Trusting the Needle in the Haystack: Cybersecurity Management of AI/ML systems

Sanjana Shukla[1], José Ignacio Parada[2] and Keri Pearlson[3]

[1] Massachusetts Institute of Technology, Cambridge MA 02142, USA
[2] Massachusetts Institute of Technology, Cambridge MA 02142, USA
[2] Massachusetts Institute of Technology, Cambridge MA 02142, USA

Abstract. Securing Artificial Intelligence/Machine Learning systems presents unique cybersecurity management issues not present in non-AI/ML, or "traditional", systems. These management issues arise from the unique components of AI/ML systems such as self-learning and the great amounts of data they need to use to train themselves. This paper presents several findings obtained from qualitative research related to key aspects of cybersecurity applied to AI/ML systems and their managerial implications. 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 proposed ways that managers can build appropriate cybersecurity plans for these systems.

Keywords: Cybersecurity, AI, ML, Management,

1 Introduction

This research focuses on investigating unique cybersecurity management issues that arise in artificial intelligence (AI) 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 recommendation and those that autonomously carry out actions resulting from recommendations. The following questions ground this work:

- What are the unique cybersecurity risks and attack vectors used to potentially harm applications that use ML?
- How should managers assess the cybersecurity risks associated with applications of ML?

AI/ML systems are designed to identify unique patterns using self-learning engines and validated 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-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 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 ML system's output. This leads to a difficult problem for cybersecurity due to the inability to identify whether an ML system's output is truly unique or has been compromised.

We also created a general AI/ML system model highlighting the key components of the system (Figure 1). In this model, training data and new input data are fed into the machine learning algorithm, which produces a prediction (or recommendation). This is then either acted upon by the business or possibly interpreted by an inference engine (such as in the example of an autonomous vehicle where the output creates an action for the car to take). Recommendations become training data as they are feed back into the model so it can 'learn' about the appropriateness and adjust the model parameters accordingly.

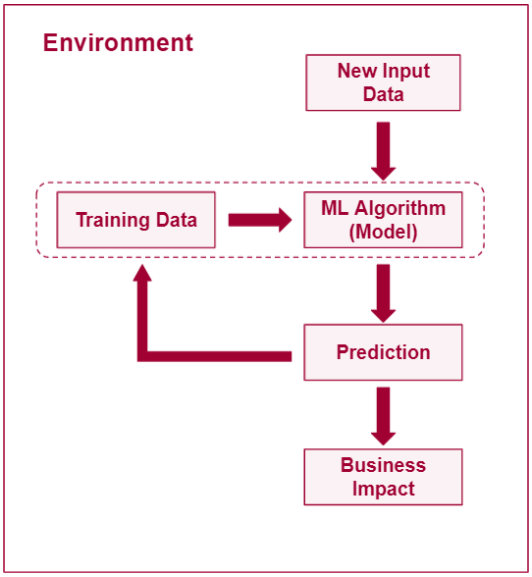

Fig. 1. Simple ML System model.

Our objective in this paper is not to explain in detail how an AI/ML system works. We provided the model in Figure 1 simply to clarify our baseline definition of an AI/ML system, which we used to narrow the scope of our research. Artificial intelligence encompasses many different technologies. We focused specifically on systems where a model is trained through an internal training process that utilizes training data to set up the system, then fine-tunes it over time with the output of the system through a feedback loop. As additional data is fed into the system, it changes

and ‘learns’ to produce better and more accurate results. At the same time, since the system auto-learns, and training is not done by a human, it can be difficult to ascertain whether the outcome of the evolved model is valid or not. It is this challenge of knowing when to accept the output of the system, and being unable to validate the results, that can get in the way of managers identifying if the system has been hacked or not.

In the previous phase of our work, we observed that within an AI/ML system’s design are three unique components that create conditions for cybersecurity management concerns: data, processes, and feedback loops. First, AI/ML systems use 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 would be biased if the same data was used to both train and validate the model. Second, AI/ML systems have both training and inference processes; ML systems are designed to be trained and then to make recommendations and possibly take action. This encompasses the learning algorithm, which uses the training datasets as inputs to train the model. The model evaluates data, and the inference process takes the output of the model and makes recommendations (and in some cases, takes action). Third, 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. Expanding on the diagram in Figure 1, the attack surfaces include the data management system, the testing, models and inference engines, the communication paths (illustrated by the arrows), the human factor, and the context or environment in which the system is used.

The paper continues in 6 major sections. In section 2 we establish our literature review. In section 3 we talk about the methodology used for this paper. In section 4 we present our findings and in section 5 we discuss them. Finally, in section 6 we propose future work and in section 7 we present our conclusions.

2 Literature Review

Prior research underpins this project and highlights the vulnerabilities in the overall system as well as with the data, the evolution of the model, the use of the model, the environment, and the context. Cybersecurity issues are not the same for AI/ML systems and traditional systems. While traditional system vulnerabilities are often due to a programmer error, AI/ML vulnerabilities are often the result of the algorithm it uses for learning and the data used to train [1].

AI/ML systems are currently seen as attackable [1]. The state-of-the-art methods are inherently vulnerable, but the environment in which they reside determine how easily it might be to launch a successful attack.

Data is seen as a potential attack vector. Protecting against data being weaponized in AI systems is one of the key goals from compliance measures. Formal validation and restricted data sharing are important controls for reducing the vulnerabilities from the data [1]. Public availability of datasets and the ability to easily construct a similar malicious dataset that might be substituted for the actual training or system data increases the opportunity for malicious actors to manipulate the data [1]. Cybersecurity professionals are responsible for ensuring that data plans are monitored and updated as necessary to reduce the potential for weaponization or manipulation of the data [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. Research by Darriaj et al suggests that it is difficult to evaluate if an AI/ML system performs exactly as planned in a live environment due to the difficulty of understanding how the algorithms evolve [2].

Table 1. AI System Cybersecurity Concerns. (Adapted from [2])

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 and 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
Trustworthiness	AI will need trust relationships being multidimensional

Previous researchers have studied cybersecurity concerns of AI systems. Table 1 summarizes a subset of these concerns.

Effective cybersecurity management of AI/ML systems ignores traditional cyber management practices recommended for non-AI/ML systems. Researchers suggest that data integrity, user authentication, and network security measures are important areas for potential attack [3]. Exploiting the limitations of network protocols such as Internet Protocol (IP), Transmission Control Protocol (TCP) or Domain Name System (DNS) are common attack vectors that apply to both AI and non-AI systems [3].

AI/ML algorithms are already being used to create better and more robust cybersecurity frameworks; AI/ML algorithms generate autonomous cybersecurity scenarios and responses [4]. This, in turn, poses the question of how cybersecure the AI/ML algorithms are that are being used to make other systems cybersecure. If they

are not, this opens the possibility of a supply chain attack similar to the SolarWinds attack [5].

Considering the prior literature review, which shows the unique aspects of cybersecurity risks in AI/ML systems, this paper focuses on using a qualitative approach to better understand how the aforementioned risks affect managerial decisions.

3 Methodology

While previous literature informed our hypothesis, our research approach began with semi-structured interviews with managers and developers of AI/ML and cybersecurity experts to understand how they manage the cybersecurity of these systems. Specifically, we sought their perspectives, opinions, and expertise on how security concerns of AI/ML systems are (and could be) managed.

We conducted 20 interviews across 15 organizations. Our interviewees consisted of four developers, nine managers, four cybersecurity experts or consultants, one regulator, and two academics. The organizations represented included: two financial services firms, two consulting firms, four IT/technology services organizations, two retail organizations, two academic affiliations, one healthcare startup, one conglomerate, and one foreign regulatory agency. The organizations 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.

Each interview was manually transcribed by at least 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 faculty and students with background in cybersecurity concepts. 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 wanting to further explore a new or otherwise interesting concept the interviewee had begun to share once the conversation was underway. The interview guide had this set of preliminary interview questions:

1. What cybersecurity vulnerabilities do you see in AI/ML systems?
2. How are 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, 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?

We used a grounded theory approach. Corbin and Strauss [6] suggest that grounded theory research follow specific procedures for data collection and analysis. The data collection and analysis are interrelated, with analysis beginning as soon as the first bit of data is collected, and the learnings from earlier interviews are used to inform the next set of interviews. In our research, this was the case. Early interviews influenced later interviews as our research team became more familiar with the subject area. Second, concepts are the basic unit of analysis, where the data collected in interviews is labeled, and subsequent responses that resemble each other share the same label. In our research, comments from the interviews were labeled as part of the coding process. Third, categories must be developed and related, grouping concepts that about the same phenomenon together in categories. That was how the themes, reported in this paper, were created.

We analyzed the interview responses and grouped the similar quotes together, resulting in categories we called 'themes.' Fourth, sampling in grounded theory is based on theoretical grounds, drawing samples of concepts and variations. In this research, the phenomenon studied was securing AI/ML systems, and as the interviews progressed, the focus was not on the AI/ML system per se, but on the security activities and concerns of the individuals who design or use the systems. Conditions such as the environment or the use case of the AI/ML system referred to by the interviewee were noted. The fifth procedure was to constantly compare as analysis is done, and that describes how our themes emerged. This helps guard against bias and provide a basis for comparing and evaluating the data collected in subsequent interviews.

The sixth procedure is that patterns and variations must be accounted for. This was noted as discussions of securing AI/ML systems happened. As patterns and variations emerged, subsequent interviewees were asked to comment on them, as appropriate, to help the researchers clarify what was 'normal' and what was 'an anomaly'. The seventh procedure is that process must be built into theory, meaning that the phenomenon studied was broken into stages, steps, or phases. Our description of the AI/ML system is a process flow with different steps starting with input data or test data, and resulting in an action or business impact. Eight is that the researchers must keep track of thoughts, formulations, and revisions of theory during the research process. Written memos were part of the research process, as interviewers recorded their impressions, revisions to the hypothesis, meta concepts arising from each discussion, and modification of themes after each interview. Ninth is that hypotheses should be developed and verified as much as possible during the research process. As noted above, interviewers tested hypotheses with interviewees, and noted their comments, agreements, and exceptions. Tenth is that grounded theory research need not work alone. In our project our research team included lead researchers present at every interview and leading discussions with the rest of the team on ideas, findings, and interpretations. In addition, a researcher with significant experience in grounded

theory participated as a consultant and advisor to assist the lead researcher on the details of this approach. Finally, the eleventh procedure is that broader conditions affecting the phenomenon must be considered, and that was accomplished by questioning the context and environment. Context in this setting meant understanding the company and business process in which the AI/ML system resides. We did not set out just to understand the specific application of the AI/ML, but the broader context of the entire system [6].

In sum, 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. We conducted interviews, coded the interview transcripts, and analyzed the codes. 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 the themes that emerged. This was in line with our interpretation of grounded theory methodology.

4 Findings

We found seven themes from the interview data. These themes came directly from the grounded theory processes and are presented here with supporting quotes from the interviews.

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

The ability to explain what happened to the data is at the heart of trusting output of any system. But AI/ML systems train themselves, which can make 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 patterns from the data, so it is difficult to determine when managers see an unexpected pattern or recommendation, 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.**"

4.2 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 preexisting models and training data sets to speed up development. Rarely do developers today start from scratch to build an AI/ML system. That means that often they use pre-trained models or training data collected and assembled by a third party. Developers neither always have insight into how these models were developed or trained nor information on how the training data sets were assembled. This opens vulnerabilities. While some commented on how open-sourced software can cause cybersecurity challenges for traditional systems, the issue is exacerbated in AI/ML systems because additional risk is created when managers do not know what has gone into the model or training data they are using. It could be rogue or tampered with prior to use, which might be even more difficult to detect.

AI/ML developers we interviewed shared that using libraries is the most common way new systems are developed today, and those components may have vulnerabilities or introduce new vulnerabilities that the developer may not know about. Some quotes that support this theme are shared below:

- 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, when models are trained using data that the developer cannot validate, the training datasets may have been modified to include biases or have been hacked in another way that would impact the training of the model. Libraries speed up the development of a system, but they can introduce unintended vulnerabilities. And the same risk can be introduced by off-the-shelf models that developers use as input to their AI/ML system.

4.3 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 developers may not be able to properly validate the data before it is used by the AI/ML system. This makes it easier for malicious data to be inserted and evade detection, effectively attacking the system's training or its inference processes. Interviewees shared how difficult it is to clean and validate the large volume of data in a timely manner, and further, to even identify malicious data inserted into a large dataset. Here are some of their comments:

- "**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 Managers need well-accepted measures of how secure an AI/ML system is

There are no well accepted measures of how secure a system is, in general, making it difficult to establish a baseline, understand when security has increased or decreased, and compare systems. But this issue is exasperated in AI/ML systems as the vulnerabilities we have already discussed make security an important contributing factor to trusting the system outputs and taking action based on their recommendations.

Effective development practices such as the secure development lifecycle (SDL) assist in the initial design of a system, but are less effective for AI/ML systems. The feedback loop which promotes automatic learning, and the unique features for inference engines that lead to actions (such as driving an autonomous car, or adjusting controls in an industrial setting) make the problem of measuring how secure the system is even more imperative. Interviewees often commented that a well-accepted measure of how secure an AI/ML system is, or even a process for validating its security, would help managers make sure their AI/ML systems met minimum standards of cybersecurity.

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. The test showed the Tesla veering into the left lane, proving that the tampering was successful [7]. What this Tesla example illustrates is that even though the autonomous vehicle (an AI/ML system) was designed such that it "could not be hacked", the developers undertook measures to ensure its training data was not biased and the model was trained using a vast number of inputs to account for many anomalies that the car was likely to encounter, all it took for a successful cyberattack to make the car veer into the wrong lane was a carefully crafted, manipulated input.

Interviewees echoed the need for well-accepted measures in quantifying how secure an AI/ML system is. When asked about measures, responses included:

- "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 whether or not 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.**"

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

One of the goals of the field of AI/ML is to build systems that operate correctly, with little to no human interaction. The vision is that the system would be initially trained, and then 'learn' from consequences of the actions suggested by and acted on by the inference engine. But our interviews suggested we are not there today.

Today, AI/ML systems require human interaction in the feedback loop to validate the output and make sure it is appropriate to introduce back into the model. While 'machine learning' is possible, developers shared that they want to see the output before it becomes input from the feedback loop to be sure the system learns appropriately. Automation can manage pattern matching and anomaly identification, but developers do not yet trust the systems enough to let output be automatically reintroduced into the model.

An example of this can be clarified by considering an AI/ML system that classifies pictures. Testers changed the pictures in subtle or obvious ways, and the system labeled images incorrectly. In one example, a picture of a bus was classified as an ostrich. To the human eye, this is obviously wrong. To make sure the classification system did not 'relearn,' a human operator would step in and 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, the 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 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 Use case significantly impacts the way managers think about its cybersecurity

The use case for an AI/ML system greatly influences the cybersecurity investment made to secure it. The use case impacts the attacks the system could face and the

potential risks should there be a successful attack. For example, safety concerns for an autonomous vehicle system are different from those of a credit evaluation system. Should the vehicle be hacked, people could be injured or worse but for a credit evaluation system, loss of life is not the impact. At worst someone would not get the credit they sought. Cybersecurity for the vehicle would be a higher priority than for the credit evaluation system.

Systems that automatically take action might need a different level of cybersecurity than something that recommends action (and likely has human intelligence evaluate the recommendation prior to taking action). The same AI/ML model can have different vulnerabilities depending on how or where it is 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 he context of the 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 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 use case is one factor in the urgency of cybersecurity, but the physical environment in which an AI/ML system works was also a theme emerging from this work. Cyber vulnerabilities differ for systems based on how they are physically housed, how the data is physically transported to and from the system, how the system is governed, how the interface to the system works, and other environmental variables. Systems that reside in the cloud have unique cybersecurity concerns than those residing in a on-site data center, for example. And cybersecurity controls used to secure systems differ in different situations, introducing different cyber vulnerabilities. In addition, the actual data scientists and other people interacting with the system factor into the cyber security posture. Here are 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

The themes emerging from the data collected in this study show that cybersecurity for AI/ML systems must be considered from additional perspectives compared to traditional IT systems. Anomalies are more difficult to identify in the models and the data. Lack of acceptable security measures make it difficult to manage the cybersecurity of these systems. Use case and environmental factors introduce cyber vulnerabilities. In this section we discuss the implications of these themes.

Managers want to trust the output of the system, but it is very difficult to differentiate between 'needle in the haystack' that is either a real, unique recommendation or finding or if it is something that hackers manipulated. To manage this vulnerability, mechanisms are needed to provide transparency as systems learn and models evolve so the user can better trust the system's output. Managers also need clear flags to help them determine whether the output of the AI/ML system is suspect.

Developers and their managers need ways to evaluate the security vulnerabilities in the open source and third-party models and data sets they use to build their systems. Vetting processes must be robust, valued, and constantly updated to account for new attack vectors (consider the SolarWinds hack in 2021 that changed the paradigm when their installation software was compromised after certification for delivery). Some companies have centralized validation teams who evaluate security considerations from third party vendors before developers are allowed to use them. Models and data sets must also undergo rigorous evaluations for cyber risks. Just because a library is popular or has been previously used and previously approved does not mean that it is still secure. Updates, maintenance, and changes can introduce new vulnerabilities. The vetting process must evolve, too.

Managers need tools to continuously monitor the AI/ML system's data to detect bias, weaponization or malicious injections in data flows. Tools must be able to keep up with the rate of the data flow, confirm the new types of data used by the system, and validate the actual values/images of the data itself. Data drift can be detected and managed before it corrupts the AI/ML system.

Another possible method for insuring integrity of data is to find new ways to encrypt the data at the source of its creation and decrypt it as part of the AI/ML system. Modern cryptography techniques offer the promise of integrity of data, regardless of the volume or rate of flow.

While there is still a need for well-accepted measures to quantify AI/ML system security, there are several steps managers can take to approximate AI/ML system's security. Some of the interviewees shared their current practices. In one case, the interviewee shared that they have a separate system auditing their AI/ML system for

security breaches. The interviewee commented, “We have a set of inputs that act as if a real user was operating the system [and we evaluate how the system acts with these inputs]. We also have user feedback, so if the user sees something that looks odd, they tell us and that would generate a flag that we would look at.” This company has found a way for the users to of the system to be part of the cyber defense of their AI/ML system, assuming that users would be able to identify if something looked wrong.

Given the immature state of validating AI/ML systems, human interaction is a necessity for systems today. While the promise of AI/ML is a fully automated system that makes all the right decisions and learns from the decisions and recommendations it does make, that reality is not the state of the technology today. When managers do not have humans in the cybersecurity loop, that might be a red flag indicating a potential cybersecurity vulnerability.

When securing AI/ML systems, managers need to 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, the interactions between them, and varying use cases require varying cybersecurity considerations.

Finally, environmental considerations for AI/ML systems are critical for reducing cyber vulnerabilities. Good cyber hygiene is the base line for the securing the environment—things like appropriate physical security, and teaching users about the importance of security as they interact with the system, download data, and take away their recommendations. But for AI/ML systems, the impact of vulnerabilities introduced by environmental decisions can be masked by the complexities of the system itself. For example, bad data governance practices can appear to be hacked or manipulated data. Inappropriate access management practices can also appear as AI/ML manipulation if the wrong people have access to the system. Further, cloud platforms, while easier in some ways to manage, can introduce new, unanticipated cyber vulnerabilities. While cloud and ‘as-a-service’ providers often offer some cyber protections, each have their own parameters, arrangements, and security options that must be set appropriately for successful AI/ML protections. Minimizing cyber vulnerabilities means properly managing the environment.

6 Future Work

While this research found actionable insights for managing cybersecurity, it also highlights a number of areas that warrant future research. In addition to the need for better measures of AI/ML security, and more structured approaches to identifying the impact of the environment in which the system operates, the data raised additional topics. We note that interviewees comments motivated many of these ideas.

- The effectiveness of regulatory and governance practices in enforcing organizations and managers to adopt new technologies. As AI/ML matures, there are situations were interviewees envisioned mandating adoption of AI/ML. Understanding the security impacts must be part of new regulations and governance practices.

- 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 that can be used for securing and transporting data, maintaining data privacy, 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.

7 Conclusion

Cybersecurity of AI/ML systems is still very immature. There are no well-accepted measures of how secure an AI/ML system is, managers may not be able to 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 solely by undertaking the same approach as securing traditional systems.

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 output is suspect. We have also found that in order to manage the cybersecurity of AI/ML systems, managers must look at both the use of the system and the environment in which the system resides to apply appropriate cybersecurity measures.

While this research has suggested themes that inform the security of AI/ML systems, there is value in building up their defenses. AI/ML and the other artificial intelligence technologies offer the promise of increased efficiency, effectiveness, and transparency for many industries and use cases. Data volume is increasing exponentially as work practices change, meta-data and meta-meta-data is created for data coming out of systems, and as use cases increase. Solving the cybersecurity

vulnerability problems of AI/ML are a must. We cannot expect to trust these systems if we cannot verify and validate the recommendations they suggest.

References

[1] M. Comiter, "Attacking artificial intelligence: AI's security vulnerability and what policymakers can do about it," *Belfer Center for Science and International Affairs, Harvard Kennedy School*, August 2019.

[2] E. Darraj 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*, pp. 132–141, 4-5 July 2019.

[3] J. Jang-Jaccard and S. Nepal, "A survey of emerging threats in cybersecurity," *Journal of Computer and System Sciences*, 10 Feb. 2014.

[4] E. Hemberg, L. Zhang, and U.-M. O'Reilly, "Exploring adversarial artificial intelligence for autonomous adaptive cyber defense," *Adaptive Autonomous Secure Cyber Systems*, pp. 41–61, 2020.

[5] M. Willett, "Lessons of the solarwinds hack," *Survival*, vol. 63, no. 2, pp. 7–26, 2021.

[6] J. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," *Qualitative Sociology, Vol. 13, No. 1*, 1990.

[7] P. Jing, Q. Tang, Y. Du, L. Xue, X. Luo, T. Wang, S. Nie, and S. Wu, "Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations," *30th USENIX Security Symposium (USENIX Security 21)*, Aug. 2021.