
CBRN AI Risks Research Sprint

Radiological & Nuclear Risk Monitoring (Track 4)

Policy Brief

Navigating Safety Measures for Nuclear Nonproliferation: AI-Enabled Early Warning Systems & Governance System For Detecting Nuclear Enrichment

Ifeoma Ilechukwu
AI Legal and Governance
Advisor

Marie-Louise Thurton
Toronto Metropolitan
University

Victor Uchenna Jonah
Software Engineer and AI
Developer

Valmik Nahata
Independent Researcher

Saahir Vazirani
Independent Researcher

With
Apart Research

September 14, 2025

Abstract

The Independent Atomic Energy Agency (IAEA) has always sought ways to create a balance between nuclear use for peaceful purposes and the prevention of weapons proliferation. Yet challenges persist in the form of clandestine uranium enrichment, covert plutonium processing and illicit trade in sensitive technologies. Recent advances in the integration of Artificial Intelligence into nuclear Command, Control and Communication (NC3) systems and procedures could help reduce errors from being made in crisis scenarios, enhance situational awareness, improve surveillance and increase operational efficiency. However, the introduction of AI systems would also amplify existing challenges and create room for adversarial manipulation where proliferators intentionally feed misleading signals to evade detection systems and reduce facility footprints. This brief critically examines the vulnerabilities and limitations in existing safeguards, recent case studies such as Iran's suspension of IAEA cooperation and a multimodal AI-enabled monitoring approach for early detection of nuclear enrichment activities. It concludes with governance approaches to reduce nuclear related misuse and policy measures to align AI capabilities with non-proliferation norms.

Introduction

The entire radiological and nuclear landscape is regulated by a network of treaties, verification mechanisms and institutions created to reduce the risks of proliferation, misuse and promote the peaceful use of nuclear energy. The centre point of this treaties is the Nuclear Non-Proliferation Treaty (NPT), a groundshifting agreement that is structured on three solid pillars such as: Non-Nuclear Weapon States (NNWS) agree not to acquire nuclear weapons, Nuclear Weapon States (NWS) commit to pursuing good-faith negotiations on nuclear disarmament and all signatory states have the right to peaceful nuclear technology, provided they adhere to strict safeguards.

Understanding the highly sensitive nature of the nuclear sector, the Independent Atomic Energy Agency (IAEA) is charged with the function of compliance through on-site inspections, material accountancy, containment and surveillance, thereby reinforcing the assurance that nuclear activities are consistent with peaceful purposes. The Agency's executive powers to inspect and demand full declaration of nuclear facilities are enforced on States that have signed the Comprehensive Safeguards Agreements and the Additional Protocol.

Sadly, these Safeguards measures are more robust on paper than in practice as there have been cases of illicit nuclear trade, smuggling nuclear materials to obtain enrichment or reprocessing technology which are difficult to track until they are already advanced. This nuclear proliferation activities in the form of enrichment and reprocessing form the core point of this paper. The risks involved are illustrated with when Iraq and North Korea engaged in clandestine nuclear activities in defiance of laid down guiding frameworks. Uranium enrichment facilities were concealed, research reactors were diverted and reprocessing plants were building grounds for producing weapons-grade plutonium. The most recent in July 2025 was Iran when the State suspended IAEA inspector access and ceased providing required reports on its nuclear operation, thereby making the IAEA lose track of the stockpiled Highly Enriched Uranium (HEU) and impacted on the continuity of knowledge upon which IAEA relies on for documentation.

This flagrant restriction was not only an affront to the authority of the IAEA but also a stark reminder of how fragile the safeguards system can become in the face of sovereign defiance with no authority to enforce compliance. This also shows the grave impact of when a state withdraws cooperation leaving room for clandestine enrichment or reprocessing and compounding global insecurity. The situation is even more heightened with the rise of Artificial Intelligence as both a solution and a threat in nuclear misuse.

Dual-Use of AI in Mitigating Nuclear Risks

The risks of radiological and nuclear proliferation are profound in uranium enrichment and plutonium reprocessing, two sensitive stages of the nuclear fuel cycle that can be exploited for weapons purposes. Advanced LLM Models and Simulation systems can be exploited by proliferators to overcome the traditional process of enrichment by reducing the number of stages required in a centrifuge cascade, shrink the physical footprint of facilities, and improve yields from fewer centrifuges. This would make clandestine enrichment more efficient, less resource-intensive, and far more difficult to detect, directly undermining the goals of non-proliferation regimes.

On the flipside, AI systems can be used as a monitoring tool or forecasting tool for early warning of proliferation activities. Our proposed monitoring tool is an AI-driven pipeline called SAFE (Sentinel AI Fusion Engine) designed to serve as an early-warning system for nuclear proliferation. It is a strategic tool that addresses the challenge of monitoring a massive volume of open-source data for potential threats. This prototype demonstrates a scalable and efficient solution to a critical global security challenge. It successfully combines web-based data with AI-powered analysis to turn raw information into a vital strategic asset.

A Quick Guide to how SAFE Works

The prototype is a fully automated pipeline built on the VectorShift platform. It operates in a continuous cycle:

Data Ingestion: It automatically pulls real-time intelligence by running targeted Google searches for keywords related to nuclear activity.

AI Analysis: The raw search results are sent to an integrated Large Language Model (LLM), which acts as a virtual national security analyst. The LLM evaluates the text for signs of clandestine activities or undeclared infrastructure.

Threat Detection: The AI's analysis is then fed into a conditional logic gate. This gate automatically filters for high-risk keywords, ensuring that only potential threats proceed down the pipeline.

Automated Alerting: If a threat is detected, the system sends an immediate email alert to a monitoring team. This provides a real-time, actionable notification that bypasses the need for manual, time-consuming data review.

Figures 1 & 2: Statistically support SAFE's initiative with anomaly detection.

Governance Measures for Nuclear Risk Reduction

Several constraints like privacy and data protection, asymmetric access to intelligence between government and intelligence agencies (e.g., Five Eyes), dual-use limitation that utilises strategic intelligence rather than probabilistic flagging, have imparted on the effective implementation of

monitoring and forecasting predictions. Our proposed governance measures include:

- Establish Clear International Protocols for AI Oversight. Develop multilateral agreements under the IAEA or UN framework to define who controls AI-enabled monitoring tools, how their outputs are verified, and under what conditions they are admissible as legitimate evidence in compliance assessments.
- Mandate Human-in-the-Loop Safeguards. Require that AI outputs be reviewed by trained analysts, with institutional checks to ensure data provenance, robust testing, and to algorithmic bias prevention and to ensure accountability in high-stakes nuclear monitoring decisions.
- Implement Standards for Transparency and Data Integrity. Create binding technical standards that mandate transparency in AI model design, dataset provenance, and auditability, reducing the risk of manipulation while building trust among states.
- Incorporate a civilian-led cooperative deterrence strategy. This would consist of multilateral cooperation among industry and research labs, and government entities, thereby enhancing collective risk awareness and containment capabilities.
- Strengthen IAEA Enforcement Authority. Enforce stricter compliance mechanisms and impose penalties for withdrawal of inspector access. Create alternative on-ground inspection processes that will be implemented immediately in cases of withdrawn or restricted inspection. Integrate a rigorous licensing and certification system that places stringent requirements on enrichment and reprocessing facilities.
- Collective Data Sharing. Establish international data-sharing protocols and encourage collaborative verification efforts to stop deception attempts.

Guidance for aligning AI with nuclear nonproliferation norms

Modern AI tools must proactively forecast, audit, and model trustworthiness and accuracy, leveraging on pretrained datasets to produce authentic results. These are proposed guidelines that can be adopted towards a safe use of AI systems being integrated into nuclear nonproliferation norms.

Guideline #1: AI Misuse Forecasting and Early Warning

Develop an AI-based forecasting layer that identifies risks by cross-referencing OSINT from satellite imagery, trade records, cloud compute patterns, and academic preprints with official reporting. The system should track and identify inconsistencies in material declarations, and forecast increases in capacity for enrichment and reprocessing via anomaly detection. Model and rate transparency based on previous audits, political volatility, cybersecurity history, and IAEA compliance. Use Cases would include R&D detection from nonsignatories and adversarial states. Risk scoring based on research publication trajectory, hiring, job postings, and cross-border collaboration and identification of underreporting patterns tied to procurement anomalies. Also,

encourage states to adopt forecasting and red-teaming exercises to identify vulnerabilities in safeguards before they are exploited.

Figure 3: Demonstrative representation of research trajectory from results for “Iran” “Nuclear” across publication platforms.

Guideline #2: AI Auditing within CBRN Workflows

Model misuse detection should be aligned with existing accounting systems, use model evaluation pipelines and red-teaming strategies to test simulation tools, lab models, and design applications. Audits could track supply chain activities including uranium mining, isotope handling, centrifuge part manufacturing. Centralize anomaly reporting in national data lakes and enable cross-agency collaboration. The system would require Human-in-the-loop governance by connecting foreign affairs, defense, intelligence, and regulatory bodies in a rapid-response task force. Escalation in the system would lead to sanctions, inspections, or diplomatic escalation.

Guideline #3: Develop a Global AI-for-Safeguards Initiative.

This will help to accelerate responsible integration of emerging technologies into verification systems. Mandate dual-use risk assessments for enrichment and reprocessing-related AI applications. Expand the scope of export controls on advanced centrifuge components, machine learning software, and dual-use sensors.

References

- Alger, J. (2008). *A Guide to Global Nuclear Governance: Safety, Security and Nonproliferation*. Nuclear Energy Futures Special Publication, Centre for International Governance Innovation.
<https://www.files.ethz.ch/isn/104891/A%20Guide%20to%20Nuclear%20Power.pdf>
- Johnson, J. (2023). *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age*. Oxford University Press, Oxford.

Louie, A. (2010). *Multimodality Imaging Probes: Design and Challenges*. National Library of Medicine. <https://doi.org/10.1021/cr9003538>

Manheim, D., Martin, S., Bailey, M., Samin, M., & Greutzmacher, R. (2025). The Necessity of AI Audit Standard Boards. *AI & Society*, 1-16.

NPT Safeguards Agreement with Iran: E3 Statement to the IAEA Board of Governors, September 2025.

<https://www.gov.uk/government/speeches/npt-safeguards-agreement-with-iran-e3-statement-to-the-iaea-board-of-governors-september-2025>

World Nuclear Association. *Safety and Security: Safeguards to Prevent Nuclear Proliferation*. <https://world-nuclear.org/information-library/safety-and-security/non-proliferation/safeguards-to-prevent-nuclear-proliferation>

Appendix I - Security Considerations

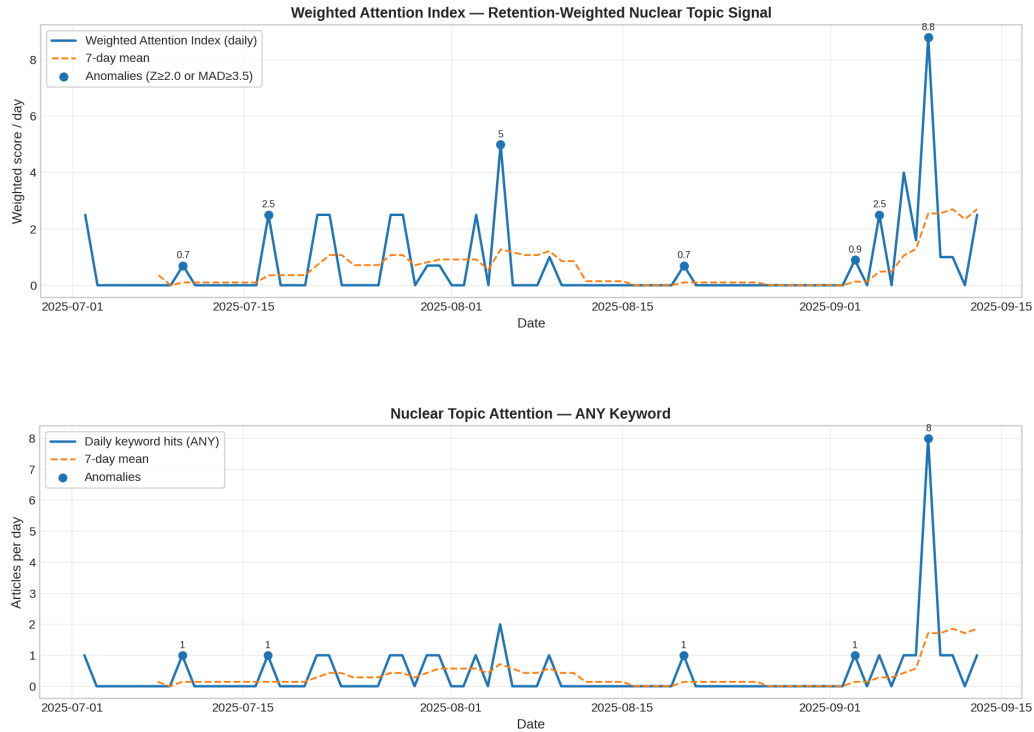
1. Classified data is excluded.
2. Gathering an international community does not guarantee transparency and HEU producers may obscure diversion from the community.
3. A significant analysis of interpretability methods that would distinguish legitimate dual-use indicators from diversion would require more time than is available in this paper.
4. Limitations on the number of pages to explain our thought process and discuss modern approaches to mitigate nuclear risks.

Appendix II

Prototype Demo: Early Warning Monitoring System Workflow for SAFE

https://drive.google.com/file/d/1-RF26YycVwx-eTvK59KbJhH9cuNev3AN/view?usp=drive_link
[nkDemo](#)

Figures 1 & 2: Statistical Anomalies Detection



<https://www.kaggle.com/code/saahirvazirani/safety-measures-nuclear-nonproliferation>

Figure 3: Research Results Demonstration

