INS International Nuclear Security
Reducing Risk of Nuclear Terrorism

Responsible Artificial Intelligence (AI) for Insider Threat

September 2025

Office of International Nuclear Security

National Nuclear Security Administration

PNNL-38436

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

RESPONSIBLE ARTIFICIAL INTELLIGENCE FOR INSIDER THREAT

Pacific Northwest National Laboratory

September 2025

Prepared by
Jessica Baweja
Jonathan Barr
Chantell Murphy

Pacific Northwest National Laboratory
Richland, Washington 99354

CONTENTS

TABLES

ABBREVIATIONS, ACRONYMS, AND INITIALISMS

AI	Artificial Intelligence
DIVA	Deep Intermodal Video Analytics
HR	Human Resources
IAEA	International Atomic Energy Agency
IARPA	Intelligence Advanced Research Projects Activity
IT	Information Technology
ITM	Insider Threat Mitigation
LLM	Large Language Models
ML	Machine Learning
NIST	National Institute of Standards and Technology
NMAC	Nuclear Material Accounting and Control
PNNL	Pacific Northwest National Laboratory
SALT	Scalable and Automated Linking Technology
SVM	Support Vector Machines
TSA	Transportation Security Administration
UAS	Unmanned Aerial Systems

EXECUTIVE SUMMARY

This report examines the application of artificial intelligence (AI) technologies for insider threat mitigation (ITM) programs in nuclear security facilities. Insider threat detection presents unique challenges due to the subtle and adaptive nature of these threats, the complex signatures involved, and the scarcity of available data for analysis. Traditional human-centered approaches, while essential, face limitations in processing large amounts of data continuously and detecting subtle patterns across multiple systems. AI technologies can potentially address these limitations by providing 24/7 monitoring capabilities, identifying complex patterns that might escape human observation, and offering consistent application of security criteria. However, the deployment of AI in nuclear security contexts introduces significant new risks, including workflow disruption, expanded attack surfaces, potential for misuse, and ethical concerns regarding privacy, fairness, transparency, safety, and security. The high-consequence nature of nuclear security decisions demands careful consideration of these risks and systematic approaches to their mitigation.

This analysis applies the US National Institute of Standards and Technology (NIST) AI Risk Management Framework to evaluate AI applications for ITM. The framework's core principles—accountability, fairness and bias mitigation, privacy and data protection, security and abuse prevention, explainability and interpretability, safety, and performance—provide an inclusive lens for assessing potential risks and developing mitigation strategies. The research involved four primary phases: literature review and capability mapping, taxonomy development, AI risk framework application, and implementation guidance development. The analysis identifies six functional categories of AI applications for nuclear security ITM:

1. **Identity and Document Verification Systems:** These systems authenticate individuals and validate documents by comparing credentials, biometric features, or documents against trusted records. Applications include facial recognition for identity verification, document-based identity verification, and fraud detection for document verification. These technologies serve as verification tools during background checks, visitor processing, or personnel onboarding while maintaining human-in-the-loop decision-making.
2. **Trustworthiness Systems:** These systems assess personnel risk levels by evaluating patterns across historical records and identifying potential concerns. They aggregate information from multiple sources to generate comprehensive risk assessments and flag concerning patterns requiring human investigation. Applications include automated risk scoring from documents and automated issue identification in background records.
3. **Behavior Observation Systems:** These systems detect anomalous activities that might indicate insider threats by establishing baselines of normal behavior and flagging significant deviations. They analyze patterns in physical movement, facility access, network activity, and digital communications. Applications include fitness-for-duty monitoring, video analytics systems, and cyber behavior analysis.
4. **Impairment Detection Systems:** These anomaly detection systems identify potential fitness-for-duty concerns by analyzing physiological and behavioral indicators of fatigue, intoxication, or other impairment. Applications include fatigue detection through multi-modal inputs and drug/alcohol impairment detection through behavioral pattern analysis.
5. **Access Control and Patrolling Systems:** These systems secure facility boundaries by authenticating personnel, screening for prohibited items, and conducting autonomous security monitoring—useful in multiple security applications, including detection of insider threats. They make real-time decisions that directly impact facility security, often with limited human intervention in the immediate decision loop. Applications include biometric access controls, automated screening at security checkpoints, and autonomous security patrols.

6. Nuclear Material Accounting and Control Systems: These systems detect deviations in nuclear materials accounting data that may indicate theft or diversion. Applications include material movement characterization, transaction anomaly detection, and synthetic data generation for training purposes.

Organizations must carefully assess their readiness across multiple dimensions before implementing AI for ITM. Technical readiness requires appropriate data quality, information technology (IT) infrastructure, and organizational capacity to integrate AI systems without creating vulnerabilities. Organizational readiness encompasses policy frameworks, personnel training, legal compliance, and coordination across multiple organizational functions. All AI applications for ITM should be implemented as decision support tools that enhance human capabilities rather than replace human judgment. The high-consequence nature of nuclear security decisions, combined with current AI limitations and sparse insider threat data, makes human oversight essential for responsible implementation.

Organizations should develop roadmaps that deploy AI applications based on organizational readiness and security benefits rather than implementing systems in isolation. Identity and document verification systems, for example, build upon established commercial technologies with clear implementation pathways. Behavior observation systems require more sophisticated infrastructure and extensive policy development.

This framework provides a foundation for risk assessment but cannot address all potential concerns for every organizational context. The rapid pace of AI advancement means this analysis represents a snapshot of current capabilities that will require regular updates. Organizations must conduct their own thorough assessments of how AI applications interact with their specific operational procedures, security measures, and regulatory requirements. The commercial AI landscape is highly dynamic, and the specific examples referenced should be understood as illustrative rather than endorsements. Organizations need specialized technical, legal, and policy expertise throughout implementation processes and should work closely with regulatory bodies to ensure compliance.

AI applications offer significant potential to enhance nuclear security ITM programs through capabilities that extend beyond traditional approaches. However, successful implementation requires careful planning, systematic risk management, and ongoing commitment to monitoring and adaptation. Organizations that invest in developing strong foundational capabilities for AI risk management will be better positioned to realize these benefits while maintaining the highest standards of nuclear security. The framework presented here provides a structured approach to evaluating and implementing AI technologies for ITM while maintaining appropriate human oversight and addressing inherent ethical and operational challenges. As AI technologies continue to evolve, the procedures and frameworks established today must be regularly updated to reflect changing capabilities, threats, and regulatory requirements. Most critically, responsible AI implementation for ITM is not a one-time effort but an ongoing process requiring continuous evaluation, adaptation, and unwavering commitment to safety and security principles in the nuclear domain.

CHAPTER 1

1. INTRODUCTION

Insider threat detection and mitigation remains a pressing challenge in nuclear security, given the complex signatures involved, the paucity of data available, and the relatively subtle and adaptive nature of insider threats. Artificial intelligence (AI) enabled technologies provide a promising new method for combatting detection of insider threats, potentially allowing for more rapid intervention and protective action. Recent work has explored the use of AI for insider threat in a variety of different ways, including through anomaly detection, natural language processing supported by large language models (LLMs), graph-based approaches, and user behavior analytics [1-5].

The introduction of AI technologies into the nuclear security enterprise, however, also introduces additional risks in the form of workflow disruption, broadening of the attack surface, and increased potential for misuse [6-11]. Furthermore, the application of AI in the high-consequence domain of nuclear security either creates or exacerbates ethical concerns regarding privacy, fairness, transparency, safety, and security of AI enabled systems [12-14]. These concerns have led to a variety of efforts by the United States Government to introduce best practices and guidance for the use of AI to help ensure that it is deployed in a secure, effective, and responsible manner [9, 15].

In this report, we present a review of the current and potential future uses of AI technology for insider threat mitigation (ITM) programs in nuclear security and develop a framework to address the risks to the organization when applying AI to ITM. We apply a broad definition of AI, including any system that can generate outputs such as predictions, recommendations, or decisions at various levels of autonomy [9]. This includes expert- or rule-based systems, machine learning (ML) systems, and more recent AI systems based on deep learning technologies, such as LLMs. We focus on AI systems that could be used in support of ITM through preventive, protective, or comprehensive measures as described by the International Atomic Energy Agency (IAEA), rather than through more general administrative support (e.g., automation of incident reports using generative AI) [16]. Using the information gathered in this review, we leverage existing knowledge on safety, security, and ethical risks of AI technologies to evaluate the risks presented by applying AI to nuclear security ITM programs. Finally, we discuss mitigation approaches for managing the risks associated with AI deployment for ITM programs.

1.1 BACKGROUND

Prevention and detection of insider threats is a complex and multidisciplinary problem, and research into more effective mitigation measures has now spanned decades [17]. Despite dedicated research, insider threat remains a critical challenge when seeking to protect nuclear facilities and material from the risk of sabotage or other threats. It remains true that, “virtually all the cases of nuclear theft in which the circumstances are known were perpetrated either by insiders or with the help of insiders,” underscoring the need for effective approaches to managing the risk of insider threat [18]. Much of the challenge results from the fact that insiders have authorized access and, in many cases, detailed knowledge of facilities and security measures that increases the likelihood of attack success [16-18]. This can make detection, delay, and response to those attacks more challenging. Indications of potential insider attacks may be subtle, spread across multiple systems, and difficult for humans (which may include physical security specialists, cybersecurity professionals, human resources personnel, and supervisors) to identify and appropriately respond [19-21].

It is no surprise, then, that researchers have explored the use of "big data," ML, and AI approaches to help predict or detect insider threats [1-5, 22]. These technologies have key advantages when addressing the challenge of insider threat: they can process vast amounts of data, detect subtle and complex patterns—even across sensors—and they can monitor systems and facilities 24/7 without fatigue. Along with these advantages, however, come significant challenges that need to be addressed.

First, it is critical that systems applied in high-consequence domains, such as the protection of nuclear material and facilities, perform effectively. AI system failure in nuclear security could have devastating impacts on employee and public safety and security, making it critical to ensure that the systems perform well at their intended functions. However, evaluation of AI system performance is an active area of research and remains challenging. This is partially due to a need for clear regulation and standards regarding the required performance of a system for high-consequence use [12, 23, 24]. System performance evaluation is especially difficult in the domain of insider threat, where insider threat training data are extremely sparse due to the small number of known incidents. As a result, practical evaluation of AI systems for insider threat detection remains extremely challenging, forcing the use of proxy metrics that may not generalize to real-world performance [1].

Recent collaboration between the Canadian Nuclear Safety Commission, the UK Office for Nuclear Regulation, and the US Nuclear Regulatory Commission has explored many of these considerations, and work is emerging to outline the regulations and standards needed to deploy AI systems in nuclear applications [23, 24]. Researchers have also explored "affirmative safety" approaches, which would ask developers to take proactive steps to demonstrate the performance and safety of their systems and outlining organizational practices to support safety [12]. While the standards and approaches are developing, organizational methods to assess the performance and safety of AI systems in a systematic, proactive, and holistic way—and identify failure modes and backup systems for those systems—is a critical component of deploying them in a responsible and effective manner.

Second, maintaining engaged human oversight of the AI system, and accountability for its outcomes, is key to ensuring that humans remain responsible for the security of the facility. There is a large body of evidence that humans may over-rely on AI systems, trusting or using system outputs even when they may not be accurate [25-28]. Relatedly, many AI systems—especially deep learning systems—lack transparency, meaning that the recommendations or decisions made by the system may not be readily explainable to a human user [11, 29-33]. The lack of sufficient transparency or explainability makes it difficult for users to understand when a system should—or should not—be relied upon. As a result, when deploying AI systems for nuclear security, it must be clear to end users how the system results should be used, how they should be verified, and when they cannot or should not be relied upon. Because the consequences of AI systems being applied to insider threat are very high (either security lapses in the case of false negatives or reputational and employment harms for false positives), human understanding of AI system outputs is paramount. Operators must, at every stage, be capable of explaining why any decisions that affect human safety, security, or livelihoods were made.

AI systems also introduce risks to privacy and security that need to be considered. The aggregation of data—especially personally identifying information on applicants or employees—for AI system training or analysis may introduce or exacerbate concerns regarding privacy [34, 35]. In addition, AI systems may inadvertently release private information through a phenomenon known as training data leakage, whereby entire segments of training data are regurgitated in system outputs [36]. Depending on the country where the system is deployed, there are complex regulations that may need to be navigated regarding the protection of private information that need to be considered when deploying AI systems for nuclear security. AI systems also expand the attack surface, introducing new threats that need to be considered, such as the threat of adversarial AI [37, 38].

Finally, AI systems may have problems with bias—both in terms of demographic characteristics when applied to people, and, in more technical terms, whereby the model demonstrates variable performance across different classes, features, or in different contexts [39, 40]. Again, the sparsity of data and signatures for insider threats may make this problem especially concerning for AI systems for ITM. Without representative, relevant, accurate, and generalizable data, issues of data or demographic bias may emerge when these systems are deployed, resulting in inaccurate, unfair, or inconsistent performance. Organizations must, therefore, consider the quality of the data on which the models were trained and evaluate the performance of the system under variable conditions to ensure that it meets reasonable standards for fair, unbiased, and robust performance.

Critically, regardless of where they are applied, AI systems should not be used for prediction of insider threat. Given the rarity of the event, the sparsity of the data, and the severity of false positives, prediction is not only unlikely to be accurate, it is also highly likely to produce damaging errors. AI systems can and should serve only as providers of information for professional evaluators (e.g., adjudicators) to make determinations regarding trustworthiness. Especially given the "black box" nature of many AI approaches, this is a critical boundary for responsible AI implementation in insider threat.

1.2 CURRENT STUDY

All of these issues emphasize the need for a framework to help organizations manage the risks of AI systems as they seek to realize their capabilities for ITM. In this study, we seek to address this need by conducting a review of the ways that AI has been or could be applied to insider threat and creating a taxonomy of those applications. Then, we apply principles from the US National Institute of Standards and Technology (NIST) AI Risk Management Framework to understand the risks of that application of AI and to identify organizational processes to mitigate them [9]. We apply this framework because a review of existing ethical frameworks and approaches suggests that it represents an emerging consensus regarding the ethical principles most relevant to effective, reliable, and trustworthy use of AI technologies: that they perform well at their tasks (i.e., valid and reliable), and are safe, secure, privacy-enhanced, fair, accountable, and explainable.

CHAPTER 2

2. METHODOLOGY

The methods applied in this work consisted of four primary phases: literature review and capability mapping, taxonomy development, risk assessment framework application, and implementation guidance development. Each phase is described in greater detail in the sections below.

2.1 LITERATURE REVIEW AND CAPABILITY MAPPING

We conducted a review of academic literature, gray literature, and commercial AI capabilities relevant to ITM. This review included peer-reviewed research papers, government reports, industry white papers, and commercially available AI-powered security products. The search focused on applications of AI, ML, and automated systems for detecting, preventing, or mitigating insider threats across various high-security domains.

To provide structure to this analysis, we mapped identified AI capabilities to the International Atomic Energy Agency's framework for ITM programs as outlined in *Preventive and Protective Measures Against Insider Threats* (Nuclear Security Series 8-G) [16]. This mapping process classified each AI application according to whether it supported preventive measures (designed to reduce the likelihood of insider threats), protective measures (designed to detect and respond to insider threats), or comprehensive measures (integrated approaches combining multiple elements).

2.2 TAXONOMY DEVELOPMENT

Using the information gathered through the literature review, we developed a functional taxonomy to categorize AI applications for ITM. The taxonomy development process involved iterative exploration of multiple organizational approaches, with the research team evaluating different categorization schemes based on their ability to capture the essential characteristics of AI technologies, their operational contexts, and their associated risk profiles.

Through collaborative analysis and discussion among team members, we converged on a taxonomy organized around six functional categories based on the primary security objectives and operational characteristics of different AI applications: (1) Identity and document verification systems, (2) Trustworthiness systems, (3) Behavior observation systems, (4) Impairment detection systems, (5) Access control and patrolling systems, and (6) Nuclear material accounting and control systems. Notably, these applications are somewhat interrelated. Trustworthiness assessments are the end goal of most personnel security focused approaches; generally, fitness-for-duty evaluations are conducted in service of a trustworthiness assessment. Behavior observation systems, generally, provide data as inputs to either fitness-for-duty evaluations or for trustworthiness assessments more broadly. However, there are AI models and systems that have targeted each level of this hierarchy (from broadly evaluating trustworthiness to granular behavior monitoring); thus, we present these different applications separately.

2.3 AI RISK FRAMEWORK APPLICATION

We applied principles from the NIST AI Risk Management Framework to systematically assess the risks associated with each category of AI applications [9]. The NIST framework's core principles of trustworthy AI systems—Accountability, Fairness and Bias, Privacy and Data Protection, Security and

Abuse Prevention, Explainability and Interpretability, Safety, and Performance (called Validity and Reliability in their approach)—provided a comprehensive lens for evaluating potential risks and challenges.

For each functional category, team members conducted independent risk assessments, examining how each NIST principle applied to specific characteristics and use cases within that category. These individual assessments were then compared and discussed among the research team to identify areas of agreement and resolve any differences in risk perception through consensus-building discussions.

For each functional category of AI applications, we developed implementation guidance made up of three components. First, we created descriptions of specific technology applications within each category, explaining their core functions, typical algorithms employed, and the fundamental security questions they address. Second, we identified prerequisites for implementation, including necessary infrastructure, policies, and organizational capabilities required for successful deployment. Third, we created "Responsible Implementation Checklists" that translate each NIST AI Risk Management Framework principle into specific organizational processes and controls tailored to the characteristics of each AI application category.

To support practical application of this research, we developed supplementary materials designed for INS partners. These included engagement materials formatted for presentation to international audiences and fact sheets summarizing key information for each group of AI technologies. These materials distill the technical analysis into accessible formats suitable for policymakers, security professionals, and organizational decision-makers considering AI adoption for ITM.

CHAPTER 3

3. RESULTS

The results of this analysis showcase a large landscape of AI applications for ITM, each offering distinct capabilities with unique implementation challenges and risks. The results are organized according to our six-category functional taxonomy, with each category addressing specific security objectives and operational requirements within nuclear facilities' ITM programs. Throughout, we provide a description of the technologies in each domain and implementation guidance to address risks identified. We provide academic or commercial examples of relevant technologies in each section; note that this is **not** an endorsement of those technologies. This information is provided only to illustrate that the categories and descriptions apply to products currently available on the commercial market or to capabilities actively being researched.

3.1 IDENTITY AND DOCUMENT VERIFICATION SYSTEMS

Identity and document verification systems employ AI to authenticate individuals and validate documents by comparing presented credentials, biometric features, or documents against trusted records. These systems analyze identity documents for consistency and authenticity, match facial features to verified images, and detect potentially fraudulent documentation. In nuclear facilities, these technologies may serve as verification tools during background checks, visitor processing, or personnel onboarding. They can help determine whether someone is who they claim to be and whether their documentation is legitimate while maintaining a human-in-the-loop approach for final verification decisions. Table 1 shows different applications of AI in this domain as well as the central question that each system addresses, how it is generally applied, common algorithm types, and the nature of the mathematical problem being addressed (e.g., classification, clustering, regression). We provide brief descriptions of each of the technologies in the sections below.

✅ **Key Benefits**
- ✓ Faster document processing
- ✓ Consistent evaluation criteria
- ✓ Improved detection of concerns
- ✓ Scalable for high-volume

Table 1: Identity and Document Verification Systems Overview

Technology Application	Central Question	Application Type (What it's trying to do)	Common Algorithm Types (How it's Built)	Mathematical Problem Type (How it Works)
Document-Based Identity Verification	Does this document belong to the applicant under investigation?	Document Processing	Random Forest Classifier, Support Vector Machines (SVM), Neural Networks	Classification
Fraud Detection for Document Verification	Is the document under review fraudulent?	Document Processing	Neural Networks, Random Forest Classifier	Classification
Facial Recognition for	Does this image belong to the applicant under investigation?	Computer Vision	Neural Networks	Classification

Identity Verification				

AI-powered facial recognition technology supports background check processes by comparing applicant photographs against images in existing records (such as driver's licenses or criminal records) to verify identity. The system generates confidence scores or classifications indicating the likelihood of a match between the provided photograph and images in official records, serving as one component of identity verification during background checks. One of the most well-known and widely-used AI-powered facial recognition software is the Clearview AI app, which has been used by law enforcement across the US to locate suspects in criminal investigations [41].

AI-powered identity verification systems support background check processes by automating the matching of records across databases to identify information relevant to a specific individual. The system generates confidence scores or classifications to indicate whether different records likely belong to the same person, serving as a preliminary screening tool that helps human analysts more efficiently review and verify background information for employment decisions. There are several examples of this automated identity verification in commercial products, such as LexisNexis' Scalable and Automated Linking Technology (SALT), Veriff Identity Verification, or Checkr's AI-powered background checks [42-44].

AI-powered fraud detection supports background check processes by automatically analyzing submitted documents for potential manipulation, forgery, or inconsistencies. The system examines various records (employment, financial, educational, etc.) and flags potential authenticity concerns for human review, serving as an initial screening tool to enhance document verification processes. AI use in fraud detection has increased in recent years in industry and its use has been explored by the US Treasury Department to understand the regulatory gaps that need to be addressed for effective implementation [45-47].

In order to implement AI-enabled identity and document verification systems, organizations must meet certain prerequisites. First, they **must have a digital document management system** with security measures in place to appropriately protect the sensitive information they contain. Relatedly, they must have or create clear **document handling and system access procedures** to protect applicant privacy and to prevent misuse of the system. Finally, before implementing an AI system for identity or document verification in background check processes, the organization must consider how this system will **integrated with their current human resources (HR) and security process**—that is, how will information be shared with appropriate parties and who will act upon information identified?

Table 2 presents the risk considerations for AI systems supporting identity and document verification as well as organizational processes that can be used to manage that risk. For these systems specifically, many of these risks are related to the sensitivity of the information collected (and protecting that information appropriately) and verifying the results prior to use in the applicant decision-making process. It is also critical to evaluate the performance of these AI systems both prior to procurement and during deployment to ensure that they continue to perform at sufficient levels across demographic and other groups.

Table 2: Implementation Checklist for Identity and Document Verification Systems

Principle	Key Considerations	Managed through…	Process
Accountability	Personnel must remain accountable for system	Human Oversight	Implement verification processes where system results are reviewed by

Principle	Key Considerations	Managed through…	Process
	decisions, with the system serving as a supporting tool rather than an autonomous decision-maker.		trained personnel before making determinations about identity or document authenticity.
Fairness & Bias	These systems may exhibit varying performance across demographic groups or document types.	Alternative Verification Methods	Maintain and regularly use multiple verification approaches to complement AI systems, ensuring individuals aren't disadvantaged by technological limitations.
Privacy & Data Protection	These systems process sensitive personal information including facial images and identity documents.	Secure Data Handling	Establish comprehensive protocols for the secure collection, storage, transmission, access, and disposal of identity documents and biometric data.
Security & Abuse Prevention	These systems may be misused through unauthorized searches, inappropriate access, or attempts to manipulate verification outcomes.	Access Controls	Restrict system access to authorized personnel with legitimate business needs, implementing role-based permissions and comprehensive audit logs.
Explainability & Interpretability	The system must provide indications for why the results are flagged, and personnel must be able to understand the role and limitations of the system.	Documentation Standards	Maintain records of all verification decisions, including system results, human determinations, and rationales for decisions that deviate from system recommendations.
Safety	These systems must support rather than impede effective background checks and procedures should avoid personnel overreliance.	Clear Procedures	Establish documented protocols for handling system limitations, uncertain results, outages, and escalation paths for concerning findings.
Performance	These systems should demonstrate reliable performance in flagging documents or images for review.	Performance Monitoring	Prior to procurement and during deployment, assess system accuracy, false positive/negative rates, and performance across different demographic groups and document types.

3.1.1 KEY TAKEAWAYS FOR IDENTITY AND DOCUMENT VERIFICATION SYSTEMS

- These systems are tools, not final decision makers,
- Regular training and audits maintain system effectiveness and fairness, and
- Maintaining human in the loop is essential for responsible use.

3.2 TRUSTWORTHINESS SYSTEMS

Trustworthiness systems use AI to assess personnel risk levels by evaluating patterns across historical records and identifying potential concerns. Here, trustworthiness refers to the individual's ability to safeguard sensitive material and information. These systems aggregate information from multiple sources to generate comprehensive risk assessments, flag concerning patterns, and identify specific issues requiring human investigation. In nuclear facilities, these technologies enhance personnel security programs by providing more consistent evaluation of complex information sets when making trustworthiness determinations. They serve as decision support tools that help security professionals identify patterns of concern that might be missed through traditional evaluation methods. Table 3 provides an overview of the ways that AI systems have been applied to the problem of evaluating personnel for trustworthiness—which, in the existing products, is generally described as personnel risk. This includes an overall risk score for applicants or employees as well as automated identification of issues (i.e., concerns) found in the individual's background.

Key Benefits

- ✓ Consistent application of evaluation criteria
- ✓ More efficient identification of potential concerns
- ✓ Enhanced pattern recognition across large volumes of information

Table 3: Trustworthiness Systems Overview

Technology Application	Central Question	Application Type (What it's trying to do)	Common Algorithm Types (How it's Built)	Mathematical Problem Type (How it Works)
Risk Scoring from Documents	What is the overall trustworthiness level of this individual based on their records?	Predictive/Scoring	Linear Regression, Random Forest, Neural Networks	Regression
Automated Issue Identification	Are there specific concerns in this person's background that warrant further investigation?	Document Processing	Neural Networks, SVM	Classification

AI-powered risk scoring systems support background check processes by automating the document review and evaluation process to assign a risk score that reflects a person's likelihood to be of concern. The system aggregates and reviews criminal records, financial statements and reports, employment history, and other records to generate an overall risk score indicating the trustworthiness of the applicant.

There are several examples of automated risk scoring in commercial application. Airbnb uses a proprietary algorithm to evaluate customers' risk for property damage or risk [48]. In recent years, researchers have explored the use of machine learning methods for credit scoring [49]. Financial technology companies have also begun to use "alternative data" (beyond those traditionally used in credit scoring). For example, some lenders use this so-called alternative data for credit decisions and ML approaches to help make more accurate lending decisions [50]. However, the use of machine learning—and especially deep learning—has raised concerns regarding transparency for decisions amongst consumers [51]. Nonetheless, there is evidence from the financial technology industry that risk scoring from documents may provide value when making consequential decisions about an applicant, suggesting promise for a risk scoring approach for ITM. Of course, the domain of financial

reliability is a much more limited one than overall risk or trustworthiness; additional research is certainly warranted to expand the application of AI systems from a single domain to a holistic judgment of the overall trustworthiness or riskiness of an individual.

AI-powered background check systems automate the process of identifying potential concerns in applicant records by analyzing different data sources (criminal, financial, employment) and flagging relevant issues for human review. The system helps streamline the background check process by automatically identifying and categorizing potential concerns that warrant further investigation. Automated identification of concerns in background checks is used by many companies who also leverage AI for their identity resolution or verification processes, including Checkr [44]. In addition, automated detection of issues during employment is a cornerstone of the US government's reformed clearance process, Trusted Workforce 2.0 [52].

In order to implement AI-enabled trustworthiness systems, organizations must meet several critical prerequisites. First, they must **establish secure data integration capabilities** that can safely aggregate information from multiple sources while maintaining appropriate access controls and data integrity. Second, organizations must **define clear risk factors and evaluation criteria** to ensure consistent and fair adjudication standards across all personnel assessments. Given that these systems process highly sensitive personal information, organizations must also implement robust data protection infrastructure that complies with applicable privacy regulations and security requirements. Finally, before deploying trustworthiness systems, organizations must **develop a comprehensive plan for integrating these capabilities with existing personnel security processes**—that is, how will automated risk assessments be incorporated into human decision-making workflows, who will have access to system outputs, and how will personnel act upon flagged concerns?

The risk considerations for AI systems supporting trustworthiness assessments are particularly complex given the high-consequence nature of personnel security decisions and the potential for algorithmic bias in risk scoring systems. This is especially concerning given that the judgment of overall trustworthiness is fundamentally subjective, allowing for the potentially strong influence of bias. Furthermore, the lack of clear and objective data regarding insider threats makes the scientific evaluation of trustworthiness especially challenging. Many of the risks of AI-based trustworthiness systems therefore relate to ensuring fair and transparent evaluation processes while protecting the sensitive personal information these systems analyze. It is also critical to validate the performance of these AI systems both prior to procurement and continuously during deployment to ensure they maintain acceptable accuracy levels and do not exhibit bias across different demographic groups or risk categories. Note that the evaluation of performance is much more straightforward for automated issue identification (where there is a clearly verifiable right and wrong) than for automated risk or trustworthiness scoring. Thus, organizations should be especially critical of systems designed to measure personnel risk to ensure that the resulting scores accurately reflect the characteristics that they wish to evaluate in a verifiable manner and fairly across demographic groups. Information on how to address some of these risks at the organizational level is displayed in Table 4.

Table 4: Responsible Implementation Checklist for Trustworthiness Systems

Principle	Key Considerations	Managed through…	Process
Accountability	These systems inform consequential decisions about personnel trustworthiness,	Decision Authority Policy	Establish a formal framework that clearly defines who has authority to make decisions

	requiring clear oversight and responsibility structures.		based on system outputs, with documentation requirements increasing with risk score severity.
Fairness & Bias	Risk scoring and issue identification may reflect historical biases in records or uneven availability of information across different groups.	Adjudication Standards	Develop standardized procedures for evaluating flagged issues and risk scores in proper context, with explicit consideration of potential information gaps or systemic factors.
Privacy & Data Protection	These systems process highly sensitive personal information from multiple sources, creating significant privacy concerns.	Data Governance	Implement strict controls on data access, usage, retention, and sharing, with formal protections against function creep or unauthorized expansions of system use.
Security & Abuse Prevention	Access to these systems could enable inappropriate investigations or misuse of sensitive personal information.	Access Controls	Create multi-level access restrictions based on need-to-know principles, with enhanced monitoring and audit requirements for high-risk functions, and implement comprehensive training for personnel on appropriate system use.
Explainability & Interpretability	The system must provide sufficient information so that personnel can understand what factors contribute to risk scores or issue flags to make informed decisions.	Adjudication Transparency	Maintain clear documentation of which factors contribute to risk scores and issue identification, ensuring decision-makers can explain the basis for system outputs.
Safety	Over-reliance on automated assessments could lead to missed issues or inappropriate conclusions about trustworthiness.	Multiple Analysis Methods	Implement complementary assessment approaches that don't rely solely on automated analysis, with manual review requirements scaling with risk level.
Performance	These systems must reliably identify genuine concerns while maintaining manageable false positive rates, which is notably challenging in this application	Regular Validation Testing	Conduct periodic testing with known outcomes to verify system performance, with more frequent and rigorous

	given limited data. Performance is likely to be a significant (and potentially insurmountable) obstacle.		validation required for high-risk applications.

3.2.1 KEY TAKEAWAYS FOR TRUSTWORTHINESS SYSTEMS

- These systems should inform but never replace human judgement in personnel security decisions,
- Clear documentation of factors contributing to risk assessment is essential for fair evaluation, and
- Strong privacy controls are critical due to the sensitive personal information processed.

3.3 BEHAVIOR OBSERVATION SYSTEMS

Behavior observation systems apply AI to detect anomalous activities that might indicate insider threats by establishing baselines of normal behavior and flagging meaningful deviations. These systems analyze patterns in physical movement, facility access, network activity, and digital communications. In nuclear facilities, these technologies provide continuous monitoring capabilities across physical and cyber domains, helping security teams identify concerning behavior patterns that would be difficult to detect through manual observation alone. They operate as persistent monitoring tools that extend security teams' awareness while requiring human analysis of alerts before actions are taken. Table 5 presents an overview of three ways that AI has been applied to behavior observation systems.

Key Benefits

- ✓ Continuous monitoring beyond human capability
- ✓ Consistent application of detection criteria
- ✓ Integration of physical and cyber indicators
- ✓ Early identification of potential insider threats

Table 5: Behavioral Observation Systems Overview

Technology Application	Central Question	Application Type (What it's trying to do)	Common Algorithm Types (How it's Built)	Mathematical Problem Type (How it Works)
Fitness for Duty and Behavior Observation	Is this individual displaying concerning behavioral patterns that warrant further investigation?	Anomaly Detection	Neural Networks, K-means Clustering	Clustering
Location and Anomalous Behavior Tracking	Is this movement or activity pattern unusual or concerning for this individual or location?	Computer Vision	Neural Networks	Classification
Cyber Behavior Analysis	Does this digital activity indicate potential insider threats or system compromise?	Anomaly Detection	Neural Networks, SVM	Classification & Clustering

AI systems that support fitness for duty analysis and behavior observation monitor behavioral patterns and apply anomaly detection approaches to identify deviations from the norm, signaling potential security threats. In this sort of an application, the AI system analyzes surveillance footage, access control records, or other sensors and outputs alerts when anomalous behavior is detected on or across systems. There is overlap between this application of AI and systems that perform location or anomalous behavior in general. Here, the focus is primarily on identifying patterns and deviations from those patterns, whether or not those occur in a physical realm. Instead, they may involve only cyber domains or correlation between cyber and physical realms. Many of these approaches focus on cyber behavior, given the relative immediacy of those data for analysis. For example, the company

Darktrace uses behavioral analytics and AI to establish baselines and identify departures from typical patterns, such as unusual logins or data transfers [53]. In one recent example of anomaly detection in nonproliferation analysis, researchers used a machine learning algorithm on a digital twin of a research reactor to detect proliferation-related anomalies, achieving a classification accuracy of 99% [54]. In general, there is evidence in both the commercial market and scientific literature that AI-based behavior observation may provide some promise for detecting concerning anomalies [1, 2, 4, 55]. However, notably, the systems can detect only observable behaviors; manual observations based on human judgment, knowledge of the individual, or experience may provide additional information beyond the sensors that AI system cannot detect. Furthermore, even high classification rates can fail when applied to low base rate systems such as insider threat. Nonetheless, there is some indication that AI systems can provide value in conducting fitness for duty or behavior observations.

AI-powered video analytics systems analyze surveillance feeds in real-time to detect potential security threats through identification of anomalous behavior, unattended objects, and other security-relevant events. In this use of AI, the system generates alerts for security personnel when it detects patterns that deviate from normal facility operations. These systems go beyond computer vision systems for intrusion detection or alarm, but focus on detection of typical operational patterns and deviations from them. There are a number of commercially available products that purport to have this capability, including Actuate, Verkada, and BriefCam [56-58]. Similarly, a previous Intelligence Advanced Research Projects Activity (IARPA) project called Deep Intermodal Video Analytics (DIVA) sought to develop technologies to detect specific activities in video feeds for forensic analysis or real-time alerts [59].

AI-powered monitoring and anomaly detection systems seek to identify and flag unauthorized transfers and prevent data loss. The system analyzes network traffic and digital communication and alerts cyber security professionals of potential data leakage, suspicious activity, or unauthorized transfers. The company Darktrace offers an AI-based anomaly detection system [53]. A variety of other commercial entities offer AI-based behavior analytics for security that could be applied in a nuclear setting [60].

Implementing AI-enabled behavior observation systems requires organizations to establish several foundational capabilities before deployment. Organizations must first **develop clear definitions of baseline "normal" behavior patterns** for their specific operational environment, as these systems rely on identifying deviations from established norms. The systems also require **integration with existing physical and cyber monitoring infrastructure** to avoid creating security gaps or redundant capabilities.

The use of AI systems for behavior observation programs also introduces novel risks to privacy, accountability, and security that must be considered before and during implementation. Given the continuous monitoring nature of these systems, organizations must **establish robust notification and escalation workflows** that define how alerts will be handled, who will respond to different types of behavioral flags, and what actions may be taken based on system outputs. Finally, the privacy implications of continuous behavioral monitoring **necessitates privacy impact assessments and appropriate legal review** to ensure compliance with applicable regulations and organizational policies before system activation. Organizational processes for addressing these concerns are described in Table 6.

Table 6: Responsible Implementation Checklist for Behavior Observation Systems

Principle	Key Considerations	Managed through…	Process
Accountability	Behavior monitoring systems must operate within clear authority structures with defined oversight responsibilities.	Alert Investigation Framework	Establish formal procedures for reviewing and investigating alerts, with clear documentation requirements and escalation paths for different alert types.
Fairness & Bias	Monitoring systems may incorrectly flag cultural differences or neurodivergent behaviors as suspicious.	Diverse Human Review Teams	Implement multi-person review protocols with diverse representation to evaluate alerts within appropriate cultural and individual contexts.
Privacy & Data Protection	Continuous monitoring creates significant privacy concerns regarding personal and potentially sensitive information.	Minimization & Purpose Limitation	Define and enforce strict data collection limitations, focusing only on security-relevant behaviors and implementing technical controls against function creep.
Security & Abuse Prevention	Behavior monitoring systems could enable targeted harassment, unauthorized surveillance, or retribution.	Anti-Harassment Safeguards	Create technical and procedural protections against targeting specific individuals, with independent oversight of monitoring patterns and usage.
Explainability & Interpretability	Security personnel must understand what specific behaviors triggered alerts to make informed judgments.	Alert Context Documentation	Require systems to provide specific behavioral indicators that triggered alerts, maintaining clear distinction between observed behaviors and algorithmic inferences.
Safety	Over-reliance on automated detection could create security vulnerabilities through missed behaviors or alert fatigue.	Balanced Monitoring Approach	Maintain traditional monitoring methods alongside automated systems, with regular exercises to ensure human observation capabilities remain effective.
Performance	These systems must reliably detect concerning patterns while maintaining manageable false positive rates.	Tunable Detection Thresholds	Implement configurable sensitivity settings with regular calibration based on operational experience and changing threat patterns.

3.3.1 KEY TAKEAWAYS FOR BEHAVIOR OBSERVATION SYSTEMS

- Clear definitions of "normal" behavior must be established with consideration for legitimate variations,
- Privacy protections and appropriate notification are essential for ethical implementation, and
- All alerts require human investigation to verify concerns before any actions are taken.

3.4 IMPAIRMENT DETECTION SYSTEMS

Impairment detection systems are a specialized case of anomaly detection where the system leverages AI to identify potential fitness-for-duty concerns by analyzing physiological and behavioral indicators of fatigue, intoxication, or other impairment. These systems monitor visual cues, movement patterns, and other observable signals that may indicate a person is not fit to perform security- or safety-critical functions. As noted earlier, they are generally considered to be in service of an overall assessment of personnel trustworthiness or reliability. In nuclear facilities, these technologies enhance existing fitness-for-duty programs by providing objective indicators to complement traditional testing methods. They serve as screening tools to identify potential impairment that may warrant further evaluation, helping maintain safety and security when personnel are performing sensitive functions.

Key Benefits

- ✓ Objective measurement of impairment indicators
- ✓ Continuous or point-of-entry screening capabilities
- ✓ Non-invasive detection methods
- ✓ Earlier identification of fitness concerns

Table 7: Impairment Detection Systems Overview

Technology Application	Central Question	Application Type (What it's trying to do)	Common Algorithm Types (How it's Built)	Mathematical Problem Type (How it Works)
Fatigue Detection	Is this individual displaying signs of fatigue that could impact safety or security?	Computer Vision	Neural Networks	Classification
Drug/Alcohol Impairment Detection	Is this individual displaying signs of substance impairment?	Multi-Modal	Neural Networks	Classification

AI systems process multi-modal inputs to analyze and identify fatigue utilizing physiological and behavioral indicators. These indicators can include facial expressions, eye movement, breathing patterns, heart rate and other fatigue indicators to assess fatigue levels, reducing the risk of errors in security-sensitive roles. Two recent reviews of research in this area suggest that wearable and AI-enabled solutions provide a strong option for fatigue monitoring and detection [61, 62]. There is also a commercially-available dashcam that incorporates drowsiness detection, although such a product would need to be adapted for use in security applications in a large-scale facility [63].

AI-powered impairment detection systems analyze surveillance footage or other inputs (e.g., speech, physiological indicators) to identify potential signs of inebriation or impairment in personnel. In this application, an AI system generates alerts when it detects behavioral patterns consistent with impairment, supporting existing fitness-for-duty programs in nuclear facilities. There are a number of research papers that have explored machine learning algorithms to support detection of impairment, especially by alcohol, as well as a small start-up company that currently sells an AI-powered device for detection of driver impairment [64-67] .

Before implementation of AI-powered impairment detection systems, organizations should first review fitness-for-duty policies and procedures to ensure that they are clear. Like other systems that inform or alter existing processes, they should discuss and determine how these systems will be integrated with existing approaches. There should be defined response procedures for detected impairment by the AI system that may different from those used for manual or human detection of

impairment. Just as with the broader fitness-for-duty program, before implementing AI for this use case, organizations should consider how they will **protect private medical information and respect the dignity of their staff**.

As with other applications of AI technology, it is important to ensure that human personnel remain accountable for real-world decisions based on system outputs. Given the sensitivity of the data that these systems may be handling, it is especially important to ensure that it is sufficiently protected from attack or from misuse. Finally, because results of these AI systems may inform critical decisions, such as the continued employment of a member of facility staff, it is critical that the results are interpretable, accurate, and fair. Table 8 outlines practices at the organizational level to address these and other risks.

Table 8: Responsible Implementation Checklist for Impairment Detection Systems

Principle	Key Considerations	Managed through…	Process
Accountability	Impairment allegations have significant consequences requiring careful handling and verification.	Multi-Step Verification	Implement structured verification protocols that require multiple forms of evidence before making impairment determinations or taking action.
Fairness & Bias	Systems may incorrectly flag medical conditions, disabilities, or other non-impairment factors.	Medical Review Process	Establish a formal process for medical professionals to review potential impairment indicators, considering legitimate medical explanations.
Privacy & Data Protection	Impairment detection involves collection of sensitive health-related information.	Medical Privacy Framework	Implement specialized privacy protections aligned with medical privacy standards, with strict separation between health data and personnel records.
Security & Abuse Prevention	Impairment allegations could be weaponized against individuals through system manipulation.	Independent Verification	Require independent confirmation of all impairment indicators through different methods, with protection against single-source allegations.
Explainability & Interpretability	Personnel must understand the basis for impairment determinations to ensure fair treatment.	Explainable/Transparent Indicators	System must identify signatures of impairment that has elevated an individual to be a concern and contributed to impairment determinations, with explanation appropriate for the individual.
Safety	Both false positives and negatives create safety risks through either removing needed personnel or allowing impaired operations.	Balanced Response Protocol	Create tiered response procedures that appropriately balance the risks of both types of errors based on the specific role and impairment indicators.
Performance	These systems must perform consistently across different	Performance Monitoring in Variable Environments	Validate system performance across varied operational conditions, lighting

	individuals, environments, and conditions.		environments, and times of day to ensure consistent detection capability.

3.4.1 KEY TAKEAWAYS FOR IMPAIRMENT DETECTION SYSTEMS

- All potential impairment alerts require verification through established fitness-for-duty protocols,
- Medical privacy considerations are essential due to the health-related nature of impairment data, and
- Systems should be calibrated to consider medical conditions that may present similar indicators to impairment.

3.5 ACCESS CONTROL AND PATROLLING SYSTEMS

Access control and patrolling systems employ AI to secure facility boundaries by authenticating personnel, screening for prohibited items, and conducting autonomous security monitoring. These systems control physical access through biometric verification, detect unauthorized items through image analysis, and extend surveillance capabilities through mobile platforms. In nuclear facilities, these technologies actively enforce security boundaries and expand monitoring coverage beyond fixed locations. They serve as front-line security systems making real-time decisions that directly impact facility security, often operating with limited human intervention in the immediate decision loop.

✅ Key Benefits

- ✓ Enhanced perimeter and internal security monitoring
- ✓ Consistent application of access control policies
- ✓ Extended surveillance coverage beyond fixed points
- ✓ Reduced personnel requirements for routine security functions

Table 9: Access Control and Patrolling Systems Overview

Technology Application	Central Question	Application Type (What it's trying to do)	Common Algorithm Types (How it's Built)	Mathematical Problem Type (How it Works)
Facial Recognition for Access Controls	Is this person authorized to access this area?	Computer Vision	Neural Networks	Classification
Automated Screening at Security Checkpoints	Does this scan contain prohibited items?	Computer Vision	Neural Networks	Classification
Autonomous Security Patrols	Is there unusual or concerning activity in this area?	Computer Vision	Neural Networks	Classification & Clustering

This application describes AI-powered biometric systems (including facial recognition and fingerprint verification) that control physical access to restricted areas within nuclear facilities by authenticating individual identities in real-time. Unlike background check applications of AI that could use facial recognition technology, this technology directly controls access decisions and requires immediate, accurate performance in operational conditions. AI-powered facial recognition for access control continues to be an active area of research, but there are also commercially available solutions that could be deployed in operations [68-71].

AI-powered systems to screen imagery at security checkpoints quickly assess items for prohibited materials with the goal of improving speed and accuracy over traditional screening methods. These systems assess images and data from security scans (e.g., X-ray) and determine the security status of each item. For example, TSA is exploring how it can use AI to enhance x-ray screening of luggage and personal items [72]. Similar technology could be deployed at security checkpoints at nuclear facilities to screen entrants for prohibited items.

In this application of AI technology, uncrewed autonomous systems (UAS; ground vehicles or aerial vehicles) conduct security patrols using various sensors and computer vision to detect potential security concerns. These systems extend surveillance coverage beyond fixed cameras, operating

independently to identify and report anomalies while following predetermined patrol routes or responding to specific areas of interest. Although not a security application, the National Aeronautics and Space Administration has explored the use of UAS for fire detection and mapping applications [73, 74]. New York has also explored the use of an autonomous security robot—created by the company Knightscope—for subway patrols, although the pilot of the system was ultimately ended in 2024 [75, 76].

Because these technologies have real-time decision-making implications, it is especially important that facilities assess their readiness for implementation. This includes the integration of these AI systems into their overall physical security plan, a robust communication infrastructure to address any system issues or alerts in real-time, and clear response protocols for when those issues arrive. It is also critical that organizations identify backup or redundant security measures before implementing AI in this real-time fashion.

In addition, the immediate impact of these systems on facility security necessitates a clear focus on measures to address risks of personnel safety, security, and accountability, and requires an especially high standard for their performance. That performance should be tested both prior to procurement and during deployment. Additionally, these systems require continuous human oversight through monitoring of outputs at regular defined intervals (e.g., audits, performance monitoring) and through unannounced review and testing. Provides more detail on the measures organizations can use to address these concerns.

Table 10: Responsible Implementation Checklist for Access Control and Patrolling systems

Principle	Key Considerations	Managed through…	Process
Accountability	These systems directly control security boundaries with minimal human intervention in immediate decisions.	Human Oversight Framework	Maintain continuous human oversight of system operations with clear responsibility chains and immediate intervention capabilities, including periodic testing and auditing.
Fairness & Bias	Access control systems may exhibit varying performance across demographic groups, potentially creating unequal barriers.	Equitable Testing	Regularly test and calibrate systems to ensure consistent performance across all demographic groups, and suspend usages if disparities are identified.
Privacy & Data Protection	Security monitoring creates significant privacy implications, especially for continuous or biometric surveillance.	Privacy by Design	Implement technical controls that limit surveillance to security-relevant areas and information, with clear notice to affected individuals.
Security & Abuse Prevention	These systems form critical security boundaries that adversaries may specifically target for compromise.	Defense in Depth	Deploy multiple, overlapping security controls with regular penetration testing and security assessments by qualified third parties.
Explainability & Interpretability	Security personnel must understand system decisions to effectively respond to events and troubleshoot issues.	Decision Traceability	The system must provide detailed logs of all security decisions with clear indication of contributing factors, enabling reconstruction of decision processes.
Safety	Autonomous security systems could create physical safety	Comprehensive Safety Controls	Implement multiple layers of safety features including physical

	hazards through unexpected movements, access restrictions, or other actions.		barriers, movement restrictions, collision avoidance, and emergency shutdown capabilities.
Performance	These systems must maintain high performance across varied environmental conditions and scenarios.	Environmental Stress Testing	Conduct rigorous testing across different weather conditions, lighting situations, operational scenarios and failure modes to ensure consistent security performance.

3.5.1 KEY INSIGHTS FOR IMPAIRMENT DETECTION SYSTEMS

- Robust backup authentication methods must be maintained for system failures or performance limitations
- Regular testing across environmental conditions is essential to ensure consistent security effectiveness
- Human oversight capabilities must be maintained despite the autonomous nature of these systems

3.6 NUCLEAR MATERIAL ACCOUNTING AND CONTROL SYSTEMS

AI-powered nuclear material accounting and control (NMAC) systems use AI to detect deviations in nuclear materials accounting data that may indicate theft of nuclear material. The anomaly detection system looks at nuclear accounting data and alerts when patterns of protracted or immediate theft occur. NMAC anomaly detection models are trained on either real or realistic but synthetic data. For example, the IAEA uses AI/ML models to create synthetic nuclear material accounting and verification data, which are then provided to external researchers to use in developing evaluation methodologies or internal training [77]. In another example, researchers experimented with clustering algorithms and graph analytics to support monitoring accounting of on-site material [78]. Especially for databases already stored in a digital format, AI presents potential for efficiency gains above and beyond traditional approaches [79].

Table 11: Nuclear Material Accounting and Control Systems Overview

Technology Application	Central Question	Application Type (What it's trying to do)	Common Algorithm Types (How it's Built)	Mathematical Problem Type (How it Works)
Material Movement Characterization	What are the normal patterns of material movement at this facility?	Dimension Reduction	Principal Component Analysis	Clustering
Transaction Anomaly Detection	Do current material transactions deviate from expected patterns?	Anomaly Detection	Neural Networks, Random Forest, Clustering	Classification & Clustering
Synthetic Data Generation	How can realistic training scenarios be created without sensitive data?	Generative Models	Generative Adversarial Networks, Variational Autoencoders	Generation

Dimension reduction approaches, such as principal or independent components analysis, have applications to NMAC in that they can help to characterize different transactions that are recorded in material databases [78]. This technique helps to summarize the types of transactions that occur, and can provide a clearer picture of how material at a facility is typically being used. In doing so, the results can serve as a baseline for understanding unusual or atypical behaviors that may be indicative of theft or diversion. Results can also help to provide resources in a more directed or thoughtful manner by discovering how material is most often used, potentially providing greater coordination. From a security perspective, however, results are most valuable as a basis for future anomaly detection.

AI shows promise as a potential tool for detecting anomalous material transactions or material movement in material accounting databases [78, 79]. By applying AI algorithms to existing databases and continually monitoring transactions, users will be able to more readily identify anomalous transactions that may indicate inappropriate use, theft, or diversion. This can reduce the burden on security personnel to identify potential subtle indications of concern in material control data. This is particularly of benefit given that protracted theft may be challenging to identify in large databases, making AI a potentially valuable tool.

Training personnel on identifying indications of material theft or diversion can be challenging given the paucity of real-world data available. Recently, the IAEA has explored the use of ML algorithms to create more realistic simulations to help support training of personnel [77]. This allows for the creation of more complex scenarios that better reflect realistic problem sets, and can facilitate information sharing with others for greater collaboration in the NMAC domain.

To implement AI systems for nuclear material accounting and control (NMAC), organizations must first maintain a digital nuclear material accounting system with high-quality historical data, as these AI applications depend on accurate baseline patterns to effectively detect anomalies in material transactions and movements. Given the serious security implications of detected anomalies, prior to implementation, organizations must develop clear response protocols that define specific investigation steps, assign responsibilities for follow-up actions, and establish timelines for resolving flagged concerns. The sensitive nature of nuclear material data also requires robust protections to ensure the integrity and security of the data. Finally, successful implementation requires personnel who are trained in both traditional nuclear material accounting principles and modern data interpretation techniques, ensuring they can effectively evaluate AI system outputs within the proper operational and regulatory context.

Once the decision to implement an AI system is made, organizations must establish processes for documenting system decisions, maintain multiple layers of security protections for the NMAC system itself, and employ strict role-based access control and audit logging to prevent system misuse or abuse. In addition, systems should be tested on a regular basis to ensure that they detect a variety of diversion methods. The system should also provide information about what triggered alerts in order to help maintain human accountability over NMAC decisions. Table 12 provides greater detail on processes that can help to address these risks.

Table 12: Responsible Implementation Checklist for NMAC Systems

Principle	Key Considerations	Managed through…	Process
Accountability	NMAC decisions have significant security and regulatory implications requiring clear responsibility structures.	Formal Alert Response Framework	Establish clear protocols defining responsibilities for investigating and responding to system alerts, with

			comprehensive documentation requirements.
Fairness & Bias	System may incorrectly flag legitimate operational variations as suspicious, potentially focusing attention inappropriately.	Operational Context Integration	Incorporate operational context into alert evaluation, ensuring normal process variations are properly distinguished from genuine anomalies.
Privacy & Data Protection	While focused on material rather than personnel, system may indirectly reveal information about individual actions.	Focused Data Usage	Implement strict controls ensuring NMAC data is used exclusively for material accounting purposes, with separation from personnel evaluation systems.
Security & Abuse Prevention	System could be targeted by adversaries seeking to mask diversion activities or create false alarms.	System Integrity Controls	Deploy multiple layers of security protections for the NMAC system itself, with rigorous access controls and comprehensive audit logging with regular human reviews of those logs.
Explainability & Interpretability	Personnel must understand specific factors triggering alerts to effectively investigate potential diversions.	Explainable/Transparent Indicators	Ensure system provides specific information about which patterns triggered alerts, with clear visualization of relevant data to support investigation.
Safety	Over-reliance on automated systems could create vulnerabilities through missed diversions or distraction via false alarms.	Alternative Detection Methods	Maintain traditional accounting methods alongside automated systems, with regular exercises to validate detection capabilities through multiple approaches.
Performance	System must reliably detect both abrupt and protracted diversion scenarios while minimizing false alarms.	Diversion Scenario Testing	Regularly test system using realistic diversion scenarios including various methods, quantities, and timeframes to ensure comprehensive detection capabilities.

3.6.1 KEY TAKEAWAYS FOR ACCESS CONTROL AND PATROLLING SYSTEMS

- Robust backup authentication methods must be maintained for system failures or performance limitations,
- Regular testing across environmental conditions is essential to ensure consistent security effectiveness, and
- Human oversight capabilities must be maintained despite the autonomous nature of these systems.

3.7 ENGAGEMENT MATERIALS AND FACT SHEETS

The information presented in this report is also summarized in a set of engagement materials designed for use with international partners. The engagement materials are intended to be used either during a standalone engagement of approximately one hour in length or as part of a larger engagement discussing insider threats and ITM. They are most suited for peer or near-peer partners as part of a mutual technical exchange, as this is an emerging technology with limited regulation or standards. Possible partners include the United Kingdom, Australia, or Belgium.

Additionally, we have created a set of fact sheets for each category of AI technology that can be left as read-behind materials for partners to consider their readiness to implement AI technology in this manner. They contain a brief version of some of the content presented here, including a short description of the technologies, benefits, prerequisites, important notes, and the "Responsible Implementation Checklists" shown in each section of this report. These fact sheets are presented in Appendix A. They should ideally be presented as part of the larger engagement to ensure that partners sufficiently understand the context and caveats regarding the material: they are intended primarily as reminders rather than comprehensive documents representing every possible risk of an AI system for each application.

CHAPTER 4

4. CONCLUSION

AI applications offer significant potential to enhance nuclear security ITM programs by providing capabilities that extend beyond traditional human-centered approaches. These technologies can process vast amounts of data continuously, detect subtle patterns across multiple systems, and provide consistent application of security criteria—advantages that are particularly valuable given the complex and often subtle nature of insider threats. However, the implementation of AI in nuclear security contexts also introduces new risks that must be carefully managed to ensure these systems enhance rather than undermine overall security posture.

This framework provides organizations with a structured approach to evaluating and implementing AI technologies for ITM while maintaining appropriate human oversight and addressing the ethical and operational challenges inherent in these applications. The six functional categories identified in this analysis—from identity verification through nuclear material accounting—represent different approaches to addressing insider threat challenges, each with distinct operational requirements and organizational considerations. Organizations must carefully assess their specific needs, capabilities, and operational context when considering which AI applications align with their security objectives.

Organizations must also adopt tailored approaches for each category of AI application they choose to pursue. Identity and document verification systems, for example, build upon well-established commercial technologies with clear performance metrics and relatively well-defined implementation pathways. Behavior observation systems, in contrast, require sophisticated organizational infrastructure, extensive policy development, and ongoing validation efforts. Nuclear material accounting applications leverage established mathematical approaches while requiring specialized expertise and regulatory compliance considerations. Each category demands careful planning and preparation to support successful deployment.

Critically, *all* AI applications for ITM should be implemented as decision support tools that enhance human capabilities rather than replace human judgment. The high-consequence nature of nuclear security decisions, combined with the current limitations of AI technology and the sparse data available for insider threat scenarios, makes human oversight essential for responsible implementation. Organizations must establish clear protocols for how AI system outputs will be used, verified, and acted upon, ensuring that human operators understand both the capabilities and limitations of these technologies.

The risk management approach outlined in this framework, based on the NIST AI Risk Management Framework principles, provides a method for addressing the challenges of AI deployment in nuclear security contexts. By systematically considering accountability, fairness, privacy, security, explainability, safety, and performance considerations, organizations can work to maximize the benefits of AI technologies while minimizing potential negative consequences. However, this requires ongoing commitment to monitoring system performance, updating procedures as technology evolves, and maintaining organizational capabilities for both AI-enabled and traditional security approaches.

Looking forward, organizations should recognize that responsible AI implementation for ITM is not a one-time effort but an ongoing process that requires continuous evaluation and adaptation. As AI technologies continue to evolve and new applications emerge, the frameworks and procedures

established today must be regularly updated to reflect changing capabilities, threats, and regulatory requirements. Organizations that invest in developing strong foundational capabilities for AI risk management today will be better positioned to adapt to future technological developments while maintaining the highest standards of nuclear security.

4.1 IMPLEMENTATION CONSIDERATIONS AND ORGANIZATIONAL READINESS

Before pursuing AI applications for ITM, organizations must carefully assess their readiness across multiple dimensions. Technical readiness involves not only the availability of appropriate data and IT infrastructure, but also the organizational capacity to integrate AI systems with existing security processes without creating vulnerabilities or operational disruptions. This includes ensuring that personnel responsible for system operation and oversight have appropriate training in both the technical aspects of AI systems and their specific applications to nuclear security contexts.

Organizational readiness also encompasses policy and procedural considerations. Many AI applications for ITM involve sensitive areas including personnel privacy, employment decisions, and regulatory compliance. Organizations must ensure they have appropriate legal and policy frameworks in place before implementation, including clear guidelines for data handling, system use limitations, and response protocols for different types of AI-generated alerts or recommendations. This may require coordination across multiple organizational functions including security, human resources, legal, and IT departments.

Different AI applications present varying operational requirements, and organizations must carefully consider whether their current capabilities are sufficient for their intended implementations. Applications such as comprehensive behavioral monitoring or trustworthiness assessments may require more extensive oversight mechanisms, validation procedures, and backup systems than applications like document verification or material accounting anomaly detection. Understanding these requirements early in the planning process helps ensure successful implementation.

Finally, organizations should consider their long-term strategic approach to AI adoption for ITM. Rather than implementing individual systems in isolation, organizations may benefit from developing comprehensive roadmaps that sequence different AI applications based on their organizational readiness and potential security benefits. This strategic approach can help ensure that foundational capabilities developed for initial AI implementations can support additional applications over time, maximizing return on investment while maintaining security effectiveness.

By following the framework presented in this analysis and carefully considering their organizational readiness across these dimensions, nuclear security organizations can work toward responsible implementation of AI technologies that enhance their ITM capabilities while maintaining the highest standards of safety, security, and ethical operation.

4.2 CAVEATS AND LIMITATIONS

Critically, the risks and considerations presented in this framework are not comprehensive and should not be considered exhaustive. Although we have applied established risk management principles from the NIST AI Risk Management Framework and drawn on a recent literature review, the specific operational contexts, regulatory environments, and organizational characteristics of individual nuclear facilities may introduce additional risks or considerations not addressed in this analysis. Organizations should view this framework as a foundation for their own risk assessment processes rather than a complete catalog of all potential concerns. Each organization must conduct its own thorough of how AI applications may interact with their specific operational procedures,

security measures, and regulatory requirements to identify additional risks or mitigation strategies that may be necessary for their unique circumstances.

Second, this document will likely become outdated relatively quickly due to the rapid pace of advancement in AI capabilities and applications. The field of AI is evolving at an unprecedented rate, with new algorithms, approaches, and commercial products emerging regularly. What represents cutting-edge or experimental technology today may become standard practice within months or years, while entirely new categories of AI applications may emerge that are not currently anticipated. Similarly, regulatory frameworks, industry standards, and best practices for AI deployment in high-consequence domains are still developing and may change significantly as governments and organizations gain more experience with these technologies. Organizations should therefore treat this analysis as a snapshot of current capabilities and considerations rather than a definitive long-term guide, and should plan to regularly update their AI strategies and risk assessments as the technology landscape continues to evolve.

Third, this analysis focuses specifically on core AI capabilities that directly support ITM functions as defined by the IAEA framework. However, there are many other AI systems and capabilities that may be relevant to ITM in more ancillary or indirect ways that are not reviewed in this document. For example, AI applications for general cybersecurity monitoring, administrative process automation, or facility management may have implications for insider threat programs even if they are not primarily designed for that purpose. Similarly, advances in AI technologies for areas such as social media analysis, communication monitoring, or predictive analytics may create new opportunities or challenges for ITM that are not fully explored here. Organizations should consider how their broader adoption of AI technologies across different operational areas may interact with their ITM strategies and may require additional analysis beyond what is provided in this framework.

Additionally, the commercial AI landscape is highly dynamic, with new vendors, products, and capabilities entering the market regularly while others may be discontinued or significantly modified. The specific commercial examples referenced in this analysis should be understood as illustrative of current capabilities rather than endorsements of particular vendors or products. Organizations should conduct their own market research and vendor evaluation processes when considering specific AI implementations.

Finally, while this framework addresses many of the key considerations for responsible AI implementation in nuclear security contexts, it cannot substitute for the specialized expertise, regulatory guidance, and organizational knowledge required for successful deployment. Organizations should ensure they have access to appropriate technical, legal, and policy expertise throughout their AI implementation processes and should work closely with relevant regulatory bodies to ensure compliance with all applicable requirements. The high-consequence nature of nuclear security applications demands careful, methodical approaches to AI adoption that prioritize safety and security above other considerations, and organizations should be prepared to invest the time and resources necessary to implement these technologies responsibly.

5. REFERENCES

[1] S. Yuan and X. Wu, "Deep learning for insider threat detection: Review, challenges and opportunities," *Computers & Security,* vol. 104, p. 102221, 2021/05/01/ 2021, doi: https://doi.org/10.1016/j.cose.2021.102221.

[2] A. D. Williams *et al.*, "A New Approach to Insider Threat Detection & Mitigation for Critical Infrastructure," in *Workshop on Research for Insider Threats*, Austin, TX, 4 - 8 December 2023.

[3] A. D. Williams, S. N. Abbott, N. Shoman, and W. S. Charlton, "Results From Invoking Artificial Neural Networks to Measure Insider Threat Detection & Mitigation," *Digital Threats,* vol. 3, no. 1, p. Article 3, 2021, doi: 10.1145/3457909.

[4] F. R. Alzaabi and A. Mehmood, "A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods," *IEEE Access,* vol. 12, pp. 30907-30927, 2024, doi: 10.1109/ACCESS.2024.3369906.

[5] E. Yilmaz and O. Can, "Unveiling Shadows: Harnessing Artificial Intelligence for Insider Threat Detection," *Engineering, Technology & Applied Science Research,* vol. 14, no. 2, pp. 13341-13346, 04/02 2024, doi: 10.48084/etasr.6911.

[6] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," *Proceedings of ACM on Human-Computer Interaction,* vol. 5, no. CSCW1, p. Article 188, 2021, doi: 10.1145/3449287.

[7] IBM. "AI Risk Atlas." https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas (accessed 14 April, 2025).

[8] P. Slattery *et al.*, "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence," p. arXiv:2408.12622doi: 10.48550/arXiv.2408.12622.

[9] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," 2023.

[10] A. Brenneis, "Assessing dual use risks in AI research: necessity, challenges and mitigation strategies," *Research Ethics,* vol. 21, no. 2, pp. 302-330, 2025, doi: 10.1177/17470161241267782.

[11] U. Anwar *et al.*, "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," doi: 10.48550/arXiv.2404.09932.

[12] A. Wasil, J. Clymer, D. Krueger, E. Dardaman, S. Campos, and E. Murphy, "Affirmative Safety: An Approach to Risk Management for Advanced Ai," *Available at SSRN 4806274,* 2024.

[13] A. A. Mughal, "Artificial Intelligence in Information Security: Exploring the Advantages, Challenges, and Future Directions," *Journal of Artificial Intelligence and Machine Learning in Management,* vol. 2, no. 1, pp. 22-34, 01/20 2018. [Online]. Available: https://journals.sagescience.org/index.php/jamm/article/view/51.

[14] C. Huang, Z. Zhang, B. Mao, and X. Yao, "An Overview of Artificial Intelligence Ethics," *IEEE Transactions on Artificial Intelligence,* vol. 4, no. 4, pp. 799-819, 2023, doi: 10.1109/TAI.2022.3194503.

[15] J. R. Biden, "Executive order on the safe, secure, and trustworthy development and use of artificial intelligence," 2023.

[16] IAEA, *Preventive and Protective Measures Against Insider Threats*. Vienna, Austria: International Atomic Energy Agency, 2020.

[17] E. Shaw, K. G. Ruby, and J. M. Post, "The insider threat to information systems," *Security Awareness Bulletin,* vol. 2, no. 98, 1998.

[18] M. Bunn and S. D. Sagan, "Introduction: Inside the Insider Threat," in *Insider Threats*: Cornell University Press, 2017.

[19] M. Bunn and S. D. Sagan, "A Worst Practices Guide to Insider Threats," in *Insider Threats*: Cornell University Press, 2017.

[20] F. L. Greitzer, L. J. Kangas, C. F. Noonan, and A. C. Dalton, "Identifying at-risk employees: A behavioral model for predicting potential insider threats," Pacific Northwest National Lab.

(PNNL), Richland, WA (United States), United States, 2010. [Online]. Available: https://www.osti.gov/biblio/1000159
https://www.osti.gov/servlets/purl/1000159

[21] F. Greitzer, J. Purl, Y. M. Leong, and D. S. Becker, "Sofit: Sociotechnical and organizational factors for insider threat," in *2018 IEEE Security and Privacy Workshops (SPW)*, 2018: IEEE, pp. 197-206.

[22] V. Koutsouvelis, S. Shiaeles, B. Ghita, and G. Bendiab, "Detection of Insider Threats using Artificial Intelligence and Visualisation," in *2020 6th IEEE Conference on Network Softwarization (NetSoft)*, 29 June-3 July 2020 2020, pp. 437-443, doi: 10.1109/NetSoft48620.2020.9165337.

[23] Canadian Nuclear Safety Commission
UK Office for Nuclear Regulation
US Nuclear Regulatory Commission. (2024). *ML24241A252, Considerations for Developing Artificial intelligence Systems in Nuclear Applications*. [Online] Available: https://www.nrc.gov/docs/ML2424/ML24241A252.pdf

[24] U.S. Nuclear Regulatory Commission. (2024). *ML24290A059, Regulatory Framework Gap Assessment for the use of Artificial Intelligence in Nuclear Applications*. [Online] Available: https://www.nrc.gov/docs/ML2429/ML24290A059.pdf

[25] E. Glikson and A. W. Woolley, "Human Trust in Artificial Intelligence: Review of Empirical Research," *Academy of Management Annals,* vol. 14, no. 2, pp. 627-660, 2020, doi: 10.5465/annals.2018.0057.

[26] K. E. Henry *et al.*, "Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system," *npj Digital Medicine,* vol. 5, no. 1, p. 97, 2022/07/21 2022, doi: 10.1038/s41746-022-00597-7.

[27] S. S. Y. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, and J. W. Vaughan, ""I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust," presented at the Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, 2024. [Online]. Available: https://doi.org/10.1145/3630106.3658941.

[28] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," *Proc. ACM Hum.-Comput. Interact.,* vol. 5, no. CSCW1, p. Article 188, 2021, doi: 10.1145/3449287.

[29] S. Casper *et al.*, "Black-Box Access is Insufficient for Rigorous AI Audits," doi: 10.48550/arXiv.2401.14446.

[30] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," presented at the Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 2020. [Online]. Available: https://doi.org/10.1145/3351095.3372852.

[31] S. Ling, Y. Zhang, and N. Du, "Impacts of AI-Generated Confidence and Explanations on Task Performance and Trust in Human-autonomy Teaming," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* vol. 0, no. 0, p. 21695067231195002, 2023, doi: 10.1177/21695067231195002.

[32] U. Bhatt *et al.*, "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty," presented at the Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, 2021. [Online]. Available: https://doi.org/10.1145/3461702.3462571.

[33] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets,* vol. 31, no. 3, pp. 685-695, 2021/09/01 2021, doi: 10.1007/s12525-021-00475-2.

[34] H.-P. Lee, Y.-J. Yang, T. Serban von Davier, J. Forlizzi, and S. Das, "Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks," p. arXiv:2310.07879doi: 10.48550/arXiv.2310.07879.
[35] K. Wach *et al.*, "The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT," (in English), *Entrepreneurial Business and Economics Review,* vol. 11, no. 2, pp. 7-30, 2023, doi: https://doi.org/10.15678/EBER.2023.110201.
[36] H. A. Inan *et al.*, "Training Data Leakage Analysis in Language Models," p. arXiv:2101.05405doi: 10.48550/arXiv.2101.05405.
[37] O.-M. C. Osazuwa and M. O. Musa, "The Expanding Attack Surface: Securing AI and Machine Learning Systems in Security Operations," *International Journal of Innovative Science and Research Technology,* vol. 9, pp. 2498-2505, 2024.
[38] A. T. Olutimehin, A. J. Ajayi, O. C. Metibemu, A. Y. Balogun, T. O. Oladoyinbo, and O. O. Olaniyi, "Adversarial Threats to AI-Driven Systems: Exploring the Attack Surface of Machine Learning Models and Countermeasures," (in English), *Journal of Engineering Research and Reports,* vol. 27, no. 2, pp. 341-362, 2025-02-13 2025, doi: 10.9734/jerr/2025/v27i21413.
[39] D. Nigenda *et al.*, "Amazon SageMaker Model Monitor: A System for Real-Time Insights into Deployed Machine Learning Models," presented at the Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington DC, USA, 2022. [Online]. Available: https://doi.org/10.1145/3534678.3539145.
[40] D. Leslie, "Understanding artificial intelligence ethics and safety," p. arXiv:1906.05684doi: 10.48550/arXiv.1906.05684.
[41] C. AI. "Clearview AI." https://www.clearview.ai/ (accessed 2 May, 2025).
[42] L. R. Solutions. "LexID® and Scalable Automated Linking Technology (SALT)." https://risk.lexisnexis.com/products/lexid-and-salt-for-government (accessed 12 February, 2025).
[43] Veriff. "Identity Verification." https://www.veriff.com/product/identity-verification (accessed 2 May, 2025).
[44] checkr. "AI Powered Core." https://checkr.com/our-technology/ai-powered (accessed 2 May, 2025).
[45] L. A. Garcia-Segura, "The role of artificial intelligence in preventing corporate crime," *Journal of Economic Criminology,* vol. 5, p. 100091, 2024/09/01/ 2024, doi: https://doi.org/10.1016/j.jeconc.2024.100091.
[46] T. Shahana, V. Lavanya, and A. R. Bhat, "State of the art in financial statement fraud detection: A systematic review," *Technological Forecasting and Social Change,* vol. 192, p. 122527, 2023/07/01/ 2023, doi: https://doi.org/10.1016/j.techfore.2023.122527.
[47] (2024). *Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector*. [Online] Available: /https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf
[48] Airbnb. "Aircover for Hosts." https://www.airbnb.com/aircover-for-hosts (accessed 2 May, 2025).
[49] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey," *Applied Soft Computing,* vol. 91, p. 106263, 2020/06/01/ 2020, doi: https://doi.org/10.1016/j.asoc.2020.106263.
[50] (2019). *WP 18-15, The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the Lending Club Consumer Platform*. [Online] Available: https://www.philadelphiafed.org/-/media/FRBP/Assets/working-papers/2018/wp18-15R.pdf?sc_lang=en
[51] (2018). *Financial Technology: Agencies Should Provide Clarification on Lenders' Use of Alternative Data*. [Online] Available: https://www.gao.gov/assets/d19111.pdf
[52] D. C. a. S. Agency. "Continuous Vetting." https://www.dcsa.mil/Personnel-Security/Continuous-Vetting/ (accessed 2 May, 2025).

[53] N. Carignan, "From Hype to Reality: How AI is Transforming Cybersecurity Practices," in *Darktrace*, ed. https://www.darktrace.com/blog/how-ai-is-transforming-cybersecurity-practices, 2025.
[54] E. Treviño *et al.*, "Autonomous anomaly detection of proliferation in the AGN-201 nuclear reactor digital twin," *Annals of Nuclear Energy,* vol. 211, p. 110990, 2025/02/01/ 2025, doi: https://doi.org/10.1016/j.anucene.2024.110990.
[55] A. D. Williams, S. N. Abbott, N. Shoman, and W. S. Charlton, "Results From Invoking Artificial Neural Networks to Measure Insider Threat Detection & Mitigation," *Digital Threats,* vol. 3, no. 1, 2021, doi: 10.1145/3457909.
[56] Verkada. "AI Video Analytics." https://info.verkada.com/surveillance-features/ai-video-analytics/ (accessed 2 May, 2025).
[57] E. Perez, "Using Video Content Analytics to Track Trends, Identify Anomalies and Accelerate Real-Time Response," in *AI and Video Analytics Blog* vol. 2025, ed: Briefcam, 2020.
[58] actuate. "Real-Time AI Video Analytics." https://www.briefcam.com/resources/blog/using-video-content-analytics-to-track-trends-identify-anomalies-and-accelerate-real-time-response/ (accessed 2 May, 2025).
[59] IARPA. "DIVA: Deep Intermodal Video Analytics." https://www.iarpa.gov/research-programs/diva (accessed 2 May, 2025).
[60] L. Stanham. "What is AI-Powered Behavioral Analysis in Cybersecurity." Crowdstrike. https://www.crowdstrike.com/en-us/cybersecurity-101/artificial-intelligence/ai-powered-behavioral-analysis/ (accessed 5 May, 2025).
[61] R. Hooda, V. Joshi, and M. Shah, "A comprehensive review of approaches to detect fatigue using machine learning techniques," *Chronic Diseases and Translational Medicine,* 2021/08/25/ 2021, doi: https://doi.org/10.1016/j.cdtm.2021.07.002.
[62] K. Kakhi, S. K. Jagatheesaperumal, A. Khosravi, R. Alizadehsani, and U. Rajendra Acharya, "Fatigue Monitoring Using Wearables and AI: Trends, Challenges, and Future Opportunities," p. arXiv:2412.16847doi: 10.48550/arXiv.2412.16847.
[63] samsara. "Stay Awake, Stay Safe: Real-time Alerts for Drowsy Drivers." https://www.samsara.com/products/safety/drowsiness-detection (accessed 2 May, 2025).
[64] A. A. Bonela, Z. He, A. Nibali, T. Norman, P. G. Miller, and E. Kuntsche, "Audio-based Deep Learning Algorithm to Identify Alcohol Inebriation (ADLAIA)," *Alcohol,* vol. 109, pp. 49-54, 2023/06/01/ 2023, doi: https://doi.org/10.1016/j.alcohol.2022.12.002.
[65] F. Amato, V. Cesarini, G. Olmo, G. Saggio, and G. Costantini, "Beyond breathalyzers: AI-powered speech analysis for alcohol intoxication detection," *Expert Systems with Applications,* vol. 262, p. 125656, 2025/03/01/ 2025, doi: https://doi.org/10.1016/j.eswa.2024.125656.
[66] P. Laptev, V. Demareva, S. Litovkin, E. Kostuchenko, and A. Shelupanov, "Machine learning-based detection of alcohol intoxication through speech analysis: a comparative study of AI models," *The European Physical Journal Special Topics,* 2025/02/27 2025, doi: 10.1140/epjs/s11734-025-01508-z.
[67] SoberRide. "SoberRide." https://www.sober-ride.com/ (accessed 2 May, 2025).
[68] H. Lee, S.-H. Park, J.-H. Yoo, S.-H. Jung, and J.-H. Huh, "Face Recognition at a Distance for a Stand-Alone Access Control System," *Sensors*, vol. 20, no. 3*,* doi: 10.3390/s20030785.
[69] R. Rameswari, S. Naveen Kumar, M. Abishek Aananth, and C. Deepak, "Automated access control system using face recognition," *Materials Today: Proceedings,* vol. 45, pp. 1251-1256, 2021/01/01/ 2021, doi: https://doi.org/10.1016/j.matpr.2020.04.664.
[70] alcatraz. "Meet the Rock: Next Level Access Control." https://alcatraz.ai/the-rock (accessed 2 May, 2025).
[71] I. Group. "Facial Recognition Access Control." https://www.idemia.com/facial-recognition-access-control (accessed 2 May, 2025).

[72] E. Graham, "TSA is Looking at Using AI to Improve Security Screening Processes," 8 May 2024. [Online]. Available: https://www.nextgov.com/artificial-intelligence/2024/05/how-tsa-looking-using-ai-improve-security-screening-processes/396398/

[73] NASA. "AI-Enabled Drone Swarms for Fire Detection, Mapping, and Modeling." https://esto.nasa.gov/firetech/ai-enabled-drone-swarms-for-fire-detection-mapping-and-modeling/ (accessed 2 May, 2025).

[74] S. P. H. Boroujeni *et al.*, "A comprehensive survey of research towards AI-enabled unmanned aerial systems in pre-, active-, and post-wildfire management," *Information Fusion,* vol. 108, p. 102369, 2024/08/01/ 2024, doi: https://doi.org/10.1016/j.inffus.2024.102369.

[75] Knightscope. "Why Choose the Knightscope K5 Autonomous Security Robot?" https://knightscope.com/products/k5 (accessed 2 May, 2025).

[76] D. M. Rubinstein, Hurubie, "Goodbye for Now to the Robot That (Sort Of) Patrolled New York's Subway," in *The New York Times*, ed. New York, 2024.

[77] P. S. Schneeweiss, T.; Baude, S., "The IAEA's innovative approach to address the challenges in the collection and analysis of safeguards relevant information.," presented at the Institute of Nuclear Materials Management (INMM) Annual Meeting, Portland, OR, 2024. [Online]. Available: https://cdn.fourwaves.com/static/media/formdata/b12613f9-28b7-4568-a069-5c7cfeff3c8e/8d5ee6c8-9afc-4bf0-812c-2f57cf07b9d9.pdf.

[78] A. Drescher, M. Adams, S. Stewart, K. J. Dayman, L. G. Worrall, and G. Westphal, "Machine Learning Approaches for Nuclear Material Accounting Data from Irradiation and Reprocessing," presented at the Conference: 61st INMM Annual Meeting - Virtual Meeting (no city), Tennessee, United States of America - 7/12/2020 4:00:00 PM-7/16/2020 4:00:00 PM, United States, 2020. [Online]. Available: https://www.osti.gov/biblio/1649128 https://www.osti.gov/servlets/purl/1649128.

[79] (2022). *Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology*. [Online] Available: https://www.iaea.org/publications/15198/artificial-intelligence-for-accelerating-nuclear-applications-science-and-technology