Cybersecurity Management of AI/ML Applications

September 20, 2020

Sanjana Shukla
George Wrenn
Dr. Keri Pearlson

Abstract

Cybersecurity is a complex management issue under the best of circumstances. Cyber leaders must make sure that systems are secure, vulnerabilities are identified and removed, valuable data and systems are protected, and plans are in place to respond and recover in the event of a breach. While the same security concerns apply to artificial intelligence (AI) and machine learning (ML) systems as traditional IT systems, the unique aspects of learning and autonomous inference engines create unique cybersecurity management requirements. This research seeks to understand the unique cybersecurity risks of AI/ML applications due to their use of data, learning algorithms, inference algorithms and feedback loops. This paper explores previous research and publicly available information about the architecture and uses of AI/ML applications, and suggests hypotheses and managerial action to manage the cybersecurity of an AI/ML system.

Background

This research focuses on investigating unique cybersecurity management issues that arise in AI and ML systems and applications. While many AI/ML applications are themselves focused on improving cybersecurity, this work does not focus on that specific use case of AI/ML. Instead, this work identifies cybersecurity threats to applications of AI/ML technologies such as those that produce recommendations and those that autonomously carry out actions resulting from recommendations. This work asks: what are the unique cybersecurity risks and attack vectors used to potentially harm applications that use AI/ML and how should managers respond?

Problem Statement

AI/ML systems are designed to find anomalies and unique patterns using self-learning engines and training/test data. Systems are trained with clean, specific data sets and outcomes are evaluated to insure the AI/ML system operates as expected. But detecting anomolies in an AI/ML system can be difficult. The conventional way of detecting anomalies in non-AI/ML systems is to use test data to create output and then to ensure the output is predictable, expected, and explanable. However, these approaches fail for AI/ML systems because of their 'black-box' nature: they are self-learning and are often designed to find unique and unexpected patterns.

The difficult problem for managing the cybersecurity of these systems occurs when the output is unexpected. Managers need to know if the output is truly unique or the result of tampering. For example, when symptoms experienced by a sick patient are entered into an AI/ML-based diagnosis system, is the unexpected result due to finding a 'needle in the haystack' or due to malicious modification or manipulation of the system? This research seeks to understand the unique cyber management considerations for AI/ML systems to ensure that the outcomes can be trusted.

Use-Cases

To set the stage for the framework and hypotheses, we begin with two AI/ML system use-cases. These hypothetical situations are compilations of real-life AI/ML applications, but are offered here to help clarify

the cyber concerns outlined later in this paper. The first use-case is a recommendation system, a medical diagnostics system, which takes in raw data in the form of various patient symptoms and conditions (e.g. body temperature, presence of cough, age, gender, medical history, etc.) and predicts the most likely diagnosis. The system's recommendation includes both the medical diagnosis and recommended next steps (e.g. that the patient should take a certain medication). In rendering a diagnosis, the system might produce a list of probable diagnoses with the likelihood and suggested treatment for each. The system then learns from input which condition was accepted by the medical professional. Note that in this kind of a system (a recommendation system), the system is only recommending; it is not carrying out the action (e.g. dispensing a medication). The decision of whether to act on the system's recommendation lies in the hands of the patient and the medical professional.

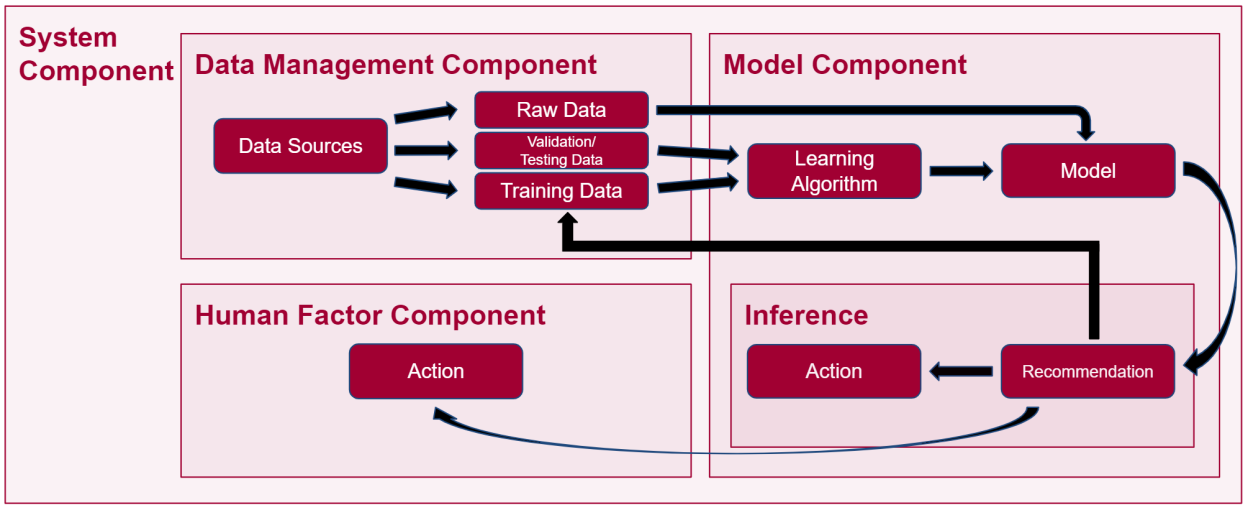
The second use-case is a autonomous vehicle, a cyber-physical system that takes in raw data in the form of images such as pedestrians, railroad crossings, traffic lights, etc. and then recommends and takes action based on the recommendation. There is likely a human override for the actions to be taken, but normal operation of this kind of system is an autonomous action taken without human interaction. An autonomous vehicle senses objects while in motion, interprets and classifies them (e.g. as another vehicle, road markings, pedestrian, etc.), and decides what action to take. For example, the system perceives a stop sign, identifies it based on previous similar examples from the learning algorithm, decides that a stop sign means stop (the recommendation), and takes the appropriate action to stop the vehicle by engaging its brakes.

While both of these use-cases take in raw data and come up with recommendations, the primary difference in this second use-case is the inference algorithm. In the second use-case, the system not only recommends an action but then engages the system in completing the action. There is an automatic response independent of human interaction.

Unique Aspects of AI/ML Systems

The model in Figure 1 is a general AI/ML system model highlighting the key components of the system.

Figure 1: General AI/ML System Model



Source: Sanjana Shukla, George Wrenn, Dr. Keri Pearlson, Cybersecurity at MIT Sloan (CAMS)
Adapted from: Gary McGraw, Ph.D., Harold Figueroa, Ph.D., Victor Shepardson, Richie Bonett
Berryville Institute of Machine Learning (BIML) [1]

Training and Testing/Validation Data

Three unique aspects of an AI/ML system create conditions for unique cyberscurity management concerns

(Figure 1). First, AI/ML systems have training and test/validation data that are critical inputs for the system's training process (described below). Training data trains the AI/ML system and fine-tunes the model parameters. Validation/test data is the data used to validate that the system produces acceptable outputs. This later data is often a sample of training data that is held back from training the model because evaluation of a model's skill would be biased if the same data was used to both train and validate the model.

Training and Inference Processes

The second unique aspect of an AI/ML system is the training and inference processes, since AI/ML systems are designed to be trained and then to make recommendations and possibly take action. The learning algorithm uses training and validation datasets as inputs and trains the model parameters. While the model evaluates data, the inference process take the output of the model and makes recommendations (and in some cases, takes action).

Building on the use-cases described above highlights some of the key features of an AI/ML system. In use-case 1, the medical diagnostic system, the training process would take symptom data where the diagnosis is known, and use it to train the model. The inference process would produce the recommendations and the likelyhood of each diagnosis. The training process would then use a second sample of data, the validation/test data to see if the resulting diagnosis are within acceptable parameters.

In the autonomous vehicle use-case, the training process would take in visual data such as traffic signs, car and other vehicle images, and other images that a vehicle might encounter during a trip (such as pedestrians). The model would be trained to recognize these inputs and process alternative recommendations. The inference process would then identify the appropriate actions for the likely output and take action such as speeding up or slowing down the car, or stopping for a red light.

Note that the inference process in a medical diagnostics system does not take any action. It only provides a recommended course of action which the human stakeholder can take (refer to the Action step in the Human Factor Component in Figure 1), However, in the second use-case, the autonomous vehicle, the learning algorithm trains the system to determine, for example, what a stop sign looks like. The model takes in visual components and determines whether or not any of these visual components are a stop sign. If the model determines a component to be a stop sign, the inference algorithm then takes this model output and engages the car's breaks to come to a halt. In a cyber-physical system, therefore, the inference process not only recommends the action but also follows through with the recommendation and performs the action (refer to the Action step in the Inference Component in Figure 1).

Feedback Loops

The third unique aspect of an AI/ML system is its feedback loop which facilitates automatic learning and reinforcement of the outputs of the recommendation and action steps. The feedback loop can be represented as:

Raw data inputs → Model → Inference algorithm → Training data.

This feedback loop conveys that the raw data entered into the system when the system is in use is then acted on by the model, creating recommendations which then become new training data. This is the automatic learning feedback loop as new data is processed and the recommendation becomes additional input so the system can learn what recommendations are more appropriate for the next set of new data.

The second feedback loop is represented as:

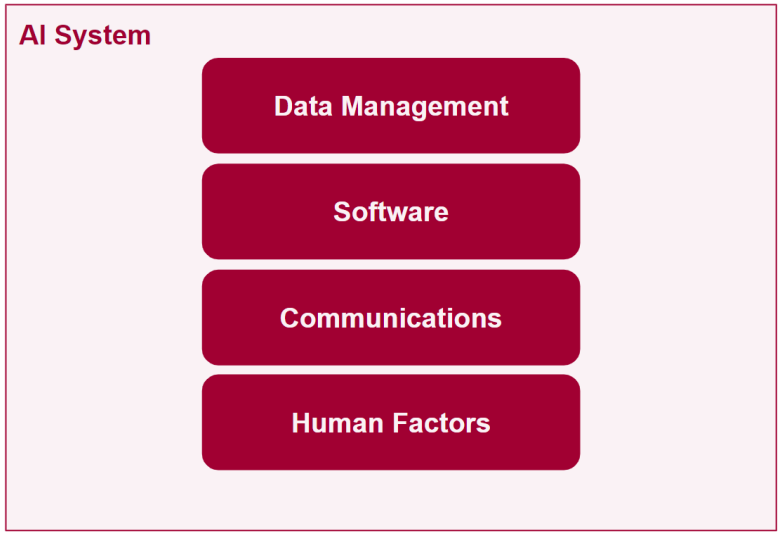
Learning algorithm → Model → Inference → Training data

This feedback loop suggests that the learning algorthim trains the model, and the resulting inference becomes new input data for the model. This helps fine-tune the model's performance during the design and evaluation of the system.

Cybersecurity Vulnerabilities in Applications of AI

In an AI/ML system, there are five major components that could serve as attack surfaces for a cyber-attack: the data management component, the software component, the communications component, the human factor component, and the overall AI/ML system or context in which the system is used (See Figure 2).

Figure 2: AI/ML System Meta View



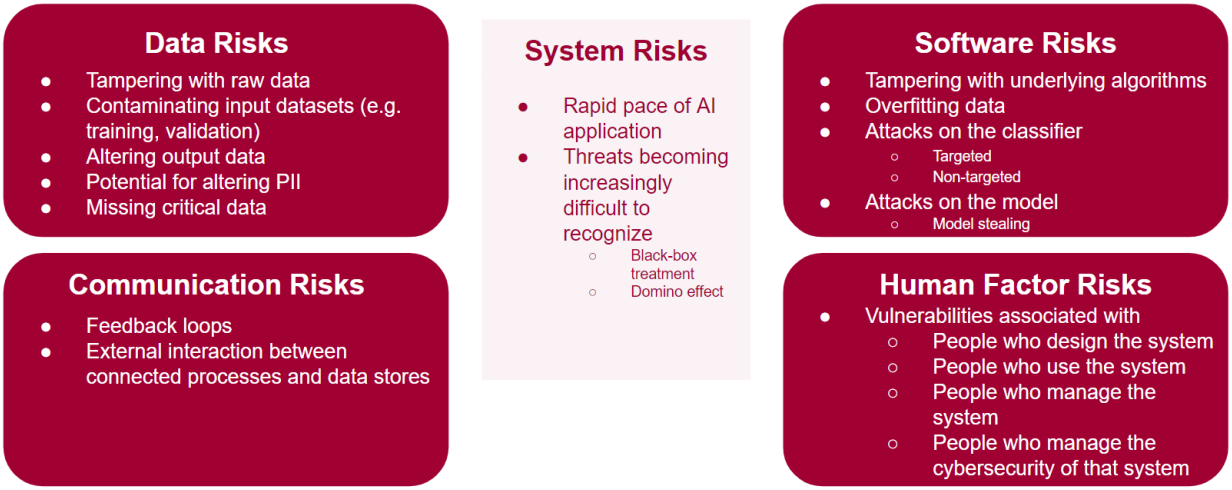
Source: Sanjana Shukla, George Wrenn, Dr. Keri Pearlson, Cybersecurity at MIT Sloan (CAMS)

A cyber-attack targeting the data component would tamper with data, altering the training, validation/test or raw data. An attack on the software component would affect the learning algorithm, the model, or the inference component (the recommendation or the autonomous action systems). The communication component opens up vulnerabilities for an AI/ML system through the feedback loops. The fourth component, the human factor component, becomes an attack vulnerability if the actions taken are altered or inappropriately modified by the people who follow up on the recommendations. Finally, the fifth component, the system's environment and context, opens up vulnerabilities such as an ability to tamper with the cloud or the on premise location of the system, the storage of the data, the networks that bring data to the system, and the data purchased from outside sources.

Figure 3 summarizes cyber vulnerabilities and attack vectors that could impact each of the components described in the meta view model (Figure 2) of an AI/ML system. Note that in this discussion, there is a difference between a *threat* and a *vulnerability*. A *threat* refers to potential actions, or potential attacks, that a malicious hacker might commit such as those described, whereas a *vulnerabiltiy* refer to attack surfaces that can be impacted in the AI/ML application.

As AI/ML systems evolve and continue to self-learn, the most concerning risk is that the system will give skewed output, inappropriate recommendations, or wrong advice, resulting in a poor or wrong decision or action such as running a red light in an autonomous vehicle or acting on a bad diagnosis in a medical recommendation system. To understand how this might happen, this section explores the risks for each of the components in the meta model shown in Figure 2.

Figure 3: Cybersecurity Risks in AI



Source: Sanjana Shukla, George Wrenn, Dr. Keri Pearlson, Cybersecurity at MIT Sloan (CAMS)

Data Management Risks

Data risks arise from tampering with the training, validation/test, or raw data. For example, if training data is altered, then an attacker will have changed the way in which the learning algorithm will learn. If the validation data is hacked, then an attacker will have changed the way in which the system validates its self-learning. If raw data is hacked or a process by which raw data is collected is hacked, then the system will have poisoned data that will produce bad recommendations. Additionally, if output data is altered, then an attacker will have changed the way in which the output data is used as a system input and the system's feedback loop will be disrupted, again impacting data used for self-learning. Impacting the output data will also impact the activity that the system is designed to do - the decision it is designed to make. This, in turn, could also impact the decision that the human ends up making, affecting the human factor component.

One type of attack vector aimed at the data management component is a *data poisoning attack*. In a poisoning attack, the attacker injects carefully crafted data samples to contaminate the AI/ML system's training data in a way that eventually impairs the system's normal functions [2]. One documented example of a poisoning attack against an autonomous vehicle is an attack against the system's model responsible for rain removal. This attack stealthily changes the way the system recognizes a speed limit sign in an image if certain conditions are met when removing raindrops in that image [3]. Consequences of mis-recognizing a speed limit sign can cause the vehicle to drive in unexpected ways.

In another example, a breach of personally identifiable information (PII) stored in a healthcare diagnostic application. PII is specifically a vulnerability because attackers can change this data or possibly steal it. The consequences of modifying PII could result in a misdiagnosis of a medical condition, and the theft of that data could have far reaching consequences for other digital assets and systems used by the patient.

Software Risks

The software risks come from attackers tampering with the underlying algorithms and model. Examples of this type of risk are inserting a malicious algorithm to alter the learning process and possibly misusing the training or valudation/test data. Another example might be an algorithm that breaks the data cleaning processes that would impact the interpretation of the data by the learning or the model itself. A third example is a hack on the inference components that change how the recommendations are made or the action are taken.

Targeted and *non-targeted misclassification attacks* are two potential attack vectors impacting the software

components. In the autonomous vehicles example, a targeted misclassification attack would misclassify a specific road sign while the rest of the road sign classification system continues to function properly. Similarly, an example of a non-targeted attack would be when a traffic sign is detected with less accuracy than previous detections [4]. In both cases, the result would be an inappropriate action taken by the autonomous vehicle.

Another example of a software vulnerability is a *model stealing attack*, where an attacker creates a similar model that functions differently from the actual model, which has results that work against the original system. Stolen AI/ML healthcare models, for example, might interpret symptoms differently than the actual model and suggest wrong or destructive recommendations.

Similarly, *evasion attacks* tamper with a system by adding an invisible (to humans) layer of data noise onto an image, leading the model to report back with high confidence that the image is something other than what it originally was. For example, a real-world simulated evasion attack took place when Google and NYU researchers tricked a system to report with almost 100% certainty that an image of a school bus was an ostrich. The goal of an evasion attack, therefore, is to cause model misclassification [5].

We also consider another example of an evasion attack. One way in which machine learning models are used by high frequency trading (HFT) systems in the financial services industry includes classifying an equity as likely to increase in price above a threshold, decrease below a threshold, or remain between the thresholds. In an evasion attack on this kind of model, the attacker seeks to degrade model performance by causing misclassification on any label, or final output, from the model. Another evasion attack may anticipate the behavior of the victim and seek to misclassify a specific label to be used by that victim. In HFT, this type of attack modifies the equity classifier until either the model no longer predicts the correct label or violates the capital constraint (e.g. violates a constraint mandating that the system not execute a trade above or below a certain value) [6].

Many other types of attacks on the AI/ML software components have been recorded by researchers. One researcher [7] suggests that potential attacks can be categorized based on the motive of the attack. *Confidentiality* attacks, *integrity* attacks, and *availability* attacks impact the basic confidentiality, integrity, and availability of the software to do what it is designed to do. For example, an availability attack might target a model designed around sentiment analysis. Over an extended period of time, an attacker publishes and promotes a series of adversarial social media messages designed to trick sentiment analysis classifiers used by system's learning algorithms. When the AI/ML trading algorithms trade incorrectly over the course of the attack, losses accrue for the parties involved, and if extensive enough, result in a possible downturn in the market [7] [8]. Likewise a *reprogramming attack* makes the model "see" things that are not actually present, leading to errors such as erroneous image analysis [9].

Communication Risks

Communication risks are best represented by the arrows in the general AI/ML system model (Figure 1). These arrows represent communication into, out of, and in between the system's data stores and processes. All of these are communication links vulnerable to interception by malicious actors resulting in the manipulation or siphoning off of data or information to be used by other parts of the AI/ML system.

But unique to AI/ML systems are the feedback loops, where information is fed back into the system for training or testing. When the feedback loop is hacked or manipulated, the learning algorithm, model, or inference engine can be impacted resulting in inappropriate recommendations and actions.

Voice recognition software in the financial services industry's trading systems presents a real-world example of communication risk. Voice recognition is the component of a voice-trading software system that recognizes an individual's voice and uses it to authenticate a user to ensure that only a trader, broker,

or other authorized party executes a trade. When voice is used as the medium for inputting information such as a trade, an attack on the speech recognition system is an attack on the communications component, resulting in compromised authorization processes. Furthermore, when the attack is done by tucking garbled voice commands into broadcasts, the communications layer of the system is compromised without the users even noticing. These voice commands that are difficult for humans to hear, interpret or understand can still trigger a device's voice control functionality [10] [11]. In trading, such an attack can affect the trading system by executing unauthorized trades, compromising the privacy of sensitive information, or setting off a series of automated actions that disrupt the markets entirely.

Human Factor Risks

The largest human factor risks occur when individuals are not trained appropriately; they subsequently pose a cybersecurity risk to a system since malicious actors leverage this ignorance to achieve a successful hack a system. For an AI/ML system, human factor risks fall into four key vulnerability types: vulnerabilities associated with people who design the system, vulnerabilities associated with people who use the system, vulnerabilities associated with people who manage the system, and vulnerabilities associated with people who manage the cybersecurity of the system. For example, if the recommendation output of the healthcare system described in the first use case is implemented incorrectly because the person receiving it has malicious intent, the overall AI/ML system will be compromised.

Another example of a human factor vulnerability arises from the black-box nature of the AI/ML system. When the people working with the system view the system as a black-box that takes raw data and produces recommendations, but the people have no idea whether the recommendations are reasonable and blindly implemet them, there is a risk. Often, an AI/ML system that makes recommendations for people to implement has an additional layer of security to insure that bad recommendations are caught before damaging action is taken. If the recommendation is unique and potentially unexpected, the person receiving the recommendation must be able to determine if this is the 'needle in the haystack' that the system uncovered or if this is the result of a manipulated or broken AI/ML system. The chance that a bad or destructive action is taken increases if the people have no knowledge on how to evaluate the appropriateness of the recommendation.

Social engineering, a type of manipulation that coaxes someone into giving up confidential information inappropriately [12], can introduce additional vulnerabilities. Although not unique to AI/ML applications, successful social engineering can cause a person to enter erronous data into a system, or use recommendations inappropriately.

System Risks

AI/ML systems face three key *system risks*. The first risk is that threats are evolving much faster, the second risk is that threats are becoming increasingly difficult to recognize, and the third risk results from the environment in which the system lives (e.g. if it lives on the cloud versus on locally hosted systems).

Technological advances in AI/ML are accompanied by threats that seem to arise at a faster pace. Malicious actors are clever and innovative, finding vulnerabilites faster than the system designers themselves. Further, specific vulnerabilities to recent advances in AI/ML applications may not yet be well-known. The pace of technology advances and associated threats makes it virtually impossible for a general manager to be informed and therefore put protections in place. To make matters more risky, new cyber vulnerabilities are often brought to light after they are found by attackers. At that point, designers plug holes and block the vulnerabilitiles hoping that no others exist. While this is not unique to an AI/ML system, it does manifest itself differently in these types of systems as this technology is increasingly complex in its design, making it difficult to stay current on new vulnerabilitiles that might be introduced.

Likewise, threats are becoming increasingly difficult to recognize. Consider this example: a doctor walks up to a healthcare diagnostics application and then inputs all relevant patient data into the system. This patient data serves as raw data to the AI/ML application. The system processes this data, and out comes an unexpected patient diagnosis. The question the doctor now has to ask is whether the output was unique but valid, or whether the system has been hacked, creating an invalid diagnosis. When the system is treated like a black-box, the system's users do not know if the system's results are valid or have been compromised. The more black-box-like, the harder it becomes to detect a breach or a threat quickly or efficiently. Threats are increasingly more difficult to recognize as the systems become more black-box-like.

The third system risk results from the larger environment in which the system lives. For example, is the system residing in the cloud or within the company's internal server? A system that lives within a company's internal firewall has different security processes than one on an external cloud, and managing cloud security might look very different than managing on-premise systems. The environment risk also extends to the policies, norms, and rules in the organization. For example, a system that operates in an untrusted environment might have policies in place to mitigate those risks, and the potential exists for these policies themselves be tampered with or altered to compromise their effectiveness.

Themes and Hypotheses for Further Study

To summarize the key findings from this work, three themes emerge, each producing hypotheses for future study.

Theme 1: There are unique cybersecurity risks to AI/ML systems.

AI/ML systems are subject to many of the same vulnerabilitiles and risks as a traditional IT system. However, given the design, architecture, and processes of AI/ML systems, they have additional cybersecurity risks that are not found in traditional IT systems.

Hypothesis 1a: There are unique cybersecurity risks to AI/ML systems pertaining to the system's data, models, communication, and human factors.
Hypothesis 1b: The primary data risks of an AI/ML system are tampering with raw, training, and validation/test data, which leads to incorrect learning and in turn an erroneous system output.
Hypothesis 1c: The primary software risks of an AI/ML system are tampered learning algorithms, model manipulations, and hacked inference algorithms, which produce erroneous output and compromise actions to be taken.
Hypothesis 1d: Feedback loops are a communications risk of an AI/ML system which compromise the system's self-learning processes.

Theme 2: The larger system in which the AI/ML system resides poses risks that must be managed.

The second theme is that the AI/ML system operates in the context of a larger system, and there are aspects of this larger system that pose risks to be managed.

Hypotheses 2a: The pace of advancements in AI/ML poses increasing number of cybersecurity risks that must be managed.
Hypotheses 2b: The increasing complexity of AI/ML applications makes it more difficult to identify compromised systems.
Hypotheses 2c: Environmental and organizational factors such as policies, norms, and rules that guide and support AI/ML applications can be compromised in a way that increases vulnerabilities and risk.

Theme 3: The way people interact with AI/ML applications can impact cybersecurity risk.

The third theme is that for humans to be able to trust the AI/ML system's output, the system must operate in a trusted environment and the people using the AI/ML system must understand enough about the way the system operates to be able to validate the recommendations and take appropriate actions.

Hypothesis 3a: A human factor risk of an AI/ML application is that it is difficult for a human stakeholder to identify a valid output from a hacked one.
Hypothesis 3b: The more opaque (black-box-like) the AI/ML application is to human stakeholders, the more likely it is that an erroneous recommendation will go undetected.
Hypothesis 3c: Human stakeholders will trust the output of an AI/ML system if it resides in a trusted environment.

Conclusion

There are unique cybersecurity management considerations for AI/ML systems. This paper presented two AI/ML use cases as context for exploring recommendation systems and autonomous cyber-physical systems. While these use-cases differ in how the action is executed, they clearly indicate similarities in the use of data, learning algorithms, models, and recommendation algorithms. Each of these components is accompanied by unique cybersecurity concerns that managers must anticipate.

While conventional cybersecurity management approaches focus on the confidentiality, integrity, and availability of systems as a fundamental framework for system security, AI/ML systems have additional cyber risks that can be classified as data, software, communications, human factors, and overall system risks. Managers using AI/ML applications must have an accompanying plan for insuring that these applications are understood to the level that recommendations can be trusted. After all, the promise of many AI/ML systems is to continuously learn so that they adapt to new and uncharted applications at the same time as they find unique and unanticipated recommendations. But to trust the output of an AI/ML system, managers must ensure that the cyber vulnerabilitiles are appropriately managed and that the AI/ML systems recommendations are reasonable. The next phase of this research will focus on ways that managers can build appropriate cybersecurity plans for their AI/ML applications.

References

[1] McGraw, G., Figueroa, H., Shepardson, V., Bonett, R. *An Architectural Risk Analysis of Machine Learning Systems: Toward More Secure Machine Learning.* Berryville Institute of Machine Learning.

[2] Lopez, M. (2019, October 3). *Women in AI: IBM's Lisa Amini Takes On AI/ML Security And Reasoning. Forbes.* Retrieved from https://www.forbes.com/sites/maribellopez/2019/10/03/women-in-ai-ibms-lisa-amini-takes-on-ai-security-and-reasoning/#d97228921b24.

[3] Ding, S., Tian, Y., Xu, F., Li, Q., Zhong, S. *Poisoning Attack on Deep Generative Models in Autonomous Driving*. Retrieved from http://www.cs.wm.edu/~liqun/paper/securecomm19.pdf.

[4] Qayyum, A., Usama, M., Qadir, J., Al-Fuqaha, A. (2019, May 29) . *Securing Connected & Autonomous Vehicles: Challenges Posed by Adversarial Machine Learning and The Way Forward*. Retrieved from https://arxiv.org/pdf/1905.12762.pdf.

[5] Kobie, N. (2018, September 11) . *To cripple AI, hackers are turning data against itself.* Wired. Retrieved from https://www.wired.co.uk/article/artificial-intelligence-hacking-machine-learning-adversarial.

[6] Goldblum, M., Schwarzchild, A., Patel, A. B., Goldstein, T. (2020, March 4). *Adversarial Attacks on Machine Learning for High Frequency Trading*. Retrieved from https://arxiv.org/pdf/2002.09565.pdf.

[7] Patel, A. (2019, November 7). *Adversarial Attacks Against AI*. F-Secure Blog. Retrieved from https://blog.f-secure.com/adversarial-attacks-against-ai/.

[8] Clouder, A. (2019, October 11). *Applications of NLP and Voice Recognition*. Retrieved from https://www.alibabacloud.com/blog/appilications-of-nlp-and-voice-recognition_595432.

[9] Metz, C., Smith, C. S. (2019, March 21). *Warnings of a Dark Side to A.I. in Health Care*. The New York Times. Retrieved from https://www.nytimes.com/2019/03/21/science/health-medicine-artificial-intelligence.html.

[10] *Voice Activated Trading: Pros and Cons*. Etna. Retrieved from https://www.etnasoft.com/voice-activated-trading/.

[11] (2016, July 11). *'Dalek' commands can hijack smartphones*. BBC News. Retrieved from https://www.bbc.com/news/technology-36763902.

[12] Davis, J. (2019, October 18). *Hackers Targeting Healthcare with Social Engineering, Email Spoofing. Health IT Security*. Retrieved from https://healthitsecurity.com/news/hackers-targeting-healthcare-with-social-engineering-email-spoofing