energy sector. These capabilities are structured as coordinated, adaptive components that support human involvement in the decision process. The objective is to improve threat response and system adaptability while preserving operator control in high-consequence environments.

The contributions of this work include the development of a multi-agentic framework for red and blue team exercises in industrial environments and the validation of the testbed to simulate and improve operations. Through rigorous evaluations of performance, the system characterizes trade-offs between reasoning depth, response latency, and token cost across multiple model configurations.

The remainder of this paper is structured as follows. The next section introduces relevant topics to understand the technologies being leveraged. The related work section reviews academic pieces relevant to our framework. The methodology section presents steps to come up with the architecture and its characteristics. The results part showcases the extensive analysis across various performance metrics. Finally, the conclusion along with future work discusses potential extensions to this research, including expansion on more attacks being performed, the integration of more agents, and addressing other sectors beyond nuclear energy.

Background

Operational Technology Systems

OT and ICS form the foundation of critical infrastructure, including power generation and distribution. These systems monitor and control physical processes through tightly coupled hardware and software components, where reliability and continuous operation are prioritized over flexibility. Unlike traditional IT environments, OT systems operate under strict performance constraints, often with minimal tolerance for downtime.

To manage complexity and enforce security boundaries, ICS environments are commonly structured using layered models such as the Purdue Model. This architecture segments systems into hierarchical levels, ranging from field devices and controllers to enterprise and external services, with communication restricted through controlled conduits. While this structure enables defense in depth, it also introduces rigidity that can limit adaptability. As industrial environments increasingly integrate Internet-facing services, the attack surface for OT and ICS has expanded significantly, exposing systems that were not originally designed for modern threat landscapes. In this context, traditional rule-based security mechanisms often struggle to provide the scalability and real-time responsiveness required to protect these heterogeneous environments (Dehlaghi-Ghadim et al. 2023).

Simulation and Virtualized ICS Environments

The development of high-fidelity virtual testbeds has become fundamental for evaluating cybersecurity solutions without endangering live production systems. Platforms such as ICSSIM provide fully containerized environments designed for cybersecurity analysis of ICS, leveraging technologies such as Docker to isolate components, including Programmable Logic Controllers and SCADA systems. These environments enable realistic network emulation and continuous telemetry generation, supporting the development and testing of data-driven detection approaches (Dehlaghi-Ghadim et al. 2023).

Similarly, frameworks such as ICS-SimLab enable rapid construction and customization of diverse industrial architectures, including smart grids and manufacturing systems. These platforms support controlled experimentation by generating both benign and malicious traffic for training and evaluation (Brown et al. 2025). While such environments provide flexibility and safety for experimentation, they are often oriented toward offline analysis workflows, limiting their ability to support adaptive and interactive security operations in real-time.

Agent-Based Systems in Cybersecurity

To address the limitations of traditional approaches, the critical infrastructure community has increasingly explored autonomous defenses powered by multi-agent systems and LLMs (Sunkara 2025; Hmimou et al. 2025). These systems introduce the ability to perceive, reason about, and respond to threats in dynamic environments, representing a shift from static rule-based detection toward adaptive and coordinated defense.

Multi-agent architectures distribute responsibilities across specialized components, enabling coordinated processes such as vulnerability discovery, protocol analysis, and response planning (Sunkara 2025). Recent work has demonstrated the effectiveness of these approaches in industrial contexts, where agent-based frameworks have been applied to identify vulnerabilities in control protocols and automate PLC code validation (Ning et al. 2025; Liu et al. 2024). These systems have shown improvements in detection accuracy and reductions in false positives, emphasizing the potential of coordinated agent behavior in complex environments (Hmimou et al. 2025). While prior work demonstrates coordination in distributed multi-agent environments, these approaches do not fully account for the operational constraints, safety requirements, and real-time demands of OT systems (Sunkara 2025).

Orchestration and Secure Agent Deployment

The deployment of multi-agent systems at scale requires mechanisms for coordination, control, and governance. Cloud-native orchestration platforms enable centralized management of agent workflows, structured communication, and integration with domain-specific knowledge sources. Platforms such as AWS Bedrock provide mechanisms for coordinating agent interactions through controlled execution patterns that support traceability and predictable behavior.

These systems also incorporate safeguards, including centralized guardrails to prevent unsafe commands from reaching industrial components and knowledge bases that ground agent reasoning in an operational context. In parallel, security approaches such as Zero Trust Architectures are being extended to agent-based systems, emphasizing continuous authentication and controlled interaction between agents. While these capabilities improve scalability and governance, their application within safety-critical OT environments remains challenging, particularly when balancing flexibility with strict operational constraints.

Human-Aware AI in Safety-Critical Systems

While multi-agent and Large Language Model (LLM) based systems show strong potential for automating cybersecurity workflows, their use in safety-critical environments introduces additional challenges (Hmimou et al. 2025). Industrial systems require not only accurate detection and response, but also transparency, predictability, and control. Fully autonomous systems may improve response speed, but they may lack sufficient awareness of physical process constraints and operational context.

As a result, there is growing interest in human-aware AI systems that incorporate human oversight into automated workflows. These approaches emphasize collaboration between human operators and intelligent agents, allowing systems to assist in decision-making while maintaining accountability and control. However, despite advances in simulation environments, agent-based security systems, and orchestration platforms, limited work has explored how these capabilities can be integrated into a unified, OT-aware architecture.

In particular, the combination of high-fidelity simulation environments with coordinated, human-aware multi-agent systems that support real-time analysis, enforce safety constraints, and maintain clear separation between analysis and execution remains underexplored. Addressing this limitation motivates the approach presented in this work.

Related Work

ICS Security and Simulation

Prior work on ICS cybersecurity has emphasized developing simulation platforms for safe, repeatable experimentation. Systems such as ICSSIM and ICS-SimLab enable researchers to model industrial environments and evaluate cyber threats without impacting live infrastructure (Dehlaghi-Ghadim et al. 2023; Brown et al. 2025). While both approaches support realistic emulation, they differ in focus. ICSSIM prioritizes flexible construction of customized testbeds, whereas ICS-SimLab emphasizes rapid deployment of domain-specific scenarios such as smart grids. Despite these capabilities, both platforms are primarily designed for offline experimentation and dataset generation. They do not inherently support continuous interaction with adaptive defense mechanisms or real-time reasoning over system state. As a result, their role is often limited to evaluation rather than active defense.

Multi-Agent Architecture for Cyber Defense

Recent research has explored multi-agent systems and LLMs for automating specific cybersecurity tasks, with an agentic framework that coordinates reasoning and task execution across distributed components. These approaches vary in scope and function across different stages of the security lifecycle. For example, the Multi-Agent LLM Fuzzing Framework (MALF) focuses on coordinated vulnerability discovery through intelligent fuzzing of industrial control protocols, emphasizing exploration of attack surfaces (Ning et al. 2025). In contrast, Agents4PLC targets system correctness by automating PLC code generation and verification, prioritizing validation over exploration (Liu et al. 2024).

Other work has proposed multi-agent architectures for threat detection and correlation, where agents process heterogeneous inputs and produce contextualized threat assessments (Hmimou et al. 2025). While these approaches demonstrate the effectiveness of agent-based reasoning, prior work shows that agentic frameworks coordinate distributed tasks; existing systems are typically limited to individual tasks such as fuzzing, validation, or detection. This specialization constrains their ability to support coordinated, end-to-end defense workflows that integrate detection, analysis, and response.

Human Oversight in Multi-Agent Systems

As multi-agent systems become more capable, prior work has increasingly focused on coordination and controlled execution across interacting agents. Cloud-based orchestration frameworks introduce structured interaction patterns that support traceability, policy enforcement, and managed communication between components. Techniques such as guardrails and knowledge grounding are used to constrain agent behavior and reduce the likelihood of unsafe or unintended outputs. Concurrently, research on the Internet of Agents has proposed security models based on continuous verification of agent behavior using interaction patterns and execution characteristics (Wang et al. 2025). While these approaches improve coordination and system-level control, they are often developed without considering environments where actions may directly impact physical processes. In such settings, it is necessary to enforce clear separation between analytical reasoning and system execution, while ensuring that human operators remain involved in decision-making workflows. Existing work provides mechanisms for constraining agent behavior, but offers limited support for structuring human oversight within coordinated, multi-agent systems operating under real-time and safety-sensitive conditions.

Cyber Defense Across Critical Domains

Existing work has applied simulation and agent-based techniques across multiple critical infrastructure domains, including energy systems, water infrastructure, and space systems. In energy systems, prior approaches have used simulation platforms to model smart grid environments and evaluate detection strategies under realistic conditions. In water systems, similar approaches have been used to study distributed control processes and improve situational awareness through correlated monitoring.

In more constrained environments such as space systems, research has focused on protocol-level vulnerabilities and control validation, leveraging approaches such as multi-agent fuzzing and automated verification (Ning et al. 2025). These efforts emphasize the importance of both vulnerability discovery and system correctness in distributed and high-stakes settings. Across these domains, existing approaches remain fragmented.

Simulation, detection, validation, and coordination are typically addressed as separate problems rather than as components of a unified system. In addition, many systems assume either fully automated operation or limited human involvement, without explicitly addressing how human oversight should be integrated into real-time decision-making. There is limited work that combines high-fidelity simulation, coordinated multi-agent reasoning, and human-aware control within a single framework. Addressing this gap is essential for enabling adaptive and trustworthy cyber defense in safety-critical environments and motivates the approach presented in this work.

Methodology

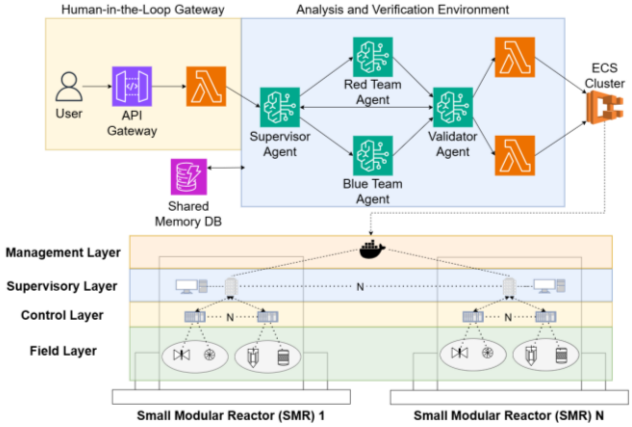

Figure 1. Multi-agent architecture, AWS technologies, and an ICSSIM environment for a digital twin of an SMR

Tiered Architectural Framework

This study adopts a multi-agent approach to cybersecurity analysis in ICS by integrating simulation environments, coordinated agent reasoning, and human-in-the-loop oversight into a unified framework. The system is organized as a layered architecture comprising access control, orchestra-

tion, agent execution, and verification components, allowing controlled interaction between modules while maintaining separation between analysis and operational environments.

At the entry point, infrastructure and telemetry integration support modular Docker-based deployments aligned with the Purdue Model, representing Small Modular Reactor environments that generate continuous telemetry for agent-driven state awareness. User interaction is managed through a controlled gateway that handles authentication, request validation, and session initiation for long-running analysis tasks. These requests are processed through centralized orchestration, where a supervisor agent maintains system context, interprets user intent, and coordinates communication across specialized agents, ensuring that all interactions pass through a single control point.

Within this framework, deterministic safety guardrails are applied to all interactions to enforce system-level constraints. Inputs are evaluated for unsafe control actions and prompt manipulation, while outputs are filtered to prevent disclosure of sensitive information or unsafe recommendations. The architecture enforces strict communication boundaries so that agents operate only through supervised channels and do not directly access or modify the target environment without explicit authorization. This supports a controlled flow of information from user request to system response while maintaining safety and accountability in cyber-physical environments.

Agent Design and System Controls

Building on this architecture, the system employs specialized agents that operate under constrained permissions and clearly defined roles aligned with different stages of the cybersecurity workflow. A vulnerability analysis agent performs read-only inspection of simulated environments, including code repositories and network activity, to identify potential weaknesses without disrupting operations. A remediation agent generates recovery strategies using retrieval-augmented methods grounded in validated industrial security standards.

These agents are coordinated through the supervisory layer and do not interact directly, enabling structured task delegation and controlled information flow. To ensure safe interaction with industrial processes, operational containment enforces a strict separation between the analysis environment and the target ICS environment. Agents are limited to observation and reporting by default and cannot execute changes unless explicitly authorized through supervisory control. In this work, models are accessed directly via the Anthropic, OpenAI, Google, and Grok APIs, enabling use of the latest model versions while maintaining the same orchestration principles.

Cost estimation based on token utilization is also incorporated to evaluate trade-offs between model performance and computational overhead. Observations indicate that higher-capacity models improve reasoning quality but require bounded resource allocation for practical deployment.

Verification and Context Traceability

To address the risks associated with automated reasoning in environments where system actions have physical consequences, the framework incorporates a multi-layer verification process to ensure the reliability of all outputs. A safety auditor agent reviews generated findings prior to finalization, operating independently to validate that recommendations align with industrial safety requirements and do not introduce unintended risks.

When potential violations are detected, an agentic refinement loop is triggered in which problematic elements are isolated and excluded while the remaining analysis proceeds without propagating unsafe conclusions. In parallel, comprehensive forensic logging captures agent interactions, reasoning traces, and system-level events. This provides an audit trail that allows human operators to review system behavior, interpret decision pathways, and maintain accountability.

Together, these components define a human-aware approach to cybersecurity in which multi-agent systems support real-time analysis while maintaining operator control in distributed nuclear environments.

Evaluation and Results

The evaluation methodology utilizes assertion-based benchmarking to automate the assessment of complex multi-agent trajectories. This approach verifies if the interactions between the Supervisor, Red/Blue teams, and the Safety Auditor meet specific security and process standards.

Metric Definitions

We define the following metrics to evaluate the effectiveness, speed, and resource density of the SMR security analysis:

- Response Success Rate (RSR): The percentage of sessions where all user-side and system-side assertions are satisfied, including valid vulnerability detection and ISA/IEC 62443 compliance (IEC 2018).
- User-Perceived Latency (UPL): The total time in seconds from the initial user request until the Supervisor delivers the audited final report.
- Verification Latency (VL): The specific duration of the Layer 4 "Safety Auditor" pass.
- Avg. Input Tokens: The average number of total input tokens used per model. This is a critical metric for understanding the communication density and potential cost of the orchestration.

To provide a concrete understanding of the RSR, it is essential to distinguish how specific security assertions are evaluated within the industrial control framework. Under the vulnerability detection assertion, a "pass" occurs when the model accurately identifies an exploit vector, such as an unauthenticated Modbus write command, and assigns the correct risk severity; a "fail" is recorded if the model misses the vector or misclassifies a critical vulnerability as low-risk.

For ISA/IEC 62443 compliance, a passing assertion requires the model to explicitly map a discovered security gap to the correct regulatory sub-section, whereas a failure involves providing vague security advice without specific regulatory citations.

Finally, in system logic assertions, the model passes by rejecting commands that violate safety-critical thresholds, such as a request to increase the primary coolant outlet temperature beyond the hardcoded safety limit, while it fails if it provides instructions to bypass these operational boundaries, potentially compromising the structural integrity of the reactor core.

Accuracy and Latency by Fleet Scale and Model

This subsection showcases the RSR and response times per model type. Each model-fleet configuration was evaluated over 25 independent sessions.

Specialists Model	SMR	RSR	UPL	VL
claude-opus-4-1-	1	0.96	58.4s	9.2s
(High Reasoning)	2	0.93	82.1s	13.5s
	3	0.91	115.6s	18.2s
claude-sonnet-4-	1	0.92	35.2s	7.8s
(Balanced)	2	0.89	51.4s	11.2s
	3	0.86	74.8s	14.5s
claude-3-5-	1	0.84	18.7s	6.5s
(High Speed)	2	0.78	26.5s	8.9s
	3	0.72	39.1s	11.3s

Table 1: RSR and latency metrics per model type and number of SMRs

Token Utilization

Tokens used as part of the input for the model were tracked per reactor by the number of SMRs (Table 2).

Specialists Model	1-SMR Input	2-SMR Input	3-SMR Input
claude-opus-4-1-20250805	498	480.5	469.67
claude-sonnet-4-20250514	503	477	459.67
claude-3-5-haiku-20241022	504	483.5	476
grok-4-0709	500	480	461.33
gemini-2.5-pro	503	482	462
gpt-4o-2024-08-06	501	480.5	473.67
gpt-4o-mini-2024-07-18	503	481.5	463.67

Table 2. Tokens used in the input per reactor count

Given that the cost can be associated with the tokens used, that value was also tracked (Figure 2).

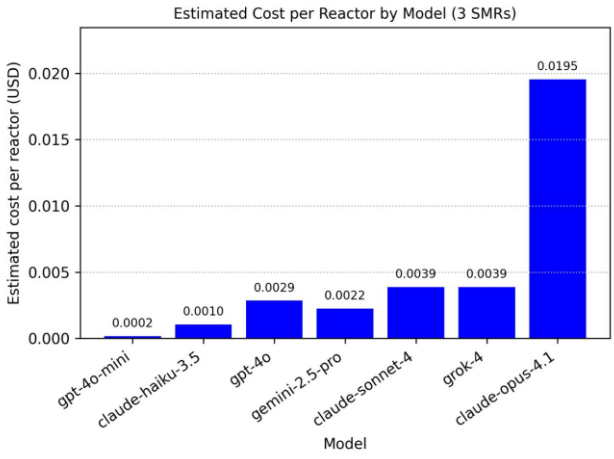

Figure 2. Estimated cost per reactor by model

Key Findings

Models that are highly reasoning, such as Claude Opus 4.1, tend to take longer to give a response, both from the user and the verification side. As an example, Claude Opus 4.1 took 115.6s in the whole loop of agents compared to Claude Haiku 3.5 with 39.1s for three SMRs. However, the accuracy is elevated as part of this effort. Let us take the case of three SMRs as well, with 0.91 compared to 0.72 with the highest reasoning and speed models, respectively. The number of SMRs has an impact on both latency and RSR as they increase due to more logs and assets needing to be taken into account by the agents (Table 1).

Delving into the usage of tokens, an interesting trend occurs where the average number of tokens consumed decreases as more SMRs are in the infrastructure. Consider the case of three SMRs, and the average input tokens being 469.67 and 476 for Anthropic's highest reasoning and speed models, respectively. This is because of prompt compression and mechanisms for the model to enhance its resource utilization as part of an analysis process. Moreover, tokens converge across models because the framework applies context summarization as fleet complexity scales, with higher-capacity models benefiting more due to larger context windows (Table 2).

Lastly, when all the popular and current models are put in the perspective of the cost associated with taking into consideration the most expensive case of using three SMRs, the model with the highest reasoning has the highest cost associated with it compared to others, such as models from companies like OpenAI (GPT), Google (Gemini), and X (Grok). For the three-reactor case, Claude Opus 4.1 costs approximately 1.95 USD cents per agent run, compared to approximately 0.10 USD cents for Claude Haiku 3.5, which is a difference of roughly 19 times. Costs for GPT, Gemini, and Grok in Figure 2 are derived from their respective public API pricing pages using equivalent token counts from the three-SMR configuration (Figure 2).

Conclusion and Future Work

This research demonstrates that a tiered, multi-agentic framework is highly effective for performing cybersecurity analysis in distributed Small Modular Reactor (SMR) environments without risking live infrastructure. By leveraging a supervisory agent to coordinate specialized vulnerability and remediation agents within a simulated ICSSIM environment, the system successfully bridges the gap between high-speed automated reasoning and the strict reliability requirements of Operational Technology (OT). The evaluation results confirm a clear trade-off between reasoning depth and operational speed: high-capacity models like Claude Opus 4.1 achieved the highest Response Success Rate (RSR), albeit with higher latency, while faster models like Claude Haiku 3.5 offered rapid response times at the expense of accuracy. Crucially, the implementation of safety auditors and deterministic guardrails ensures that human operators remain the final decision-makers, preserving the necessary oversight for safety-critical nuclear systems.

Future work will focus on expanding the technical capabilities of the specialized agents to address a broader spectrum of cyber-physical threats. While the current study validated the system using specific red and blue team exercises, future iterations aim to incorporate more complex attack vectors and a wider variety of specialized agents, such as dedicated protocol analysts or forensic specialists. There is also an intent to refine the agentic refinement loop to further reduce false positives and improve the handling of real-time telemetry from larger SMR fleets. By enhancing prompt compression techniques and resource utilization, which already showed promising results in reducing token consumption as system complexity scaled, the framework can be optimized for even more cost-effective and scalable deployments. The human-aware design principles embedded in the architecture, including controlled agent interfaces, forensic logging, and supervisory escalation, are evaluated structurally in this work; empirical assessment of operator cognitive load and human factors is reserved for further iterations.

Beyond the nuclear sector, there is a significant opportunity to adapt this multi-agent architecture to other critical infrastructure domains, including water treatment facilities, smart grids, and aerospace systems. Future research will explore how the separation between analysis and execution can be maintained across these diverse OT environments, each with its own unique safety constraints and communication protocols. Additionally, the team plans to conduct deeper studies into human-aware AI to better understand how to minimize the cognitive load on operators during high-stress security events. This involves developing more intuitive interfaces for the supervisory agent's reporting, ensuring that the forensic logs and reasoning traces provided to humans are both actionable and transparent in a real-time defense context.

Ethical Statement

The framework presented in this work is designed exclusively for use within isolated simulation environments and is not intended for deployment against live operational systems. The red team capabilities embedded in the architecture carry inherent dual-use potential; a system capable of identifying exploit vectors in nuclear OT environments could, if misapplied, inform adversarial operations rather than defensive ones. We mitigate this risk through strict simulation-only boundaries, deterministic guardrails that reject unsafe commands, and human-in-the-loop authorization for all non-read-only actions. Responsible deployment of this framework requires institutional oversight, access controls aligned with IEC 62443 security levels, and restriction to credentialed security researchers operating within sanctioned exercise environments.

References

Al Bari, S. 2026. *Nuclear power in the US*. IBISWorld Industry Report 22111b. Melbourne, Australia: IBISWorld.

Brown, J.; Pham, D.-S.; Soh, S.-T.; Motalebi, F.; Eswaran, S.; and Almashor, M. 2025. ICS-SimLab: A containerized approach for simulating industrial control systems for cyber security research. In *Proceedings of the IEEE Conference on Communications and Network Security (CNS)*. doi:10.1109/CNS66487.2025.11195055.

Dehlaghi-Ghadim, A.; Balador, A.; Moghadam, M. H.; Hansson, H.; and Conti, M. 2023. ICSSIM — A framework for building industrial control systems security testbeds. *Computers in Industry* 148 (June): 103906.

Haseltine, C. A.; and Albert, L. A. 2025. Cybersecurity threat modeling for small modular reactor stations. In *Proceedings of the IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. doi:10.1109/COINS65080.2025.11125731.

Hmimou, Y.; Tabaa, M.; Khiat, A.; and Hidila, Z. 2025. A multi-agent system for cybersecurity threat detection and correlation using large language models. *IEEE Access*. doi:10.1109/ACCESS.2025.3602681.

International Electrotechnical Commission (IEC). 2018. *IEC 62443: Industrial Communication Networks - Network and System Security.* Geneva, Switzerland: IEC.

Liu, Z.; Zeng, R.; Wang, D.; Peng, G.; Wang, J.; Liu, Q.; Liu, P.; and Wang, W. 2024. Agents4PLC: Automating closed-loop PLC code generation and verification in industrial control systems using LLM-based agents. arXiv:2410.14209.

MarketLine. 2025. *Global nuclear energy*. Manchester, UK: MarketLine.

Ning, B.; Zong, X.; and He, K. 2025. MALF: A multi-agent LLM framework for intelligent fuzzing of industrial control protocols. arXiv:2510.02694.

Pati, A. K. 2025. Agentic AI: A comprehensive survey of technologies, applications, and societal implications. *IEEE Access*. doi:10.1109/ACCESS.2025.3585609.

Rodiles Delgado, B. G.; Aguayo, J.; Gomez Rivera, A. O.; and Tosh, D. 2024. Reconfigurable Network Slicing Orchestration in Network Function Virtualization Compatible Operational Technology Environment. In *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*.

Rodiles Delgado, B. G.; Casio Iracheta, B. E.; Solano, A. X.; Martínez, M. D.; Tosh, D.; and Servin, C. 2025a. Network Slicing for Dynamic DMZ and Federated Learning in Operational Technology Environment. In *2025 Resilience Week (RWS)*.

Rodiles Delgado, B. G.; Estrada Aguirre, L. D.; Marfo, W.; Servin, C.; and Tosh, D. 2025b. Network Slicing for Federated Learning in Operational Technology Environment. In *2025 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*.

Rodiles Delgado, B. G. 2025c. *Enhancing Security and Resiliency in Operational Technology Environments Through Network Slicing and Federated Learning*. Master's thesis, Department of Computer Science, The University of Texas at El Paso, El Paso, TX.

Sunkara, G. 2025. Multi-agent AI systems for coordinated cybersecurity in smart cities. *International Journal for Research Publication and Seminar* 16(3). doi:10.36676/jrps.v16.i3.287.

Wang, Y.; Pan, Y.; Guo, S.; and Su, Z. 2025. Security of Internet of Agents: Attacks and Countermeasures. *IEEE Open Journal of the Computer Society* 6: 1611-1624. doi:10.1109/OJCS.2025.3589638.