Cybersecurity Management of AI Systems

May 2021

Sanjana Shukla (shuklas@mit.edu)
Ignacio Parada (iparada@mit.edu)
Dr. Keri Pearlson (kerip@mit.edu)

MIT Sloan School of Management
Cybersecurity at MIT Sloan (CAMS) Group

Abstract

We use the term "AI" to encompass both artificial intelligence (AI) and machine learning (ML) systems, and use the term "AI/ML" accordingly. Securing AI/ML systems presents unique cybersecurity management issues not present in non-AI/ML, or "traditional," systems. These unique management issues arise from the unique components of AI/ML systems that are absent from traditional systems. This work is a continuation of our research where we seek to identify the unique cybersecurity concerns that arise in the development and use of AI/ML systems as well as propose ways that managers can build appropriate cybersecurity plans for these systems.

1. Literature Review

Publicly available research on this topic is limited, however existing work echoes the findings we have arrived at independently through our work.

First, upon comparison with cybersecurity issues arising in traditional systems, "Unlike traditional cybersecurity vulnerabilities, the problems that create AI attacks cannot be 'fixed' or 'patched.' Traditional cybersecurity vulnerabilities are generally a result of programmer or user error. As a result, these errors can be identified and rectified. In contrast, the AI attack problem is more intrinsic: the algorithms themselves and their reliance on data are the problem" [1]. Also, "among the state-of-the-art methods, there is currently no concept of an 'unattackable' AI system" [1] and the differences between AI and traditional systems have "significant ramifications for policy and prevention" [1]. Furthermore, an "important lesson from traditional cybersecurity policy is the superiority of foresight and pre-deployment planning over reactionary remedies" [1]. Among the recommendations proposed in the above referenced publication as part of its AI Security Compliance measures, one key proposal includes the review and updating of "data collection and sharing practices to protect against data being weaponized against AI systems. This includes formal validation of data collection practices and restricting data sharing" [1]. Another key aspect includes "[d]etermining the ease of attacking a particular system" [1] where "[t]he degree of vulnerability can be determined by characteristics such as public availability of datasets, the ability to easily construct similar datasets, and other technical characteristics that would make an attack easier to execute. One example of an application that could be particularly vulnerable to attack is a military system that automatically classifies an adversary's aircrafts. The dataset for this task would likely consist of collected radar signatures of the adversary's aircraft. Even if the country collected the data itself, stored it perfectly and safely with encryption, and had flawless intrusion detection – all of which would guarantee that the adversary could not get this data and use it to formulate an attack – the adversary could still execute a successful attack by building a similar dataset itself from scratch, which could easily be done because the adversary clearly has access to its own aircraft" [1]. As yet another notion echoed in our own work, a third key aspect of the proposal in the aforementioned paper is the proposition that "[t]he damage that an attack can precipitate should be assessed in terms of the likelihood of an attack and the ramifications of the attack" [1].

Furthermore, the importance of data security for AI/ML is echoed in publications on the topic. "One key element for AI will be the mass stores of data which will require technical oversight and protection. The ability of AI using and creating mass data will be a security concern where cybersecurity professionals will need to ensure mass data plans are monitored and updated as necessary." [2] Another key cybersecurity risk arising from the use of AI/ML applications is the potential for the system to deviate from its original design or intention [2]. Overall, "with our lack of understanding of AI and the algorithms created, it is understandable to have a lack of cognizance of the AI system performing precisely as planned in a live environment" [2]. A selected set of previously published cybersecurity concerns arising in AI systems have been reproduced from the publication "Artificial Intelligence Cybersecurity Framework: Preparing for the Here and Now with AI" and are identified in Figure 1 below.

Type	Description of AI Cybersecurity Issue
AI Design	Integrity of algorithms and output – bias or external bias
Code	Secure code analysis of AI code/functions/AI generated code & functions
Privacy	AI data lakes – privacy issues with mass data collected
Privacy	Algorithms could result in exposure of sensitive data
AI Design	Variables added to an AI system causing undesirable outcome

Trustworth-iness	AI will need trust relationships being multidimensional

Figure 1: AI Cybersecurity Concerns

Source: Emily Darraj (Capitol Technology University), Char Sample (ICF Inc., SABSA Institute), and Connie Justice (Purdue School of Engineering and Technology) [2]

It is also important to note that the effective cybersecurity management of AI/ML systems does not forego or omit traditional cyber management practices recommended for non-AI/ML systems. For example, “the usual security countermeasures to prevent unauthorized access through user authentication, techniques to preserve data integrity through cryptography and network defense mechanisms are active area of interest in the field” [3] of security for embedded systems, defined as “any system that has a microprocessor” (including smart objects) “with the exception of PCs, laptops, and other equipment readily identified as a computer” [4]. Furthermore, “[o]ne of the most common network attacks occurs by exploiting the limitations of the commonly used network protocols Internet Protocol (IP), Transmission Control Protocol (TCP) or Domain Name System (DNS)” and in general, “malware can be inserted at any point in the system life cycle” [3]. This reinforces the fact that AI systems, also with unique components and unique cybersecurity management issues, still face many of the same threats present to non-AI systems and in many ways require the effective implementation of the same security measures. We also note the existence of literature that is further tackling the intersection of artificial intelligence and cybersecurity such as adversarial artificial intelligence for cybersecurity, researched in the context of autonomous vehicles by Erik Hemberg and Una-May O’Reilly at the Massachusetts Institute of Technology. [5]

Lastly, we reference the model shown below in Figure 2, from which we adapted our definition of a machine learning system. Figure 2 illustrates a generic ML system and consists of 9 components, each numbered and labeled for clarity.

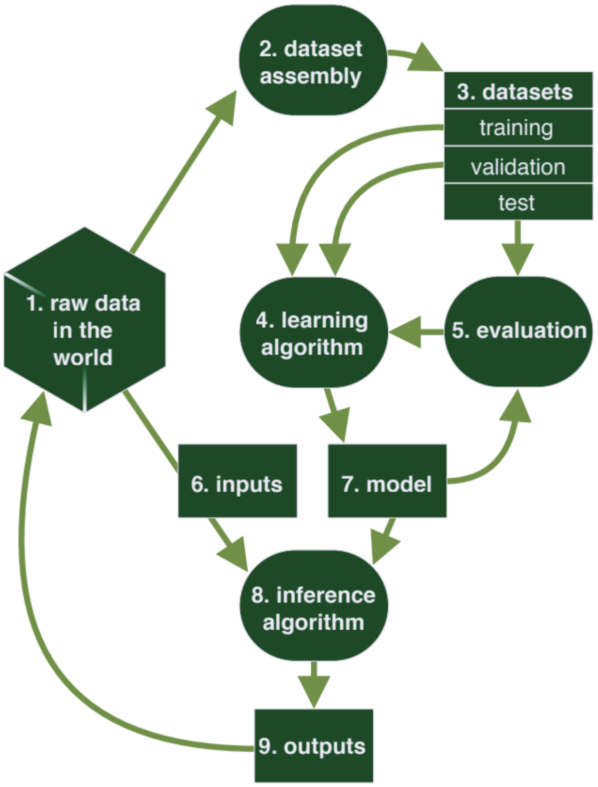


Figure 2: Generic ML System

Source: Gary McGraw, Ph.D., Harold Figueroa, Ph.D., Victor Shepardson, Richie Bonett Berryville Institute of Machine Learning (BIML) [6]

The next section discusses the background of our research, including our previous work on the topic.

2. Background

This research focuses on investigating unique cybersecurity management issues that arise in artificial intelligence (AI) and machine learning (ML) systems and applications. While many AI/ML applications are themselves focused on improving cybersecurity, this work does not focus on that specific application 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. We also note that AI is very broad, and we are focused specifically on systems that undergo self-learning. We ask the following questions to ground this work:

1. What are the unique cybersecurity risks and attack vectors used to potentially harm applications that use AI/ML?
2. How should managers assess those cybersecurity risks associated with applications of AI?

In the previous phase of this research project, we recognized that 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 ensure the AI/ML system operates as expected. However, detecting anomalies 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 outputs, and then to ensure the output is predictable, expected, and explainable. 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. Managers want to simply trust the AI/ML system's output. This leads to a difficult problem for cybersecurity due to the inability to identify whether an AI/ML system's output is truly unique or has been compromised.

We also devised a general AI/ML system model highlighting the key components of the system (Figure 1), reproduced here for reference.

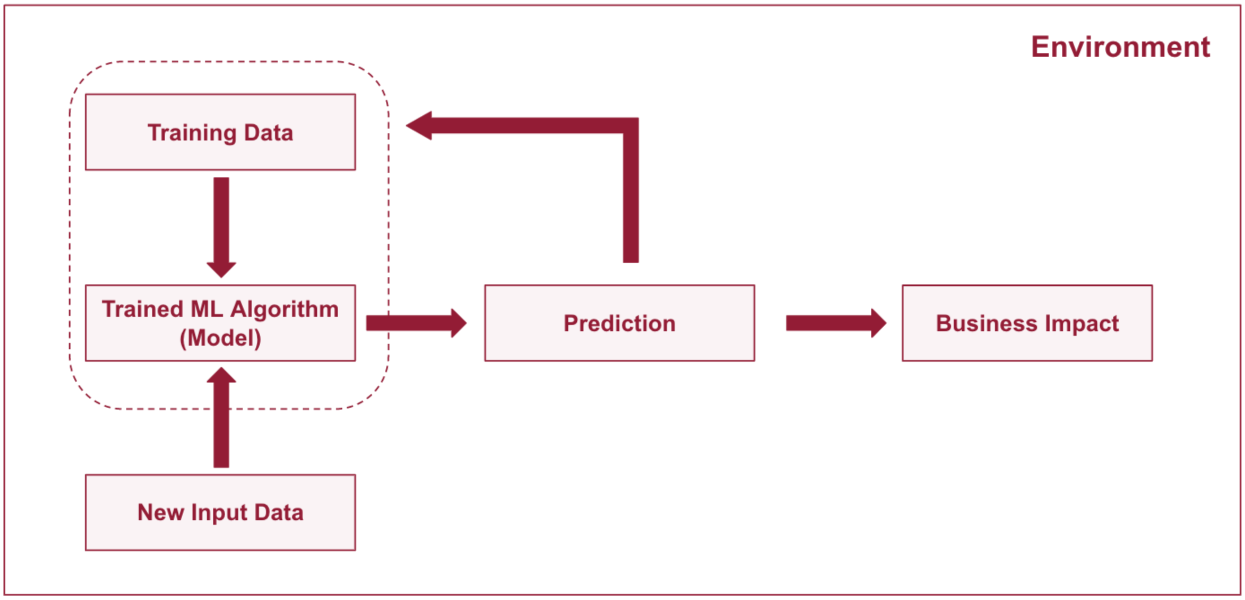


Figure 3: General ML System Model

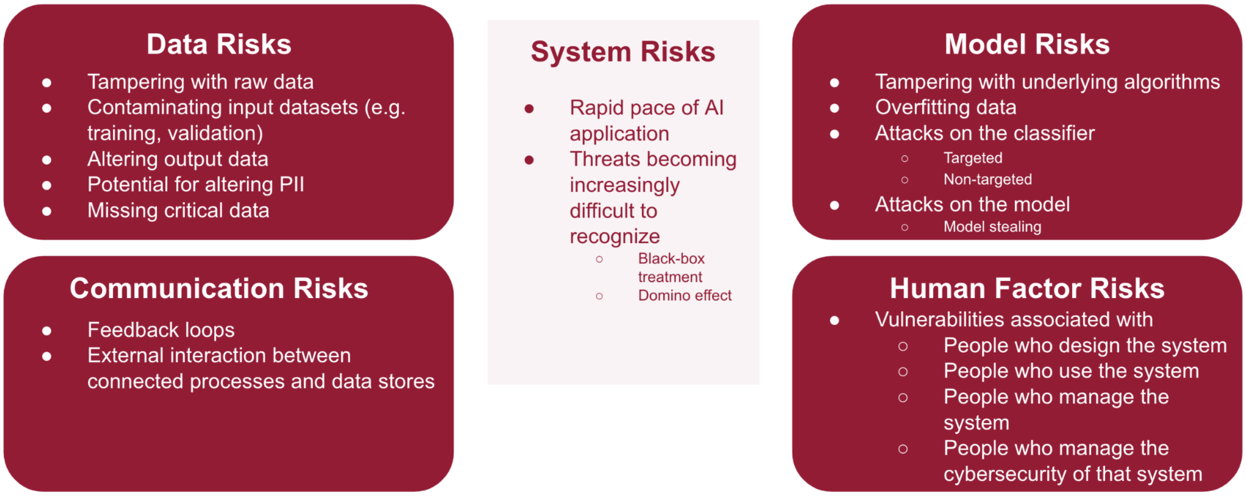
Source: Sanjana Shukla, Ignacio Parada, Dr. Keri Pearlson, Cybersecurity at MIT Sloan (CAMS)

Figure 1 illustrates an oversimplified machine learning model. Our objective in this paper is not to explain in detail how a machine learning system works or detail each AI/ML system that has been developed. We provide the above figure to clarify our baseline definition of an ML system, which we used to narrow the scope of our research. As previously mentioned, AI is a broader concept, and we focused specifically on systems where the model is trained through an internal training process that intakes training data and is fine-tuned over time. As evident, since this training is not done by a human much of the time, it is difficult to completely decipher how the model creates its prediction. Furthermore, usually these systems continue to evolve as more and more data is fed into them as inputs. This compounds the challenge of completely understanding these systems' inner workings.

In the previous phase of our work, we observed that within an AI/ML system's model are three unique aspects of these systems that create conditions for unique cybersecurity management concerns. First, AI/ML systems have training and test/validation data that are critical inputs for the system's training process. Training data trains the AI/ML system and fine-tunes the model parameters. Validation/test data is 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. Second, AI/ML systems have training and inference processes, since AI/ML systems are designed to be trained and then to make recommendations and possibly take action. This unique component encompasses the learning algorithm, which uses training and validation datasets as inputs and trains the model parameters. It also includes the model, which evaluates data, and the inference process, which takes the output of the model and makes recommendations (and in some cases, takes action). Third, AI/ML systems have feedback loops, which facilitate automatic learning and reinforcement of the outputs of the recommendation and action steps.

We also identified five major components in an AI/ML system that could serve as attack surfaces for a cyber-attack. These components are evident in Figure 1 above and include the data management component, the model component, the communication component (illustrated by the intra-system arrows), the human factor component, and the overall AI/ML system, which encompasses the context in which the system is used as well as the system's environment.

In our previous research, we also identified a number of cybersecurity risks that corresponded to each component identified. This figure has been reproduced here, with one modification, for reference.



Data Risks

- Tampering with raw data
- Contaminating input datasets (e.g. training, validation)
- Altering output data
- Potential for altering PII
- Missing critical data

System Risks

- Rapid pace of AI application
- Threats becoming increasingly difficult to recognize
  - Black-box treatment
  - Domino effect

Model Risks

- Tampering with underlying algorithms
- Overfitting data
- Attacks on the classifier
  - Targeted
  - Non-targeted
- Attacks on the model
  - Model stealing

Communication Risks

- Feedback loops
- External interaction between connected processes and data stores

Human Factor Risks

- Vulnerabilities associated with
  - People who design the system
  - People who use the system
  - People who manage the system
  - People who manage the cybersecurity of that system

Figure 4: Cybersecurity Risks in AI/ML Systems
Source: Sanjana Shukla, George Wrenn, Dr. Keri Pearlson, Cybersecurity at MIT Sloan (CAMS)

The next section discusses our research process and methodology for this work.

3. Methodology

Our research approach to inform these hypotheses entailed conducting interviews with managers and developers of AI/ML systems as well as cybersecurity experts to understand how they manage the cybersecurity of these systems. Specifically, we sought their perspectives, opinions, and guidance on how security concerns of AI/ML systems are (and could be) managed. Our research spanned a period of four months, during which time we conducted 20 interviews across 15 organizations. The conversations ranged from at least 30 minutes to over one hour. Upon the conclusion of these interviews, we had 60 pages of interview transcripts generated. Each interview was manually transcribed by at least one interviewer, and most interviews were transcribed by two interviewers (one primary interviewer and partial transcriber and one primary transcriber and partial interviewer). A select number of these conversations were transcribed by three interviewers. The interviewers were some combination of the co-authors of this white paper. This white paper does not reflect the intimate details of any organization or interviewee, and all findings reported here have been presented in aggregate with any quotes or direct references anonymized.

The interviews followed a semi-structured approach, with the questions evolving as we gained experience conducting the conversations or wanted to further explore a new or otherwise interesting concept the interviewee had begun to share once the conversation was underway. Before pursuing this research, the preliminary set of interview questions consisted of:

1. What cybersecurity vulnerabilities do you see in AI/ML systems?
2. How is managing cybersecurity concerns in an AI/ML system different from a non-AI/ML system?
3. Where do the biggest cyber vulnerabilities come from (e.g. training data, learning engine, model, inference engine, output data, other components not identified here)?
4. How do you measure the cybersecurity risk of your AI/ML systems?
5. What kind of investments does your team make to resolve/minimize risks from AI/ML systems?

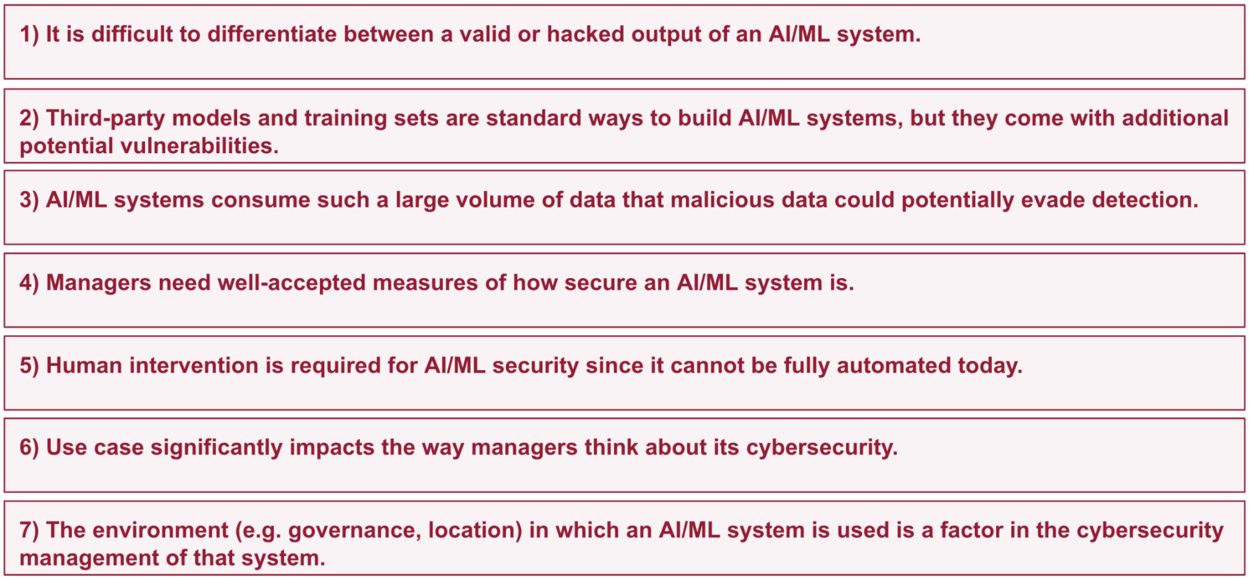
In terms of interviewee and organizational demographics, our interviewees consisted of: 4 developers, 9 managers, 4 cybersecurity experts or consultants, 1 regulator, and 2 academics. The organizations represented included: 2 financial services firms, 2 consulting firms, 4 IT/technology services organizations, 2 retail organizations, 2 academic affiliations, 1 healthcare startup, 1 conglomerate, and 1 foreign regulatory agency. The organizations that we selected our interviewees from were chosen because AI/ML applications played an important role as either the organization’s product offering (e.g. a health diagnostics product which relied on an AI/ML system), in the organization’s importance in its regulatory role, or in its internal processes.

Due to the qualitative nature of this research, we used a grounded theory approach. To reference the Oxford Research Encyclopedias for a definition of grounded theory, “grounded theory methodology is one of the most widely used approaches to collect and analyze data within qualitative research. It can be characterized as a framework for study design, data collection, and analysis, which aims at the development of middle-range theories. The final result of such a study is called a ‘grounded theory,’ and it consists of categories that are related to each other.” [7] Findings and analysis of our initial interviews informed our data collection process in subsequent interviews by iterating and reframing the kinds of questions we were asking interviewees. As previously unidentified themes emerged during our conversations, so did the discussion questions. In this way, “during the research process data collection and analysis alternate[d] and interact[ed].” [7] To further analyze our interviews and identify shared remarks between interviewees and emerging themes, we undertook an interview coding-based analytical process which allowed for a step-by-step development of categories that are grounded in data. By identifying which categories emerged from the coding, we derived a new set of themes conveying the key takeaways from our work. This was in line with the grounded theory approach, in which “category development entails comparisons at all stages, for example, of different cases during sampling, of different data pieces, and of different codes and categories during analysis.” [7]

In summary, Grounded Theory Methodology is a qualitative research approach, using which we conducted interviews, coded the interview transcripts, and analyzed the codes to construct our findings, which we name themes, described in the next section. These codes allowed us to break down the interviews into bite sized pieces of data, or mini-takeaways throughout the conversation. As mentioned above, we then grouped together the codes to identify any themes that emerged.

4. Findings

We found seven themes, synonymous to key takeaways, that arose from our analysis of the interviews. This section provides an explanation for each of these themes:



1) It is difficult to differentiate between a valid or hacked output of an AI/ML system.

2) Third-party models and training sets are standard ways to build AI/ML systems, but they come with additional potential vulnerabilities.

3) AI/ML systems consume such a large volume of data that malicious data could potentially evade detection.

4) Managers need well-accepted measures of how secure an AI/ML system is.

5) Human intervention is required for AI/ML security since it cannot be fully automated today.

6) Use case significantly impacts the way managers think about its cybersecurity.

7) The environment (e.g. governance, location) in which an AI/ML system is used is a factor in the cybersecurity management of that system.

Figure 5: Themes from the Data

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

4.1 Theme 1: It is difficult to differentiate between a valid or hacked output of an AI/ML system.

Explainability is at the heart of trusting an AI/ML system.This theme captures the idea that because AI/ML systems train themselves, then it is hard to explain how the system reached its output. If we cannot explain its output, then we do not know if the AI/ML system has been hacked or if the system has evolved beyond what is managerially justifiable.

Overall, AI/ML systems are designed to find unique solutions, so it is difficult to determine when managers see that unique solution if it is truly the appropriate output or if somehow the system has been tampered with through bad training, manipulated models, or bad data.

Some of the quotes that support this theme are:

- “From the human operator perspective, these systems tend to be somewhat buggy. What I have been focusing on is ***how should the operator even delineate if this is the system just acting up or if it is being attacked by something?***”

- “Interpretability of the activities that are going on is key. ***If no one has a sense of whether the system is proceeding normally, how can they understand if we are being hacked?***”
- Question: How to trust the system’s output and when not to?
  Answer: “You want these places where you’re actively looking for where the machine and human differ. It seems idiosyncratic and how to deal with its scale. You want these systems to find the outputs you don’t expect. You want it to take all that’s going on and

figure it out, because why would we even use a system if we didn't want to use it to figure it out? ***These are fundamentally hard problems***."

What we are taking away from these quotes is that managers want to trust the output of the system, but it is very difficult to figure out when you see that needle in the haystack if it is a real, valid result or something that hackers manipulated.

4.2 Theme 2: Third-party models and training sets are standard ways to build AI/ML systems, but they come with additional potential vulnerabilities.

Developers like to use libraries of models and training data sets to speed up development. Rarely do developers today start from scratch to build an AI/ML system, and developers do not always have insight into how these models and training data sets were pre-trained or built, and this opens up vulnerabilities. There is literature and discussion about how open sourced software causes cybersecurity challenges, for example, in 2017, Equifax had a major cybersecurity incident because Apache Struts, an open-source tool, had a vulnerable patch. But it is important to note that this issue is exacerbated in AI/ML systems because in addition to risks such as this, managers do not understand how the model has been trained or even the model itself.

Similarly, for AI/ML developers today, using libraries is the most common way new systems are developed, and those components may have vulnerabilities or introduce new vulnerabilities that at first glance the developer may not know about.

For previously-built models:

- "[For example,] the weights that get stored, which are usually pretrained. ***I worry about the possibility for manipulation there***. How do I know that when I download one of the models available (e.g. Google's model), [it hasn't] been manipulated in some way?"

For libraries:

- "The library side is a real risk… As I look at a ResNet image classification, ***how could that library have manipulated a model?*** That's less common, but still something that makes me nervous. Libraries getting used to do this work can be very opaque [...]"

In summary, models are trained using data that the developer cannot validate. Training datasets may have biases or have been hacked in a way that is not detectable by the user. So even though libraries definitely speed up the development of the system, they can introduce unintended vulnerabilities which can be exacerbated by the fact that it is an AI/ML system.

4.3 Theme 3: AI/ML systems consume such a large volume of data that malicious data could potentially evade detection.

The volume of data required by an AI/ML system is so large, the speed of incoming data is so great, and the data can carry so much noise that managers cannot track each piece of data. This makes it easier for malicious data to be input and evade detection, which can be used to attack the system's training or its inference processes. It is very difficult to clean that amount of data and validate every piece of it to make sure the data is valid and not hacked. Even if managers could clean and validate each piece of data, it is not clear whether they could identify malicious data inserted into each dataset.

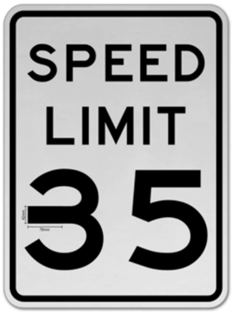


Figure 6
Source: Alex Woodie. Datanami. [8]

One example of malicious data inputs resulting in an issue is an adversarial attack, which is defined as a way of manipulating a machine learning model by feeding the system a specially crafted input. As Figure Y demonstrates, it is evident to the human eye that the image displays a number reading “35,” albeit an evidently altered “3”. However, researchers at McAfee demonstrated that a Tesla Model S perceived this was an 85 miles per hour speed limit sign [8].

In another case, a skin cancer detection algorithm mistakenly classified every skin image that contained ruler markings as indicative of melanoma. This was because most of the images of malignant lesions contained ruler markings, and it was easier for the machine learning models to detect those than the variations in lesions themselves [9].

The interviews conducted as part of this research provided further support in evidence of this theme. A select number of quotes are as follows:

- “***Volume*** of data is a concern, since ML/AI systems train on ***a large volume of data which is harder to protect*** and maintain than a small volume of data.”
- “The ***speed*** with which data changes (e.g. the data changing every 5 minutes or so) ***makes it hard to maintain*** and protect it.”
- “The inference algorithm is offline testable, but the ***new data streams can rapidly evolve***, which is a concern.”
- “The speed and volumes of data we are talking about will multiply exponentially… ***Data integrity and accuracy is critical.*** We can build the algorithm [...] in the system, but how can I ensure that the data I’m putting into the application is the best at that specific point in time? ***Data integrity concerns me.***”
- “If someone is very sophisticated, then ***they can launch an attack based on faulty data***, and then [use the] input data to tamper with the model.”

4.4 Theme 4: Managers need well-accepted measures of how secure an AI/ML system is.

Measuring the cybersecurity of an AI/ML system is in general difficult. Managers need well-accepted measures for general systems, AI/ML or not. While there are good development practices such as the secure development life cycle (SDLC), for AI/ML systems, with their feedback loop, automatic learning, system training, and other unique processes absent in non-AI/ML systems, the opportunity for hacking going undetected is greater than in non-AI/ML systems. As a result, having a well-accepted measure of how secure an AI/ML system is, or even a process for validating its security, would help managers manage the cybersecurity of the AI/ML systems they are responsible for.

We can consider the example of hackers tricking a Tesla into veering into the wrong lane. Keen Labs, a top cybersecurity research group in China, developed two kinds of attacks to tamper with Tesla's autopilot lane-recognition technology. The researchers created a "fake lane" by placing three miniscule square stickers at an intersection. The researchers hypothesized that the Tesla algorithm would detect these stickers and interpret them as a continuation of the right lane, and they were shown to be correct as the Tesla veered into the left lane, proving that the tampering was successful [10]. What the Tesla example illustrates is that even though the AI/ML system, in this scenario an autonomous vehicle, was designed such that it could not be hacked, and even though the developers of the AI/ML system undertook measures to ensure its training data was not biased and its model was trained on a vast number of inputs, along with a number of other measures, all it took for a successful cyberattack was for the vehicle to encounter a carefully crafted and placed input, which resulted in the car veering into the wrong lane.

We also heard our respondents echo the need for well-accepted measures in quantifying how secure an AI/ML system is:

- "The KPIs, tooling and standards used for measuring risk in AI is, ***at best, an immature and disparate discipline***."
- When our research team asked one interviewee on whether their organization had any metrics in place to ensure that their AI systems were secure, the interviewee responded with: "***I think nobody has asked me that yet.***"
- "***I don't think that culture exists.*** I think nobody has asked me that yet. But if they were, ***I'd give a qualitative answer not a quantitative one.*** In data science, there are metrics around algorithm performance. On the engineering side, there are metrics around latency time, etc. ***There's no standardization around cybersecurity.***"
- "So far, ***we treat [AI systems] like other automated systems.*** [For example,] if your GPS is off, how do you detect it's off? You compare what that one sensor is telling you to other sensors. You look at the paper map and compare [it] to what the system is telling you. ***We don't have that done yet for this system.***"
- "Other engineering disciplines are used to not having perfect measurements. IT is different. It's binary… I think computer scientists are ill-equipped. They don't take

statistics or lab classes. ***People don't think statistically or probabilistically, so your testing mechanisms aren't set up properly.***"

In terms of managerial implications (i.e. what managers can learn from the managers interviewed as to how they are approximating the security of AI/ML systems), we note that while there is still a need for well-accepted measures to quantify how secure an AI/ML system is, there are a number of steps managers can take to approximate their AI/ML system's security. A number of these steps were reiterated by our interviewees and reflect the practices they follow in their organizations:

- "We also have a ***system auditing our system***. [...] Here, ***we have a set of inputs and the inputs would act as if a real user was operating***… We also have ***user feedback***, so if they see something that looks odd they'd tell us and that would flag something we would look at."
- "If you use questionnaires it's easy to have poisoned data; people know what to check 'yes' for. ***You need continuous checking*** in that case, [so] that [there] is some kind of rating of the security over a longer period of time. ***You need [a] rating for the security of third parties.*** [This] might be ***more important if you have an AI system.***"
- "Another concern is getting very robust on the inputs and outputs, and how we track [those inputs]. [...] ***We can look at distributions over time.***"

4.5 Theme 5: Human intervention is required for AI/ML security since it cannot be fully automated today.

There is a general belief that the security of an AI/ML system can be automated, but that is not possible right now. AI/ML systems require a human in the loop in order to be secure. While it would be nice to be able to fully automate the security of an AI/ML system, it takes a combination of automated and human intelligence to ensure a system is secure. The automated part can do pattern matching and identify anomalies, but today there is a need for human intelligence to ensure that the system is not making blatantly obvious classification errors.

An example of this is an AI/ML system that classifies pictures, but through an adversarial attack that changes the pictures in subtle ways, it can label obvious images as something different, for example, a picture of a bus as an ostrich. To the human eye, this is obviously wrong, and a human operator would be able to step in and then course correct the system.

There was also a concern voiced by an interviewee that if you automate the system and all the checks and all the processes around it, the threat actually becomes much greater. Other responses include:

- "We also have a ***human auditing layer***, where we will use analytics tools and summary statistics."
- "There's ***no substitute to having a group of folks*** whose job is to make sure there's no bias in the system."
- "Whenever you have a solution that helps automate or build something, a threat in terms

of impact moves from a linear aspect to an exponential in terms of a function. If I compromise a static access control, if the wrong port is open (say port 88 instead of port 80), that's fine because that's a very predictable and easy problem to solve. When it comes to AI systems, however, ***you can automate all of these processes around it, and the threat becomes far, far greater*** because this thing learns over time because the number of tasks you trust it to do is so much more severe."

- "Attacks have to be in such a way for them to impede the system that the human is fooled. So in the end, when I send a technician to your place, the technician will have a set of particular instructions. ***For someone to do a successful attack, you have to get past the human intelligence***."

4.6 Theme 6: Use case significantly impacts the way managers think about its cybersecurity.

It is very important to consider the use case when thinking about cybersecurity of AI/ML systems. Not only in terms of the attacks the system could be under, but also what are the potential risks if the system is successfully hacked. For example, concerns for an autonomous vehicle system are different from those of a credit evaluation system.

It is also important to consider that AI/ML can exacerbate risks by opening new attack vectors, which are dependent on context, making the threat bigger as mentioned by one of our interviewees in the previous theme. In that sense, looking at the cybersecurity of AI/ML is contextually dependent.

One way of looking at this is thinking about the way the system is used. Systems that automatically take action might have a different level of cybersecurity needs than something that recommends action. But also which are the users of the system and who might be the attackers. The same AI/ML model can have different vulnerabilities depending on how or where it is going to be used. With respect to this theme, our interviewees shared:

- "***You need a 'fit for purpose' idea.*** [...] A one size fits all [approach], instead of understanding context of environment I'm in, doesn't work."
- "AI systems are very application dependent so ***security is different depending on applications***."
- ***The way they will attack AI is completely different*** from legacy systems. For that, ***we need to look at the use case*** – how do we manage risk from use case perspective? Then, control data? Then, protect algorithm?
- All tie together with physical safety. ***If action can hurt a person*** or another physical system, ***it's more complicated***.

4.7 Theme 7: The environment (e.g. governance, location) in which an AI/ML system is used is a factor in the cybersecurity management of that system.

The environment in which an AI/ML system works impacts its cybersecurity vulnerabilities both in severity as in breadth. Although for all software systems it is clear that, for example, residing in the cloud implies different cyber vulnerabilities from those when in a customer premises, for AI/ML systems there are extra concerns that do not occur in non AI/ML systems or vulnerabilities that do exist become exacerbated.

Another big aspect to consider when thinking of the environment is that while intuitively many people think that it only consists of the system's hardware (e.g. server), in reality the environment also consists of the regulations governing the AI/ML system, its storage location, and the data scientists interacting with it just to name a few. Since AI/ML systems are so new, there can be vulnerabilities in all the different system's components that are yet to be discovered or exploited. Also, the people that interact with these systems are not yet trained on the ways that they can be attacked.

Here are some insights from our interviewees:

- "***Environment is based not only on where the data is hosted, but also the context of the threats you're trying to secure against.*** For example, don't just think about the data center or the air gap network... Here, you're operating with the idea that whatever is in your environment is closed off in your firewalls and there's no threat."
- "[...] [We can think about] cloud infrastructure (e.g. public vs. private cloud infrastructure), remote sites (e.g. branch office, a building in Cambridge for example with its own internal network), [and] ***the environment can also be two guys in a pickup truck with a modem using 4G or LG to connect to the internet.***"

5. Discussion

Now that we have discussed our themes, which were our key takeaways, and the evidence we encountered in support of each of these themes, we provide further recommendations to managers based on our research. These recommendations, or managerial implications, are discussed for each theme in this section.

5.1 Theme 1: It is difficult to differentiate between a valid or hacked output of an AI/ML system.

We recommend that managers utilize mechanisms to better understand the cybersecurity considerations for their AI/ML systems, so they can more easily trust the system's output. We also note that managers need clear flags to help them determine whether the output of the AI/ML system is suspect.

5.2 Theme 2: Third-party models and training sets are standard ways to build AI/ML systems, but they come with additional potential vulnerabilities.

We recommend that managers and their organizations utilize a greater level of vetting when considering third-party models or libraries that have AI/ML components for internal use. This also applies to vetting third-party vendors or any external stakeholders who engage with the AI system. Furthemore, we recommend that managers keep in mind that just because a library is popular or has been previously used and previously approved does not mean that it is secure. Popularity does not translate into security. Previous use does not mean that after an update or a change, a library is still secure.

5.3 Theme 3: AI/ML systems consume such a large volume of data that malicious data could potentially evade detection.

In terms of the managerial implications of this theme (i.e. in terms of actions that managers of AI/ML systems can take and the priorities they can set), we recommend that managers continuously monitor the AI/ML system's data, the way that managers would try to detect if there was bias present in the datasets to detect if the system's data has been hacked. While not a complete solution, continuous monitoring is an entry point for building situational awareness about the system. Furthermore, while continuous monitoring for data drift management is used to alert managers if and when the system's users begin exhibiting a different behavior than expected, continuous monitoring of data can also be used to identify whether the system's data has been tampered with.

5.4 Theme 4: Managers need well-accepted measures of how secure an AI/ML system is.

To reiterate the findings for this theme, in terms of managerial implications (i.e. what managers can learn from the managers interviewed as to how they are approximating the security of AI/ML systems), we note that while there is still a need for well-accepted measures to quantify how secure an AI/ML system is, there are a number of steps managers can take to approximate their AI/ML system's security. For these steps, please reference the quotes for this theme in the previous section.

5.5 Theme 5: Human intervention is required for AI/ML security since it cannot be fully automated today.

With respect to the managerial implications of this theme, we recommend that managers and AI/ML organization leads require having humans in the loop when it comes to monitoring the AI/ML system's security, at least until the security of these systems can be automated. We also note that if managers do not have humans in the cybersecurity loop, that might be a red flag indicating a potential cybersecurity threat to the AI/ML system.

5.6 Theme 6: Use case significantly impacts the way managers think about its cybersecurity.

In terms of the managerial implications of this theme, we recommend that managers develop cybersecurity risk classification methods that classify system cybersecurity risk based on use cases (i.e. at a use case to use case level) and understand their systems at the use case level instead of understanding only specific parts or certain components of the system. Cybersecurity management of AI/ML systems requires the managerial understanding of each component of the AI system, and varying use cases require varying cybersecurity considerations.

5.7 Theme 7: The environment (e.g. governance, location) in which an AI/ML system is used is a factor in the cybersecurity management of that system.

With respect to the managerial implications of this theme, when managing the cybersecurity of AI/ML systems, we recommend managers assess the environment in which the AI/ML system resides, understanding it as a whole and considering not only its hardware, but regulations, users, etc.. This also connects with the managerial implications of the previous theme, where we discussed the need for a cybersecurity risk classification method.

6. Conclusion

Cybersecurity of AI/ML systems is still a very immature discipline. There are no well-accepted measures of how secure an AI/ML system is, managers cannot tell the difference between data that is hacked versus data that has not been hacked, and human intervention is necessary to secure these systems. As managers think about investing in AI/ML systems for their organizations, AI/ML system security cannot be achieved by undertaking the same approach as securing non-AI/ML or “traditional” systems.

We have found that managers require mechanisms to better understand the cybersecurity plans for the AI/ML systems so they can trust the output arising from these systems. Managers also need clear flags to watch for to determine if the output is suspect (other than their gut feelings). We have also found that in order to manage the cybersecurity of AI/ML systems, managers must look at the environment in which the system resides and apply appropriate cybersecurity measures.

Overall, the biggest implication from our work is that AI/ML systems are different from non-AI/ML systems from a cybersecurity perspective. As a result, managers need to re-evaluate the cybersecurity measures in place for securing traditional when dealing with AI/ML systems.

We also highlight a number of future insights that arose from our conversations that warrant future research. The insights discussed here were shared by our interviewees, but were not mentioned with enough frequency to be aggregated and distilled into a theme. These points include:

- The effectiveness of regulatory and governance practices (e.g. requiring the use of new technology that was purchased in recent history, as opposed to continued use of legacy systems) in enforcing organizations and managers to adopt new technologies.
- The impact of organizational culture in ensuring that cybersecurity is at the forefront of AI/ML systems development as opposed to an after-thought that is bolted on after the system has been designed and developed.
- The applications for emerging cryptographic methods in securing data and maintaining data privacy by tracking the data pipeline, securing the data source, and monitoring the flow of data through each component of an AI/ML system and across other systems in an organization.
- Whether bias in systems is a similar problem for AI/ML trust and security, and whether solutions to preventing bias in AI/ML systems can be applied to securing AI/ML systems.
- Cybersecurity risks that arise in the supply chain in which AI/ML systems play a role.
- Further research on tangible measures of determining how secure AI/ML systems are.

Furthermore, one future project might be to devise a well-accepted framework to determine the cybersecurity of AI/ML systems. Another future project might be to devise a framework for determining when human interaction is needed in securing an AI/ML system, as opposed to when the cybersecurity of AI/ML systems can be automated. A third future project might be to undertake a systems-view of an AI/ML system, which, based on STAMP research here at the Massachusetts Institute of Technology, would entail modeling AI/ML systems including its environment and components (including software, hardware, human, etc.) and understanding where the biggest cybersecurity risks are in them from a systems design perspective.

6. References

[1] Comiter, Marcus. “Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It.” Belfer Center for Science and International Affairs, Harvard Kennedy School, 2019, www.belfercenter.org/publication/AttackingAI.

[2] Darraj, Emily, et al. “Artificial Intelligence Cybersecurity Framework: Preparing for the Here and Now With AI.” 18th European Conference on Cyber Warfare and Security (ECCWS 2019) Coimbra, Portugal, 4-5 July 2019, by Tiago Cruz and Paulo Simoes, Curran Associates, Inc., 2019, pp. 132–141.

[3] Jang-Jaccard, Julian, and Surya Nepal. “A Survey of Emerging Threats in Cybersecurity.” Journal of Computer and System Sciences, Academic Press, 10 Feb. 2014, www.sciencedirect.com/science/article/pii/S0022000014000178.

[4] “Embedded Systems.” Embedded Systems - an Overview | ScienceDirect Topics, www.sciencedirect.com/topics/computer-science/embedded-systems.

[5] Hemberg, Erik, et al. “Exploring Adversarial Artificial Intelligence for Autonomous Adaptive Cyber Defense.” Springer Link, 5 Feb. 2020.

[6] 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.

[7] Ruppel, Paul Sebastian, and Günter Mey. “Grounded Theory Methodology.” Oxford Research Encyclopedia of Communication, 29 Mar. 2017, oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-522.

[8] Woodie, Alex. “Hacking AI: Exposing Vulnerabilities in Machine Learning.” Datanami, 29 July 2020, www.datanami.com/2020/07/28/hacking-ai-exposing-vulnerabilities-in-machine-learning/.

[9] Dickson, Ben. “What Is Machine Learning Data Poisoning?” TechTalks, 7 Oct. 2020, bdtechtalks.com/2020/10/07/machine-learning-data-poisoning/.

[10] Brewster, Thomas. “Hackers Use Little Stickers To Trick Tesla Autopilot Into The Wrong Lane.” Forbes, Forbes Magazine, 1 Apr. 2019, www.forbes.com/sites/thomasbrewster/2019/04/01/hackers-use-little-stickers-to-trick-tesla-autopilot-into-the-wrong-lane/?sh=3df0bcd77c18.