Chapter 16
Risk Management

Alexandre K. Ligo, Alexander Kott, Haley Dozier, and Igor Linkov

1 Introduction

Risk management is an important topic in research and practice of cybersecurity (Hubbard & Seiersen, 2016; Oltramari & Kott, 2018). One situation of interest involves the assessment of risks that a certain system or mission is exposed to, followed by an analysis of possible strategies to mitigate those risks. For a given mitigation strategy, one can evaluate how much of the risks assessed initially are eliminated or reduced. However, we must not forget to account for new risks that might be introduced by mitigation strategy itself.

A. K. Ligo (✉)
Environmental Laboratory, US Army Corps of Engineers, Engineer Research and Development Center, Concord, MA, USA

Engineering Systems & Environment, University of Virginia, Concord, MA, USA

US Army, University of Virginia, Concord, MA, USA

A. Kott
Environmental Laboratory, US Army Corps of Engineers, Engineer Research and Development Center, Concord, MA, USA

Engineering Systems & Environment, University of Virginia, Concord, MA, USA
e-mail: alexander.kott1.civ@army.mil

H. Dozier
Information Technology Laboratory, The U.S. Army Engineer Research and Development Center, Vicksburg, MS, USA
e-mail: Haley.R.Dozier@erdc.dren.mil

I. Linkov
Environmental Laboratory, US Army Corps of Engineers, Engineer Research and Development Center, Concord, MA, USA
e-mail: Igor.Linkov@usace.army.mil

A. Kott (ed.), *Autonomous Intelligent Cyber Defense Agent (AICA)*, Advances in Information Security 87, https://doi.org/10.1007/978-3-031-29269-9_16

This chapter is a discussion of risks that might emerge when AICA are adopted as part of a defense strategy. These risks can be associated with AICA inherent complexity. The concept and reference architecture of AICA was developed by NATO for military missions. Earlier in this book, Norlander notes that the military and other critical domains require extraordinary awareness and management of risk. This is because in these domains a successful cyberattack can result in death, injuries, or catastrophic material damage – a well-known example is the impact that the Stuxnet malware caused on Iran's nuclear program and its probable weapon capability (Kott & Linkov, 2019). In contrast, Norlander argues that in commercial operations objectives such as operational reliability, availability, and high technical performance at the lowest possible cost have priority over risk mitigation. Nevertheless, even in such commercial applications the use of AICA-like defenses may become essential. For example, intrusion detection and prevention systems tend to be increasingly autonomous given the rise in sophistication and frequency of cyberattacks, as well as the potential financial loss these attacks cause. Manual or semi-automated defenses will not be able to respond in required time, scale, and accuracy.

The inherent complexities of AICA in military missions and AICA-like systems in commercial applications introduce new kinds of risk. Norlander's chapter argues that AICA fits the definition of a cognitive system as one that can" modify its pattern of behavior on the basis of past experience in order to achieve specific anti-entropic ends". This would introduce specific risks may be related to AICA malfunction or AI bias, unintended effects arising from swarm-like behavior, communications or coordination failures among agents, or even attacks targeting AICA themselves. In this chapter we introduce the types of new risks, their consequences, and possible ways to mitigate them while preserving the AICA mission.

2 Types of Risks Introduced by AICA

Vast amounts of historical data about cyber activity are increasingly available. These data include logs of login attempts, domain resolution or webpage requests, application programming interface (API) calls, network traffic, and other activities. It is expected that AICA make use of these data to enhance AI algorithms by training machine learning (ML) models that detect future attacks (Kott & Theron, 2020). The enhanced AI capability translate into unsupervised actions that bring both opportunities and new risks. Some of these risks include flawed AICA actions due to wrong AI predictions. "Black box" AI models (discussed earlier in this book by Fitzpatrick) make it hard to prevent AI errors (Linkov et al., 2020). Likewise, data that are biased or contaminated with measurement errors may also result in wrong AICA predictions or actions (see Drasar's chapter on perception). Moreover, intentional hacking or destruction of the AICA themselves is also a risk.

Another type of new risks is related to collective AICA action. First, multiple agents may be required to cooperate with each other to achieve the scale or scope required for a given defense. Communications failures due to packet loss, poor signal-to-noise ratio, or network congestion impair coordination and action. Second,

communications between AICA may be intentionally corrupted by malicious agents. Finally, a group of AICA might exhibit swarm behavior that differ from the action of individual agents in unpredictable ways.

3 Consequences of Risks Introduced by AICA

Risks arising from AICA may have harmful consequences of functional, safety, security, ethical, or moral nature. Such consequences can be imposed on parties who do not benefit from the AICA actions or do not agree to accept the respective risks (Morgan, 2017). Types of consequences from AICA-specific risks include:

- Functional consequences: AICA might inadvertently impair the system's mission or functionality. One example is AICA needlessly shutting down service to avert an attack.
- Safety consequences: AICA might injury or kill system's operators or communities. For example, AICA take action against a cyberattack on an oil refinery, but the defense might inadvertently disable critical control systems and cause an explosion killing residents nearby (Ligo et al., 2021a).
- Security consequences: AICA might inadvertently create vulnerabilities that enable unauthorized access or data breaches, with consequences similar to the breach of Equifax data in 2017 that followed from vulnerabilities in Apache software (Federal Trade Commission, 2022).
- Ethical, moral or unfair consequences: AICA algorithmic biases might result in defenses that produce questionable results or prioritize certain groups over others. This includes considerations about whether AICA should maximize benefits for immediate stakeholders over social welfare at large. For example, should a self-driving car prioritize the safety of its occupants even if it exposes nearby pedestrians to increased risk?

The awareness of the nature of AICA-related risks or the possible consequences of these risks does not make AICA safer or more effective. Moreover, mitigating these risks is likely a challenging task. Nevertheless, understanding the nature and consequences of new AICA-related risks is a required step towards an evaluation of the net benefit of deploying AICA. In other words, are the risks mitigated by AICA more important than the new ones that are introduced? A different but related question is how these new AICA-related risks can be mitigated, which you increase the net value of AICA.

In the next sections we discuss possible mitigation strategies in deploying AICA that enhance cybersecurity and cyber-resilience while minimizing new risks. In particular, we discuss the human role in the design and control of defenses, as well as design or algorithmic strategies. While this discussion is non-exhaustive, it provides possible directions of research in risk management for AICA.

4 Human-Centric Approaches with Real-Time Cooperation

The natural remedy to mitigate the novel risks of harm caused by AICA is to have them team up with humans. This collaboration is essential not only from a risk perspective but also to ensure effective mission accomplishment, as noted by Norlander previously in this chapter for military operations – where he articulates the concept of Joint Cognitive Systems (JCS) for the interaction between humans and AICA.

However, having AICA depend on real-time human action may not help and in fact may cause other problems for certain cyber-defense scenarios. The vision for AICA includes their ability to respond faster than humans, or at a larger scale. Hence, human intervention may be detrimental to the autonomous defense. For example, intrusion prevention systems (IPS) may be able to autonomously avert data breaches in a fraction of a second. However, this is not the case if the IPS is part of a semi-automated workflow when human operators are required to review alerts or approve blocking of requests and addresses. Moreover, even a well-trained and alert human operator may slow down defenses against large scale attacks that target several points of the system simultaneously.

Another problem is that a human taking over during an attack (after AICA initiated maneuvers) may not have the level of situational awareness required for adequate defense and ruin it (Kott et al., 2014). Consider the related and perhaps familiar context of autonomous driving described in (Ligo et al., 2021a). The Society of Automotive Engineers defines a five-level scale of vehicle automation (Automated Vehicles 3.0 – Preparing for the Future of Transportation, 2018). In all but level 5, a human driver is expected to take control over the machine during an emergency. Consider the scenario of a self-driving car in level 3 or 4, thus having a human driver in stand-by, when a child runs into the street from between parked cars. If the human tries to retake control to swerve and miss the child in its path the vehicle could override the driver. If the vehicle senses the child and begins a collision-avoidance maneuver, then any human operator action may ruin the automated system's plan for avoiding a collision, or the person's reaction time may be dangerously longer than the time taken by the machine. If neither the human nor the vehicle does anything, there will be a dead child and liability for all involved. As long as the probability of error by the vehicle is sufficiently low, the best course of action is that the human driver does **not** interfere with the autonomous operation after the collision-avoidance maneuver starts.

The car example has similarities with autonomous cyber-defenses teaming with humans. For example, both types of systems require quick and accurate decision making. If machine action (either assisted by human or not) is not effective, negative consequences from AICA may follow. However, there is a key difference between driving and AICA action. Autonomous cars are designed to replace a human ability – driving. Therefore, a trained human driver can usually take the wheel and achieve currently acceptable driving performance if enough time is available. On the other hand, autonomous cyber-defenses may need to perform "superhuman" defenses with respect to response times, volume of data processed, or scale

of response. These attributes of AICA make it impossible for humans to intervene appropriately.

Therefore, humans should avoid interfering with the operation of AICA after they determine and start a course of action. This is especially true in situations when there is not enough time for the human to acquire situation awareness, decide, and respond.

Should we *never* have human-in-the-loop in real time? If humans should not interfere with AICA when a planned course of action is underway, are there any exceptions? There is no single solution that satisfies every situation. If AICA take risky or harmful action but the human alternative is not safer nor less damaging, then there is no value in overriding the AICA. However, in practice evaluating which action is preferred – machine or human – is not straightforward. Perhaps there is no time to evaluate because a cyberattack is already underway, or there is not enough information, or the AICA course of action is not entirely explainable. In these situations, it is not entirely clear when humans should override AICA, it at all.

5 Human-Centric Approaches with Data-Driven Intervention

With unknown risks and challenges of determining human-machine cooperation during cyber-defense operation, it is beneficial to consider some form of "offline" cooperation, or ways in which modelers can shape AICA behavior *before* agents are deployed. There are at least a few general approaches for such offline intervention. One is related to the data engineering processes involved in training machine learning models.

Machine learning algorithms are often categorized with three general types: unsupervised, supervised, and reinforcement learning. Unsupervised machine learning refers to the type of algorithms that identify patterns in data. For example, unsupervised learning algorithms such as *k-means* clustering could be applied to historical data from cyberattacks to learn classes of malware with respect to their signatures, impacts or other features that might be present in the data. This might be useful when AICA respond uniquely to different types of malware.

In contrast, supervised machine learning is a type of algorithms that depend on previously labeled data that represent an outcome of interest. These labels are often provided by humans to enable the algorithm to train a model that represents the relationship between features in the data and the outcome. For example, in email spam detection features may include the relative frequency of upper-case letters, number, symbols or other clues that distinct spam from legitimate messages. In this example the outcome is whether a given message is spam. The goal of the algorithm is to use labeled data (i.e., messages that were previously classified as spam or not spam, typically with human assistance) to fit the model's parameters (Goodfellow et al., 2020).

In general, the majority of AI algorithms is based on supervised learning. This prevalence is likely to be true in AICA as well, as supervised learning algorithms are building blocks of cybersecurity and autonomy to monitor user activity and traffic to detect malware and attacks. This is a major opportunity for humans to shape AICA behavior and mitigate their specific risks. Data scientists and engineers provide labeled training samples that ideally represent the population of individuals, or in our case, cyber-events of interest.

However, this opportunity is highly dependent on the availability of labeled data that is representative of the future scenarios of AICA action. Quality data is scarce or expensive. For AICA-induced risks of functional, safety and security consequences, it is probably unknown the exact extent to which insufficient data increase such risks. Moreover, regarding AICA-specific risks of ethical, moral or unfair consequences, there is a growing body of literature on how biased data can lead to algorithmic bias, or ML models that produce outcomes that are racist or otherwise exacerbate inequality (Ligo et al., 2021b; Linkov et al., 2020; Vincent, 2018). These biases are again caused by labelled examples that are insufficient or not representative.

Another challenge is measuring how much of AICA-specific risk is mitigated with improved labeling. In today's systems, the influence of labeled data on performance of machine learning models is assessed and the data updated on a regular basis. For example, an intrusion detection system may include a supervised learning algorithm trained with historical data from attacks. The trained model will probably have high classification precision – able to detect most of the intrusions with a small number of legitimate users flagged as malware (false positives). However, it is not uncommon for the precision of these classification systems to fade over time. This is because malware and attack characteristics change over time, as do legitimate applications, causing the number of misclassifications to increase over time (false positives of legitimate use being classified as intrusion and false negatives of attacks being classified as normal use). As AICA become more autonomous, it is likely that an increasingly greater number of more sophisticated supervised learning models will be deployed. This will imply that AICA will require more and more up-to-date labeled data to re-train the algorithms more frequently than today's spam or fraud detection systems. Research will be needed to fully understand how much new labeled data and at what frequency will be needed by AICA, and how much risk mitigation can be achieved per byte of fresh data.

Besides, no amount of labeled data can account for "black swans." Taleb defines those as events that are so unlikely and impactful that they are impossible to predict (Taleb, 2007) – think about 9/11. People naturally collects lots of data and derive conclusions after black swans occur, but their unique nature prevent the use of data about past black swans to accurately predict the next one. For example, the emergence of the Internet is a life-changing but singular data point – knowing its history does not allow to predict when the next life-changing technology will emerge nor what its impact will be. Likewise, AICA based on supervised machine learning is good only to defend against attacks that are similar to previous ones.

6 Human-Centric Approaches Based on Algorithm Design

Because of the challenges mentioned in the previous section, human intervention should go beyond providing labeled data for supervised machine learning. A second and perhaps more direct approach for human control of AICA relates to resilience by design (Kott et al., 2021) and refers to the choice and development of machine learning algorithms themselves.

One possibility is reinforcement learning (Sutton & Barto, 2020), which is a promising choice for AICA algorithms (Cam, 2020). Reinforcement learning (RL) is the category of machine learning algorithms that interacts with an environment or simulation in a recurrent way. This typically involves models that perform a series of tasks over time while managing a balance between long and short-term objectives. In this way, an RL-based AICA can try a certain course of action and measure the outcome based on how well the objectives were met and relative to how "important" those objectives are thought to be. If the outcome contributes to a long-term goal (for example, avert a cyberattack or restore service), then a relatively strong reward input is fed back to algorithm as a signal that the current course of action should be kept. On the other hand, if the outcome does not contribute to the long-term goal (e.g., there is no significant restoration), then the reward is relatively lower or negative to signal that the algorithm needs to change its course of action. As a cyber-defense example, consider a combat scenario in which AICA are deployed to defend a series of targets. The long-term and highest rewarded goals of such a scenario would be for all targets to remain intact as well as for the mission to be carried by the targets as planned. Additional goals may be set for desirable, but less important, outcomes (e.g., minimizing resource use) and when met, can be marginally rewarded.

Cam provides a conceptual model that is applicable to AICA, in which RL is used to predict actions from attackers and enable agents to counterattack appropriately (Cam, 2020). However, the proposed model does not include mitigation of specific AICA-related risks. Nevertheless, RL opens the possibility for design choices in which the optimization of the long-term goal of the algorithm could include the minimization of AICA risks. If probability or consequence of these risks can be measured over time, then they can be incorporated into the reward of the RL algorithm to minimize long term AICA risk over time.

One possible design choice might be to have an RL-based agent to control AICA-specific risk as a separate agent from the AICA themselves. In other words, this would be a design of agents controlling other agents – AICA performing the main cyber-defense mission and coexisting with other agents specialized in monitoring AICA courses of action, estimating risk, and acting either to change AICA operation or to remedy whatever damage the AICA cause. This hypothetical architecture highlights another data challenge. Data about rare cybersecurity events is… well, rare. Data from autonomous agents that allow the inference of the incremental risk and negative impact of the agent's actions should be even scarcer. Furthermore, to

our knowledge human-labeled data of actions, risks, and negative impacts of agents is probably non-existent.

Another possibility of human intervention with AICA to mitigate risk through algorithmic design could be inspired by generative adversarial networks (GAN). Conceptually, a GAN is a pair of "competing" machine learning algorithms (Goodfellow, 2016). One is a "generative" neural network that is trained to determine its parameters to approximate an unknown distribution of examples that are fed to the generative system; it then generates synthetic examples that are as similar as possible to the original data. The other algorithm is a "discriminative" neural network that is trained to classify whether given examples come from the original data or are synthetic examples output by the generative algorithm. The result of the classification by the discriminative system are fed back to the generative algorithm for improvement. The two networks are then trained simultaneously. Ideally, the networks interact until the generative model outputs examples for which the discriminative network would assign the same probability as for real examples, meaning that the discriminative network can no longer differentiate the output from the generative network from the original data.

We hypothesize that algorithms like GANs could be used for incremental risk mitigation. Data would be provided from a set of possible AICA courses of action that result in acceptable functional, safety, security, ethical, and moral consequences. AICA would play the role of the generative part of this GAN-like system, meaning that AICA would approximate acceptable courses of action as close as possible. On the other hand, a discriminative algorithm would be fed both the data on acceptable actions and data about AICA actions and try to discriminate the origin of the fed data. The output of the discriminative algorithm would be fed back to AICA in order to re-adjust their actions and make them as similar as possible to the acceptable courses of action (Ligo et al., 2021a). Once again, this concept implies a data challenge. Available data on acceptable actions needs to be collected and curated (probably by humans) to be fed both the AICA and the discriminative algorithms.

7 Simulation of Strategies

Most of the strategies discussed so far for mitigation of AICA-related risks involve gathering, labeling and/or curation of data by humans at some degree. AI algorithms in cybersecurity, computer vision, natural language processing and other applications are based on deep learning algorithms, which are particularly known to demand massive amounts of data (Goodfellow et al., 2016). What is worse is that data about risks, consequences and/or acceptable courses may simply not exist. Moreover, strategies based on historical data will not work for novel threats and situations.

This limitation in data urges the exploration of other opportunities. One general way to manage risk is to anticipate outcomes by simulation. AICA-specific risks could then be inferred by a simulation of outcomes that are synthetized and labeled

for supervised learning algorithms of AICA. There are advantages and disadvantages with this approach. One advantage is that while real data is limited to historical events that were recorded and labeled, synthetic data is limited only by human imagination – new attacks, disasters or accidents can be conceived and simulated. Disadvantages of simulated data include simulation models that are simplistic or unrealistic representations of systems or attackers. For example, simulations of cyber events can be simple tabletop exercises where the scale and complexities of the real system, attacks and AICA are not considered. These exercises are useful to review human procedures, but the data resulting from the simulation may not be useful to train AICA's supervised learning models.

Another possibility is if AICA have a built-in (or have remote access to) a simulation system that estimates risks and likely outcomes of a given course of action *before* AICA triggers that action. Estimation the optimal course of action is likely to be extremely complex, as noted by Ma in the chapter about recovery planning using simulation. Therefore, it is probable that a simulation of outcomes needs to be a digital twin – a high-fidelity and probably expensive representation of both the AICA and its environment (i.e., the system being defended). In any case, the reliability of the simulated outcomes is a risk in itself – a wrong estimate of risks and outcomes would result in overconfidence about the chosen course of action, which may lead to negative consequences.

The feasibility of use of simulation with AICA depends on a trade-off between fidelity, scenario complexity, and computational cost. A "physical" example of the advantages and disadvantages of simulation for risk assessment and reduction is demonstrated in the Operational Analysis community through the use of simulation software, such as the Advanced Framework for Simulation, Integration, and Modeling, or AFSIM (Clive et al., 2015; Dozier, 2021). AFSIM is a framework that can be leveraged to develop and visualize either high or low fidelity combat engagements (Fig. 16.1). For example, in a simulation an air combat platform can be represented simply as a point in space traveling along a vector or as a

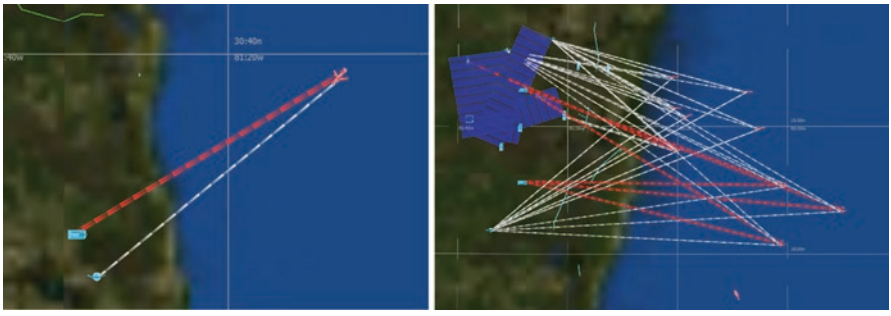

Fig. 16.1 An example of a simple (left) and more complex (right) simulation within AFSIM involving air and ground units. The computational expense of the simulation in the right figure is much higher due to many factors including the number of platforms in the engagement, missile tracking, communications between platforms, radar, and routing

six-degree-of-freedom (6-DOF) model with the ability to realistically change speed, altitude, and direction using the AFSIM physics engine. With high fidelity models, AFSIM users are able to gain an accurate assessment of the success of the simulated mission, but this level of fidelity comes with a high computational cost. The expense of complex, high fidelity simulations and models prohibits the use of simulation in real-time, and therefore limits "on the spot" engagement outcome evaluation. Therefore, when simulation results are required quickly, lower fidelity simulations with less accurate outcomes must be utilized.

8 Software-Centric Strategies: Constraints to AICA Algorithms

We have discussed the use of several types of machine learning algorithms as ways for humans to intervene with AICA at design time, aiming to mitigate the risk of negative consequences arising from AICA actions. Another form of human intervention through algorithmic design might involve an explicit design of constraints. The obvious analogy is Asimov's three laws of robotics (Asimov, 2004): (1) robots may not injure humans; (2) robots must obey orders given by humans unless they violate (1); and (3) robots must protect themselves unless the protection violates (1) or (2). The analogy might look silly when considering the complexity of AI systems. However, it is illustrative of the use of rules explicitly defined by humans, as opposed to rules learned by AICA and derived from the data available, sometimes in a non-explainable way.

Constraints are imposed at the design phase in such a way that if the behavior learned from data by the AICA violates those rules, then the agent's course of action is aborted or reversed. For example, the AICA might learn to shut down an oil pipeline in the event of unauthorized access. But what if that line is critical for heating to a certain community in a cold day? An explicit rule could cancel or remedy the action executed by AICA.

While the idea of constraints may look simple, rule-based programming can be challenging and has limitations. Defining rules for every single condition is impractical for certain applications. Consider a search engine, for example. If one implemented it exclusively with rules like "if the search term is X then return Y", they would need to code an "if" statement for every possible search term. This is impractical to code and maintain because the number of "if" statements would be in the order of billions, if not trillions (it is estimated that Google processes 1.2 trillion searchers per year) (Internet Live Stats, 2022). Nevertheless, it may be possible to design generic case-based rules or principles that can be coded to limit the degrees of freedom for the courses of action, preventing AICA to learn or execute actions that violate pre-defined functional, safety, security, ethical, or moral limits for the outcomes. Of course, no rule is able to avoid outcomes that are unknown, but this problem is present with any of the other approaches as well.

9 Summary

In this chapter we discussed how AICA may introduce new risks. These risks might overshadow the cyber-defense improvement brought by the intelligent agents. Types of new risks include flawed AICA actions caused by faulty algorithms or training data that is biased or tampered, or flaws arising from collective AICA behavior that is not observed from individual agents. These AICA-introduced risks may produce harmful outcomes of functional, safety, security, ethical, moral or equity nature. Such consequences demand mitigation strategies that prevent AICA risks to surpass their benefits.

An intuitive approach is to consider human cooperation and oversight of autonomous agents. However, human intervention in real-time during AICA action or operation is not recommended in some situations because it may make the harm worse. This includes situations in which humans cannot respond within the time or scale required to absorb or recover from the attack or disaster, or when humans cannot acquire the situational awareness required for the action.

There are options of human-centered strategies that allow humans to shape AICA behavior before they choose and execute a course of action. One option is to provide labeled data for the training of supervised learning algorithms of AICA that mitigates risk. One challenge is to determine the amount of training data required to mitigate risk, or even gather historical data that is representative of cyber-defense scenarios that are relevant for AICA training. Another challenge is how to measure risk mitigation itself, including the determination of how frequently the assessment of AICA-related risks should be executed. Finally, training AICA exclusively on historical data restricts their behavior to what has already happened in the past and is of no help to mitigate risks that are totally new.

A second strategy is to focus on the choice and design of machine learning algorithms such as reinforcement learning and generative adversarial networks applied to AICA. Again, one likely challenge of this approach is the availability of data about risks and outcomes of each algorithm. Simulation might be possible approach to overcome the data limitations of both strategies above, as it may be able to help estimate risks (historical or not) and possible mitigation strategies before AICA perform any action on production systems. However, simulation approaches must consider the trade-off between fidelity and computational cost of simulation scenarios.

A third strategy is to focus on general algorithmic rules or principles that constrain AICA actions (e.g. Asimov rules of robotics), regardless of ML training. This could leverage the power of AI and machine learning while minimizing risks by explicitly constraining the space of possible outcomes.

AICA represent a necessary, and perhaps unique, response to cyber threats that have been increasing in frequency, scale, and autonomy. Therefore, AICA-related risks should not be an obstacle to their deployment. Rather, effective risk mitigation strategies must be developed and implemented such as the benefits of AICA can be fully experienced.

References

Asimov, I. (2004). *I, robot*. Random House Publishing Group.

Cam, H. (2020). Cyber resilience using autonomous agents and reinforcement learning. In T. Pham, L. Solomon, & K. Rainey (Eds.), *Artificial intelligence and machine learning for multi-domain operations applications II* (Vol. 11413, p. 35). SPIE. https://doi.org/10.1117/12.2559319

Clive, P. D., Johnson, J. A., Moss, M. J., Zeh, J. M., Birkmire, B. M., & Hodson, D. D. (2015). Advanced Framework for Simulation, Integration and Modeling (AFSIM). In *International conference on scientific computing CSC'15*.

Dozier, H. (2021). Machine-assisted Mission engineering: An exploration of reinforcement learning with the advanced framework for simulation, integration, and Modeling (AFSIM). *ERDC/ITL SR, 21*(4), 28.

Federal Trade Commission. (2022). *Equifax data breach settlement*. https://www.ftc.gov/enforcement/refunds/equifax-data-breach-settlement

Goodfellow, I. (2016). NIPS 2016 tutorial: Generative adversarial networks. In *Conference on neural information processing systems*. http://arxiv.org/abs/1701.00160

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. *MIT Press.*. https://www.deeplearningbook.org

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144. https://doi.org/10.1145/3422622

Hubbard, D. W., & Seiersen, R. (2016). *How to measure anything in cybersecurity risk*. Wiley.

Internet Live Stats. (2022). *Google search statistics*. https://www.internetlivestats.com/google-search-statistics/

Kott, A., & Linkov, I. (2019). Cyber resilience -of systems and networks. In A. Kott & I. Linkov (Eds.), *Cyber resilience of systems and networks*. Springer. https://doi.org/10.1007/978-3-319-77492-3

Kott, A., & Theron, P. (2020). Doers, not watchers: Intelligent autonomous agents are a path to cyber resilience. *IEEE Security and Privacy, 18*(3), 62–66. https://doi.org/10.1109/MSEC.2020.2983714

Kott, A., Buchler, N., & Schaefer, K. E. (2014). Kinetic and cyber. In A. Kott, C. Wang, & R. F. Erbacher (Eds.), *Cyber defense and situational awareness*. Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-11391-3_3

Kott, A., Golan, M. S., Trump, B. D., & Linkov, I. (2021). Cyber resilience: By design or by intervention? *Computer, 54*(8), 112–117. https://doi.org/10.1109/MC.2021.3082836

Ligo, A. K., Kott, A., & Linkov, I. (2021a). Autonomous Cyberdefense introduces risk: Can we manage the risk? *Computer, 54*(10), 106–110. https://doi.org/10.1109/MC.2021.3099042

Ligo, A. K., Rand, K., Bassett, J., Galaitsi, S. E., Trump, B. D., Jayabalasingham, B., Collins, T., & Linkov, I. (2021b). Comparing the emergence of technical and social sciences research in artificial intelligence. *Frontiers in Computer Science, 3*, 653235. https://doi.org/10.3389/fcomp.2021.653235

Linkov, I., Galaitsi, S., Trump, B. D., Keisler, J. M., & Kott, A. (2020). *Cybertrust: From explainable to actionable and interpretable artificial intelligence*. IEEE Computer.

Morgan, M. G. (2017). *Theory and practice in policy analysis*. Cambridge University Press.

Oltramari, A., & Kott, A. (2018). Towards a reconceptualisation of cyber risk: An empirical and ontological study. *The Journal of Information Warfare, 17*(1), 1–22. http://www.nist.gov/cyberframework/

Sutton, R. S., & Barto, A. G. (2020). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable* (Vol. 2). Random House.

U.S. Department of Transportation. (2018). *Automated Vehicles 3.0 – Preparing for the Future of Transportation*.

Vincent, J. (2018). *Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech*. The Verge. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai