Final Take-home Assignment

BY HAN THAO NGUYEN
#991575791

AI IN CYBERSECURITY

I. Overview

This assignment has provided me an great opportunity to explore how artificial intelligence is not just a buzzword but a practical and necessary tool for modern cybersecurity. As the Canadian Centre for Cyber Security (CCCS) and other reports have made clear, the growing volume of data and the sophistication of threats mean that human security analysts cannot keep up on their own. From my perspective, the key is to strategically implement AI to augment human capabilities, not replace them.

II. Mapping CCCS Controls to AI Use Cases

I believe that two of the most critical CCCS controls that AI can significantly enhance are "Security Event Logging and Monitoring" and "Automated Threat Detection and Response." Without AI, these controls would quickly become overwhelming and ineffective in today's threat landscape.

1. Security Event Logging and Monitoring

The CCCS emphasizes the importance of logging and monitoring to detect security incidents. I think this is a foundational requirement, but simply collecting logs is not enough. The sheer volume of data generated by a large organization, such as a major Canadian telecommunications company, is too great for humans to sift through manually. My personal experience working as a Technology Analyst for a

municipal government at the Region of Peel has shown me the scale of this problem firsthand, as we manage thousands of devices and countless events daily.

In my co-op experience with TD, I was responsible for creating tools to data mine and filter outliers across six provinces. I quickly learned that the old way of doing things, where a single account search could take at least 30 minutes, was completely unsustainable. This is where anomaly detection comes in. An AI model can learn the "normal" behavioral baseline for a user or a system and flag anything that deviates from it. For example, a bank could use an **autoencoder** to analyze massive databases of debit and credit card accounts. The autoencoder would be trained on millions of normal transactions, learning to efficiently compress and recreate the data. If a series of suspicious transactions suddenly occur, the autoencoder would fail to compress this abnormal pattern, resulting in a large "reconstruction error." This would immediately flag the event, allowing an analyst to intervene in real time. This proactive detection of an outlier is exactly what the CCCS's guidance on "Advancing Cyber Threats in AI" suggests, allowing defenders to catch "unknown, novel, or stealthy attacks" that would otherwise go unnoticed.

2. Automated Threat Detection and Response

The CCCS control for "Automated Threat Detection and Response" is another area where AI is indispensable. I think a perfect Canadian use case for this is in healthcare, where the primary threat vector for ransomware is often a phishing email. My previous work on the Operation Defend the North tabletop exercise (Computer Security) highlighted this vulnerability, which is a key weakness in many hospitals.

To counter this, a supervised classification model can be trained to automatically detect and respond to threats like phishing. The model would be fed a large dataset of emails labeled as either "SPAM" or "Not SPAM". Features for this model would include the sender's address, the presence of suspicious links, and keywords like "urgent" or "invoice." The model could then be integrated into an email

security system. When a new email arrives, the model would classify it as either benign or a threat. If it's classified as a threat with a high confidence score, an automated response can be triggered, such as quarantining the email and notifying the security team. I believe that focusing on metrics like **precision and recall** is crucial here. As our class notes emphasize, a high performing model must have high precision to avoid flagging legitimate emails as SPAM and high recall to ensure no malicious emails make it to the user's inbox. This automated process significantly reduces the burden on human analysts and provides a scalable defense that is critical for protecting sensitive patient data.

III. Policy Meets Practice: AI Risks in a Canadian SOC

While the CCCS provides solid guidance, there are significant technical vulnerabilities that can challenge an AI driven security operations center (SOC). The CCCS threat report mentions that attackers are using AI to bypass defenses, and this raises serious concerns about the resilience of our own AI models.

1. Technical Vulnerabilities and Manifestations

The most prominent vulnerabilities include adversarial input, model evasion, and model poisoning. These are not theoretical risks; they have tangible manifestations in real operational settings.

a. Adversarial Input

I reckon this is the most immediate risk. An attacker can make subtle changes to a data point to cause a model to misclassify it. This would manifest in an EDR (Endpoint Detection and Response) tool. An attacker could slightly alter the code of a known malware variant, just enough to make its features look "normal" to a trained

EDR model. The model, thinking it's a benign file, would not generate an alert, allowing the malware to execute.

b. Model Evasion

This is a broader, more sophisticated version of adversarial input. Attackers can conduct reconnaissance on an organization's security controls to build their own models that help them create payloads designed to bypass the target's AI detection. The result is a cyberattack that flies "under the radar" of both an EDR and a SIEM (Security Information and Event Management) tool.

c. Model Poisoning

This is a long term, devastating attack. An attacker could subtly corrupt the training data that an organization uses to build its AI models. Over time, the model would be poisoned, learning to ignore certain types of malicious behavior. This would manifest in a SIEM tool as a steady decline in alerts for a specific type of attack, lulling security teams into a false sense of security. My TD experience made me realize the inherent risk of a single point of failure with AI. If an attacker gains access to the AI model itself, they could essentially compromise all the sensitive data it was trained on and a wide range of sensitive data. This would be catastrophic. This is similar to my experience as a junior C# .NET developer, where we were responsible for a progressive disease prediction model for cystic fibrosis. If that model had been poisoned or the security of the model itself was breached, the impact would have been a catastrophic privacy violation, not just a security breach. It's not just about defending against outside attackers, but also securing the AI itself as a point of critical failure.

2. Recommendations and Improvements

I think the CCCS recommendations are a great starting point, but they are not enough on their own. Policy is only as good as its implementation. Given the rising sophistication of attackers, I believe a crucial improvement is to integrate **explainability** into every AI model deployed in a SOC.

Explainable AI (XAI) tools like SHAP and LIME, which we discussed in the Feature Selection & Data Mining lecture (Module 7), are essential because they help analysts understand *why* an AI made a certain decision. This is critical for building trust in the technology and reducing analyst fatigue. For instance, if a network anomaly detector, perhaps an **Isolation Forest** (Anomaly Detection - Outliers, 2025), flags an event, the output shouldn't just be "suspicious." It should also provide a reason, such as, "Suspicious due to a file transfer from an unknown IP address to an unusual country, bypassing normal firewall rules." This empowers the analyst to make an informed decision and to know exactly what they are looking for in their investigation. My experience assisting with audit evidence at TD Bank (Basel Data Governance) taught me the importance of being able to prove our security controls were effective. You can't prove an AI model is working if you can't explain its decisions to an auditor or a regulator. Explainable AI makes the model's logic auditable and defensible. I reckon we should also look at incorporating AI into access control approvals. At TD, an intern handled the processing of access requests, and non technical employees often didn't know the exact data schema they were requesting access to, only the "data owner's" label for it. AI could be used to recognize messages asking for approval to a database and help with the signing on/off process. Right now they have to fill in a request access form but many of them don't know exactly what they are getting access to, i.e., the schema, but they know the label of the data. AI could streamline this process that is currently rigid and prone to human error, which is a major security risk.

Furthermore, my co op experience as a junior C.NET developer on a predictive model for cystic fibrosis taught me the critical importance of data privacy and security. We had to make sure the training and test data was completely sanitized, masked, and randomized to meet strict regulations like HIPAA. This project highlights a key vulnerability - if the AI model itself is a single point of failure and is compromised, it could expose not just financial data but incredibly sensitive health information. As mentioned in our lectures, this is a major concern with AI systems,

as they must be "treated as untrusted actors" with proper confinement controls to prevent them from leaking the very sensitive data they were trained on.

IV. Adversaries and AI Evasion in the Canadian Threat Landscape

From recent threats to Canadian organizations, I believe an attacker would bypass AI powered detection by using a combination of zero day exploits and compromised, legitimate credentials. And the **MOVEit breach** is a perfect example of this (Anomaly Detection - Outliers, 2025). The vulnerability was unknown to security vendors and their AI models, allowing attackers to exploit it for an extended period before a patch was released and detection signatures were created. This is a classic "unknown unknown" scenario that a standard supervised learning model would have missed because it had never seen the attack before. Similarly, in the **SickKids hospital ransomware incident** (CCCS, 2025), an attacker (LockBit) might gain access through a valid but compromised affiliated account. An AI model trained to detect brute force attacks would likely miss this login because it appears to be a legitimate user logging in from a known location. To counter this type of evasion, I believe a Canadian specific defensive strategy must combine both technical and procedural elements.

1. A Zero Trust Framework with Behavioral Analytics (Technical)

The most effective technical strategy is to adopt a **Zero Trust framework** across the organization. This framework operates on the principle of "never trust, always verify." Every user, device, and network connection must be verified continuously, regardless of where it is. This is especially important for organizations with a complex infrastructure, such as a large Canadian bank or a government

agency like the Region of Peel. The rigid access controls at TD were a good start, but a full Zero Trust model would go beyond that.

Within this framework, AI can be used to monitor user behavior for subtle anomalies. The goal is to move beyond simple event based detection to a deeper understanding of behavior. My experience working as a cashier for Walmart taught me this principle. We were trained to spot shoplifters not by looking in their bags to pick out easily stolen items (like signature-based detection), but by observing their behaviors like how they acted, where they looked, and their patterns of movement - in conjunction with the identified most “wanted” item by thieves. We learned to clock them without tipping them off. This same concept applies in cybersecurity. Instead of just flagging a login from a new location, an AI model would look for a series of behaviors that, together, are highly suspicious. For example, a trusted employee's account logging in from a known location, but then trying to access a highly sensitive database they've never accessed before. This "collective anomaly" is a powerful indicator of a compromised credential and can be detected by an AI model even if each individual event looks normal (Anomaly Detection - Outliers, 2025). This approach counters the attacker's evasion strategy by focusing on the "how" of the attack, not just the "what."

2. Continuous Incident Response Drills (Procedural)

Technically, a Zero Trust model is a powerful defense, but it’s only as good as the team behind it. A critical procedural strategy is to conduct regular, realistic tabletop exercises, just like the one we've been practicing. These drills, which should be tailored to specific threats like a healthcare ransomware scenario, help security teams and leadership understand and respond to the human and operational aspects of a breach. A Canadian specific defense would also involve clearly defined communication plans with regulators and government agencies like the CCCS. A coordinated, verifiable response is essential. By ensuring that roles, responsibilities, and escalation paths are clearly defined and tested, a Canadian organization can reduce the impact of an attack, even when an AI has been bypassed.

References

Canadian Centre for Cyber Security. (2025). *National Cyber Threat Assessment 2025-2026*. Retrieved from

https://www.cyber.gc.ca/en/guidance/national-cyber-threat-assessment-2025-2026

Hirji A. (2025). *Anomaly Detection - Outliers (Week 11).* Sheridan College. Retrieved from

https://slate.sheridancollege.ca/d2l/le/content/1387461/viewContent/17997358/View

Hirji A. (2025). *Feature Selection & Data Mining (Module 7, Week 9 & 10).* Sheridan College. Retrieved from

https://slate.sheridancollege.ca/d2l/le/content/1387461/viewContent/17992550/View

Hirji A. (2025). *Practice SPAM Detection (Week 13).* Sheridan College. Retrieved from

https://slate.sheridancollege.ca/d2l/le/content/1387461/viewContent/18011991/View