United States
Department of Energy
National Nuclear Security Administration
International Nuclear Security

AI for Insider Threat Mitigation: Capabilities, Applications, and Responsible Implementation

Jessica Baweja, Jon Barr,
Chantell Murphy

21 October 2025

PNNL-SA-216591

INS International Nuclear Security
Reducing Risk of Nuclear Terrorism

What is an insider?

- A person who has, or had, authorized access to an organization’s facilities, information, materials, personnel, resources, or systems
- Examples:
  - Employees
  - Contractors
  - Vendors
  - Former employees
  - Inspectors

International Atomic Energy Agency (2020) Preventive and protective measures against insider threats, Implementing Guide, IAEA Nuclear Security Series No.8G (Rev.1). IAEA, Vienna

What is an insider threat?

- A person who uses their access, authority, or knowledge—intentionally or unintentionally—to do harm to an organization.
- Insider threats may commit acts of:
  - Espionage
  - Sabotage
  - Theft
  - Workplace violence
  - Harassment

Insider Threats and Nuclear Security

"In every case of theft of nuclear materials where the circumstances of the theft are known, the perpetrators were either insiders or had help from insiders." [1]

"The insider threat remains one of the greatest challenges faced by the nuclear security community." [2]

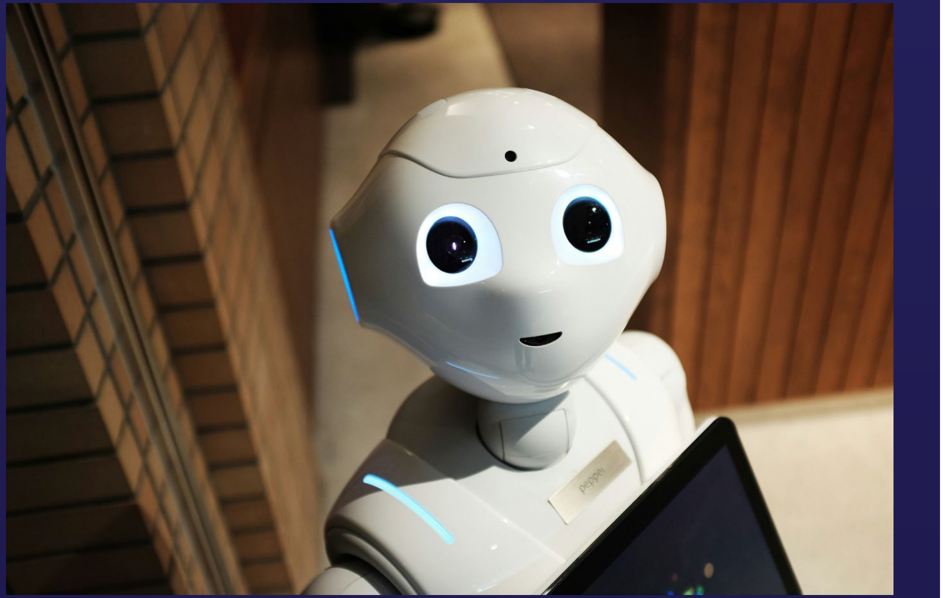
"We usually lack good and unclassified information about the details of such nuclear incidents." [3]

[1,3] Source: Sagan, S., & Bunn, M. (2014). Worst Practices Guide to Insider Threats: Lessons from Past Mistakes. American Academy of Arts & Sciences
[2] Source: Lisa E. Gordon-Hagerty, Former U.S. DOE Under Secretary for Nuclear Security and NNSA Administrator, 2019

Insider Threat Challenge

- Insiders have authorized access, knowledge of systems, and authority
- Traditional security approaches often focus on external threats
- Insiders can act maliciously, be manipulated, or create vulnerabilities through negligence
- Monitoring is resource-intensive and prone to inconsistency
- Complex data environments make pattern detection difficult for human analysts

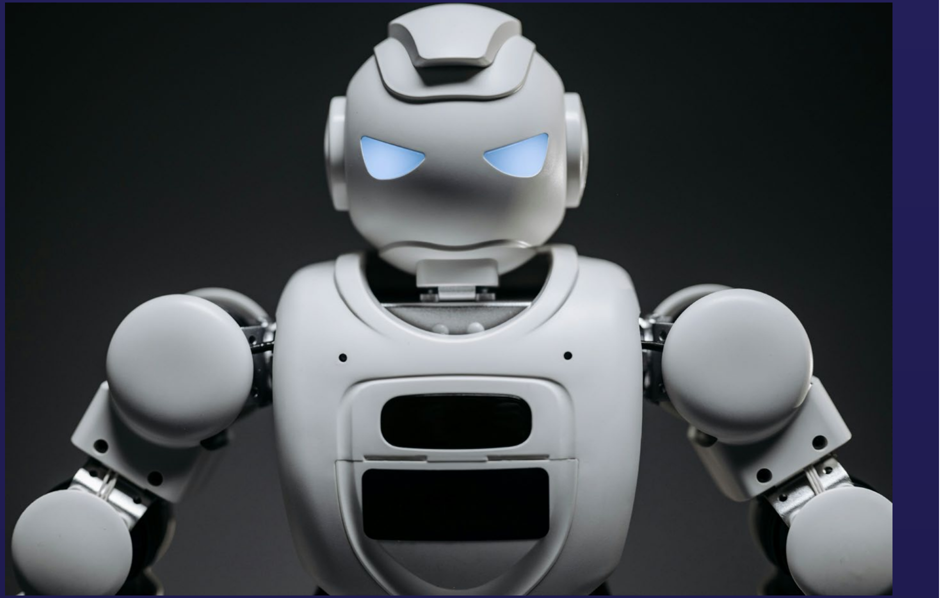
Artificial Intelligence (AI) and Insider Threat Mitigation: Enhancements



- AI can process more data than human analysts
- AI can potentially detect subtle patterns across different data sources
- AI systems can maintain consistent monitoring without fatigue
- AI is better able to integrate information across large datasets than human analysts, potentially identifying threats earlier

Hashimoto, D. A., Rosman, G., Rus, D., & Meireles, O. R. (2018). Artificial intelligence in surgery: promises and perils. *Annals of surgery*, *268*(1), 70-76.
Williams, A. D., Abbott, S. N., Shoman, N., & Charlton, W. S. (2021). Results from invoking artificial neural networks to measure insider threat detection & mitigation. *Digital Threats: Research and Practice (DTRAP)*, *3*(1), 1-20.

AI and ITM: Key Considerations

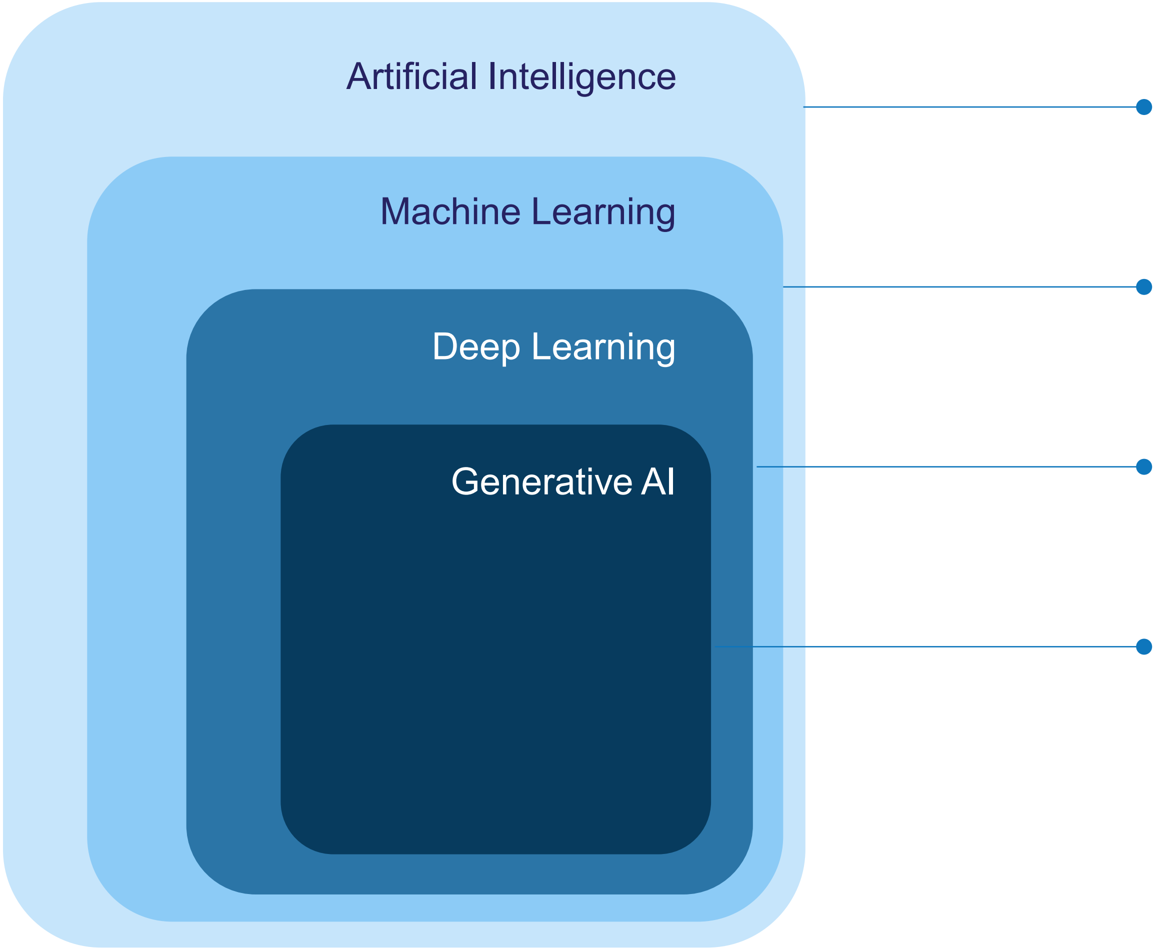


- AI systems may be biased, leading to unfair outcomes to different groups
- Aggregating data for use in AI systems may increase privacy or security concerns
- AI systems may not have sufficient transparency or explainability to support high-consequence decisions
- Human oversight remains critical for effective nuclear security
  - What happens if the system makes an error?
  - What happens if the system stops working?
  - How can we maintain accountability for AI decisions or recommendations as we integrate it into nuclear security?

Baier, L., Jöhren, F., & Seebacher, S. (2019). Challenges in the Deployment and Operation of Machine Learning in Practice ECIS 2019 proceedings . 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden, June 8-14, 2019. Research Papers, Stockholm/Uppsala.
King, J., & Meinhardt, C. (2024). Rethinking privacy in the AI era: Policy provocations for a data-centric world. Stanford Institute for Human-Centered Artificial Intelligence.
Pluff, A., & Nair, S. (2023). "Don't Blame the Robots"-Artificial Intelligence Bias & Implications for Nuclear Security. Stimson Center.

Artificial Intelligence

Foundational Definitions



Artificial Intelligence (AI): A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.

Machine Learning (ML): A set of techniques that can be used to train AI algorithms to improve performance at a task based on data.

Deep Learning (DL): A subset of machine learning, which is essentially a neural network with three or more layers.

Generative AI (GenAI): The class of AI models that emulate the structure and characteristics of input data to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.

Department of Energy Generative Artificial Intelligence Reference Guide, version 2. 14 June 2024. https://www.energy.gov/sites/default/files/2024-06/Generative%20AI%20Reference%20Guide%20v2%206-14-24.pdf

Types of Machine Learning

Supervised

Model learns from labeled data and is used to predict labels for new, unseen data

Unsupervised

Model learns from unlabeled data to discover patterns or structures

Reinforcement

Model learns optimal actions to maximize a reward signal

Common AI Methods in Everyday Life

	Central Question	Goal	Everyday Example
CLASSIFICATION	Which category does this item belong to?	Sort items into predefined categories	Email spam filter
REGRESSION	What numerical value can we predict?	Predict specific numbers based on factors	Real estate price estimation
CLUSTERING	Which items naturally group together?	Discover natural groupings—usually without labels	Netflix movie recommendations
DIMENSION REDUCTION	How can we simplify complex data?	Represent complex information more simply	Music streaming genre categories

INS International Nuclear Security
Reducing Risk of Nuclear Terrorism

de Oliveira, E. C. L., da Costa, K. S., Taube, P. S., Lima, A. H., & Junior, C. D. S. D. S. (2022). Biological membrane-penetrating peptides: computational prediction and applications. *Frontiers in Cellular and Infection Microbiology*, *12*, 838259.

Common AI Methods in Insider Threat Mitigation

	ITM Example	Central Question	Goal
CLASSIFICATION	Flagging unusual file downloads	Is this access pattern normal or suspicious?	Identify potentially malicious activities by comparing to known patterns
REGRESSION	Employee risk score calculation	What is this person’s current risk level?	Quantify potential threat level based on behavioral indicators
CLUSTERING	Grouping similar employee behaviors	Which employees exhibit similar work patterns?	Discover behavioral norms and identify outliers
DIMENSION REDUCTION	Simplifying complex user activities	What are the key patterns in this employee’s behavior?	Transform daily actions into meaningful behavioral indicators

AI Applications in Insider Threat Mitigation

- Identity & Record Verification
- Trustworthiness Assessment
- Behavior Observation
- Impairment Detection
- Access Control & Security Monitoring
- Nuclear Material Accounting and Control

Identity & Record Verification: Overview

Overview

Help authenticate individuals and validate documentation by comparing identity data (e.g., facial images, records, or documents) to trusted sources to detect fraud and evaluate legitimacy.

Common Applications

Facial Recognition for Identity Verification

Compares submitted images to official records to assess identity match.

Document-Based Identity Verification

Links applicant data across databases using confidence scores to identify matching records.

Fraud Detection for Documents

Flags inconsistencies and manipulation in submitted documents using AI.

✅ Key Benefits

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

Identity & Record Verification: Common Methods

	AI Application	Central Question	Inputs	Outputs
CLASSIFICATION	Document-Based Identity Verification	Does this document belong to the applicant under investigation?	Digital records (e.g., employment, financial) and identity information	Result indicating whether the record belongs to the applicant
CLASSIFICATION	Fraud Detection for Document Verification	Is the document under review fraudulent?	Digital personal records (e.g., employment financial)	Result indicating whether the record is fraudulent
CLASSIFICATION	Facial Recognition for Identity Verification	Does this image belong to the applicant under investigation?	Photographs with verified and unverified authenticity	Result indicating whether the photograph is of the applicant

Trustworthiness Systems

Overview

Use AI to assess personnel risk levels and identify potential concerns by evaluating patterns across historical records

Common Applications

Risk Scoring from Documents

Aggregate information from multiple sources to generate a trustworthiness score for personnel

Automated Issue Identification

Analyze applicant records to flag specific concerns requiring human investigation

✅ Key Benefits

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

Trustworthiness Systems: Common Methods

	AI Application	Central Question	Inputs	Outputs
REGRESSION	Risk Scoring from Documents	What is the overall trustworthiness level of the applicant?	Digital personal records (e.g., employment records, financial records)	Overall risk score indicating the trustworthiness of the applicant
CLASSIFICATION	Automated Issue Identification	Are there specific concerns warranting investigation?	Digital personal records (e.g., employment records financial records)	Issues identified in personal records (e.g., unpaid debts, criminal history)

Behavior Observation Systems

Overview

Apply AI to detect anomalous activities that might indicate insider threats by establishing baselines and flagging significant deviations

Common Applications

Fitness for Duty and Behavior Observation

Identify concerning behavioral patterns requiring further investigation.

Video Analytics Systems

Analyze video feeds to detect unusual physical movements or activities.

Cyber Behavior Analysis

Monitor network activity to detect unusual digital behaviors that may indicate data theft attempts or compromised credentials.

✅ Key Benefits

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

Behavior Observation Systems: Common Methods

	AI Application	Central Question	Inputs	Outputs
CLASSIFICATION	Fitness for Duty and Behavior Observation	Is this individual displaying concerning behavioral patterns that warrant further investigation?	Personnel behavior data, access patterns, communications metadata, HR indicators	Behavioral anomaly alerts, risk trend indicators, potential concern reports
CLASSIFICATION	Video Analytics Systems	Is this movement or activity pattern unusual or concerning for this individual or location?	Surveillance video feeds, defined security zones, normal movement patterns	Real-time alerts for unusual movements, unauthorized access attempts, abandoned objects, or suspicious activities
CLASSIFICATION	Cyber Behavior Analysis	Does this digital activity indicate potential insider threats or system compromise?	Network traffic data, user activity logs, file access records, data transfer patterns	Alerts for unusual data access, anomalous login patterns, unauthorized data transfers, or potential credential compromise

Trusted Workforce 2.0 - AI-Powered Personnel Vetting

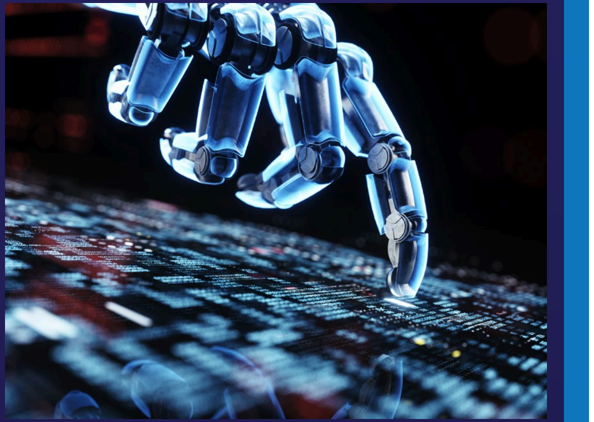
- What is Trusted Workforce 2.0?
  - US cross-government initiative launched in 2018 to transform the personnel vetting process
  - Addresses critical challenges:
    - Record-high backlog of 725,000 investigations
    - Outdated, paper-based processes taking hundreds of days
    - Vulnerabilities exposed by 2015 OPM data breach affecting 22 million records

The AI/ML Innovation:

- Transition from periodic reinvestigations to continuous vetting
- Automated record checks across multiple data sources
- AI models identify potential concerns in real-time

https://federalnewsnetwork.com/management/2025/04/trusted-workforce-2-0-ushers-in-new-era-of-personnel-vetting-but-big-challenges-remain/

How Trusted Workforce 2.0 Leverages AI/ML



- Key AI/ML Components:
  - Combines information from multiple databases and sources
  - Provides automated alerts to security personnel of new concerns in an individual’s records
- Real-World Results:
  - Reduced backlog from 725,000 to 200,000 cases within two years
  - Eliminated resource-intensive periodic reinvestigations
  - Created scalable system handling millions of clearance holders
  - Detected potential security concerns in near real-time rather than on 5–10-year cycles

Implementation & Lessons Learned

- Technical Challenges:
  - Development of National Background Investigation Services (NBIS) system
  - Data quality and integration across multiple sources
  - Balancing automation with necessary human judgment
- Policy Innovations:
  - Revised questionnaires to reflect modern realities (e.g., marijuana use, mental health)
  - Three-tier investigative model based on position risk
  - Data-centric approach to security clearance mobility
- Human Element:
  - Strong leadership commitment
  - Cross-agency collaboration with "egos left at the door"
  - Balance between technology and human expertise

Impairment Detection Systems

Overview

Leverage AI to identify potential fitness-for-duty concerns by analyzing physiological and behavioral indicators.

Common Applications

Fatigue Detection

Analyze facial expressions, eye movements, and other physiological indicators to identify signs of fatigue.

Drug/Alcohol Impairment Detection

Monitor behavioral patterns and physiological indicators to identify potential substance impairment.

✅ Key Benefits

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

Impairment Detection Systems: Common Methods

	AI Application	Central Question	Inputs	Outputs
CLASSIFICATION	Fatigue Detection	Is this individual displaying signs of fatigue that could impact safety or security?	Video of facial expressions, eye tracking data, posture information, work duration data	Fatigue level assessments, alertness warnings, rest recommendations
CLASSIFICATION	Drug/Alcohol Impairment Detection	Is this individual displaying signs of substance impairment?	Video inputs, speech patterns, movement coordination data, physiological measurements	Impairment probability alerts, specific behavioral indicators identified, recommended verification actions

Access Control & Security Monitoring

Overview
Employ AI to secure facility boundaries through automated authentication, prohibited item detection, and autonomous surveillance.

Common Applications

Facial Recognition for Access Control
Authenticate individuals through biometric matching to control entry to restricted areas.

Automated Security Screening
Analyze scanner imagery to detect prohibited items at security checkpoints.

Autonomous Security Patrols
Independently monitor facility areas to detect and report security anomalies.

✅ 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

NNSA
National Nuclear Security Administration

Access Control & Security Monitoring: Common Methods

	AI Application	Central Question	Inputs	Outputs
REGRESSION	Facial Recognition for Access Controls	Is this person authorized to access this area?	Facial images, authorized personnel database, access level information	Real-time access decisions, authentication logs, confidence scores
CLASSIFICATION	Automated Screening at Security Checkpoints	Does this scan contain prohibited items?	X-ray or scanner images, prohibited item database, item characteristic patterns	Item classification results, threat detection alerts, confidence scores by item type
ACTION REWARD REINFORCEMENT LEARNING	Autonomous Security Patrols	Is there unusual or concerning activity in this area?	Mobile sensor data, surveillance video, environmental readings, defined patrol routes	Security anomaly alerts, patrol reports, video evidence of incidents, location-specific security status

INS International Nuclear Security
Reducing Risk of Nuclear Terrorism

Nuclear Material Accounting and Control

Overview

Utilize AI to detect potential diversion of nuclear materials by identifying anomalous patterns in accounting data.

Common Applications

Material Movement Characterization

Dimension reduction techniques to establish baseline patterns of normal material movement and usage.

Transaction Anomaly Detection

Identify unusual material transactions or sequences that may indicate theft or diversion.

Synthetic Data Generation

Create realistic training scenarios for personnel and algorithm development without compromising actual facility data.

✅ Key Benefits

- ✓ Enhanced sensitivity to subtle material diversions
- ✓ Earlier detection of potential protracted diversion scenarios
- ✓ Improved pattern recognition across complex data
- ✓ Reduced false alarms through better baseline modeling

Nuclear Material Accounting and Control: Common Methods

	AI Application	Central Question	Inputs	Outputs
DIMENSION REDUCTION	Material Movement Characterization	What are the normal patterns of material movement at this facility?	Historical material transaction data, process information, facility layout data	Baseline material movement profiles, transaction type categories, usage pattern visualization
CLASSIFICATION	Transaction Anomaly Detection	Do current material transactions deviate from expected patterns?	Material transaction data, inventory records, established baseline patterns	Diversion possibility alerts, anomaly classification, pattern deviation reports
REGRESSION	Synthetic Data Generation	How can realistic training scenarios be created without sensitive data?	Pattern templates, facility characteristics, known diversion scenarios	Synthetic transaction datasets, training scenarios, evaluation benchmarks

INS International Nuclear Security

Reducing Risk of Nuclear Terrorism