INL/RPT-22-66111
Revision 0

Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data

December 2020

A report prepared for the NNSA Office of International Nuclear Security Emerging Threats and Technologies Working Group

Shannon Eggers, Char Sample

Idaho National Laboratory

INL is a U.S. Department of Energy National Laboratory
operated by Batelle Energy Alliance, LLC

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

INL/RPT-22-66111
Revision 0

Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data

A report prepared for the NNSA Office of International Nuclear Security Emerging Threats and Technologies Working Group

Shannon Eggers, Char Sample

December 2020

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
National Nuclear Security Administration
Office of International Nuclear Security
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Page intentionally left blank

EXECUTIVE SUMMARY

Artificial intelligence (AI) applications driven by machine learning (ML) are transformational technologies within the international nuclear security regime. Advancements realized by AI—faster and improved data insights, more efficient and automated processes, reductions in human error—enable nuclear security applications such as behavior analysis for insider threat mitigation, source tracking of stolen nuclear material, and facial recognition software for physical protection. In addition to the advantages, however, there are also inherent vulnerabilities and threats associated with its use and risk mitigations must be built into any AI/ML-enabled systems.

This work provides a background on AI and ML and different data types used in the field, including open-source intelligence information (OSINT) that is discoverable by AI tools and application data that are used by AI tools for decision-making and automation. Current and potential AI applications and vulnerabilities related to their use within the nuclear security regime are also discussed.

Key findings include:

i. Data harvesting of OSINT can be performed offensively by adversaries for reconnaissance and future attack development as well as defensively for proactive identification of data and system security lapses. Nuclear security regimes must apply information security standards and best practices to reduce these vulnerabilities.

ii. AI systems often use large amounts of data, whose quality are very important. Poor quality data can yield, at best, inaccurate results leading to improper decisions or, at worst, inaccurate (and dangerous) results leading to nuclear sabotage, theft of nuclear material, or personnel injury.

iii. While the use of AI within nuclear security regimes is currently limited, many new applications could be developed introducing new vulnerabilities in the regime. It is imperative that AI application development follows best practices in AI application and data security to reduce the introduction of unmitigated threats and vulnerabilities.

The principal recommendation for INS is to include information on the unique and enhanced data vulnerabilities introduced by AI/ML applications in appropriate cyber trainings, workshops, and technical exchanges.

Page intentionally left blank

ACKNOWLEDGEMENTS

This work was funded by the U.S. Department of Energy, National Nuclear Security Administration, Office of International Nuclear Security (INS) under DOE Idaho Operations Office, Contract DE-AC07-05ID14517. This report was initially published in December 2020 as part of the INS Emerging Threats and Technologies Working Group "*Artificial Intelligence—Applications to and Implications for Nuclear Security"* compendium.

Page intentionally left blank

CONTENTS

FIGURES

TABLES

ACRONYMS

AI	artificial intelligence
CISA	Cybersecurity and Infrastructure Security Agency
CUAS	counter unmanned aircraft system
ETSI	European Telecommunications Standards Institute
FIPS	Federal Information Processing Standards
ICT	information and communications technology
IEC	International Electrotechnical Commission
IIoT	industrial internet of things
ISO	International Organization for Standardization
IVMS	integrated video management system
ML	machine learning
NIST	National Institute of Standards and Technology
NMAC	nuclear materials accounting and control
OSINT	open-source intelligence information
OT	operational technology
PIDAS	perimeter intrusion detection and assessment system
PPS	physical protection system
SIEM	security information and event management
UAS	unmanned aircraft system
USB	Universal Serial Bus

Page intentionally left blank

Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data

1. INTRODUCTION

The field of AI is rapidly transforming the world with wide-ranging innovations across all industries, including international nuclear security. AI applications have the capability to advance the security of nuclear materials and facilities worldwide to help reduce nuclear proliferation and the risk of nuclear terrorism. Existing and near-term AI security solutions include technologies such as behavior monitoring for insider threat identification and enhanced security tools for information and communications technology (ICT) and operational technology (OT) environments. Mid- to long-term AI solutions include improved data-fusion applications for physical protection and nuclear materials accounting and control (NMAC) as well as advancements in unmanned, autonomous aerial and underwater vehicles used for perimeter defense.

While there are vast opportunities for use of AI within the nuclear security regime, there is also a growing potential for misuse or compromise of these tools. Adversaries may use AI for their own purposes, such as probing system vulnerabilities and collecting open-source data for development of future attacks. They may also directly attack AI applications or models as well as compromise the underlying data used to train and operate them in order to adversely affect the model behavior. Protecting the models and data against adversarial AI is paramount to ensure the applications behave and perform as designed. Some protections can be accomplished with standard information security best practices, but some threats require more sophisticated approaches.

This paper contains a brief primer on AI applications, data, and their vulnerabilities; potential current to long-term horizon AI applications for use in the international nuclear security regime; and recommendations for AI protections, including standards and best practices.

2. TECHNOLOGY BACKGROUND

2.1 Artificial Intelligence and Machine Learning

AI is a broad field concerned with making computers or machines perform tasks that would normally require human intelligence [1]. As illustrated in Figure 1, AI mimics human intelligence, including the ability to sense, reason, engage, and learn through applications such as voice recognition, natural language processing, computer vision, robotics and motion, planning and optimization, and knowledge capture [1].

Systems achieve AI capabilities by using ML algorithms, which are mathematical models that learn patterns in data for use in data analytics and decision-making. In contrast to AI, ML can only perform based on what it was trained to do; it cannot adapt the learning process itself nor can it apply the results to understand a problem. As shown in Figure 1, there are three types of machine learning models: supervised, unsupervised, and reinforcement learning. These models use training data during the learning phase to teach the model, validation data to validate and verify

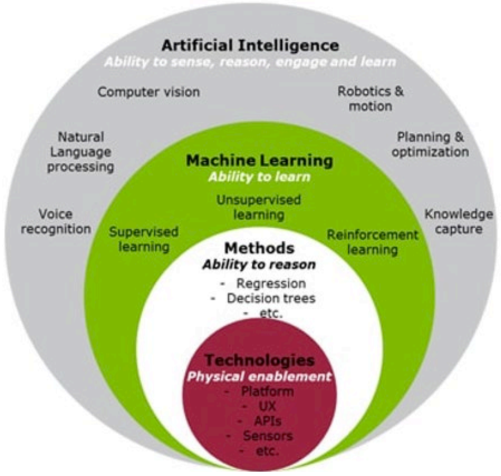

Figure 1. Taxonomy of AI and ML [1].

the behavior of the model, and inference data during the operational phase for data analytics.

Training data are collected before an ML model is built; it is assumed that this training data include all the phenomena the model will need to interpret while operational. The quality of training data directly impacts model performance—better data lead to better performance. For supervised ML, the training data are labeled to reflect some notion of ground truth, such as known malware signatures in antivirus software or pictures of flying drones labeled as drones. In unsupervised ML, training data are unlabeled and uncategorized, such as infrared data from sensors in perimeter intrusion detection systems. Once an ML model (supervised or unsupervised) is fully trained and deployed for operations, it will ingest new data for inference and comparison or classification. The resulting output from this mathematical comparison depends on the application and algorithm design. Table 1 provides additional examples of model categories and where they arise in applications used within the nuclear security regime.

Table 1. Real-world machine learning applications.

Application	ML Type	Training Data	Learning Process	Inference Data	Model Operation and Output
Antivirus Software	Supervised learning	Known virus signatures	Model learns to identify virus signature using training data	Computer files	As computer files are scanned, they are mathematically compared to the training set. Files are quarantined if a match to a virus signature is found.
Perimeter intrusion detection	Unsupervised learning	Sensor data captured during normal operations, including known anomalies such as weather events	Model learns to identify ‘normal’ sensor data	Real-time sensor data captured after training period	As new sensor data are acquired by the model, mathematical or statistical functions are performed to determine if the new data are normal or abnormal. If data are anomalous, the detection system will provide an alert that there may be an intruder.
Source detection by radiation detector on unmanned vehicle	Positive reinforcement learning	Simulated scenario with known fixed-position source with detector mounted on vehicle	If detector moves closer to source, the model is rewarded, else the model is punished	Real-time radiation in environment as sensed by detector	As the vehicle moves, the algorithm learns so the vehicle continuously moves closer to the radiation source until found.

Contrary to supervised and unsupervised learning, reinforcement learning models are developed using a reward/punishment rule set in a trial-and-error approach aimed at maximizing the reward. Instead of just training on data, the model is trained by receiving a reward after correct behavior or a punishment after incorrect behavior. Over time, the model learns the behavior required to maximize the reward.

2.2 DATA TYPES

Data used in AI systems can be categorized as endpoint, communication, configuration, monitoring, or meta data as described in Table 2. Data can also be categorized as OSINT and application data.

Table 2. AI data type descriptions [2].

Data Type	Description
Endpoint	Operational or security-related information generated by ICT, OT, or industrial internet of things (IIoT) edge devices, such as sensors, programmable logic controllers, cameras, or computers
Communication	Generated as part of the network transmission process
Configuration	Settings in ICT, OT, or IIoT devices
Monitoring	Generated during monitoring activities, such as system logs, alerts, and indications
Metadata	Describes other data

2.2.1 Open-Source Intelligence

OSINT is sensitive information about an organization, facility, or system that can be used by an adversary for opportunistic activities (see Table 3), including development of future attacks. The acquisition of OSINT datasets and information are adversarial techniques identified in MITRE's ATT&CK framework, a knowledge base of adversary tactics and techniques based on real-world observations (IDs T1247, T1266, T1277) [3]. In fact, the Department of Homeland Security's Cybersecurity and Infrastructure Security Agency (CISA) recently published an alert warning that they have consistently observed Chinese threat actors use publicly available, open-source data to plan and execute cyber operations [4]. OSINT is also a technique used in national defense [5].

Table 3. OSINT examples and adversarial misuse cases.

OSINT Example	Potential Adversarial Misuse
Facility layout and hardware, software, and firmware design information for digital systems and ICT or OT architectures	Enables development of physical, cyber, or hybrid attacks against the facility infrastructure and systems
Type, quantity, quality, and location of nuclear material or radioactive material	Enables theft of nuclear or radioactive material
Sensitive transport information, such as schedules, routes, and vehicles	Enables theft of nuclear or radioactive material
Personnel information, including phone numbers, email addresses, and work location	Identifies targets for social engineering campaigns

Examples of OSINT sources include:

- Facility, company, or organizational websites including vendor, manufacturing, contractor, and service provider websites
- Government, regulatory, or state agency public records
- Search engines such as Shodan [6]; the OSINT Framework website provides an in-depth hierarchy of free OSINT tools and resources [7]
- Curated public sources, such as wikis
- Bug reports and vulnerability data repositories [8-12]; these repositories can be automatically cross-referenced against target devices to fully elucidate weaknesses and vulnerabilities
- Social media
- News and media sources
- Scholarly publications or conferences.

Often facility owners, vendors, and regulators are unaware of the breadth and depth of sensitive information about their organizations that can be found in open data sources.

2.2.2 AI Application Data and Data Quality

The type of data used in AI applications depends on the purpose. For instance, data-driven AI/ML perimeter intrusion detection and assessment systems (PIDAS) or behavioral anomaly detection applications for insider threat mitigation may use endpoint data while AI/ML-based security information and event management (SIEM) solutions may use both communication and monitoring data. Additionally, data can be in many different formats, including text, numerical, categorical, time series, images, videos, and audio. It is also common for AI applications to use data-fusion techniques that combine multiple data types.

High-quality data are a key requirement for an effective, robust, and reliable AI application. Data must be correct, complete, and unbiased, otherwise the AI system will not behave as expected. Incorrect training data may result in inaccurate model design and/or learning, leading to unexpected behavior and confidence reduction in the model. Similar results occur if the validation data used during the development phase does not correctly represent the inference data used during the operational phase. Further, incorrect or corrupt inference data in the operational phase may skew the results over time. Bias may be a result of under- or over-sampling of data and lead to inaccurate model behavior and reduction in model confidence. To reduce bias, training and validation data must include a large enough dataset over the possible range of inputs.

2.3 Vulnerabilities in OSINT

Data harvesting is the use of AI/ML tools to programmatically search through online content to identify relevant information (i.e., OSINT) based on user inputs or requirements, such as key terminology. A quick internet search reveals that many data harvesting or data mining tools exist for OSINT, both free and for cost. Open-source frameworks and code libraries, rented hardware, and leaked versions of AI tools are available [13]. These tools may use combinations of AI/ML and natural language processing to acquire, cleanse, process, and store data. Results may be categorized and stored in a database that can be readily queried by the user.

AI/ML data harvesting tools can be used by both sides in cyber warfare. Adversaries can use the tools to discover OSINT within the international security regime, including potential vulnerabilities about their target, to assist with development of cyber campaign strategies and attacks. These tools can be automated to enable fast and up-to-date acquisition of information. In addition to adversaries using AI/ML tools for reconnaissance to identify and collect OSINT about an organization or their activities, adversaries may also use AI/ML for opportunistic cyber campaigns targeting their digital technology.

In addition to the recent CISA alert regarding Chinese use of OSINT for execution of cyber operations [4], an alert was released in 2018 documenting the use of open-source and network reconnaissance by the Russian government [14]. As identified in this report, seemingly innocuous information posted to company websites may contain sensitive data or information. It is postulated in this report that threat actors downloaded a small photo from a publicly accessible corporate website that, when enlarged, displayed control systems equipment models and status information [14]. The adversaries were able to use this information as part of a larger campaign to target the company's industrial control system. In general, acquiring OSINT datasets and information is a standard approach for target reconnaissance to help the adversary identify vulnerabilities and develop cyber campaign strategies. Once vulnerabilities are identified, separate AI/ML tools can as be used to automatically generate exploits based on those weaknesses.

Alternatively, stakeholders on the opposite side of the campaign (e.g., nuclear facility owners, government organizations, vendors) can use these same AI/ML tools to identify their vulnerabilities. Searching for OSINT about their own organization and operations is necessary so that measures can be implemented to remove or protect the data. Additionally, if sensitive information regarding their operations is found, then organizations should recognize that adversaries may have already collected this information and thus should take protective measures to limit adverse consequences from the exposure. In addition to using AI/ML tools to identify OSINT, they can also be used for vulnerability and defense automation to enforce security policies to ensure sensitive information is not accessible from external locations or by autonomous AI programs.

2.4 Vulnerabilities in AI Application Data and Models

Incorrect, incomplete, and biased data in AI/ML systems may result in unexpected model behavior and confidence reduction. While these data quality issues may be the result of improper selection, design, or just bad data, data may also be corrupted by adversaries. Malicious compromise of AI/ML data or models is intended to cause degradation or failure of AI/ML systems. A draft National Institute of Standards and Technology (NIST) report identifies 11 unique attack techniques in adversarial ML [15]. A similar set of intentional failures are provided in [16].

Direct data corruption attacks include data poisoning and data evasion attacks. Data poisoning occurs when an adversary injects or manipulates data in the training dataset. Contamination of the training dataset can skew model learning and cause misclassification or incorrect predictions during the inference phase. Similarly, imperceptible perturbations input during data evasion attacks in the operational phase can also result in misclassification or inaccurate predictions. As shown in Figure 2, the addition of a small amount of noise during the operational phase can cause misclassification by an image recognition program [17]. These evasion attacks often require knowledge of the model. In addition to data poisoning and data evasion attacks, there are many other less direct attacks that can result in model failure. Most of these attacks can be prevented by properly securing data [16].

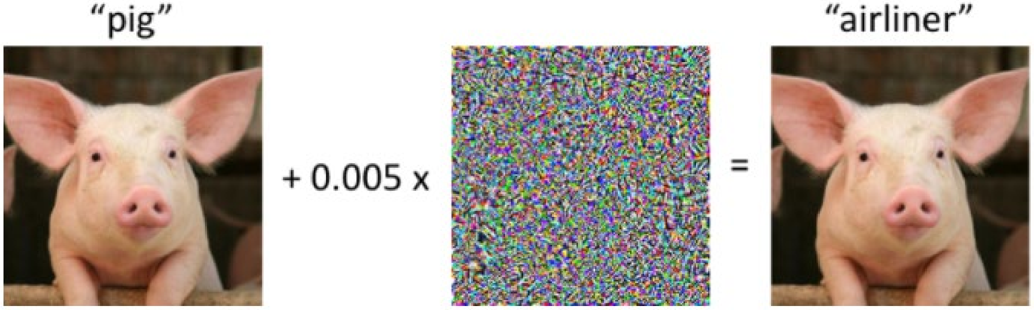

Figure 2. Example of a data evasion attack in which a small 0.005 perturbation of every pixel (every pixel is in the range [0,1]) causes misclassification by an image recognition program [17].

AI models themselves are also vulnerable to adversarial attacks [15,16]. Algorithms can be manipulated or reprogrammed, or sensitive information about the model can be leaked, enabling development of more sophisticated attacks such as data evasion and data poisoning [15]. Model tampering can lead to alteration of the learning process as well as result in the introduction of an algorithm backdoor in which the model performs correctly until a specific trigger condition is met [15,16]. For instance, a facial recognition application may perform well on most inputs but result in a planned misclassification when a specific feature is identified (e.g., unobtrusive sticker located on a face). This could result in an adversary gaining access to a facility when the application mistakenly misclassifies the image as an employee.

The risk associated with cyber attacks can be challenging to determine as it is based on characterizing the adversarial threat, vulnerability associated with the system or environment, adverse impact resulting from the attack, and likelihood that an attack will occur and be successful. While there were no studies found by the authors that rank the cyber risk specific to attacks on AI/ML data or systems, it is important to recognize that these attacks are possible today, even if their likelihood of occurring is much less than other, simpler cyber or physical attacks. In the 3- to 10-year horizon, as AI systems are deployed in greater number and as adversaries become more sophisticated, it is likely that AI-based attacks will become more probable and more frequent.

3. AI SYSTEMS AND POTENTIAL VULNERABILITIES IN THE NUCLEAR SECURITY REGIME

3.1 AI Security in Nuclear Facilities

Like traditional cyber security, AI/ML cyber security goals include maintaining confidentiality, integrity, and availability of the system. Confidentiality is assurance that sensitive or confidential data and proprietary model design are not leaked or stolen; integrity is assurance that model data are not poisoned and models or functions are not compromised; and availability is assurance that there is no interruption of model operation, no degradation of performance, or loss of function. A cyber attack is defined as any attack by any method via cyber space intended to disrupt, disable, destroy, or maliciously control digital systems or infrastructure; or to destroy the integrity of data or steal controlled information [18].

Cyber attacks in nuclear facilities occur through five threat vectors: wired, wireless, portable media (e.g., Universal Serial Bus (USB) drives, maintenance laptops), insiders, and supply chain. If a network is not properly segregated from the internet by secure architecture, an adversary may be able to penetrate the network enabling access to AI systems. Similarly, if wireless is used in a facility or wireless features are enabled on a device, an adversary can potentially gain access to an AI system. Even if wired and wireless access is prevented, digital systems can still be directly compromised during maintenance activities (configuration changes, updates) when portable media are connected to a network or device. Further, insiders such as employees or contractors may intentionally or unintentionally access and compromise a system. And finally, a system may be compromised during the acquisition process—AI software or system information can be corrupted or stolen throughout all phases of the supply chain lifecycle [19].

AI systems are potentially vulnerable to compromise via each of these five threat vectors. Regardless of how an adversary gains access to an AI system, they can launch attacks against AI data and applications. Therefore, it is important for nuclear facilities to minimize the vulnerabilities associated with these vectors. Corporate networks, plant networks, and/or control networks should be segregated with proper ICT security controls (as defined in the next section) and wireless communications should be disabled or secured. Protections against compromise via portable media include disabling or removing unused access connections or ports and establishing administrative procedures to prohibit or securely control their use. Insider mitigation programs can assist with reducing the risk from insiders, as can behavior monitoring tools (which are discussed later). Finally, activities, such as developing awareness into supply chain vulnerabilities [19], adding cyber security vendor requirements to procurement contracts [20], and enhancing software validation and verification processes can help reduce supply chain cyber risk.

3.2 AI Applications in the International Nuclear Security Regime

The following sections describe current and future AI/ML applications in the international nuclear security regime and potential consequences if the data or the applications are corrupted. In addition to data and model corruption attacks, all AI/ML applications are susceptible to data or model theft attacks. These confidentiality or privacy attacks result in loss of sensitive data that can be maliciously used by adversaries for illegal purposes or development of future sophisticated attacks.

Some of the applications outlined are not fully matured and active in the field but may potentially be developed and implemented within the next 3 to 10 years. Additionally, the likelihood of cyber attacks against these systems may be relatively low, but it is still important to recognize the threats, vulnerabilities, and impacts associated with these AI technologies to ensure implementation of appropriate defenses and protections.

3.2.1 Cyber Security

Cyber security crosscuts the entirety of the international nuclear security regime due to the use of ICT, OT, and IIoT technologies. Since any malicious compromise of data used to build, train, test, and run AI/ML models is a cyber attack, cyber security plays a key role in securing the regime against AI/ML threats and vulnerabilities.

Boundary security devices found in traditional ICT solutions within the regime, such as host intrusions detection systems, network intrusion detection systems, firewalls, antivirus software, and SIEM systems, typically use AI/ML-based technologies for anomaly detection, prediction, and prognostics. Endpoint detection and response platforms use advanced capabilities of AI/ML algorithms for malware identification, behavioral analysis, and exploit prevention to protect nuclear organizations against sophisticated and targeted attacks [21].

While not all traditional ICT security devices use AI/ML, secure architecture technology for cyber security defense, detection, and response is (or should be) used within all organizations in the international nuclear security regime. The use of AI/ML for cyber defense may expand the attack surface, as most AI/ML-based defenses have not considered adversarial attacks against them [21], thus all ICT AI/ML-based security devices used within the regime are potentially vulnerable to data poisoning and evasion attacks. These attacks may skew the AI/ML models to misclassify data, enabling adversaries to stealthily enter secure architectures and/or launch undetected cyber attacks against the network and computer systems. Attacks that enable the adversary to maintain a long-term presence result in extended loss of sensitive data and information. In addition, undetected malware and advanced persistent threats can lie in wait until triggered.

In addition to using traditional ICT tools, AI/ML tools can be used for preventing and detecting attacks within OT environments, such as digital instrumentation and control systems. Expert systems designed with data-driven and/or physics-based algorithms to detect and predict anomalies using process data or expected system behavior, respectively, provide additional defense-in-depth protections for OT environments in a nuclear facility. Similar to ICT security devices, compromise of OT security devices by data poisoning or evasion attacks may allow adversaries to gain entry and/or launch cyber attacks in OT networks without detection by AI/ML security tools. Cyber attacks that either compromise or bypass these AI systems and impact the digital instrumentation and control systems in a nuclear power plant or research and test reactor may result in radiological sabotage impacting the health and safety of the public; financial loss due to equipment damage, loss of generation, and/or repair costs; and intangible loss such as reputation or industry perception.

3.2.2 Facility Operations

Nuclear power plants and research and test reactors may also use AI/ML tools for condition monitoring—detection, prediction, and prognostics of equipment degradation and failure. In the future, big data applications that fuse historical process data, work management history, and corrective action program data may improve insight into equipment and reactor operational status. Data poisoning, evasion, or logic attacks against these AI/ML tools could result in unanticipated equipment failure or degradation.

In addition to condition monitoring, other AI/ML applications have been proposed to improve performance and efficiency at nuclear power plants. For instance, computer-based procedures that automatically prompt operators to perform the next procedure step based on current plant conditions have been proposed by Oxstrand and Le Blanc [22]. In the future, AI/ML-based technology may proceed through procedures automatically using expert systems, performing steps in sequence using autonomous control systems, with or without hold points requiring operator interaction. Attacks against these AI/ML systems could result in incorrect or out-of-order plant actions leading to an adverse or unknown reactor state.

Furthermore, it is anticipated that, in addition to advanced displays and diagnostics using AI/ML and expert systems, advanced reactors and microreactors will use autonomous, AI/ML-based control systems with limited requirements for human interaction. Similar to autonomous, computer-based procedures, attacks against AI/ML reactor control systems could result in adverse impacts to the reactor, including challenges to safety systems resulting in radiological release.

3.2.3 Insider Threat Mitigation

The objectives for insider threat mitigation are to promote best practices, understand weakness, and provide recommendations on insider threat countermeasures for the protection of nuclear materials and facilities. Insider threats to nuclear facilities include physical and cyber threats. To assist with reducing both cyber and physical threats from insiders, insider mitigation programs may use AI/ML-based behavioral recognition programs to identify suspicious employee behavior. These programs monitor employee computer-based actions, such as file browsing, file usage, file downloads, USB usage, and application/ system logins as well as physical actions such as facility entries and exits, to identify normal and abnormal behavior. Data poisoning or evasion attacks could mask behavior of an insider such that abnormal or suspicious behavior is not detected [13].

3.2.4 Detection and Response

Current AI/ML applications within physical protection systems (PPS) include facial recognition software and abnormal behavior identification. Near-term and future PPS applications include voice recognition and natural language processing applications. Additionally, PPS data-fusion algorithms can combine data from multiple PIDAS, such as integrated video management systems (IVMS), microwave systems, and infrared systems, to ensure high probability of intrusion detection while eliminating nuisance alarms. An example of a deliberate motion algorithm using sensor fusion was presented at the Light Water Reactor Sustainability Physical Security Stakeholder meeting in 2019 [24].

In NMAC systems, equipment used for monitoring nuclear material includes flowmeters, mass spectrometers, tank-level indicators, nondestructive assay equipment, scales, video surveillance, and radiation monitoring and contamination control equipment [23]. Future AI/ML-based systems can use the data from these diverse and disparate systems to detect, predict, and respond to anomalies, including unanticipated removal of material. These data may be transmitted to central reporting locations for monitoring at the facility or company level, or at the State level by entities such as the State's competent authority.

In conjunction with PPS, future tools used in the sabotage area may be able to model a facility's security defenses, vulnerabilities, and response capabilities. These tools may use AI/ML to integrate sensors (e.g., automatic identification sensors, radar, PIDAS, IVMS), vehicles, aircraft, UAS, unmanned underwater vehicles, personal devices (phones, tablets), and geographic information systems data to model a physical location's vulnerabilities based on defensive posture against simulated threat events.

While these sabotage applications may assist with vulnerability analysis, design basis threat, target set, and vital area identification to improve defensive capabilities, they may also be used as command and control platforms for event response including monitoring, anomaly detection, alerting, and response coordination between organizations (e.g., protective forces, law enforcement agencies, emergency response). Border and maritime security solutions as well as response tools may use data-fusion applications incorporating IVMS, facial recognition, natural language processing, and permanent or portable radiation detection systems. These AI systems may also incorporate global positioning and radiation detection systems on UAS and unmanned underwater vehicles for detecting, tracking, and locating stolen or diverted nuclear material [25-27]. Similarly, AI systems could be used with tracking transport of known nuclear material shipments with geographic information-based systems to identify routes to maintain compliance with best practices and with vehicle global positioning systems, UAS, and radiation detection systems (permanent and portable) to track shipments during transit as well as during theft events.

Compromise of these AI systems could have far-reaching consequences. For instance, data poisoning or evasion attacks against these PPS or NMAC intrusion detection systems could result in undetected adversarial entry from situations such as failure-to-alert due to masking of the intrusion pathway, or generation of too many alerts resulting in failure to identify the true intrusion. In addition, compromised AI/ML facial or voice recognition software could result in failure to identify an adversary or misidentification of a friendly as an adversary [13]. Further, compromise of AI/ML-based NMAC systems may result in inaccurate data for nuclear and radiologic materials and failure of the system to identify loss or removal of material.

Data poisoning or evasion attacks against intrusion vulnerability assessment modeling applications for sabotage prevention could result in adversaries impacting simulation scenarios and assessment results leading to weakened or ineffective defenses. Adversarial AI could also cause loss of control with nuclear shipments or loss of security control at border or maritime facilities enabling adversaries to move unauthorized nuclear material across (or into) States without detection. Additionally, compromise of response systems could result in loss of tracking capability and/or misdirection of response forces. Compromise of emergency and/or security response coordination applications could impact necessary communications and collaborative actions between response forces.

3.2.5 Perimeter Defense Equipment

In addition to PPS, remotely operated [28] and autonomous weapon systems as well as ground, aquatic, and CUAS aerial robots or drones have been proposed for perimeter defense in the physical protection mission space. Cyber attacks against AI/ML weapons systems could result in friendlies being injured or intruders not being stopped. Similarly, cyber attacks against AI/ML-enabled CUAS could result in masking of the algorithm such that intruders remain undetected.

4. RECOMMENDED COURSE OF ACTION

4.1 STANDARDS AND GUIDANCE

As reflected in Table 4, standards and best practices for securing OSINT and other sensitive or private data exist at both the domestic and international level. Not included in the table are a wide range of NIST standards related to cryptography which can be used to ensure data integrity. There are currently no standards or guidance documents specifically focused on data security in nuclear facilities issued from the Nuclear Regulatory Commission, Nuclear Energy Institute, International Atomic Energy Agency, CSA Group, NIST, or ISO/IEC. Since the NIST standards are free and available for use, they are often used internationally as well as domestically. At minimum, nuclear organizations should ensure they follow NIST, International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), or similar information protection standards or guidelines to ensure OSINT cannot be acquired by adversaries and that AI application data are protected against theft or compromise.

While many States have published regulations and guidance for cyber security within nuclear facilities, this guidance does not generally include information specific to AI or ML security [41-45]. Additionally, standards organizations, such as NIST, ISO/IEC, and the European Telecommunications Standards Institute (ETSI), have not yet published cyber security standards related to AI. In 2019, NIST released a plan recommending that the U.S. government commit to standards development for building trustworthy AI systems [46]. Further along than NIST, ISO/IEC has numerous AI standards in various stages of development [47]. Similarly, ETSI has an industry specification group developing standards focused on using AI to enhance security, mitigating against attacks that leverage AI, and securing AI itself from attack. States should continue to monitor these organizations to stay current in recommended AI security practices.

Table 4. OSINT-related standards.

U.S.-based Standards	**Description**
Federal Information Security Management Act [29]	Requires federal agencies to protect sensitive data.
NIST SP 800-18 [30]	Requires federal organizations to develop and implement a system security plan to protect information systems.
Federal Information Processing Standards (FIPS) 199 [31]	Establishes requirements to inventory and classify information systems by security categories.
FIPS 200 [32]	Establishes requirements to implement minimum security controls.
NIST SP 800-53 [18]	Establishes requirements to implement minimum security controls.
NIST SP 800-171 [33]	Provides guidance for protecting controlled unclassified information in nonfederal systems.
NIST SP 800-122 [34]	Provides guidance for protecting the confidentiality of personally identifiable data.
NIST SP 800-209 [35]	Provides guidance for protecting storage infrastructure.
NIST Data Security Website [36]	Provides guidance documents and practice guides for protecting, detecting, and responding to data integrity and data confidentiality attacks.
International-based Standards	**Description**
ISO 27001 [37]	Provides guidance on establishing, implementing, and maintaining information security management systems.
ISO 27002 [38]	Provides guidance on implementing security controls defined in ISO 27001.
ISO 27034 series [39]	Provides guidance on application security.
ISO 27040 [40]	Provides guidance on storage security.

NIST has issued SP 1800-7, *Situational Awareness for Electric Utilities*, which describes the use of SIEM solutions for aggregating data and performing anomaly detection on ICT and OT systems. Of note in this document is the acknowledgment that, while SIEMs are useful for identification of cyber attacks, these applications can also become an attack vector if left unprotected. Since an adversary can modify or delete SIEM data, alter analysis processes, or alter data in transit, it is necessary to provide for data control, verification, and integrity protection.

4.2 Best Practices

Data protection strategies are often designed based on the data type and state of the data—whether the data are at rest, in use, or in motion [2]. In addition, data security technologies sometimes overlap with boundary security technologies. In some cases, the technologies are independent and require integration. In cases where integration is required, such as encryption solutions, these technologies may enhance one aspect of security (privacy and integrity) at the cost of another (data verification).

The goals for AI/ML data and application protections are to ensure reliability, discretion, robustness, and resilience. However, there is not a ‘one-size-fits-all’ solution to ensuring security of AI/ML data and models, especially since there are so many variations of algorithms. For instance, supervised machine learning uses a static training set that has different vulnerabilities than the dynamic training sets generated in unsupervised ML.

Ideally, the AI/ML model should be resilient by using techniques to identify anomalous behavior and prevent manipulation outside of normal boundaries [48]. For example, AI applications developed for the nuclear security regime must be able to reject malicious training data or user input that does not meet these acceptable behaviors and would result in a negative impact on model behavior. Multiple resilience-centric security solutions exist [48,49]. In addition, Microsoft, Google, and others have suggested integrity-based security solutions for AI/ML [50-53]. Additional tools in the future will provide added validation that an application used in the nuclear security regime is trustworthy, robust, reliable, and resilient [48].

As AI systems are developed to improve and optimize facility operations, physical protection, material security, and other processes in the nuclear security regime, they must be designed with security in mind to mitigate consequences from adversarial misuse or compromise. Best practices, such as AI data and application hardening, and monitoring for indicators of compromise and anomalous behavior, should be baseline requirements.

Additionally, it is important to identify if or when a human-in-the-loop is needed to validate the system design, operation, and/or output. In these instances, it is necessary to recognize that humans may also perform malicious or inadvertent actions that compromise the system. For example, if facial recognition software flags a person as an intruder, is a human required to confirm the assessment? And, if a human is needed for validation, what are the consequences if a wrong choice by the human is made? Similarly, if an AI-enabled UAS is used to track stolen or diverted nuclear material, when are humans deployed to confirm the UAS has tracked the material correctly? What human interaction is required if the UAS mistakenly tracks an incorrect item? These hypothetical scenarios highlight the importance of including human-in-the-loop security considerations along with AI data and model protection best practices in an AI system's security strategy.

5. CONCLUSIONS

AI systems developed for cyber security, physical protection, insider mitigation, NMAC, transport, sabotage, and response have great potential for advancing global nuclear security by reducing nuclear proliferation and the risk of nuclear terrorism. However, it must be recognized that these systems are themselves susceptible to adversarial cyber attacks. AI systems must be developed and deployed following best practices in AI security in order to provide assurance that the system is secure, trustworthy, and reliable and is less likely to be successfully attacked.

It must also be recognized that adversaries can use AI systems to identify OSINT for development of future sophisticated physical and/or cyber attacks. Not only must AI data and models be protected from compromise and corruption, but this open-source sensitive data be must also be protected from adversarial discovery. Use of standards and best practices for implementing data and information system protection can help prevent access of this sensitive data by adversarial AI systems.

6. REFERENCES

[1] van Duin, S. and N. Bakhshi, "Part 1: Artificial intelligence defined: The most used terminology around AI," ed: Deloitte.

[2] "Industrial internet of things volume G4: Security framework," Industrial Internet Consortium (IIC), 2016.

[3] "MITRE ATT&CK." The MITRE Corporation. Accessed on: April 28, 2020. Available: https://attack.mitre.org/.

[4] "Alert AA20-258A: Chinese ministry of state security-affiliated cyber threat actor activity," Cybersecurity and Infrastructure Security Agency (CISA), September 2020.

[5] "ATP 2-22.9, MCRP 2-10A.3, Open-Source Intelligence," Department of the Army, United States Marine Corps, June 2017.

[6] "Shodan: The world's first search engine for internet-connected devices." Shodan. Accessed on: September 29, 2020. Available: https://www.shodan.io/.

[7] Nordine, J. "OSINT Framework." Accessed on: October 5, 2020. Available: https://osintframework.com/.

[8] "National Vulnerability Database (NVD)." National Institute of Standards and Technology. Accessed on: September 29,, 2020. Available: https://nvd.nist.gov/.

[9] "Common Weakness Enumeration (CWE)." The MITRE Corporation. Accessed on: September 29, 2020. Available: https://cwe.mitre.org/.

[10] "Common Vulnerabilities and Exposures (CVE)." The MITRE Corporation. Accessed on: September 29, 2020. Available: https://cve.mitre.org/.

[11] "Common Vulnerability Scoring System (CVSS)." FiRST Accessed on: September 29, 2020. Available: https://www.first.org/cvss/.

[12] "National Cyber Awareness System." Cybersecurity and Infrastructure Security Agency. Accessed on: September 29, 2020. Available: https://us-cert.cisa.gov/ncas.

[13] "Artificial intelligence: Using standards to mitigate risks," Analytics Exchange Program (AEP), 2018.

[14] US-CERT, "TA18-074A: Russian government cyber activity targeting energy and other critical infrastructure sectors," Revised March 16, 2018.

[15] Tabassi, E., K.J. Burns, M. Hadjimichael, A.D. Molina-Markham, and J.T. Sexton, "Draft NISTR 8269: A taxonomy and terminology of adversarial machine learning," National Institute of Standards and Technology, October 2019.

[16] Kumar, R.S.S., J. Snover, D. O'Brien, K. Albert, and S. Viljoen, "Failure modes in machine learning," Microsoft, November 2019.

[17] Madry, A. and L. Schmidt, "A brief introduction to adversarial examples," in "Gradient Science," July 2018 2018.

[18] Initiative, J.T.F.T., "SP 800-53 Revision 5. Security and Privacy Controls for Information Systems and Organizations," National Institute of Standards and Technology, 2017.

[19] Eggers, S., "A novel approach for analyzing the nuclear supply chain cyber-attack surface," Nuclear Engineering and Technology (In press, 2020).

[20] "Department of Homeland Security: Cyber Security Procurement Language for Control Systems," Department of Homeland Security, September 2009.

[21] Brundage, M. *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," arXiv preprint arXiv:1802.07228 (2018).

[22] Oxstrand, J. and K. Le Blanc, "INL/JOU-15-37135: Supporting the future nuclear workforce with computer-based procedures," Nuclear Future 12 34-39.

[23] "IAEA Nuclear Security Series No. 25-G: Use of Nuclear Material Accounting and Control for nuclear security purposes at facilities," International Atomic Energy Agency, Vienna, 2015.

[24] Osborn, D., J. Lord, and H. Werner, "SAND2020-0764, Light Water Reactor Sustainability Program: September 2019 Physical Security Stakeholder working group meeting," Sandia National Laboratories, January 2020.

[25] Aleotti, J. *et al.*, "Detection of nuclear sources by UAV teleoperation using a visuo-haptic augmented reality interface," Sensors 17 (10) (2017) 2234.

[26] Magocs, B., M. Cook, and J. Gerka, "Design of a Multi-Tier Unmanned Aerial Vehicle Search Algorithm for Locating Radioactive Sources in a Large Complex Region," in *INMM 2018 Annual Meeting*, 2018.

[27] Allard, Y. and E. Shahbazian, "Unmanned underwater vehicle (UUV) information study," OODA Technologies Inc Montreal, Quebec Canada, 2014.

[28] Hallbert, B., K. Leonard, M. Farmer, C.A. Primer, and R. Szilard, "Light Water Reactor Sustainability Program integrated program plan," Idaho National Lab, INL/EXT-11-23452-Rev006, 2019.

[29] *Federal Information Security Act (FISMA) of 2002,* 107th United States Congress, Amended 2014.

[30] Swanson, M., J. Hash, and P. Bowen, "SP 800-18 Revision 1. Guide for Developing Security Plans for Federal Information Systems," National Institute of Standards and Technology, 2006.

[31] "Federal Information Processing Standards Publication (FIPS PUB) 199. Standards for Security Categorization of Federal Information and Information Systems," Department of Commerce, 2004.

[32] "Federal Information Processing Standards Publication (FIPS PUB) 200. Minimum Security Requirements for Federal Information and Information Systems," Department of Commerce, 2006.

[33] Ross, R., V. Pillitteri, K. Dempsey, M. Riddle, and G. Guissanie, "SP 800-171 Revision 1 . Protecting Controlled Unclassified Inforamtion in Nonfederal Systems and Organizations," National Institute of Standards and Technology, 2016.

[34] McCallister, E., T. Grance, and K. Scarfone, "SP 800-122. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)," National Institute of Standards and Technology, 2010.

[35] Chandramouli, R. and D. Pinhas, "SP 800-209. Security guidelines for storage infrastructure," NIST Special Publication (2020).

[36] "Data security." National Institute of Standards and Technology. Accessed on: September 10, 2020. Available: https://www.nccoe.nist.gov/projects/building-blocks/data-security.

[37] "ISO/IEC 27001:2013, Information technology - security techniques - information security management systems - requirements," International Organization for Standardization/International Electrotechnical Commission, October 2013 2013.

[38] "ISO/IEC 27002:2013, Information technology - security techniques - Code of practice for information security controls," International Organization for Standardization/International Electrotechnical Commission, 2013.

[39] "ISO/IEC 27034 Series, Information technology - security techniques - application security," International Organization for Standardization/International Electrotechnical Commission, 2011.

[40] "ISO/IEC 27040:2015, Information technology - security techniques - storage security," International Organization for Standardization/International Electrotechnical Commission, 2015.

[41] "Regulatory Guide 5.71, Cyber security programs for nuclear facilities," U.S. Nuclear Regulatory Commission, January 2010.

[42] "N290.7-14, Cyber security for nuclear power plants and small reactor facilities," CSA Group, 2014.

[43] "IAEA NST047, Computer Security Techniques, Step 12," International Atomic Energy Agency, Vienna, 2018.

[44] "IAEA Nuclear Security Series No. 13, Nuclear Security Recommendations on Physical Protection of Nuclear Material and Nuclear Facilities (INFCIRC/225/Revision 5)," International Atomic Energy Agency, Vienna, 2011.

[45] "IAEA Nuclear Security Series No. 17, Computer Security at Nuclear Facilities," International Atomic Energy Agency, Vienna, 2011.

[46] "Artificial intelligence: AI standards." National Institute of Standards and Technology. Accessed on: August 11, 2020. Available: https://www.nist.gov/topics/artificial-intelligence/ai-standards.

[47] "Standards by ISO/IEC JTC 1/SC42: Artificial Intelligence." ISO/IEC. Accessed on: September 29, 2020. Available: https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0.

[48] Marshall, A., R. Rojas, J. Stokes, and D. Brinkman, "Securing the future of artificial intelligence and machine learning at Microsoft," vol. September 29, 2020, ed: Microsoft.

[49] "Protecting the protector: Hardening machine learning defenses against adversarial attacks," vol. September 29, 2020, ed: Microsoft Defender ATP Research Team.

[50] "Homomorphic encryption." Microsoft. Accessed on: September 18, 2020. Available: https://www.microsoft.com/en-us/research/project/homomorphic-encryption/.

[51] Sinha, R., S. Gaddam, and R. Kumaresan, "LucidiTEE: A TEE-Blockchain System for Policy-Compliant Multiparty Computation with Fairness."

[52] McMahan, B. and D. Ramage, "Federated learning: Collaborative machine learning without centralized data," ed: Google AI Blog.

[53] Guevara, M. "Google developers blog: Enabling developers and organizations to use differential privacy." Accessed on: September 29, 2020. Available: https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html.