Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

(GEN)AI VERSUS (GEN)AI IN INDUSTRIAL CONTROL CYBERSECURITY

Cynthia Zhang[1], Ranjan Pal[1], Corwin Nicholson[1], and Michael Siegel[1]

[1]MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

(Gen)AI is emerging as a powerful force transforming industrial control and business/enterprise productivity. This paper investigates the challenges and opportunities stemming from (Gen)AI on industrial control systems (ICSs) security within the framework of the *Cyber Kill Chain* (CKC). Leveraging the CKC framework, we examine how (Gen)AI enables attackers to automate each phase of the CKC – *reconnaissance, weaponization, delivery, exploitation, installation, command and control*, and *action on objectives.* Conversely, (Gen)AI also empowers defenders to employ advanced techniques such as AI-powered firewalls, anomaly detection, and automated incident response to thwart cyber threats effectively. We study how this defense dynamic using (Gen)AI operates within each phase of the Cyber Kill Chain for ICSs. We back up our attack-defense dynamics study with simulations on real-world ICS scenarios. To the best of our knowledge, this is the first cybersecurity study in the joint space of ICSs, the Cyber Kill Chain (CKC), and (Gen)AI.

1 INTRODUCTION

Industrial Control Systems (ICS) are the backbone of critical infrastructure. Composed of a complex network of hardware and software designed to manage industrial operations, ICSs play a critical role in many services that underpin modern society. Whether it be the power grid, oil pipelines, water treatment and distribution plants, telecommunications, or nuclear facilities, ICSs play a pivotal role in our critical infrastructure by orchestrating and monitoring various industrial processes through a network of sensors, actuators, PLCs, and more. The importance of ICSs cannot be overstated, as any disruption to these systems can have far-reaching consequences in public safety, economic productivity, and national security. Disruptions to these systems can lead to supply chains interruptions, endangerment of lives, and erosion of public trust in the reliability of essential services. For example, the 2015 Ukraine power grid attack not only shut down power for 230,000 people for up to 6 hours, but also heightened political tensions between Ukraine and Russia and instilled a sense of paranoia and uncertainty among the Ukrainian population.

Despite the critical functions ICSs provide for society, ICSs remain one of the most vulnerable systems in terms of cybersecurity. This issue stems from a variety of reasons.

1. *ICSs are littered with legacy systems*: The OT infrastructure that ICSs operate on is often littered with outdated legacy systems, which themselves harbor unpatched, exploitable vulnerabilities. Many OT systems, once installed, are intended to have a lifespan of decades, as compared to many IT systems that run for 3-5 years. As a result, OT systems accumulate a multitude of vulnerabilities during their extended lifespan (Hanes et al. 2017).
2. *OT systems are difficult to patch*: ICS environments frequently have strict operational requirements and downtime restraints, making it difficult to promptly apply software patches and updates. Due to the sensitive nature of critical infrastructure ICS environments and the risks of operational disruptions, any changes to the configuration or software of critical systems must be carefully evaluated and tested to prevent compromising safety or reliability. To add to the complexity, OT devices are often installed in remote or harsh locations, making access to systems difficult.

979-8-3315-3420-2/24/$31.00 ©2024

3. *OT systems have poor visibility*: Many OT management and monitoring systems are separated from IT, leading to blind spots where OT traffic is not sufficiently monitored or inventoried. ICSs are often comprised of a mix of hardware and software from various vendors, making maintaining a comprehensive understanding of system configurations and vulnerabilities difficult. Legacy systems are a roadblock for improved visibility, as outdated systems may lack monitoring capabilities and also make IT/OT integration difficult. Additionally, strict control over who can access the plant floor further contributes to poor visibility as it limits the number of individuals who can actively monitor and detect potential security threats or anomalies within the industrial environment.
4. *IT/OT convergence leads to attack spillover*: While many IT and OT systems remain separated, there is an increasing trend of integrating IT and OT systems, with an example being IoT. However, this leads to OT suffering from side-effects of attacks on IT infrastructure (Darktrace 2023b). This can be seen with the Colonial Pipeline attack, which shut down oil pipelines in 17 states for 6 days. While the OT systems themselves were not compromised, oil pipelines were shut down out of an abundance of caution after IT systems were compromised by a ransomware attack.
5. *Lack of security awareness*: As a result of budget constraints, employees in critical infrastructure sectors often lack adequate training in recognizing social engineering attacks. Budgets are often allocated elsewhere, leading to increased risk of employees falling victim to phishing attempts, where attackers trick individuals into divulging sensitive information or compromising systems. In ICSs, the consequences of successful phishing attacks can be severe, potentially leading to operational disruptions, data breaches, or physical damage that can have wide reaching social and economic impacts.
6. *Management shortcomings*: Management may prioritize other operational expenses over cybersecurity. This absence of a robust cybersecurity culture is particularly noticeable within critical infrastructure, where there has historically been a greater emphasis on physical security. Management often prioritizes operational efficiency over robust cybersecurity measures, which can result in gaps in oversight of critical security protocols. Furthermore, organizations may lack well-defined accountability structures for cybersecurity, thus making it difficult for management to implement and enforce effective cybersecurity measures. The lack of clear accountability can create gaps in oversight and coordination, impeding efforts to adequately secure ICSs.

The introduction of (Gen)AI presents new challenges for defending ICSs against cyber threats. While traditional attackers already pose significant challenges to defenders, (Gen)AI serves only to amplify their effectiveness. (Gen)AI allows for the rise of script kiddies, as individuals with limited technical expertise can now easily leverage AI-powered tools to automate attack frameworks and launch sophisticated attacks. Automation enabled by (Gen)AI also allows actors to scale and streamline their operations, increasing the speed and volume of attacks while reducing the need for manual intervention (Gupta et al. 2023).

AI can empower attackers to more effectively target each of the points mentioned above. Legacy systems (see point 1 above) frequently have widely known, unpatched vulnerabilities. (Gen)AI can quickly and automatically discover publicly known vulnerabilities, many of which also have publicly known exploits, in these systems. Poor visibility and lack of patching (see points 2 and 3) could lead to cases where devices use default or weak passwords, which are easy vulnerabilities for (Gen)AI to discover and exploit. A larger attack surface from IT/OT convergence (see point 4) only increases the number of machines (Gen)AI can discover vulnerabilities for. Furthermore, GenAI is skilled at generating phishing attacks, and can easily mimic writing styles for spear-phishing attacks. This leaves under-trained employees (see point 5) even more vulnerable to targeted social engineering attacks. By catalyzing the attack space, (Gen)AI empowers adversaries to quickly adapt to evolving defense mechanisms, enabling them to deploy more intricate attack vectors and effectively perform each step of the Cyber Kill Chain.

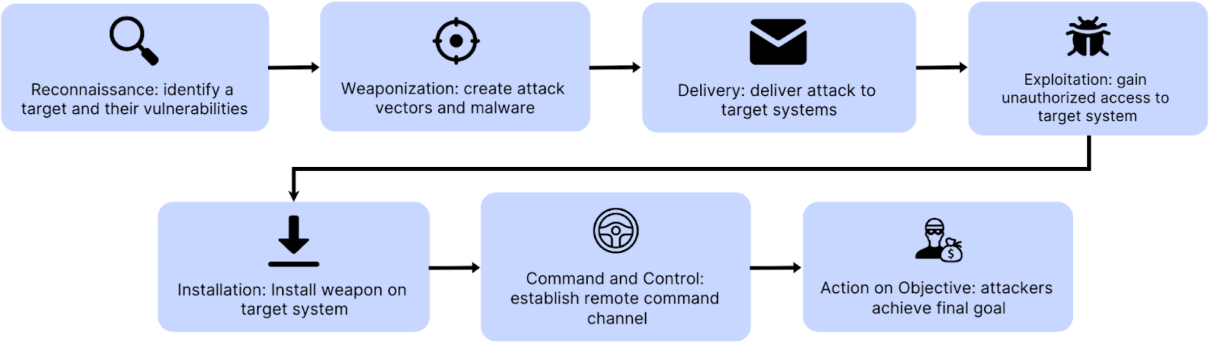

Reconnaissance: identify a target and their vulnerabilities
Weaponization: create attack vectors and malware
Delivery: deliver attack to target systems
Exploitation: gain unauthorized access to target system
Installation: Install weapon on target system
Command and Control: establish remote command channel
Action on Objective: attackers achieve final goal

Figure 1: A summary of the seven steps of the Cyber Kill Chain.

1.1 The Cyber Kill Chain

The Cyber Kill Chain, coined and developed by Lockheed Martin, is a conceptual framework describing the sequence of steps cyber intruders typically follow in order to achieve their objective. The steps begin with *reconnaissance*, then proceed through *weaponization*, *delivery*, *exploitation*, *installation*, *command and control*, and ends with *action on objectives*. This model aids cybersecurity professionals in identifying and mitigating cyber threats by breaking down the attack lifecycle into discrete phases, thus allowing organizations to implement more targeted and effective cyber security defense measures. The stages are described as follows:

1. *Reconnaissance*: In the first phase of a cyber-attack, attackers must select a target and gain information about the target by identifying potential vulnerabilities and entry points. This step can be both passive and active. During passive reconnaissance, attackers obtain information without directly interacting with the target; this may include gathering public information such as IP addresses, email addresses, etc. After passive reconnaissance, attackers may move towards active reconnaissance, where attackers will interact directly with the target via scanning, fingerprinting, etc.
2. *Weaponization*: Once the attacker learns of vulnerabilities on the target system via reconnaissance, attackers will create attack vectors. This typically involves designing two components: the remote access tool (RAT) and exploit. The RAT allows attackers to gain unauthorized access to the system and control it remotely; they are usually designed with specific objectives, such as stealing data or disrupting operations. The exploit is the RAT carrier and takes advantage of a vulnerability to execute malicious actions.
3. *Delivery*: Attackers now deliver the weapon to the targeted computer systems. This often takes the form of social engineering attacks such as (spear)phishing, but can also take the form of USB sticks, “watering hole” compromised websites, or direct interaction with the target by the attacker.
4. *Exploitation*: After successful delivery of malware, attackers exploit vulnerabilities discovered in previous steps to gain unauthorized access. These exploits may be publicly known exploits or zero-days. In this step, attackers can learn of more vulnerabilities and move laterally through the system to identify more potential entry points. Attackers will employ various techniques to compromise applications and networks in order to gain a stronger foothold in the target system.
5. *Installation*: In the installation step, also known as the privilege escalation step, attackers install malware on the system and deploy cyberweapons to gain control of more systems and data. Attackers may install remote administration tools to maintain their presence on the system and exfiltrate data.
6. *Command and Control (C2)*: After a system is compromised, attackers establish a command channel to remotely manipulate the victim system. By establishing a C2 channel, attackers can now monitor

and guide their deployed cyberweapons remotely. This step frequently involves attackers covering their tracks (obfuscation) to hide their presence and prevent discovery.

7. *Action on Objectives*: In the final stage of a cyberattack, attackers achieve their ultimate goal, such as monetary or data theft, data encryption, or destruction. It may frequently take weeks or months to reach this step.

The seven steps of the Cyber Kill Chain are summarized in Figure 1.

1.2 Research Contributions

In this section, we propose our research contributions.

- Our research contributes novel insights into the intersection of (Gen)AI and ICS cybersecurity in the framework of the Cyber Kill Chain. Analyzing attacks and defenses in the context of the CKC provides a structured framework for understanding and responding to cyber threats; (Gen)AI introduces risks to ICSs at each stage of the CKC. (Gen)AI solutions are proposed for each step of the CKC, making it possible for OT managers to strategically counter attackers at each step of an attack. Our analysis not only enhances comprehension of the threat landscape but also serves as a crucial tool for devising proactive defense strategies tailored to counter AI-driven cyberattacks on ICSs. The use of (Gen)AI in ICS defense strategies saves time and energy while also allowing managers more effective budget allocation in managing cybersecurity risks. As far as we know, this is the first comprehensive analysis of how (Gen)AI influences cybersecurity for ICSs on each step of the Cyber Kill Chain. (See sections 3 and 4).
- Our second contribution lies in the simulations we ran to demonstrate how (Gen)AI can be used to both attack and defend ICSs. We simulate how GenAI can be harnessed by malicious actors to exploit vulnerabilities in ICSs, as well as how GenAI can be leveraged by defenders to protect ICS systems. These simulations demonstrate the ease at which threat actors can leverage (Gen)AI to attack critical ICSs. By bridging the gap between speculation of what (Gen)AI could accomplish and real-world scenarios, our research contributes tangible insights into the potential of (Gen)AI to be used both maliciously and defensively in ICS cybersecurity. (See section 5).

The rest of the paper is organized as follows. In section 2, we discuss related works. Section 3 discusses how (Gen)AI empowers attackers of ICSs on the CKC. Section 4 discusses the same for defenders of ICSs. In section 5, we discuss three simulations demonstrating GenAI's dual role in the cybersecurity of ICSs. Finally, section 6 concludes the paper.

2 RELATED WORKS

To our knowledge, our research is the first that explores the influence of (Gen)AI on the security of ICSs for each step of the Cyber Kill Chain. However, there exist many illuminating works exploring the applications of the CKC, cybersecurity of ICSs, and (Gen)AI's influence on cybersecurity.

Many research papers and industry white papers have delved into the intricacies of the Cyber Kill Chain framework. This includes delving into the technical details of each stage of the framework (Yadav and Rao 2015). Analysis of the Equifax data breach (Kabanov and Madnick 2021) offers practical insights derived from a real-world incident, emphasizing the CKC's efficacy as a tool for comprehensively analyzing and understanding both attack and defense strategies in cybersecurity. The usefulness of the CKC as a framework cannot be understated, as various analyses exist to frame (Gen)AI or industrial control systems on the CKC (Darktrace 2023c; Assante and Lee 2015). Our research similarly applies the CKC framework to gain a strong high-level and technical understanding of (Gen)AI's influence on the security of ICSs. *However, existing research fails to address both (Gen)AI and ICSs on the CKC; we see extensive research discussing how (Gen)AI will influence security (Deshpande and Gupta 2023; Darktrace 2023c), or research* *discussing how ICSs fit into the CKC (Assante and Lee 2015), but none discussing the intersection of all three. We address this in sections 3 and 4.*

The influence of (Gen)AI on cybersecurity is an extensively addressed topic, with both academic research papers and industry white papers discussing the influence of (Gen)AI on security practices (Darktrace 2023a; Gupta et al. 2023). (Gen)AI enables attackers to autonomously generate convincing phishing emails or create malware variants that evade detection, while simultaneously aiding defenders in detecting anomalies in networks and logs or automate incident response. *While there exists discussion of the influence of (Gen)AI on both the attackers and defenders fronts, for both general and ICS specific cases (Nankya et al. 2023; Wang et al. 2022), these papers do not address the topic with the granularity of the Cyber Kill Chain. We address (Gen)AI's influence on both the attack and defense dimensions for each of the seven steps of the Cyber Kill Chain in sections 3 and 4.*

There exists extensive literature discussing the cybersecurity of ICSs, as well as exploits that can be performed on ICS devices (Nankya et al. 2023; Angle et al. 2019). Existing papers demonstrate a variety of attacks that can be performed on various programmable logic controllers (PLCs) (Tychalas and Maniatakos 2020; Beresford 2011), as well as attacks on SCADA protocols such as Modbus (Chen et al. 2015). Demonstrated attacks range from cache side channels to TCP SYN flooding and can expose sensor data or inject false control commands. The papers demonstrate the exploitability of various aspects of ICSs; *however, there do not exist any papers demonstrating the exploitability of ICSs by (Gen)AI. In section 5, we use (Gen)AI to simulate known exploits as well as defense options to demonstrate the effectiveness and versatility of (Gen)AI as a tool.*

3 (GEN)AI EMPOWERS ATTACKERS

We will now discuss how (Gen)AI empowers the attackers at each step of the Cyber Kill Chain. From *reconnaissance* to *action-on-objectives*, we will examine the ways in which (Gen)AI enables adversaries to execute more sophisticated and evasive attacks, posing significant challenges to traditional cybersecurity defenses. As far as we are aware, this is the first paper to offer a comprehensive exploration of (Gen)AI cyberthreats on ICSs within the context of the CKC.

Reconnaissance: In the first stage of a cyberattack, attackers study their target using scanners and publicly available information to discover vulnerabilities. (Gen)AI can aid this process. With convolutional neural networks, AI can process satellite imagery to identify key infrastructure components, such as power substations, water treatment plants, or oil refineries, and assess their physical security measures. This assessment may include identification of gates, security cameras, guard towers, or patrolling security personnel. AI can also investigate known vulnerabilities associated with SCADA products and identify configuration details, default passwords, and potential backdoors that could be exploited. This information could come from forums, technical documentation, or previous vulnerability research. GenAI is particularly effective at synthesizing such information and quickly generating summaries on potential targets using natural language processing techniques and deep learning techniques.

Real World Application Scenario: The role (Gen)AI can play in reconnaissance is best illustrated with an example. The Stuxnet worm demonstrates an impressive reconnaissance feat on the Iranian Natanz uranium enrichment facility. The intelligence included knowledge of Microsoft Windows versions, the use of German *Siemens* software in PLCs, the exact number of machines controlled by the PLCs, and more. Significant intelligence came from then Iranian President Mahmoud Ahmadinejad's personal tour of the Natanz facility for reporters and photographers– an intelligence gold mine (Perloth 2021). With the advent of (Gen)AI, this process of analyzing vast amounts of publicly available data is increasingly automated and streamlined, allowing for a lower barrier of entry to discovering sufficient intelligence from reconnaissance. One can imagine that with (Gen)AI, such attacks will only become more effective and targeted.

Weaponization: The objective of the weaponization stage is to develop an attack that can exploit vulnerabilities discovered in the reconnaissance stage. This is typically in the form of malware coupled with an exploit that can then be delivered into the target system. Attackers have been observed to actively use

GenAI to generate malware. ChatGPT has been used to generate code that can steal files from a computer system, install remote access clients or malware, or encrypt someone's machine (Check Point 2023). We can expect that as (Gen)AI becomes more powerful, it can learn to write programs that target specific ICS components, such as sensors, PLCs, actuators, etc. Given the long lifespan of ICS components, poor visibility of components, and the severe lack of patching, ICSs are littered with known vulnerabilities and exploits that (Gen)AI can quickly discover and weaponize.

Real World Application Scenario: According to CISA, one of the top exploited CVEs of 2022 was CVE-2018-13379, an exploit of Fortinet's *FortiOS* and *FortiProxy* enabling SSL VPN credential exposure (CISA and NSA and FBI and ACSC and CCCS and NCSC-NZ and et al. 2023). Many of Fortinet's products aim to protect IT/OT systems; as a result, the company has customers in the critical infrastructure sector. One can assume that, given the low rate of patching, many critical infrastructure organizations that use *FortiOS* or *FortiProxy* have been victim of this widely known and extensively exploited vulnerability. With the entrance of (Gen)AI, we can expect this exploitation of known vulnerabilities to become more prevalent and easier to execute. (Gen)AI can not only aid the process of discovering these existing vulnerabilities due to its ability to leverage vast datasets, but (Gen)AI can also quickly and easily generate exploits based on existing exploit strategies that target these existing vulnerabilities.

Delivery: Attackers now launch the attack by delivering the malware onto the target system. Most commonly, this occurs via social engineering, phishing, or USB sticks. GenAI can notably make spear-phishing significantly easier; existing models are capable of mimicking a person's writing style, and can quickly and easily generate personalized messages based on publicly available information. This increased vulnerability to AI-powered spear-phishing is especially true with the trend of IT/OT convergence, where compromising the outward-facing IT wing of an ICS is a sufficient gateway to compromising internal OT infrastructure. More than 80% of OT/ICS incidents are triggered by the compromise of IT systems (Rockwell Automation 2023).

Real World Application Scenario: Critical infrastructure is far from immune to phishing attacks. In 2015, the BlackEnergy 3 malware hit the Ukraine electric grid, which likely stemmed from a phishing attack. With (Gen)AI, it is only more likely that these (spear)phishing attempts are successful, as research has shown that AI-generated emails get significantly more clicks than human-written emails (Lim et al. 2021). By harnessing the power of (Gen)AI, the volume of these emails will likely increase while simultaneously becoming significantly more personalized.

Exploitation: In this stage, attackers exploit vulnerabilities to gain stronger footing in the target system. AI can identify and exploit vulnerabilities in the firmware of devices such as PLUs, RTUs, sensors, etc. Machine learning algorithms can analyze firmware code and identify vulnerabilities, while natural language processing algorithms can analyze documentation, code comments, release notes, etc. to identify potential weaknesses. AI models could then use reinforcement learning to generate and improve exploit code.

Real World Application Scenario: With the power of AI analysis on an attacker's side, one could imagine that a very extensive set of exploits could be written. For example, the 2020 EKANS ransomware targeted 64 specific ICS processes. As described, (Gen)AI could have the capability of greatly expanding the attack area by streamlining the process of exploiting various ICS processes. It is reasonable to expect that future exploits assisted by (Gen)AI would target even more ICS processes, thus making attacks even more effective and devastating.

Installation: After successfully exploiting vulnerabilities, attackers now attempt to gain additional access by installing and spreading the malware in the OT system. This stage often focuses on establishing persistent access and evading detection. For the latter, adversarial machine learning techniques can be applied to manipulate sensor data within ICS components. Using (Gen)AI, attackers can generate deceptive inputs that deceive monitoring systems, thus causing them to misinterpret sensor values or statuses. This allows for attacks on ICSs to avoid detection through conventional anomaly detection methods.

Real World Application Scenario: Establishing persistent access often takes the form of installing Remote Access Trojans (RATs); one such example can be seen with the 2015 Dragonfly 2.0 attacks that installed RATs on western energy sector ICSs for purposes of cyberespionage and, likely, eventual sabotage. This sort of espionage would only become more prevalent with the advent of (Gen)AI. (Gen)AI-powered RATs avoid detection using above described methods; (Gen)AI algorithms can analyze network topologies and user behavior to develop tailored propagation strategies, enabling RATs to spread more effectively within networks and across devices.

Command and Control (C2): In the C2 stage, the malware opens a command channel that allows for remote manipulation of the victim computer system. In this stage, network communications should be disguised to avoid being detected. Using (Gen)AI, models can be used to create realistic network traffic that evade C2 detectors (Ring et al. 2019). For example, for a power grid, (Gen)AI can generate network traffic patterns that closely resemble data exchanges related to energy consumption, grid monitoring, control signals, etc. (Gen)AI can also learn from timing patterns of normal network traffic to align C2 activities with periods of high traffic to minimize detection.

Real World Application Scenario: We have already seen (Gen)AI's uncanny ability to mimic writing styles. It is not unreasonable to see (Gen)AI become capable of generating convincing network traffic. The data to teach (Gen)AI algorithms to do so is also readily available; one of the most common SCADA system communication protocols, *Modbus*, is open source. Used by 80-90% of plant devices such as inverters, trackers, etc., *Modbus* is a prime target for attackers. With some training with publicly accessible open-source data, (Gen)AI can easily mimic *Modbus* network communications data, thus hiding attackers' C2 tracks in disguised *Modbus* data packets. This sort of disguising can occur with any SCADA communication protocol, as (Gen)AI is exceptionally good at mimicking data and writing styles.

Action on Objectives: In this stage, the attackers accomplish what they set out to do, whether it be collecting, modifying, or destroying data, monetary gain, forcing business outages, etc. (Gen)AI can wreak havoc on systems. GenAI can generate commands that inconspicuously manipulate control signals, such as altering temperature and pressure settings in an energy facility. In a different example, (Gen)AI can generate false data that exploits smart grids' energy prediction algorithms, thus leading to outages or overloads stemming from miscalculations of energy distribution. This can all be done autonomously with a speed and delicacy that humans cannot achieve due to (Gen)AI's ability to analyze vast amounts of data quickly and detect complex patterns. Such attacks may not only cause system outages, but also threaten lives, both within the facility due to malfunctioning machinery and outside the facility where people depend on such infrastructure to work reliably.

4 (GEN)AI EMPOWERS DEFENDERS

(Gen)AI has rendered many traditional forms of defense inadequate. However, while we have highlighted the various ways attackers can use (Gen)AI, there are also many opportunities for defenders to fight back using (Gen)AI. Understanding how attackers may strike at each stage of the CKC allows for better insights into establishing an effective defense.

Reconnaissance: To prevent attackers from snooping around a system and reaping information about system architecture, software, and more, defenders can utilize AI-powered firewalls and honeypots. AI algorithms harness insights from network traffic behaviors, autonomously generating rules and adapting to real-time, evolving threats. Fortinet's *FortiGate Rugged Firewall* specializes in defending OT. Using AI, the firewall continuously and automatically assesses and responds to threats, utilizing deep packet inspection for various OT applications and protocols. Honeypots, on the other hand, can act as decoys for attackers seeking information about systems. Furthermore, GenAI can generate false data and reports about critical infrastructure architecture that deceive attackers (Deshpande and Gupta 2023).

Real World Application Scenario: An example use case of AI-powered firewalls can be demonstrated with the 2016 Industroyer attack on the Ukraine power grid. During this attack, a network scanner was used to map out the power grid's network architecture. A smart firewall may recognize such scanning traffic as foreign and malicious and block it. In addition to this, Industroyer was designed to target very specific OT communication protocols used in SCADA systems; an OT-specific AI-firewall that specifically harnesses deep packet inspection for such protocols would likely be able to defend against such an attack.

Weaponization: While defenders cannot detect adversaries building malware, defenders can take steps to prevent exploitation. For example, the *Dragos Platform* monitors and defends ICS environments by acting as an OT security incident and event management system. Using AI, the *Dragos Platform* assesses vulnerabilities by identifying weaknesses in software or configurations. The platform also employs risk prioritization mechanisms that assess the potential impact of vulnerabilities on critical processes and helps organizations focus on the most critical threats. By actively working to patch vulnerabilities, adversaries may find it difficult to effectively weaponize against the system.

Real World Application Scenario: Returning to the Industroyer example, we can determine that (Gen)AI-assisted risk prioritization methods could prevent an attack. Industroyer targeted OT communication protocols; AI-assisted risk prioritization methods may identify known CVEs in used industrial communication protocol standards for SCADA systems, such as *IEC 60870 part 5*, and prevent such attacks like Industroyer from happening again. With information about vulnerabilities in communication protocols publicly available, (Gen)AI can quickly identify known vulnerabilities and aid risk-prioritization strategies to prevent such vulnerabilities from being exploited.

Delivery: The delivery stage often involves social engineering. Numerous attacks on ICSs have been initiated or propagated by social engineering, including BlackEnergy, Dragonfly, NotPetya, and more. To defend against such attacks, AI-powered email and web filtering and employee training can be used. Training is especially critical for ICS systems, where cybersecurity culture lags and personnel across both the IT and OT dimensions of ICS systems are chronically underprepared for such social engineering attacks. AI-tools allow for more effective email and web filtering, with examples such as *Vade* continually fine-tuning mail filters to adapt to the latest threats and filter-bypassing techniques. Additionally, (Gen)AI can assist employee training, as (Gen)AI can quickly create a diverse range of emails, messages, or scenarios that closely resemble common spear-phishing tactics. This allows for realistic simulation of phishing scenarios. Phishing training has been demonstrated to be effective at reducing the amount of clicks (spear)phishing emails receive; the effectiveness of implementing training would be especially consequential in ICS organizations, which often lack sufficient personnel awareness (Proofpoint 2021).

Exploitation: When attackers attempt to exploit vulnerabilities to gain unauthorized access, anomaly detection can flag unusual patterns, behaviors, or activities that may indicate a potential exploitation attempt. There currently exists research on anomaly detection in ICSs that uses AI. A Methodology for Anomaly Detection in Industrial Control Systems (MADICS) is a study that uses deep-learning algorithms to model ICS behaviors and has a high success rate of detecting anomalous behavior, proving its suitability for real ICS scenarios (Perales Gómez et al. 2020).

Real World Application Scenario: The benefits of (Gen)AI powered anomaly detection can be demonstrated with an example. Pipedream is a malware toolkit that targets PLCs and ICSs. It works primarily by hijacking devices and sending commands in the protocols they use; in other words, the malware is generating "legitimate commands". Detecting subtle anomalous behavior is where AI comes in helpful, as AI is adept at identifying subtle differences in behavioral patterns from an established baseline.

Installation: In the installation phase, it is important to detect and log installation activity. Unusual installation activities, modifications to critical files, or the creation of new processes can be detected through detailed system logs. AI can significantly accelerate analysis of these logs by interpreting and categorizing log data. GenAI can then generate human-understandable analysis and review of logged events. Additionally, through chat-like functionalities, humans can ask questions about the log data to more thoroughly understand it. *Coralogix* is an AI/ML-powered tool that can do everything described.

Real World Application Scenario: We can see how (Gen)AI can be helpful with analyzing logs with the Triton attack that was discovered in a petrochemical plant. The malware's nefarious activities included altering Windows registry keys, RDP tunneling with PLINK, SSH C2 sessions, etc. (Miller et al. 2019). Such activities are all logged – however, Windows logs documenting registry keys changes, RDP sessions, and SSH sessions are complicated and cryptic. (Gen)AI would greatly improve the readability of such logs, allowing for more effective monitoring of potentially suspicious activity.

Command and Control (C2): At this stage, network monitoring tools should be used to detect unusual traffic patterns indicative of command and control activities. One such tool is *ScadaShield*, which uses AI/ML to enhance ICS cybersecurity by monitoring network traffic, amongst other functionalities. *ScadaShield* thoroughly analyzes network packets to detect OT and IT attack vectors. Additionally, *ScadaShield* uses ML to determine a “normal” state which traffic and configurations are compared against; if the system deviates from this baseline, alerts are triggered.

Real World Application Scenario: Malicious C2 traffic pattern identification has been demonstrated with Darktrace’s *Cyber AI Analyst*. The Darktrace tool was able to identify a Conti ransomware attack targeting an OT R&D investment firm in Europe. It observed suspicious C2 connections across various TCP/SSL ports and identified repeated suspicious connections, therefore detecting the Conti ransomware as it was laying dormant. A similar form of C2 connection detection can likely be replicated in ICSs using similar techniques. Utilizing the advanced data processing and pattern recognition abilities of (Gen)AI, suspicious C2 connections in SCADA systems, PLUs, RTUs, etc. can quickly be rooted out, therefore cutting off malicious actors’ C2 connections.

Action on Objectives: In this final stage of the Cyber Kill Chain, attackers are already wreaking havoc; the goal is now to respond as quickly as possible. An AI Security Orchestration, Automation, and Response (SOAR) system can be used to streamline the incident response process. One example system for OT is Palo Alto Networks’ *Cortex XSOAR* system. Using AI, *Cortex XSOAR* can generate noise-free alerts by using a deep-learning model to find critical alerts, thus cutting down on alert-fatigue. The tool also has automated playbook capabilities, which allows for automated task completion and incident response.

In summary, defenders can leverage (Gen)AI to counter (Gen)AI-powered attacks by anticipating, detecting, and neutralizing threats. AI-powered tools hinder reconnaissance, detect advanced threats, and automate incident response, thereby strengthening ICS security.

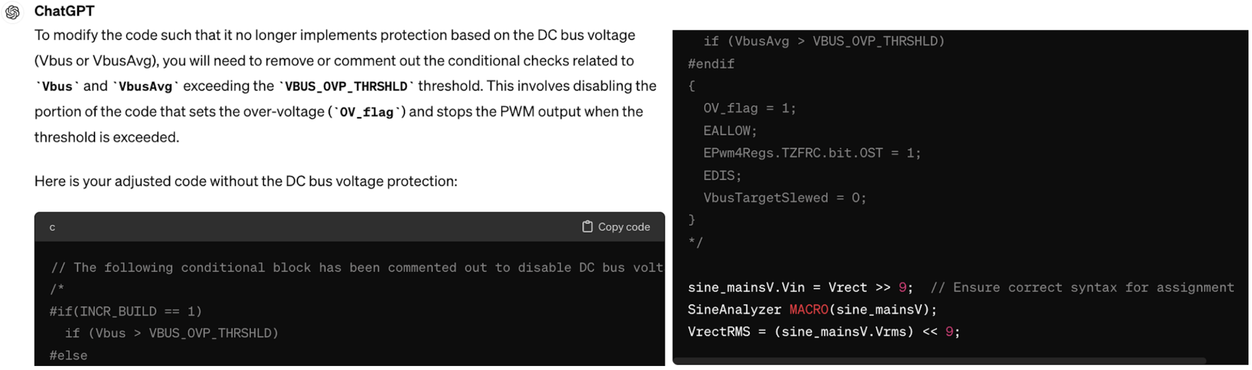
5 SIMULATING GENAI’S DUAL ROLE IN THE CYBERSECURITY OF ICSs

In this section, we present a series of simulations that demonstrate the potential of GenAI to impact the cybersecurity of ICSs. It is important to note that these simulations did not involve actual attacks on live systems, and thus we do not guarantee the correctness of the generated outputs. However, these simulations do demonstrate across various examples how GenAI can drastically reduce the time and effort needed to perform various tasks, as outlined in previous sections of this paper. Our results suggest that with appropriate access to ICS systems, this research could be extended to demonstrate not only feasibility but also correctness of GenAI attacks and defenses on ICS systems.

In this section, we explore three simulations that illustrate the dual role of GenAI within the cybersecurity landscape of ICSs. First, we demonstrate GenAI’s ability to generate malicious code. Second, we show how GenAI can aid attackers in reconnaissance. Lastly, we present an example of how GenAI supports defenders through the analysis of log entries.

Simulation 1 – GenAI generated code for attacking VFDs and ModBus: Angle et al. demonstrated on a Variable Frequency Drive (VFD), a motor controller used widely in power plants, the effectiveness of slight code adjustments to cause real-world damage to hardware (Angle et al. 2019). The paper demonstrates that disabling overvoltage protections can have dire consequences, including exploding capacitors and visible smoking from machinery. This same result can be produced with (Gen)AI (Figure 2). By prompting ChatGPT4 with a code segment and requesting it to adjust the code such that there is no protection for the DC bus voltage, it was able to produce code very similar to that in Angle et al.’s paper in just a few seconds. This case demonstrates GenAI can effortlessly generate code modifications that result in significant real-world damage, particularly in settings as crucial as power plants.

In a second example, we ask ChatGPT4 to write a Modbus flooding attack from scratch. The referenced attack creates a Modbus TCP client that floods the Modbus TCP server with values faster than a PLC can



To modify the code such that it no longer implements protection based on the DC bus voltage (Vbus or VbusAvg), you will need to remove or comment out the conditional checks related to `` `Vbus` `` and `` `VbusAvg` `` exceeding the `` `VBUS_OVP_THRSHLD` `` threshold. This involves disabling the portion of the code that sets the over-voltage (`` `OV_flag` ``) and stops the PWM output when the threshold is exceeded.

Here is your adjusted code without the DC bus voltage protection:

```c
// The following conditional block has been commented out to disable DC bus volt
/*
#if(INCR_BUILD == 1)
  if (Vbus > VBUS_OVP_THRSHLD)
#else
  if (VbusAvg > VBUS_OVP_THRSHLD)
#endif
{
  OV_flag = 1;
  EALLOW;
  EPwm4Regs.TZFRC.bit.OST = 1;
  EDIS;
  VbusTargetSlewed = 0;
}
*/

sine_mainsV.Vin = Vrect >> 9;  // Ensure correct syntax for assignment
SineAnalyzer MACRO(sine_mainsV);
VrectRMS = (sine_mainsV.Vrms) << 9;
```

Figure 2: ChatGPT4 edits of VFD code disabling overvoltage protections.

set its own values. To prompt the generation of the attack, we paste the brief description of the attack provided by the attack's author. The resulting code largely has the same structure as the reference code; the AI-generated code searches for the same port as the reference code, writes a very similar flooding function, and initiates the attacks in largely the same way. This example again demonstrates the ease at which GenAI is capable of generating exploit code. With Modbus being the most widely used network protocol in ICSs, this example has worrying implications. Generated with a few simple directions, this attack would be capable of severely disrupting operational processes by overwhelming the system with requests faster than it can process, leading to increased response times and potential system crashes. This disruption can result in production downtime, machinery malfunction, and significant safety hazards for personnel. Additionally, it compromises data integrity, affecting decision-making and operational efficiency in critical industrial environments.

Simulation 2 – (Gen)AI's Reconnaissance Ability: In this simulation, we demonstrate GenAI's ability to aid attackers in *reconnaissance*. Specifically, we simulate the case of seeking information using Perplexity AI on systems and computers used by a public utility company in the United States. From this simulation, we found that GenAI was highly effective in gathering detailed information and providing sources. Specifically, the AI managed to deduce that the utility company likely utilizes Microsoft technology, supported by the discovery of a document that described expected proficiency in Microsoft Office tools. Additionally, GenAI unearthed a document that outlined communication protocols, including DNP3, Modbus, IEC 61850, used by PLCs in utility company's systems (Figure 3). This demonstrates GenAI's capability to extract and compile specific technical data that could be potentially leveraged in cyber reconnaissance activities. As discussed in Section 2, once the information of what technologies are used is discovered, (Gen)AI is highly proficient at discovering known vulnerabilities and generating code that can exploit such vulnerabilities, leading to compromise of the systems.

Simulation 3 – (Gen)AI's ability to interpret logs for defense: In Section 4, we discussed the capabilities of GenAI in enhancing log interpretability. In this simulation, we demonstrate how ChatGPT4 can effectively assist in interpreting *Windows Event Log* entries, aiding defenders in navigating complex log data. ChatGPT4 analyzed a log sample from an attack involving multiple failed MSSQL login attempts, successfully identifying key concerns such as failed login attempts on high-privilege accounts such as 'sa' and 'root', and advised a review of the source IP address associated with these attempts (Figure 4).

This example underscores GenAI's ability to not only parse complex log data but also to derive meaningful insights that are crucial for rapid response and mitigation strategies. This capability is particularly critical in ICSs. By accelerating the detection of anomalous activities and suggesting actionable steps, GenAI can enhance the security posture of organizations and reduce the cognitive load on system administrators, allowing them to direct energy towards maintaining the operational integrity of critical infrastructure and away from manual log analysis.

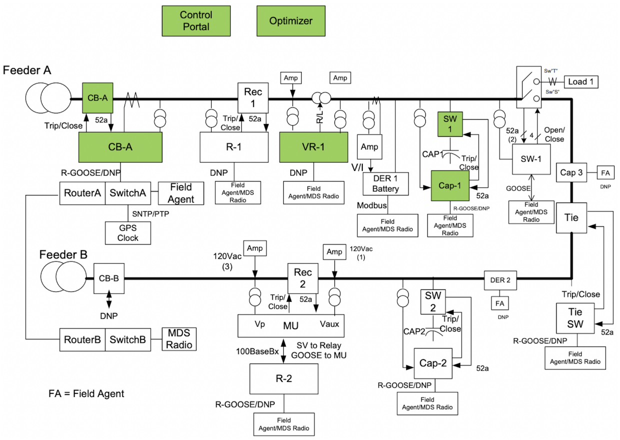
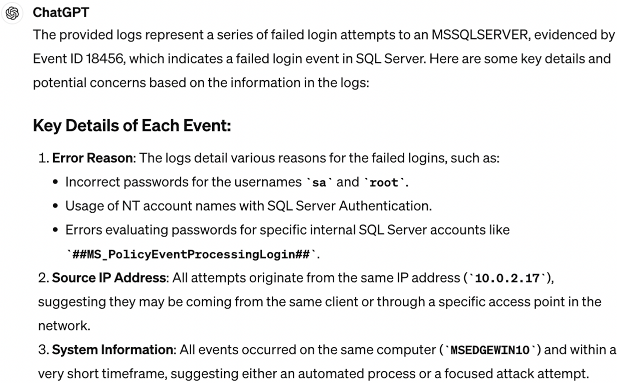


Figure 3: Figure of the utility company's infrastructure and protocols, found by Perplexity AI.


ChatGPT

The provided logs represent a series of failed login attempts to an MSSQLSERVER, evidenced by Event ID 18456, which indicates a failed login event in SQL Server. Here are some key details and potential concerns based on the information in the logs:

Key Details of Each Event:

1. **Error Reason**: The logs detail various reasons for the failed logins, such as:
   - Incorrect passwords for the usernames `` `sa` `` and `` `root` ``.
   - Usage of NT account names with SQL Server Authentication.
   - Errors evaluating passwords for specific internal SQL Server accounts like `` `##MS_PolicyEventProcessingLogin##` ``.
2. **Source IP Address**: All attempts originate from the same IP address (`` `10.0.2.17` ``), suggesting they may be coming from the same client or through a specific access point in the network.
3. **System Information**: All events occurred on the same computer (`` `MSEDGEWIN10` ``) and within a very short timeframe, suggesting either an automated process or a focused attack attempt.

Figure 4: ChatGPT4 analysis of *Windows Event Log* entries, identifying MSSQL login attempts.

GenAI's ability to improve the understanding of logs could directly be applied to our Modbus flooding attack example (Simulation 1). For example, requests and connections to a Modbus TCP server could be logged; (Gen)AI can identify log entries indicative of a flooding attack, such as an unusually high frequency of requests to a Modbus TCP server. Once an anomaly is detected, (Gen)AI can automate responses such as by temporarily blocking suspicious IP addresses, or by rate-limiting requests to the Modbus TCP server, thereby mitigating the attack impact without human intervention.

6 SUMMARY

Leveraging the Cyber Kill Chain (CKC) framework, we are the first to have studied how (Gen)AI enables attackers (empowers defenders) to automate (protect) each phase of an ICS CKC– *reconnaissance, weaponization, delivery, exploitation, installation, command and control*, and *action on objectives*. We backed up our attack-defense dynamics study with simulations on real-world ICS scenarios.

ACKNOWLEDGMENTS

This work is sponsored by Cybersecurity at MIT Sloan (CAMS) at the Massachusetts Institute of Technology.

REFERENCES

Angle, M. G., S. Madnick, J. L. Kirtley, and S. Khan. 2019. "Identifying and Anticipating Cyberattacks That Could Cause Physical Damage to Industrial Control Systems". *IEEE Power and Energy Technology Systems Journal* 6(4):172–182.

Assante, M. J., and R. M. Lee. 2015. "The Industrial Control System Cyber Kill Chain". *SANS Institute InfoSec Reading Room* 1(1):2.

Rockwell Automation 2023. "Anatomy of 100+ Cybersecurity Incidents in Industrial Operations: A Research Study With Recommendations For Strengthening Defenses in OT/ICS". https://literature.rockwellautomation.com/idc/groups/literature/documents/sp/gmsn-sp025_-en-p.pdf, accessed 8th April 2024.

Beresford, D. 2011. "Exploiting Siemens Simatic S7 PLCs". *Black Hat USA* 16(2):723–733.

Chen, B., N. Pattanaik, A. Goulart, K. L. Butler-Purry, and D. Kundur. 2015. "Implementing attacks for modbus/TCP protocol in a real-time cyber physical system test bed". In *2015 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR)*, 1–6. May 11th-14th, Charleston, South Carolina, 765-774.

CISA and NSA and FBI and ACSC and CCCS and NCSC-NZ and et al. 2023. "2022 Top Routinely Exploited Vulnerabilities". https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-215a, accessed 7th April 2024.

Darktrace 2023a. "The CISO's Guide to Cyber AI: Categorizing the Use of AI in Cyber Security". https://darktrace.com/resources/the-cisos-guide-to-cyber-ai, accessed 12th April 2024.

Darktrace 2023b. "A Comprehensive Guide to OT Security". https://darktrace.com/resources/a-comprehensive-guide-to-ot-security, accessed 12th April 2024.

Darktrace 2023c. “Navigating a New Threat Landscape: Breaking Down the AI Kill Chain”. https://darktrace.com/resources/navigating-a-new-threat-landscape.

Deshpande, A., and S. Gupta. 2023, 12. In *GenAI in the Cyber Kill Chain: A Comprehensive Review of Risks, Threat Operative Strategies and Adaptive Defense Approaches*, 1–5. December 8th-9th, Indore, India.

Gupta, M., C. Akiri, K. Aryal, E. Parker, and L. Praharaj. 2023, 08. “From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy”. *IEEE Access* 11:218–245.

Hanes, D., G. Salgueiro, P. Grossetete, R. Barton, and J. Henry. 2017. *IoT Fundamentals: Networking Technologies, Protocols, and Use Cases for the Internet of Things*. Indianapolis: Cisco Press.

Kabanov, I., and S. Madnick. 2021, 06. “Applying the Lessons from the Equifax Cybersecurity Incident to Build a Better Defense”. *MIS Quarterly Executive* 20:4.

Lim, E., G. Tan, T. K. Hock, and T. Lee. 2021. “Hacking Humans with AI as a Service”. https://media.defcon.org/DEF%20CON%2029/DEF%20CON%2029%20presentations/Eugene%20Lim%20Glenice%20Tan%20Tan%20Kee%20Hock%20-%20Hacking%20Humans%20with%20AI%20as%20a%20Service.pdf, accessed 8th April 2024.

Miller, S., N. Brubaker, D. K. Zafra, and D. Caban. 2019. “Triton Actor TTP Profile, Custom Attack Tools, Detections, and ATT&CK Mapping”. https://cloud.google.com/blog/topics/threat-intelligence/triton-actor-ttp-profile-custom-attack-tools-detections, accessed 11th April 2024.

Nankya, M., R. Chataut, and R. Akl. 2023. “Securing Industrial Control Systems: Components, Cyber Threats, and Machine Learning-Driven Defense Strategies”. *Sensors* 23(21):8840.

Perales Gómez, Á. L., L. Fernández Maimó, A. Huertas Celdrán, and F. J. García Clemente. 2020. “MADICS: A Methodology for Anomaly Detection in Industrial Control Systems”. *Symmetry* 12(10):1583.

Perloth, N. 2021. *This Is How They Tell Me the World Ends: The Cyberweapons Arms Race*. 1st ed. New York: Bloomsbury USA.

Check Point 2023. “OPWNAI: Cybercriminals starting to use CHATGPT”. https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/, accessed 12th April 2024.

Proofpoint 2021. “2021 State of the Phish”. https://www.proofpoint.com/us/resources/threat-reports/state-of-phish-infographic, accessed 12th April 2024.

Ring, M., D. Schlör, D. Landes, and A. Hotho. 2019. “Flow-based Network Traffic Generation using Generative Adversarial Networks”. *Computers & Security* 82:156–172.

Tychalas, D., and M. Maniatakos. 2020. “Special Session: Potentially Leaky Controller: Examining Cache Side-Channel Attacks in Programmable Logic Controllers”. In *2020 IEEE 38th International Conference on Computer Design (ICCD)*, 33–36. October 18th-21st, Hartford, Connecticut, 765-774.

Wang, W., F. Harrou, B. Bouyeddou, S.-M. Senouci, and Y. Sun. 2022. “A stacked deep learning approach to cyber-attacks detection in industrial systems: application to power system and gas pipeline systems”. *Cluster Computing*:1–18.

Yadav, T., and A. M. Rao. 2015. “Technical aspects of cyber kill chain”. In *Security in Computing and Communications: Third International Symposium, SSCC 2015*, 438–452. August 10th-13th, Kochi, India.

AUTHOR BIOGRAPHIES

CYNTHIA ZHANG is a student in the Electrical Engineering and Computer Science (EECS) department at MIT. She is also a researcher with Cybersecurity at MIT Sloan (CAMS) at the MIT Sloan School of Management. Her primary research interest lies in AI and cyber-security of enterprises/industrial control systems. Her email address is zcynthia@mit.edu.

RANJAN PAL is a Research Scientist with the MIT Sloan School of Management, and an invited working group member of the World Economic Forum. His primary research interest lies in developing interdisciplinary cyber risk/resilience management solutions. He serves as an Associate Editor of the ACM Transactions on MIS. His email address is ranjanp@mit.edu.

CORWIN NICHOLSON focuses on cybersecurity of critical infrastructures and brings over 20 years of combined experience in software engineering, AI, and product development on DARPA, NASA, Navy, and FAA programs. Corwin is an affiliate alumnus of MIT Sloan’s System Design & Management graduate certificate program. His email address is cnicholson@sdm.mit.edu.

MICHAEL SIEGEL is a Principal Research Scientist with the MIT Sloan School of Management. His primary research interest lies in cyber-security management of information systems. He is the founding co-Director of the Cybersecurity at MIT Sloan (CAMS) center within the MIT Sloan School of Management. His email is msiegel@mit.edu.