Carnegie Mellon University
Software Engineering Institute

A RISK MANAGEMENT PERSPECTIVE FOR AI ENGINEERING

Brett Tucker

June 2020

Adopting artificial intelligence (AI) technology in an organization represents significant change. And when implementing any significant change, the organization will likely encounter risk; a fact that has been recognized by both public sector organizations, such as the United States Department of Defense (DoD) and private sector organizations such as McKinsey [U.S. DoD 2018, McKinsey 2019]. Fortunately, risk can be controlled to limit the potential impact to the organization.

OCTAVE FORTE is a process model that organizations can use to identify and mitigate risks. Its steps include establishing risk governance, appetite, and policy; identifying risks, threats, and vulnerabilities; and forming and implementing an improvement plan.

This paper focuses on a few of these steps in the context of adopting AI technology. These steps provide a starting point for organizations investigating the adoption of AI and exploring the associated risks.

1 Background

CERT Resilience Management Model (CERT-RMM), the foundation for a process improvement approach to operational resilience, defines the practices needed to manage operational resilience. Two definitions from CERT-RMM are particularly relevant to the discussion of risk in this paper [SEI 2016].

> **Risk** – The possibility of suffering harm or loss (From a resilience perspective, risk is the combination of a threat and a vulnerability [condition], the impact [consequence] on the organization if the vulnerability is exploited, and the presence of uncertainty.)
>
> **Condition** – A term that collectively describes a vulnerability, an actor, a motive, and an undesirable outcome (A condition is essentially a threat that the organization must identify and analyze to determine if exploitation of the threat could result in undesirable consequences.)

Risks can come to fruition and have impact if certain conditions exist, and some of these conditions might be interdependent. The next section discusses risk conditions and how organizations can control them to successfully manage risk.

2 Control the Conditions to Control the Risk

Organizations that adopt AI often encounter the following risk conditions:

- ill-defined problem statement
- lack of expertise
- model-system-data disconnection
- unrealistic expectations
- data challenges
- lack of verifiability

We describe each of these risk conditions in the following sections.

2.1 Ill-Defined Problem Statement

Typically, organizations encounter evolving needs or poorly defined requirements that obfuscate the actual problem the AI system is designed to solve. Organizations must constantly adapt to shifting environmental conditions, including internal or external threat actors who change their tactics, changes in the value of assets, and a change in venue. These shifts can complicate the problem to be solved and often lead to the risk condition involving an **ill-defined problem statement**.

In a cybersecurity environment (like most environments), vulnerabilities arise and technologies evolve. Some of these changes occur when the organization doesn't know about them. These changes—individually or in concert—demand nimble models that represent how those changes affect the organization, and these models require a form of data collection that strains the limits of current practice. From the onset, the organization must be aware of the many changes it faces in its environment. Also, the organization's risk managers must educate project managers about the uncertainty that arises when the organization faces undefined or improperly defined risks.

The organization must decompose each problem statement into smaller pieces and refine the requirements for addressing them. Theoretically, problem statement decomposition can dilute risk exposure; a wrong decision affecting a small piece of the problem has less impact than a wrong decision affecting the entire problem. However, breaking up the problem requires more projects to complete. Because all of these pieces and their projects also apply to the whole problem, they have to be integrated to solve the overall problem statement. In terms of analytic and processing rigor, this integration puts greater demands on the teams providing the solution.

A risk appetite statement[1] helps an organization understand the amount and type of risks it is willing to accept. It should identify the organization's (1) tolerance for risk, (2) how much impact of a realized risk is acceptable, and (3) how much the likelihood of risk realization requires mitigation. The organization should tune the impact and likelihood sections of its risk appetite statement so that it only

[1] An organization's *risk appetite* refers to the general amount of risk it is willing to take when seeking to achieve its strategic objectives [ISO 2018].

adopts AI technology using a measured approach. In addition, the organization should periodically pause to consider whether the AI solution is reasonable in the given situation.

2.2 Lack of Expertise

An organization may be unable to assemble the expertise it needs to enable the proper use, selection, development, deployment, maintenance, testing, etc. of AI-related technology. Tasks such as defining the problem, developing the model, collecting data, and constructing systems require skills and expertise that may not be readily available. An organization that is unprepared to adopt AI with its current assets presents the **lack of expertise** risk condition. This risk condition may be interdependent with other risk conditions; once the risk condition is recognized, its interdependence with other risk conditions usually becomes evident.

This risk condition also can render the organization ineffective in framing problems and crafting appropriate models that solve them. However, most organizations have similar risks that already exist on their risk register,[2] usually for other technical fields where the organization lacks the talent needed to realize its goals. The solutions to risk conditions like these should be a proactive talent strategy. Organizations must cultivate talent by fostering educational opportunities, identifying opportunities that provide experience, and following rigorous hiring practices. All of these actions take time and discipline to follow while adhering to a specific talent strategy.

It can sometimes be faster to hire someone than to train an existing workforce member. However, while there are many new candidates to consider, their talent and experience may be lacking. Developing expertise requires solutions from HR and workforce development. HR must vet new-hire candidates according to a repeatable and reliable process to hire qualified candidates. Once these candidates are hired, workforce development will arrange for needed training to develop additional skills. Clearly, talent strategy activities require time and other precious resources.

Some organizations might consider hiring consultants to fill a talent gap, but such a reactive approach can be costly and adds a supply-chain risk for services that also robs the organization of a degree of control. The organization can use the impact section of its risk appetite statement as a business case to justify whether it should assume the costs of the talent strategy.

2.3 Unrealistic Expectations

Some customers may be uninformed or uneducated about AI technology; therefore, they may not understand what AI technology can do. This situation can lead to the **unrealistic expectations** risk condition. Customers of AI technology must be educated to understand that the science of AI relies on mathematical modeling that enables automated, risk-based decisions. Since risk is probabilistic, there is always a chance of error.

---

[2] A *risk register* is a document that lists the organization's risks and information about them.

In fact, these limitations must be made clear to all stakeholders. Most importantly, customers and providers must understand and consider the consequences of errors occurring and determine if the impacts of those consequences fit into their risk appetite. Of course, they can decide to use non-AI solutions where practicable.

Some organizations identify the risks of using AI technologies based on misinformation and unrealistic expectations. Organizations that promote technology innovation or have a lot of early adopters are familiar with this risk condition. They not only recognize the potentially great benefits of working with unproven technologies, but they also understand the risks. To ensure success, organizations can tune their risk appetite and their readiness to adopt cutting-edge technologies to adjust to the potential for unrealistic expectations.

2.4 Model-System-Data Disconnection

When designing an AI system, it must emulate the conditions of its environment, and it must be fed the appropriate data to operate as expected. For example, suppose an organization develops an AI system that makes risk-based decisions from data gathered from sensors fitted in the organization's network. Such a system must be able to sense, collect, and compute the needed data to make the decision. The risk condition that involves the disconnection of **model, system, or data** can result in a system that doesn't meet its requirements. This risk condition can trigger a risk event with an undesired consequence, such as an AI system that produces poor decisions.

When developing an AI system, organizations must have a proactive and disciplined process for requirements exploration and secure development operations with a flexible and nimble software architecture. Such an approach is not new; a development solution that addresses other innovation-related risks may already be in place, and it can be applied to AI technology. Agile software development is an example of how developers can build a system that is flexible enough to continually adapt to changing conditions while maintaining model, system, and data alignment. An organization can temper the impact of this risk condition by defining its confidence in the system requirements at the outset of the project, thereby identifying the risks to the model, system, and data.

2.5 Data Challenges

Ensuring that the information being used for the AI system is sound relies on adhering to a model[3] where (1) data is assumed to be accurate and precise or (2) the data lake is so large that aberrant data points are overwhelmed by the volume of other, more accurate data points. Effective models such as these rely on large volumes of data. However, when models rely on the *relevance* and *accuracy* of data, the **data challenges** risk condition comes into play.

[3] A *model* is a conceptual representation of a process, system, or other entity.

Data relevance pertains to how applicable the data being used is to the model being used to deliver the desired information. Data relevance also depends on the model, system, and data being aligned with one another. In this regard, the model may experience concept drift, meaning that the real-world conditions being modeled shift in a way that invalidates the model. That same drift can happen with the data collected for the model (e.g., sensor fidelity might be inadequate or sensors might provide too much noise and need significant grooming).

Data accuracy pertains to how correct the data is. The accuracy of data depends on many factors, including how it is collected, the fidelity of the sensors collecting it, and the environment where the data is collected from. *Data poisoning* is one way that data accuracy can be compromised. It can be intentional (when a threat actor intentionally injects data that misleads neural networks) or unintentional (when sensors are flawed or collection requirements are unclear).

Data must be cleansed and labeled; this maintenance can be costly and tax organizational resources. Threats to these data maintenance efforts include the following:

- poisoned data
- biased interpretation of data
- faulty data collection
- low volume of data

Organizations must take proactive steps to limit the likelihood of these threats, which can have negative consequences.

To ensure that a model for an AI system can consistently make decisions as designed, the organization must build strategies for how to collect, use, and maintain the data the system uses. The organization must also consider other actions, such as determining appropriate refresh rates, whether or not to expunge old data, and how well the system accommodates change.

Although more reactive than strategic, AI users must continuously evaluate the verifiability[4] and efficacy of AI system results. The organization must accept that (1) not all data is perfect, and (2) making data usable consumes significant resources. The risk investment—in terms of collection, grooming, and evaluation—must be outweighed by the return of accurate, timely, and automated risk decisions.

2.6 Lack of Verifiability

Depending on the model and the data used by the system to make decisions, it may be a challenge for a user to **verify the results** of the system. It is critical that users confirm that the risk-based decisions made by the AI system are appropriate. This verification, if lacking, may question the potential bias and overall trust the users have in the system or even AI technology.

[4] See Section 2.6 for more details.

For the following reasons, this risk condition makes it challenging to verify results:

1. Interpretability of the results may be as important as knowing what results to expect. People often have trouble processing complex problems and may rely on AI technology for good information.
2. It is difficult for the organization to modify and tune the AI model when errors are identified.
3. It is difficult for the organization's risk appetite to tolerate model corrections and provide the sought-after benefits without risking that their stakeholders will lose trust in AI.

There are consequences and impacts to this risk condition. Because decisions made by the AI system can be difficult to verify, stakeholders can lose trust in the AI technology. This cause and effect begs the question of how often AI can fail before the organization abandons it. To reduce the risk of abandoning AI technology, the organization should undertake the following:

- Provide education about AI.
- Temper organizational expectations.
- Define the scope of where AI technology will be applied, starting with small problems with limited potential negative impacts.

These challenges are not exclusive to AI; therefore, the organization should find other scenarios with risks related to using other new and emergent technologies. Finding these scenarios also helps identify risks that share interdependencies with AI risks. Once risks are identified, risk managers can apply existing mitigations to the AI risks.

3 Risk and New Technology

The common theme among all of these risk conditions is that they are not exclusively encountered with AI systems. It is prudent to explore other circumstances where there are risks related to using new and emergent technologies. As an additional benefit, this exploration can identify AI risks that share interdependencies with other technologies. When these risks are identified, risk managers can increase the scope of existing mitigations to also account for the AI risk conditions presented in this paper.

An analogy of how multiple risk conditions can combine and be interdependent is manning and training a damage-control workforce on a battleship. The workforce can mitigate many risks including fires, flooding, injury, and other battle damage. To mitigate these risks, the following must be in place:

- The staffing plan must identify the correct talent.
- The ship must be designed to withstand the risk of damage coming to fruition.
- The Navy must adjust its risk appetite to tolerate a warship going into harm's way with confidence that their damage-control workforce can withstand the challenges.

AI technology may not be employed just yet in situations with the possibility of such extreme risk impacts, but that time is inevitable.

4 Dealing with the Consequences

Given the current state of AI technology, adopting it is subject to probabilistic error. Models can break down, irrational actors can act, data can be corrupted, and conditions can shift. Despite the proactive measures an organization takes, these risk events might happen. Organizations should plan for a measured response when these risk events occur to mitigate the impacts.

The risks of adopting AI technology can differ from the risk of adopting other new technology or innovations. AI can be granted the power to take significant actions without the knowledge or release authority of the organization. Therefore, the likely consequence of risk events is that the AI system will deliver unacceptable decisions or actions.

When analyzing the potential consequences of these risks, the impacts all tend to be the same type regardless of the risk conditions that brought about the consequences, but the magnitude of the "pain' may vary. For example, data challenges, model break down, and lack of talent can all lead to bad decisions. The risk manager must consider all the possible risk conditions that can lead to the negative consequence. The good news is that the risk manager can apply reactive planning[5] to all of the risks related to AI technology to avoid duplication of effort. To be successful, the risk manager must maintain a broad perspective that analyzes how these risks apply across the organization.

The risk conditions described in this paper can result in one consequence in AI technology: a system that makes bad decisions or takes inadvisable actions. Although the impact of each consequence varies by context, risk managers must conduct a business impact analysis (BIA)[6] to learn the extent of the pain experienced when a technology fails.

The similarity of AI-related risks to other new-technology risks may provide helpful BIA information. This information is critical, especially when viewed through the lens of the organization's risk appetite. The organization can also use this BIA information to tune its risk tolerance, making it easier to decide where to apply AI technology.

Early adopter organizations that seek to adopt AI should evaluate scenarios where confidentiality, integrity, or availability (CIA) are lost. Previous risk analyses related to the services and assets planning to use AI can be used to make early conclusions about what to do if a catastrophe occurs. In the end, an organization can use proactive measures that systematically introduce AI to limit risk exposure while adopting the new technology.

[5] *Reactive planning* is a plan for how to react when something occurs.

[6] BIA is an important step in risk management since it quantifies how much "pain" the organization "feels" when risks are realized.

5 Conclusion

AI technology continues to evolve, and organizations have an increased appetite for automated solutions to the challenges they face. AI may eventually satisfy that appetite and deliver systems that make risk-based decisions in a cybersecurity context. Until then, organizations must (1) routinely identify the uncertainties related to AI technology and (2) understand AI benefits and the ramifications of AI technology failings.

Initial steps for adopting AI technology should include the following:

1. The organization should establish a standardized risk management policy and procedures for implementing that policy. Doing so ensures consistency when adopting new technologies amid the related uncertainties.
2. The organization should establish a governance structure where risk-based decisions, such as adopting new technologies, can be made. If a risk governance structure is not yet established, the organization may opt to use other decision-making bodies such as a technology council.
3. The organization's risk program must work with executives to understand and communicate the willingness of the organization to take risks so that a reasonable risk appetite bounds the scope of decisions.

Proactively examining how to control the risk conditions related to adopting AI can help organizations introduce AI technology with minimal risk exposure. The SEI's CERT Division is evolving the OCTAVE model to help organizations with enterprise risk management. OCTAVE Allegro helps risk managers identify, analyze, and prioritize information-security-related risks. OCTAVE FORTE is the next evolution of OCTAVE that identifies all types of risk experienced across the organization in a way that resonates with executives. FORTE provides an approach for strategically introducing new technologies and prioritizing related risks. The SEI will publish OCTAVE FORTE in 2020. For updates, visit the SEI's website at https://www.sei.cmu.edu/about/divisions/cert/.

Bibliography

[Churilla 2020]

Churilla, Matthew. "Toward Machine Learning Assurance." Cybersecurity Developmental Test (CyberDT) Cross-Service Working Group (XSWG).

[Cohen 2020]

Cohen, Benjamin. Three Risks in Building Machine Learning Systems [blog post]. *SEI Blog*. May 2020. https://insights.sei.cmu.edu/sei_blog/2020/05/three-risks-in-building-machine-learning-systems.html

[Deloitte 2018]
Deloitte. *AI and Risk Management Innovating with Confidence.* Centre for Regulatory Strategy EMEA. 2018. https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovate-lu-ai-and-risk-management.pdf

[HBR 2020]
Harvard Business Review. *The Case for AI Insurance.* The Harvard Business Review. 2020. https://hbr.org/2020/04/the-case-for-ai-insurance

[ISO 2018]
International Organization for Standardization. *ISO 31000:2018 Risk Management Guidelines*. International Organization for Standardization. 2018. https://www.iso.org/standard/65694.html

[McKinsey 2019]
McKinsey & Company. *Confronting the Risk of Artificial Intelligence.* McKinsey Quarterly. 2019. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence

[NASEM 2019]
The National Academies of Science, Engineering, and Medicine. *Implications of Artificial Intelligence for Cybersecurity, Proceedings of a Workshop.* National Academies Press. 2019. https://www.nap.edu/read/25488

[SEI 2019]
Software Engineering Institute. *AI Engineering: 11 Foundational Practices*. Software Engineering Institute. 2019. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=633647

[SEI 2016]
Software Engineering Institute. *CERT Resilience Management Model (CERT-RMM) Version 1.2.* Software Engineering Institute. 2016. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=508084

[U.S. DoD 2018]
United States Department of Defense. *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity.* United States Department of Defense. 2018. https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

Contact Us

Software Engineering Institute
4500 Fifth Avenue, Pittsburgh, PA 15213-2612

Phone: 412/268.5800 | 888.201.4479
Web: www.sei.cmu.edu
Email: info@sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM20-0432