RESEARCH UPDATE

FRONTIER MODEL FORUM

Research Update: Frontier AI and Nuclear Security

DATE

January 30, 2026

Introduction

The rapid progress of AI presents potential dual-use concerns for the security of civilian nuclear materials and facilities, referred to below as "nuclear security." Advanced AI systems have the potential to enhance nuclear security, including by improving predictive maintenance within nuclear energy facilities.[1] But frontier AI may also introduce novel risks to the nuclear security ecosystem or heighten existing ones. For instance, the novel capabilities of frontier AI may enable malicious actors to overcome physical, technical or logistical barriers that have previously made it difficult to acquire nuclear materials or sabotage nuclear facilities.

Over the past year, the Frontier Model Forum (FMF) carried out preliminary research into those risks in collaboration with nuclear security experts. This update provides a high-level summary of our initial findings. It highlights the risk of frontier AI acting as a *possible accelerant* to existing threats to critical nuclear infrastructure rather than creating entirely new ones, for example by lowering the barrier for malicious actors to plan high-consequence attacks. In addition, the research update also aims to advance an awareness of why frontier AI risk management in the nuclear security domain is uniquely complex, especially given the highly regulated environment, the sensitivity of nuclear information, and the attendant uncertainty about which frontier AI safeguards should be developed and deployed.

Unlike biosecurity and cybersecurity, the nuclear domain benefits from inherent physical barriers and mature security frameworks that substantially mitigate frontier AI-related risks to materials and facilities. Yet proactively assessing and addressing risks at the intersection of frontier AI risks and nuclear security is nonetheless essential. Based on extensive engagement with the nuclear security

community, this research update aims to inform greater understanding of those risks and establish a foundation for future collaboration on frontier AI risk management between the frontier AI and nuclear security communities.

Key Risks in Nuclear Security

To better assess the risks of frontier AI with respect to civilian nuclear security, the FMF partnered with the Vienna Center for Disarmament and Non-Proliferation (VCDNP) to host a series of virtual and in-person workshops in 2025. Bringing together leading experts from the AI and nuclear security communities, the workshops aimed to identify the risks that frontier AI may pose to the security of nuclear materials and facilities and to consider potential mitigations for jointly addressing them. The discussions were narrowly scoped to the potential for frontier AI to uniquely uplift malicious human actors to achieve harmful, high-consequence outcomes in the nuclear domain.[2] At the onset, experts highlighted that one key challenge in assessing frontier AI risks to nuclear security is the difficulty of making progress without extensive nuclear expertise and access to highly classified information. However, this problem can be effectively avoided by focusing instead on issues that are upstream from weapons development: namely, the physical and cyber security of nuclear facilities and materials.

Across the convenings, nuclear security experts consistently reiterated the following outcomes as the most severe and high-risk:

- **Theft of nuclear material** in a form and quantity potentially usable for a nuclear device.[3] This requires **access to nuclear material,** specifically the physical stores of high-quality nuclear material held inside nuclear facilities.
- **Successful sabotage of nuclear facilities** resulting in severe harm to people and the environment.[4] This requires **access to critical systems** within nuclear facilities that, if tampered with, could result in a massive radiation release.

For frontier AI risk management, the most salient question is whether frontier AI capabilities can meaningfully increase the ability of a malicious actor to steal nuclear material or sabotage a nuclear facility. As experts repeatedly stressed, however, answering that question requires understanding the existing operational and technical barriers that have historically prevented malicious actors from accessing nuclear material and the key critical systems of nuclear facilities. This in turn depends on an awareness of existing approaches to nuclear security, which are briefly outlined below.

Existing Approaches to Nuclear Security

Conventional approaches to nuclear security have expanded over time. At its outset, the civilian nuclear security industry primarily focused on the physical protection of nuclear facilities and materials.[5] However, as more information about nuclear facilities became digitized and made publicly available online, data security and cyber security became of increasing importance to securing nuclear materials.[6]

Nuclear security has long rested on two foundational strategies: “Defense in Depth” and “Graded Approach.” The former refers to multiple, independent, and layered physical and cyber security systems that protect any single target, such that an adversary must overcome several distinct barriers to succeed – i.e., no single point of failure.[7] The latter means that the most sensitive targets, such as weapons-usable nuclear

material, receive the most stringent levels of security compared to lower-consequence targets.[8]

The primary goal of each strategy is to prevent attacks before they can be executed. Proactive measures such as restricting sensitive data about facility layouts, transport schedules, and security system configurations aim to prevent attacks by making it more difficult for a would-be attacker to select a target and develop an effective attack plan. If an attack nonetheless occurs, the strategies also stress the importance of detection, delay, and response. Detection relies on physical controls such as video cameras to monitor facility entry points, as well as cyber controls to monitor for online or digital intrusions. To delay an ongoing attack, facilities also incorporate multiple physical and cyber barriers, such as access restrictions to sensitive areas and air-gapping critical digital control systems. This layered defense is designed to be resilient and to extend the window for an effective response.[9]

For frontier AI risk management, a key implication is that the effectiveness of these approaches depends in part on limiting malicious actors' access to information about how specific nuclear materials and facilities are secured.

Potential Frontier AI Capabilities of Concern

As noted below, it is inherently difficult to determine whether frontier AI models and systems could substantially increase nuclear risks. Any elaboration of potential capabilities of concern should therefore be taken as preliminary and in need of further research.

Experts in both frontier AI and nuclear security nonetheless identified three potential areas where AI capabilities could theoretically erode long-standing security bottlenecks. Notably, the capabilities described here are not uniquely germane to nuclear risks, and are relevant to risks and benefits in other domains, including other types of critical infrastructure:

- **Information Synthesis for Vulnerability Discovery**: Frontier AI's ability to rapidly aggregate, process, and synthesize vast quantities of disparate, publicly available data could help adversaries identify potential vulnerabilities in nuclear facilities. This includes making novel inferences from complex information about a facility's operations, physical layout or broader supply chain, thereby lowering the expertise and resources required to develop a viable attack plan.

- **Sophisticated Attack Planning and Execution**: AI might assist malicious actors in generating and refining complex attack plans. This capability could be used to model and simulate different scenarios, helping adversaries devise strategies to overcome layered security measures, deduce the physical layout of a sensitive facility, or identify novel vectors of physical or cyber attack that are not immediately obvious.

- **Advanced Human and System Manipulation**: AI might be used to craft highly convincing disinformation or sophisticated social engineering campaigns targeting personnel. By generating realistic but false communications or operational data, an adversary could seek to deceive staff, manipulate systems, or create confusion during a security event to exploit vulnerabilities. This capability could be particularly useful for identifying and leveraging potential insider threats.

Note that each of these bottlenecks depend on a heightened ability to retrieve and synthesize information about nuclear facilities and personnel. While high transparency with the public about nuclear safety and operations has long been a feature of the civilian nuclear sector, it may have posed less risk in the past when information often had to be actively requested (rather than being readily available online) and manually aggregated and processed. The nuclear security community should increasingly consider how information about nuclear materials and facilities may be retrieved, aggregated, and synthesized.

Future work on the impact of frontier AI on nuclear security should also consider examining threat pathways that result from the capabilities above.

Challenges for AI-Nuclear Risk Management

Experts raised several key considerations when moving forward with frontier AI risk management processes. These considerations stem from industry characteristics that may not be present in other domains and may therefore result in a different set of risk management processes.

First, the nuclear sector is a highly regulated and security-conscious industry, where much of the critical information about nuclear security protocols, facility vulnerabilities, and threat scenarios is either sensitive or classified, depending on the country and facility in question. This restrictive informational environment, while important for security, creates a significant barrier for private sector companies like AI developers to identify and address any risks created at the model or system-level. Without access to realistic and detailed information about the threats and vulnerabilities, it can be difficult to build an understanding of the potential pathways to harm, design effective evaluations to test for dangerous capabilities, and establish meaningful thresholds for what constitutes a risk.

Second, as noted above, many of the risks associated with AI and nuclear security stem from circumventing critical infrastructure security systems, including both the physical and virtual barriers to preventing attacks. This presents complications for frontier AI risk assessment because, unlike adjacent risk domains like biology or chemistry, it can be difficult to identify domain-specific knowledge that is harmful in isolation from the threats to critical infrastructure.[10] Since the principal bottleneck for a malicious actor is access to high-quality nuclear material rather than hazardous nuclear knowledge, it may be more appropriate to focus most heavily on stress-testing or strengthening the physical or cyber security measures of nuclear facilities and materials instead of frontier AI capabilities.

Finally, international guidance on nuclear security currently recommends the implementation of strong measures preventing the illegitimate access and misuse of nuclear materials. While frontier AI may diminish the relative strength of these measures, there is substantial uncertainty about what further measures may be needed to fill these gaps, and further, which industry is best-placed to implement them. For example, one useful defence against AI-enabled threats to nuclear security might be to limit the information that models can collate and process. To address these challenges, safeguards could be applied at the AI model- or system-level (e.g. by removing certain information from models[11] or implementing classifiers to flag keywords), or at the ecosystem level, by strengthening the existing information security practices to limit the sensitive public information available to models.

Conclusion

The intersection of frontier AI and nuclear security poses complex and evolving challenges, as well as many open questions. The primary risks to be managed may not stem from frontier AI creating novel threats, but instead from its potential to act as a powerful accelerant, empowering malicious actors to more effectively exploit existing information and circumvent established security protocols. The initial analysis presented here, based on preliminary expert discussions, underscores that mitigating these risks may require a holistic approach combining technical safeguards at the model or system level with strengthened security measures across the nuclear ecosystem.

Sustained, collaborative research is essential to stay ahead of this threat. This research update represents an early step in that process. Future work, including through the FMF's AI-Nuclear Workstream, may focus on several key areas:

- **Developing Shared Threat Models:** Working with government and civil society partners to build a more granular, shared understanding of the most plausible AI-enabled threat scenarios.
- **Creating Shared Evaluation Methodologies:** Designing domain-specific evaluations to test AI models for dangerous capabilities related to nuclear security without using sensitive or classified information.
- **Facilitating Greater Information-Sharing:** Establish more opportunities for the AI and nuclear security communities to collaborate, share relevant threat information, and inform risk assessment and mitigation strategies.

Effective risk management for AI-nuclear will depend on building durable collaborations between the frontier AI and nuclear security communities. Nuclear security experts are uniquely able to inform AI developers about credible threat scenarios without revealing sensitive details, while AI developers can best help the nuclear security community understand the actual capabilities and limitations of frontier models. Sustained engagement between communities, building on approaches that have proven effective in bridging the frontier AI and biosecurity ecosystems, will be essential for developing effective, evidence-based risk assessments and mitigation frameworks and will help drive clarity about threat models and outcomes of concern. The FMF aims to further collaborate with the nuclear security community in the future.

FOOTNOTES

1. For predictive maintenance, see Lin, L., Walker, C., & Agarwal, V. (2025). Explainable machine-learning tools for predictive maintenance of circulating water systems in nuclear power plants. *Nuclear Engineering and Technology*, *57*(9), Article 103588. https://doi.org/10.1016/j.net.2025.103588. Beyond nuclear energy security, advanced AI may also augment the use of nuclear technology for medical imaging. See Cheng, Z., Wen, J., Huang, G., & Yan, J. (2021). Applications of artificial intelligence in nuclear medicine image generation. *Quantitative Imaging in Medicine and Surgery*, *11*(6), 2792–2822. https://doi.org/10.21037/qims-20-1078.
2. This document does not address the potential vulnerabilities of, or assessment and mitigation measures for, applications of frontier AI models and systems used in nuclear facilities (e.g. for command, control, and communications).

3. This refers primarily to "Category I" nuclear material, which consists of 2 kg or more of unirradiated plutonium or uranium-233, or 5 kg or more of unirradiated uranium-235 enriched to 20% U-235 or more. See Annex II in Amendment to the Convention on the Physical Protection of Nuclear Material, INFCIRC/274/Rev.1/Mod.1 (Corrected), International Atomic Energy Agency (1979), https://www.iaea.org/sites/default/files/publications/documents/infcircs/1979/infcirc274r1m1c.pdf for more detail.
4. Nuclear facilities are defined as "a facility (including associated buildings and equipment) in which nuclear material is produced, processed, used, handled, stored or disposed of." See IAEA (2022), Nuclear Safety and Security Glossary https://www-pub.iaea.org/MTCD/Publications/PDF/IAEA-NSS-GLOweb.pdf
5. This is an approach often referred to as "guns, gates, and guards." See University of Tennessee, Knoxville. (n.d.). *What is nuclear security?* Nuclear Engineering. https://nuclear.utk.edu/what-is-nuclear-security/
6. International Atomic Energy Agency. (2016, December 9). *Guns, guards, gates and geeks: Romania strengthens computer security at nuclear installations*. https://www.iaea.org/newscenter/news/guns-guards-gates-and-geeks-romania-strengthens-computer-security-at-nuclear-installations
7. International Atomic Energy Agency. (2013). *Objective and essential elements of a state's nuclear security regime* (IAEA Nuclear Security Series No. 20). https://www-pub.iaea.org/MTCD/Publications/PDF/Pub1590_web.pdf, p. 11.
8. International Atomic Energy Agency. (2011). *Nuclear security recommendations on physical protection of nuclear material and nuclear facilities* (IAEA Nuclear Security Series No. 13). https://www-pub.iaea.org/MTCD/Publications/PDF/Pub1481_web.pdf, p. 14. See also Immonen, E. (2023). *Graded approach to nuclear safety - State of the practice* (VTT Research Report No. VTT-R-00996-22). VTT Technical Research Centre of Finland, p. 4.
9. For a good overview of detection, delay, and response, see National Research Council. (2002). *Making the nation safer: The role of science and technology in countering terrorism* (Chapter 4). The National Academies Press. https://doi.org/10.17226/10415.
10. Both nuclear security and frontier AI security experts repeatedly stressed that many of the risks associated with the misuse of frontier AI in the nuclear security domain are similar to misuse risks related to critical infrastructure broadly. As such, stress-testing and securing the physical and cyber security access points and safeguards at nuclear facilities should therefore be considered key elements of risk management.
11. Although data filtering shows promise, the effectiveness of data-level mitigations to reduce frontier risk continues to be an open research question.