National Nuclear Security Administration

United States
Department of Energy
National Nuclear Security Administration

International Nuclear Security

L'IA pour l'atténuation des menaces internes : capacités, applications et mise en œuvre responsable

Jessica Baweja, Jon Barr,
Chantell Murphy

21 octobre 2025

PNNL-SA-216591

INS International Nuclear Security
Reducing Risk of Nuclear Terrorism

Qu’est-ce qu’un agresseur interne ?

- Une personne qui a, ou a eu, l’accès autorisé aux installations, informations, matières, personnel, ressources ou systèmes d’une organisation
- Exemples :
  - Employés
  - Entrepreneurs
  - Fournisseurs
  - Anciens employés
  - Inspecteurs

Agence internationale de l’énergie atomique (2020) Mesures de prévention et de protection contre les menaces internes, Guide d’application, no 8 de la collection Sécurité nucléaire de l’AIEA (Rév.1). AIEA, Vienne

Qu’est-ce qu’une menace interne ?

- Une personne qui se sert de son accès, de son autorité ou de ses connaissances – volontairement ou involontairement – pour porter atteinte à une organisation.
- Les menaces internes peuvent donner lieu à des :
  - Actes d’espionnage
  - Actes de sabotage
  - Vols
  - Violence sur le lieu de travail
  - Harcèlement

Menaces internes et sécurité nucléaire

« Dans chaque cas de vol de matières nucléaires où les circonstances du vol sont connues, les auteurs sont des agresseurs internes ou ils ont reçu l'aide d'agresseurs internes ».[1]

« La menace interne demeure l'un des plus grands défis auquel la communauté de la sécurité nucléaire est confrontée ».[2]

« En général, il nous manque des informations non classifiées pertinentes sur les détails de tels incidents nucléaires ».[3]

[1,3] Source : Sagan, S., & Bunn, M. (2014). Worst Practices Guide to Insider Threats: Lessons from Past Mistakes (Guide des pires pratiques en matière de menaces internes : enseignements des erreurs passées). American Academy of Arts & Sciences
[2] Source : Lisa E. Gordon-Hagerty, ancienne sous-secrétaire à la sécurité nucléaire du Department of Energy américain et administratrice de la NNSA, 2019

Défis posés par les menaces internes

Les agresseurs internes disposent d'un accès autorisé, de connaissances des systèmes et d'une autorité

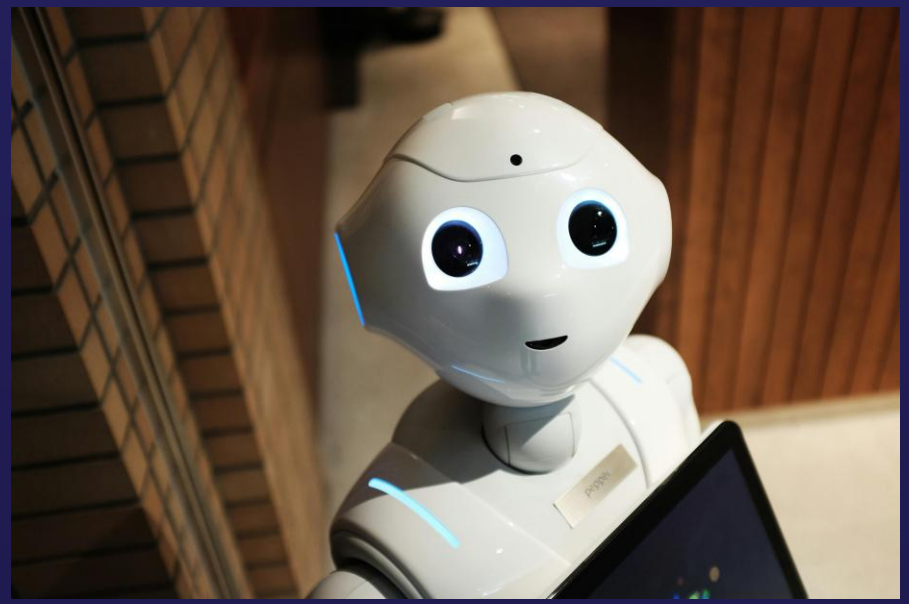
Les approches de sécurité traditionnelles se concentrent souvent sur les menaces externes

Les agresseurs internes peuvent agir de manière malveillante, être manipulés ou créer des vulnérabilités par négligence

La surveillance fait appel à beaucoup de ressources et est sujette aux incohérences

Les environnements de données complexes rendent la détection de schémas difficile pour les analystes humains

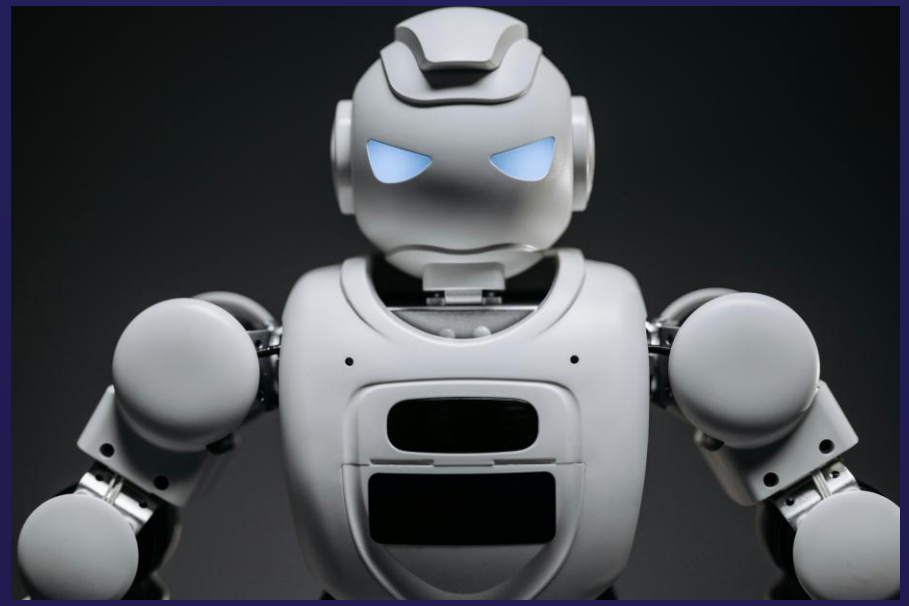
L’intelligence artificielle (IA) et l’atténuation de la menace interne : optimisations



- L’IA peut traiter davantage de données que les analystes humains
- Potentiellement, l’IA peut détecter des schémas subtils parmi différentes sources de données
- Les systèmes d’IA peuvent maintenir une surveillance constante sans éprouver de fatigue
- L’IA est meilleure que les analystes humains lorsqu’il s’agit d’intégrer des informations sur des ensembles de données, en identifiant potentiellement les menaces plus tôt

Hashimoto, D. A., Rosman, G., Rus, D., & Meireles, O. R. (2018). Artificial intelligence in surgery: promises and perils. (L’intelligence artificielle en chirurgie : promesses et dangers). *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 (Résultats de l’utilisation de réseaux neuronaux artificiels pour mesurer la détection et l’atténuation des menaces internes). *Digital Threats: Research and Practice (DTRAP)*, *3*(1), 1-20.

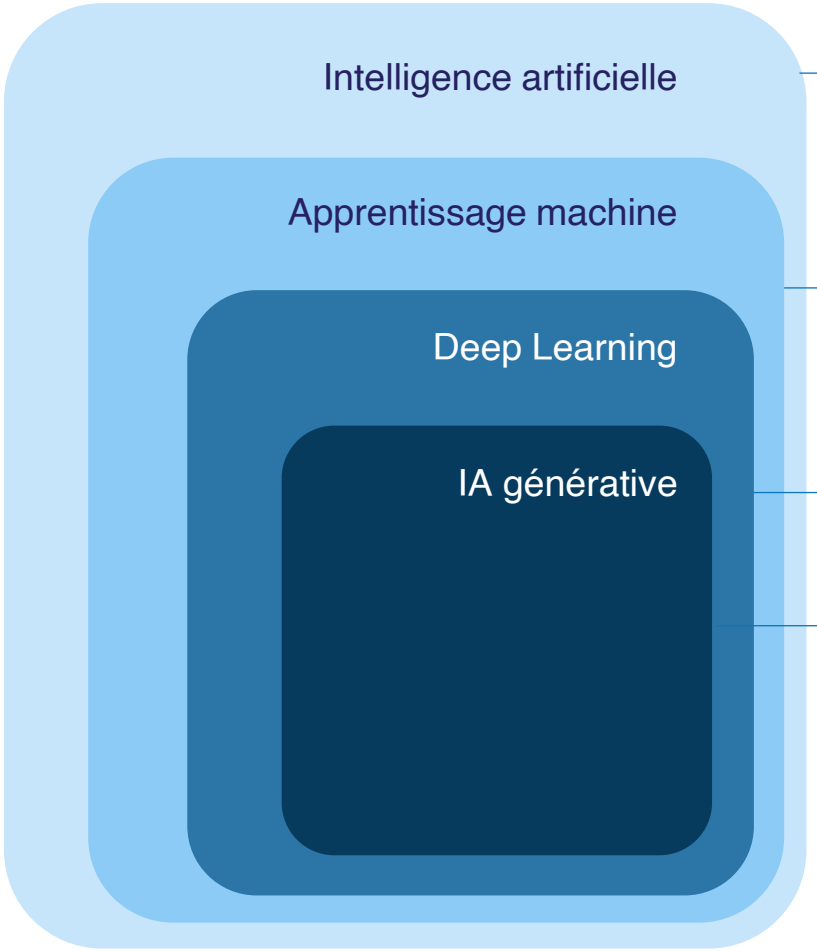
IA et atténuation de la menace interne : considérations clés



- Les systèmes d'IA peuvent être biaisés, ce qui peut entraîner des résultats injustes pour différents groupes
- L'agrégation de données destinées à être utilisées dans des systèmes d'IA peut accroître les préoccupations en matière de confidentialité ou de sécurité
- Les systèmes d'IA peuvent ne pas offrir suffisamment de transparence ou d'explications pour étayer des décisions ayant des conséquences importantes
- La supervision humaine reste essentielle pour une sécurité nucléaire efficace
  - Que se passe-t-il si le système fait une erreur ?
  - Que se passe-t-il si le système cesse de fonctionner ?
  - Comment pouvons-nous maintenir la redevabilité des décisions ou des recommandations de l'IA lorsque nous l'intégrons à la sécurité nucléaire ?

Baier, L., Jöhren, F., & Seebacher, S. (2019). Challenges in the Deployment and Operation of Machine Learning in Practice (Défis liés au déploiement et à l'exploitation de l'apprentissage machine en pratique), débats ECIS 2019. 27e Conférence européenne sur les systèmes d'information (ECIS), Stockholm et Uppsala, Suède, 8-14 juin 2019. Articles de recherche.
King, J., & Meinhardt, C. (2024). Rethinking privacy in the AI era: Policy provocations for a data-centric world. (Repenser la vie privée à l'ère de l'IA : provocations politiques pour un monde centré sur les données). Institut Stanford pour l'intelligence artificielle centrée sur l'humain.
Pluff, A., & Nair, S. (2023). « Don't Blame the Robots » (« Ce n'est pas la faute des robots ») - Les biais de l'intelligence artificielle et leurs implications pour la sécurité nucléaire. Stimson Center.

Intelligence artificielle
Définitions fondamentales



Intelligence artificielle (IA) : un système basé sur la machine qui peut, pour un ensemble donné d'objectifs définis par l'humain, faire des prédictions, des recommandations ou prendre des décisions influençant des environnements réels ou virtuels.

Apprentissage machine (Machine Learning - ML) : un ensemble de techniques qui peuvent être utilisées pour entraîner les algorithmes de l'IA, afin d'améliorer les performances d'une tâche basée sur des données.

Deep Learning (DL) : un sous-ensemble d'apprentissages machine, qui est en fait un réseau neuronal avec au moins trois couches.

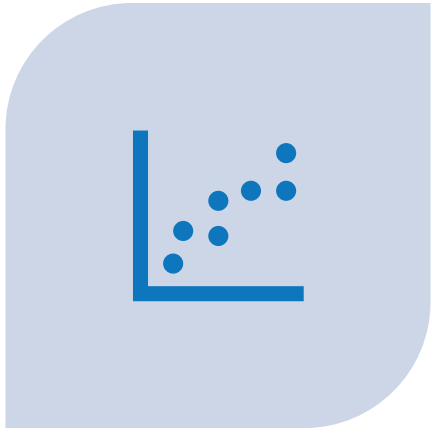
IA générative (GenAI) : la catégorie de modèles d'IA qui émulent la structure et les caractéristiques des données d'entrée pour générer du contenu synthétique dérivé. Cela peut inclure des images, des vidéos, de l'audio, du texte et d'autres contenus numériques.

Guide de référence sur l'intelligence artificielle générative du Department of Energy américain, version 2. 14 juin 2024.
https://www.energy.gov/sites/default/files/2024-06/Generative%20AI%20Reference%20Guide%20v2%206-14-24.pdf

Types d’apprentissage machine

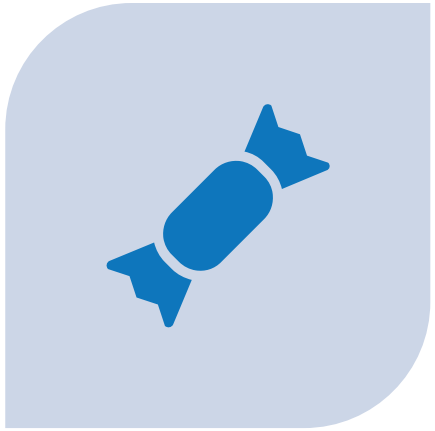
Supervisé

Le modèle apprend à partir de données étiquetées et est utilisé pour prédire les étiquettes de nouvelles données inconnues



Non supervisé

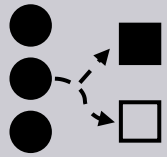
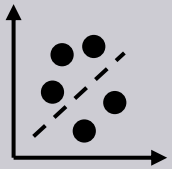
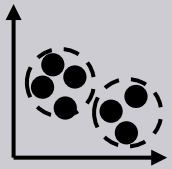
Le modèle apprend à partir de données non étiquetées pour découvrir des schémas ou des structures



Renforcement

Le modèle apprend les actions optimales, afin de maximiser un signal de récompense

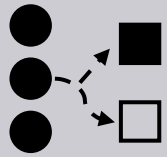
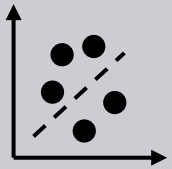
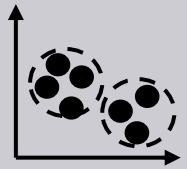
Méthodes d'IA courantes dans la vie quotidienne

	Question centrale	Objectif	Exemple de tous les jours
 CLASSIFICATION	À quelle catégorie cet élément appartient-il ?	Trier les éléments en catégories prédéfinies	Filtre de spam pour courrier électronique
 RÉGRESSION	Quelle valeur numérique pouvons-nous prédire ?	Prédire des nombres spécifiques en fonction de facteurs	Estimation de prix immobiliers
 REGROUPEMENT	Quels éléments se regroupent naturellement ?	Découvrir les regroupements naturels – d'habitude sans étiquettes	Recommandations de films Netflix
 RÉDUCTION DE LA DIMENSION	Comment pouvons-nous simplifier les données complexes ?	Représenter les informations complexes plus simplement	Catégories de genres de streaming musical

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 (Peptides biologiques pénétrant les membranes : prédiction computationnelle et applications). *Frontiers in Cellular and Infection Microbiology*, *12*, 838259.

Méthodes d’IA courantes pour atténuer les menaces internes

	Exemple d’atténuation	Question centrale	Objectif
 CLASSIFICATION	Signaler des téléchargements de fichiers inhabituels	Ce schéma d’accès est-il normal ou suspect ?	Identifier les activités potentiellement dangereuses en les comparant aux schémas connus
 RÉGRESSION	Calcul du score de risque pour les employés	Quel est le niveau de risque actuel de cette personne ?	Quantifier le niveau de menace potentiel en fonction d’indicateurs comportementaux
 REGROUPEMENT	Regrouper des comportements similaires d’employés	Quels employés présentent des schémas de travail similaires ?	Découvrir les normes comportementales et identifier les anomalies
 RÉDUCTION DE LA DIMENSION	Simplifier les activités complexes d’utilisateur	Quelles sont les tendances clés du comportement de cet employé ?	Transformer les actions quotidiennes en indicateurs comportementaux ayant du sens

INS International Nuclear Security
Reducing Risk of Nuclear Terrorism

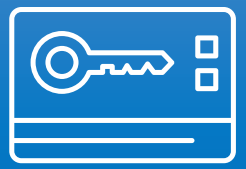
Applications de l’IA dans l’atténuation des menaces internes

Vérification de l’identité et de dossiers

Évaluation de l’intégrité

Observation du comportement

Détection d’altérations



Contrôle de l’accès et surveillance de sécurité

Comptabilité et contrôle des matières nucléaires

INS International Nuclear Security
Reducing Risk of Nuclear Terrorism

Vérification de l’identité et de dossiers : vue d’ensemble

Vue d’ensemble
Aider à authentifier les individus et à valider les documents en comparant les données d’identité (par exemple, les images faciales, les dossiers ou les documents) à des sources fiables, afin de détecter les fraudes et d’évaluer la légitimité.

Applications courantes

Reconnaissance faciale pour la vérification de l’identité
Comparer les images soumises aux registres officiels, afin d’évaluer la correspondance de l’identité.

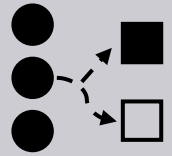
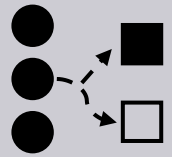
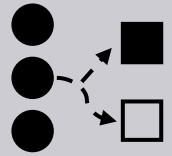
Vérification de l’identité basée sur des documents
Relier les données des candidats entre différentes bases de données en employant des scores de confiance, afin d’identifier les dossiers correspondants.

Détection de la fraude pour les documents
Signaler les incohérences et les manipulations dans les documents soumis en se servant de l’IA.

✅ Avantages clés

- ✓ Traitement plus rapide des documents
- ✓ Critères d’évaluation cohérents
- ✓ Détection améliorée des préoccupations
- ✓ Évolutif pour le traitement à grande échelle

Vérification de l’identité et de dossiers : méthodes courantes

	Application d’IA	Question centrale	Intrants	Extrants
 CLASSIFICATION	Vérification d’identité basée sur des documents	Ce document appartient-il au candidat examiné ?	Dossiers numériques (p. ex., emploi, finances) et informations d’identité	Résultat indiquant si le dossier appartient au candidat
 CLASSIFICATION	Détection de la fraude pour la vérification des documents	Le document examiné est-il frauduleux ?	Dossiers numériques personnels (p. ex., emploi, finances)	Résultat indiquant si le dossier est frauduleux
 CLASSIFICATION	Reconnaissance faciale pour la vérification de l’identité	Cette image appartient-elle au candidat examiné ?	Les photographies avec authenticité vérifiée et non vérifiée	Résultat indiquant si la photographie est celle du candidat

Systèmes de contrôle de l’intégrité

Vue d’ensemble

Utiliser l’IA pour évaluer les niveaux de risques personnels et identifier les préoccupations potentielles en évaluant les tendances parmi les dossiers historiques

Applications courantes

Notation du risque à partir des documents

Agréger des informations à partir de plusieurs sources, afin de générer un score d’intégrité pour le personnel

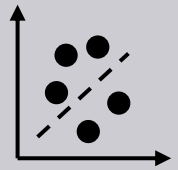
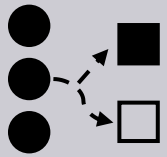
Identification de problèmes automatisée

Analyser les dossiers de candidats pour signaler des préoccupations spécifiques nécessitant un examen humain

✅ Avantages clés

- ✓ Application cohérente des critères d’évaluation
- ✓ Identification plus efficace des préoccupations potentielles
- ✓ Reconnaissance de tendances améliorée sur de larges volumes d’informations

Systèmes de contrôle de l'intégrité : méthodes courantes

	Application d'IA	Question centrale	Intrants	Extrants
 RÉGRESSION	Score de risque établi à partir de documents	Quel est le niveau d'intégrité global du candidat ?	Dossiers numériques personnels (p. ex., dossiers professionnels, financiers)	Score de risque global indiquant l'intégrité du candidat
 CLASSIFICATION	Identification de problème automatisée	Y a-t-il des préoccupations spécifiques justifiant une enquête ?	Dossiers numériques personnels (p. ex., dossiers professionnels, financiers)	Problèmes identifiés dans les dossiers personnels (p. ex., dettes impayées, casier judiciaire)

Systèmes d’observation du comportement

Vue d’ensemble

Appliquer l’IA pour détecter les activités anormales qui pourraient être révélatrices de menaces internes, en établissant des bases de référence et en signalant les écarts importants

Applications courantes

Aptitude au service et observation du comportement

Identifier les modèles comportementaux nécessitant un examen plus approfondi.

Systèmes d’analyse vidéo

Analyser les flux vidéo pour détecter des mouvements ou activités physiques inhabituels.

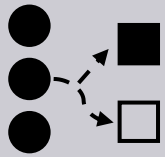
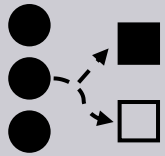
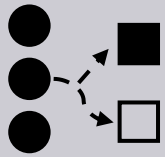
Analyse du cyber-comportement

Surveiller l’activité de réseau pour détecter des comportements numériques inhabituels qui pourraient indiquer des tentatives de vol de données ou des identifiants compromis.

✅ Avantages clés

- ✓ Surveillance continue au-delà de la capacité humaine
- ✓ Application cohérente des critères de détection
- ✓ Intégration des indicateurs physiques et cyber
- ✓ Identification précoce des menaces internes potentielles

Systèmes d'observation du comportement : méthodes courantes

	Application d'IA	Question centrale	Intrants	Extrants
 CLASSIFICATION	Aptitude au service et observation du comportement	Cet individu présente-t-il des schémas de comportement préoccupants qui justifient un examen plus approfondi ?	Données comportementales sur le personnel, modèles d'accès, métadonnées de communications, indicateurs RH	Alertes d'anomalie comportementale, indicateurs de tendances des risques, rapports de préoccupations potentielles
 CLASSIFICATION	Systèmes d'analyse vidéo	Ce schéma de mouvement ou d'activité est-il inhabituel ou préoccupant pour cet individu ou ce lieu ?	Flux de vidéosurveillance, zones de sécurité définie, schémas de mouvements normaux	Alertes en temps réel pour les mouvements inhabituels, tentatives d'accès non autorisés, objets abandonnés ou activités suspectes
 CLASSIFICATION	Analyse du cyber-comportement	Cette activité numérique est-elle révélatrice de menaces internes potentielles ou d'une compromission du système ?	Données de trafic du réseau, journaux d'activité d'utilisateur, archives d'accès aux fichiers, schémas de transferts de données	Alertes en cas d'accès aux données inhabituel, tendances de connexion anormales, transferts de données non autorisés ou compromission potentielle d'identifiants

Trusted Workforce 2.0 – Vérification du personnel optimisée par l'IA

L'innovation IA/ML :

- Passage d'une réévaluation périodique à une vérification continue
- Vérifications automatisées des antécédents à partir de plusieurs sources de données
- Les modèles d'IA identifient les problèmes potentiels en temps réel

- Qu'est-ce que Trusted Workforce 2.0 ?
  - Initiative intergouvernementale américaine lancée en 2018 pour transformer le processus de vérification du personnel
  - Répond à des défis critiques :
    - Retard record de 725 000 enquêtes
    - Processus obsolètes, basés sur le papier, prenant des centaines de jours
    - Vulnérabilités exposées par la violation des données de l'OPM en 2015, affectant 22 millions d'enregistrements

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

Comment Trusted Workforce 2.0 tire parti de l'IA/ML

- Composants IA/ML clés :
  - Combine les informations provenant de bases de données et sources multiples
  - Envoie des alertes automatiques au personnel de sécurité en cas de nouveaux problèmes dans les dossiers d'un individu
- Résultats dans le monde réel :
  - Réduction du nombre de dossiers en attente de 725 000 à 200 000 en deux ans
  - Élimination des ré-investigations périodiques exigeant beaucoup de ressources
  - Création d'un système évolutif capable de traiter des millions de détenteurs d'autorisations
  - Détection des problèmes de sécurité potentiels en temps quasi réel plutôt que sur des cycles de 5 à 10 ans

Mise en œuvre et enseignements tirés

- Défis techniques :
  - Développement du système national d'enquête sur les antécédents (National Background Investigation Services - NBIS)
  - Qualité et intégration des données issues de sources multiples
  - Équilibrage entre automatisation et jugement humain nécessaire
- Innovations des politiques :
  - Questionnaires révisés pour refléter les réalités modernes (p. ex. consommation de marijuana, santé mentale)
  - Modèle d'enquête à trois niveaux, fondé sur le risque lié au poste
  - Approche centrée sur les données pour la mobilité des habilitations de sécurité
- Élément humain :
  - Engagement fort de la direction
  - Collaboration entre agences sans « ego »
  - Équilibre entre technologie et expertise humaine

Systèmes de détection des altérations

Vue d'ensemble

Exploiter l'IA pour identifier les préoccupations potentielles liées à l'aptitude au service, en analysant les indicateurs physiologiques et comportementaux.

Applications courantes

Détection de la fatigue

Analyser les expressions faciales, les mouvements oculaires et d'autres indicateurs physiologiques, afin d'identifier les signes de fatigue.

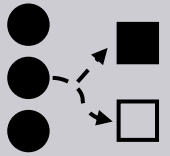
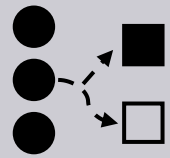
Détection des altérations liées à la consommation de drogues/d'alcool

Surveiller les comportements et les indicateurs physiologiques, afin d'identifier les altérations potentielles liées à la consommation de substances psychoactives.

✅ Avantages clés

- ✓ Mesure objective des indicateurs de déficience
- ✓ Capacités de contrôle continu ou au point d'entrée
- ✓ Méthodes de détection non invasives
- ✓ Identification précoce des préoccupations d'aptitude au poste

Systèmes de détection des altérations : méthodes courantes

	Application d'IA	Question centrale	Intrants	Extrants
 CLASSIFICATION	Détection de la fatigue	Cet individu présente-t-il des signes de fatigue qui pourraient impacter la sûreté ou la sécurité ?	Vidéo des expressions faciales, données d'oculométrie, informations sur la posture, données sur la durée de travail	Évaluations du niveau de fatigue, avertissements sur le niveau de de vigilance, recommandations de repos
 CLASSIFICATION	Détection des altérations liées à la consommation de drogues/d'alcool	Cet individu présente-t-il des signes d'altérations liées à la consommation de substances psychoactives ?	Entrées vidéo, schémas vocaux, données de coordination des mouvements, mesures physiologiques	Alertes de probabilité d'altération du comportement, indicateurs comportementaux spécifiques identifiés, actions de vérification recommandées

Contrôle d’accès et surveillance de sécurité

Vue d’ensemble
Utiliser l’IA pour sécuriser les limites des installations par le biais de l’authentification automatisée, la détection des objets interdits et la surveillance autonome.

Applications courantes

Reconnaissance faciale pour le contrôle d’accès
Authentifier les personnes par le biais de la reconnaissance biométrique, afin de contrôler l’accès aux zones restreintes.

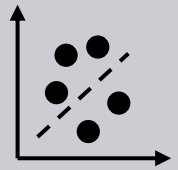
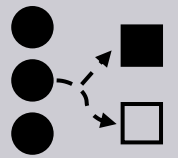
Contrôle de sécurité automatisé
Analyser les images du scanner pour détecter les objets interdits aux points de contrôle de sécurité.

Patrouilles de sécurité autonomes
Surveiller de manière indépendante les zones de l’installation pour détecter et signaler les anomalies de sécurité.

✅ Avantages clés

- ✓ Surveillance améliorée de la sécurité périmétrique et interne
- ✓ Application cohérente des politiques de contrôle d’accès
- ✓ Couverture de surveillance étendue au-delà des points fixés
- ✓ Besoins en personnel réduits pour les fonctions de sécurité de routine

Contrôle d'accès et surveillance de sécurité : méthodes courantes

	Application d'IA	Question centrale	Intrants	Extrants
 RÉGRESSION	Reconnaissance faciale pour les contrôles d'accès	Cette personne est-elle autorisée à accéder à cette zone ?	Images faciales, base de données du personnel autorisé, informations sur le niveau d'accès	Décisions d'accès en temps réel, journaux d'authentification, scores de confiance
 CLASSIFICATION	Contrôle automatisé aux points de contrôle de sécurité	Ce scan montre-t-il des objets interdits ?	Images radiographiques ou scannées, base de données des objets interdits, schémas caractéristiques des objets	Résultats de classification des objets, alertes de détection des menaces, scores de confiance par type d'objete
ACTION  RÉCOMPENSE APPRENTISSAGE PAR RENFORCEMENT	Patrouilles de sécurité autonomes	Y a-t-il une activité inhabituelle ou préoccupante dans cette zone ?	Données issues des capteurs mobiles, vidéo-surveillance, mesures environnementales, itinéraires de patrouille définis	Alertes d'anomalies de sécurité, rapports de patrouille, preuves vidéo des incidents, statut de sécurité spécifique à un lieu

Comptabilité et contrôle des matières nucléaires

National Nuclear Security Administration

Vue d'ensemble

Utiliser l'IA pour détecter les détournements potentiels de matières nucléaires en identifiant les anomalies dans les données comptables

Applications courantes

Caractérisation des mouvements de matières

Techniques de réduction de la dimension pour établir des modèles de référence de l'utilisation et des mouvements normaux des matières.

Détection des anomalies dans les transactions

Identifier les transactions ou séquences inhabituelles de matières qui pourraient indiquer un vol ou un détournement.

Génération de données synthétiques

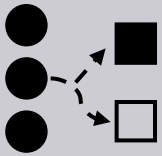
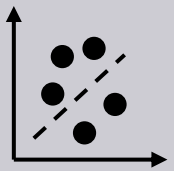
Créer des scénarios de formation réalistes pour le personnel et le développement d'algorithmes sans compromettre les données réelles des installations.

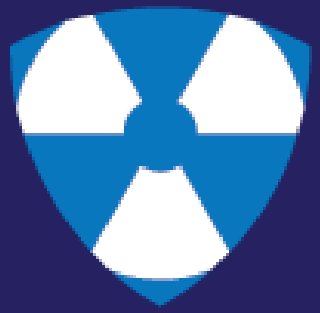
✅ Avantages clés

- ✓ Sensibilité accrue aux détournements subtils de matières
- ✓ Détection plus précoce des scénarios potentiels de détournement prolongé
- ✓ Reconnaissance améliorée des schémas dans des données complexes
- ✓ Réduction des fausses alertes grâce à une meilleure modélisation de référence

INS International Nuclear Security
Reducing Risk of Nuclear Terrorism

Comptabilité et contrôle des matières nucléaires : méthodes courantes

	Application d'IA	Question centrale	Intrants	Extrants
 RÉDUCTION DE LA DIMENSION	Caractérisation des mouvements de matières	Quels sont les schémas de mouvement des matières normaux dans cette installation ?	Données historiques sur les transactions de matières, informations sur les processus, données sur la configuration de l'installation	Profils de référence des mouvements de matières, catégories de types de transactions, visualisation des schémas d'utilisation
 CLASSIFICATION	Détection d'anomalie dans les transactions	Les transactions actuelles impliquant des matières s'écartent-elles des modèles attendus ?	Données sur les transactions de matières, registres d'inventaire, schémas de référence établis	Alertes de possibilité de détournement, classification des anomalies, rapports sur les écarts par rapport aux schémas
 RÉGRESSION	Génération de données synthétiques	Comment créer des scénarios de formation réalistes sans données sensibles ?	Modèles de schémas, caractéristiques des installations, scénarios de détournement connus	Ensembles de données synthétiques sur les transactions, scénarios de formation, critères d'évaluation



INS International Nuclear Security

Reducing Risk of Nuclear Terrorism