ITAI 3377

Kaylee Auguillard

Cheyenne Hathaway

Daniel Wright

**Cybersecurity Plan for AI-integrated Smart Factory**

**System Design and Vulnerability Identification**

**System Selection and Design:**

A smart factory integrates artificial intelligence and industrial internet of things to automate and optimize manufacturing processes. This can include several components, such as sensors, AI driven analytics to process data in real time to detect a number of issues, and even robots, all in an effort to reduce downtime and improve efficiency. These proposed systems would rely on continuous data collection, machine learning models, and network communication to function.

The key components of this hypothetical system would combine emerging technologies with long-existing technology.This would include IoT devices such as robotic arms, conveyor belts, and CNC machines that perform automated tasks. In addition, Industrial Control Systems (ICS) like PLCs and SCADA would manage operations, while AI models predict failures and optimize workflows. Sensors provide real-time monitoring of temperature, pressure, and vibration, ensuring all machinery is performing as expected. This also allows for detection of manufacturing issues early, before too many defective items are created. It also means a reduction of downtime, as mentioned, because issues with the machinery will be detected earlier, allowing for repairs before the issue is too serious.

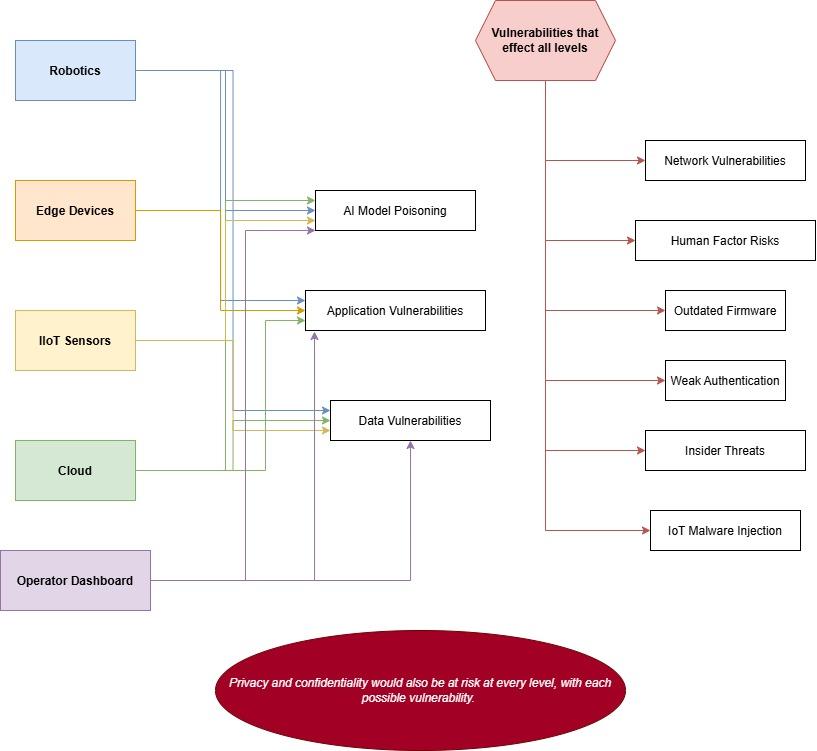
This would require that the network infrastructure connect devices through a mix of Ethernet, Wi-Fi, and 5G to enable fast data transmission. Edge computing processes data locally before sending it to cloud storage for deeper analysis. As seen with the smart city project earlier in the semester, this will also allow for adherence to privacy and confidentiality constraints, as the data can be processed and deleted, with only the relevant information being sent to the cloud for deeper analysis. The data flow follows a structured path: sensors collect data, AI models process it, and insights are delivered to operators through dashboards.

**Vulnerability Assessment:**

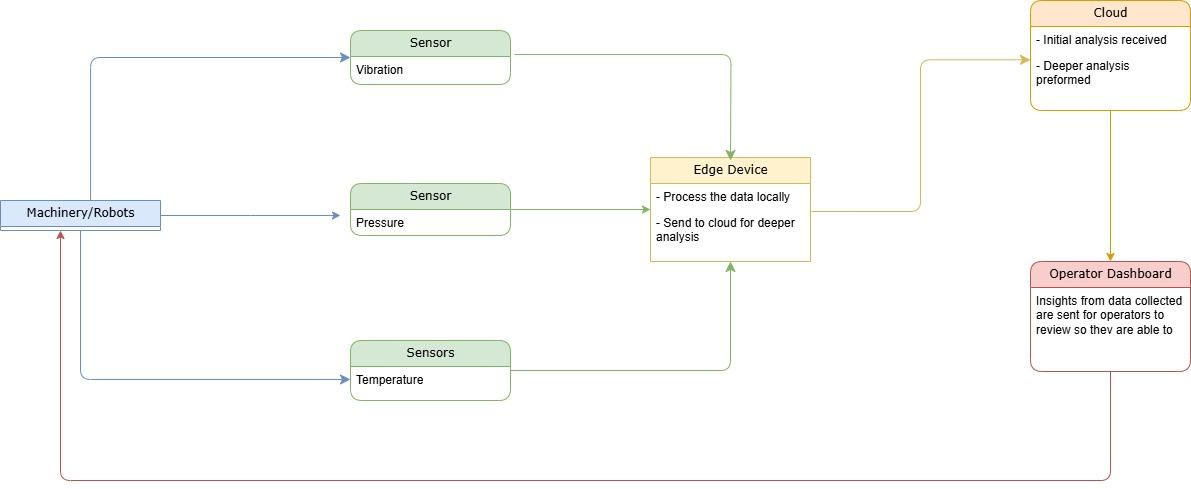
Even with the added security of local processing, there are still several vulnerabilities to keep in mind, and stay ahead of:

* **Outdated firmware**, which leaves machines open to exploits.
* **Weak authentication methods** that can allow unauthorized access
* **Malware** injected into IoT devices, leading to system- wide disruptions and/or unauthorized data extraction.
* **Network vulnerabilities** arise from weak authentication protocols and unencrypted communications, making it easier for attackers to intercept data. Without network segmentation, a breach in one area could spread across the entire system, compromising critical operations.
* **Data vulnerabilities** include improper access controls that allow unauthorized users to modify or steal sensitive information.
* **AI model poisoning** is another risk, where attackers manipulate training data to cause inaccurate predictions, leading to faulty decision-making.
* **Application vulnerabilities** may include hard coded credentials, making systems an easy target for attackers who gain access to the source code.
* **Unsecured interfaces between applications and devices** can allow unauthorized data extraction, leading to information leaks.
* **Human factor vulnerabilities** involve weak passwords, social engineering attacks like phishing, and a general lack of security awareness among employees.
* **Insider threats** are also a risk, where employees with access to critical systems may unintentionally or intentionally compromise security.

A diagram is below, showing what parts of the system the different threats can affect:



And here is a diagram showing where data is shared:



**Defense Strategy Development**

**Defense Measures:**

Despite the many vulnerabilities this system has, there are as many ways to keep the system secure. First, secure by design principles should be implemented, including a zero-trust model where every system request has to be verified. Additionally, the least privilege access approach ensures that users and devices only have the minimum permissions necessary to perform tasks. These models will address potential vulnerabilities to the network and data, as well as assist in simple human user error, as people who are not thoroughly trained on security protocols will not have access. For the same reason, it can address potential insider threats, as users would not have access to any part of the system unless absolutely necessary.

The next weakness to address is weak authentication issues. To get ahead of this potential problem, authentication and access control should include multi-factor authentication (MFA) and role-based access control (RBAC) to restrict unauthorized access. Strong password policies and automated session timeouts will further protect sensitive areas of the system. This again aids in the human factor risks, as well as possible insider threats.

To further address possible data security weaknesses, encryption and data protection measures should involve end-to-end encryption. Secure API gateways should enforce authentication, and regular backups should be maintained to prevent data loss.

Added network security will require proper segmentation to ensure that different parts of the factory operate in isolated zones. Firewalls, intrusion detection/prevention systems (IDS/IPS), and VPNs should be used to filter traffic and secure remote access. This will also aid in possible malware threats. If a zone is compromised, it is already isolated, so the risk of it spreading is far lower. It also helps to address insider threats, as users will have access only to the necessary zones, aiding the least privilege access model.

For more protection against malware, secure software development should focus on regular patching of vulnerabilities and enforcing secure coding practices. Additionally, all firmware and apps must be kept up to date to ensure any security patch is in place as soon as possible. As part of this, security monitoring and incident response should include real-time logging and SIEM (Security Information and Event Management) tools. An incident response plan should be documented and tested regularly to ensure quick containment of security breaches.

AI-specific security measures must protect AI models from poisoning attacks by verifying data sources. Anomaly detection should be integrated to identify unusual patterns that could indicate malicious activity. AI models should be validated and retrained using clean datasets to prevent adversarial attacks to reduce the threat of model poisoning.

Finally, physical security measures should be in place, such as restricting access to network equipment and using tamper-proof hardware. Surveillance cameras and biometric authentication can prevent unauthorized individuals from accessing critical systems to prevent or reduce the harm of intentional insider threats, as well as threats of simple user error.

**Implementation Plan:**

To implement the proposed security measures, first a security assessment is necessary. A full evaluation of the system will show all vulnerabilities, and which risks have the potential to pose an immediate threat. This step will help determine which security controls need immediate implementation. From there, assuming there is not a need to prioritize one threat over another, here is the proposed implementation plan:

1. Authentication and encryption setup – Enforce MFA, RBAC, and encrypted data storage. Ensuring strong authentication mechanisms will significantly reduce unauthorized access attempts.
2. Network segmentation and monitoring – Implement VLANs, IDS/IPS, and firewalls to limit lateral movement. Real-time network monitoring tools should be deployed to detect and respond to suspicious activities at this stage as well.
3. Secure software updates – Develop a process for regular patching and scanning for vulnerabilities. As part of this process, AI models should be periodically retrained and validated to prevent data poisoning, and any third party apps that are used need to be checked for pending updates.
4. Employee training and awareness – Conduct security awareness training on phishing, password security, and social engineering attacks. Employees should be able to recognize and report potential security threats
5. Physical security- If not already in place, tamper proof seals on all hardware should be installed, and code or keycard access should be required to access equipment.
6. Penetration testing – Simulate attacks to test the effectiveness of security measures. This process will help identify weaknesses and refine the defense strategy.

**Penetration Testing Simulation**

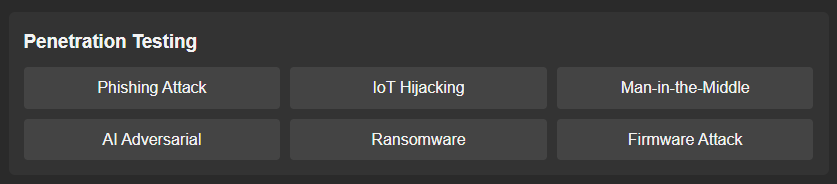
**Attack Simulation:**

Here is the proposed outline for a testing simulation to check the proposed security, and proposed improvements should the security plan fail:

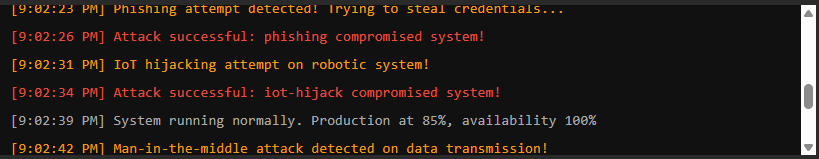
* Phishing attacks will be tested by sending simulated phishing emails to evaluate how employees respond. This will help determine the effectiveness of training and identify areas for improvement. If the phishing attempt succeeds, any employee who fell for the phishing attempt will have to complete further training on company security measures and why they are in place.
* IoT hijacking simulations will involve attempting to exploit outdated firmware or weak authentication on smart devices. The goal is to see if attackers can gain control over factory machinery or extract sensitive data. If the hijacking is successful, there will be an immediate prioritization of updating the firmware, be it in the form of a patch or third party software update.
* Man-in-the-middle (MitM) attacks will test network security by intercepting data transmissions. If successful, this could expose weak encryption or lack of secure communication protocols. This would call for encryption policies to be evaluated for improvements, and system communication to be updated.
* AI adversarial attacks will be simulated by introducing manipulated input data to the AI models. The objective is to evaluate how well the models can detect and mitigate adversarial threats. This is a test of the model itself, so if successful, the model will go through more training on trusted data before being tested again.
* Ransomware simulations will involve encrypting test data and observing how well the system can detect and recover from an attack. This will test the effectiveness of backup and incident response strategies. If this attack is successful, the system to detect anomalies may require an update, and the policies in place around human response when an anomaly is detected may require re-evaluation.

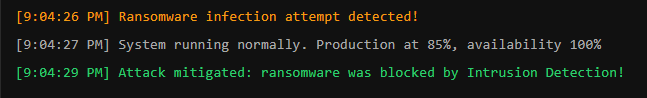
Also included are screenshots of the simulation we created. First is the simulation dashboard. Here we can enable or disable proposed security measures to show why they are necessary. It also tracks the production rate, system availability, anomalies detected, and data breaches:



Here are the simulated attacks we can run to test the system:

And finally, here is where the system notices the attack and determines if the attack was successful.





Website for the simulation: https://hccai.ngrok.io/#simulation

**Further Assessment and Improvement:**

After each attack simulation, the team will analyze how well the defense mechanisms performed and identify any weaknesses. Documenting attack success rates and security lapses will help refine strategies. These assessments will identify weak points in authentication, encryption, or access control. The weakness will be noted, and additional security measures will be proposed. If a specific attack was successful, a plan for strengthening defenses will be developed, depending on what kind of attack was successful. The final penetration testing report will summarize key findings, lessons learned, and recommendations for future improvements. This will ensure that security measures evolve and remain effective.

**Sources:**

* Microsoft. (n.d.). *Security design principles*. Microsoft Learn. Retrieved April 1, 2025, from<https://learn.microsoft.com/en-us/azure/architecture/guide/security/security-start-here>
* StrongDM Team. (2024, March 18). *Zero trust vs. the principle of least privilege: What's the difference?* StrongDM. Retrieved April 1, 2025, from<https://www.strongdm.com/what-is/zero-trust-vs-principle-of-least-privilege>
* Melecio, N. (n.d.). *Enhancing cybersecurity for IoT and AI systems: A comprehensive overview*. ScaleUp Lab Program. Retrieved April 1, 2025, from<https://scaleuplab.gatech.edu/enhancing-cybersecurity-for-iot-and-ai-systems-a-comprehensive-overview/>​
* DataGuard. (n.d.). *10 key challenges and cybersecurity solutions for smart factories in manufacturing*. Retrieved April 1, 2025, from<https://data-guard365.com/manufacturing/10-key-challenges-and-cybersecurity-solutions-for-smart-factories-in-manufacturing/>