VCDNP Vienna Center for Disarmament and Non-Proliferation

December 2025

Artificial Intelligence and Nuclear Security Governance: Addressing the Risks of Frontier AI

Dr. Sarah Case Lackner
Zaheed Kara

Authors

Dr. Sarah Case Lackner is a Senior Fellow at the VCDNP. Her work focuses on nuclear security and its interactions with AI and other emerging and disruptive technologies. Among other positions, she was a Senior Nuclear Security Officer at the International Atomic Energy Agency (IAEA), serving as the Scientific Secretary for the Nuclear Security Guidance Committee and the Director General's Advisory Committee on Nuclear Security. She also served as Co-Scientific Secretary of the 2022 Conference of Parties to the A/CPPNM.

Zaheed Kara is an AI safety researcher focused on risk management for frontier AI. He is particularly interested in the design and implementation of Frontier AI Frameworks to mitigate CBRN risks and has experience designing and leading multi-stakeholder AI-biosecurity and AI-nuclear security work in the AI industry. Zaheed was previously a senior research associate at the Centre for International Governance Innovation (CIGI), focused on global policy solutions for frontier AI, and a policy analyst in the Government of Canada's AI Hub, working on international AI policy.

About the VCDNP

The Vienna Center for Disarmament and Non-Proliferation (VCDNP) promotes international peace and security by conducting research, facilitating dialogue, and building capacity on nuclear non-proliferation and disarmament.

The VCDNP is an international non-governmental organisation, established in 2010 by the Federal Ministry for European and International Affairs of Austria and the James Martin Center for Nonproliferation Studies at the Middlebury Institute of International Studies at Monterey.

Our research and analysis provide policy recommendations for decision-makers. We host public events and facilitate constructive, results-oriented dialogue among governments, multilateral institutions, and civil society. Through in-person courses and online resources on nuclear non-proliferation and disarmament, we train diplomats and practitioners working in Vienna and around the world.

Acknowledgements

This paper presents independent research emerging from a workshop series funded by the **Frontier Model Forum**. The findings and views expressed are those of the authors.

VCDNP
Vienna Center for Disarmament and Non-Proliferation

Andromeda Tower, 13/1
Donau-City-Strasse 6
1220 Vienna
Austria

vcdnp.org
info@vcdnp.org
@VCDNP
VCDNP

New capabilities in AI are built on recent advancements in machine learning techniques combined with access to an unprecedented scale of data and computational resources.

Background

The field of artificial intelligence (AI)[1] has made dramatic progress over the last decade. AI models – computational systems that use machine learning to recognise patterns in data and make predictions or decisions about novel data – have burst into the public consciousness in the last few years as tools that have the potential to simplify and automate many aspects of industry, business, and everyday life. Their new capabilities are built on recent advancements in machine learning techniques combined with access to an unprecedented scale of data and computational resources, enabling them to learn from bodies of recorded information rather than being explicitly programmed.

However, while frontier AI has the potential to provide substantial benefits, its ever-increasing capabilities might also be misused by criminals and other malicious actors. For example, an AI model might prove to be a valuable resource for such actors seeking to construct and target a weapon of mass destruction (WMD).[2] The scientific, policy, and AI communities recognise this risk.[3]

1 For a brief introduction to AI models and systems for policymakers, see Case-Lackner, S and Kara, Z, Artificial Intelligence Models and Systems, Emerging Tech Brief No. 1, October 2025 (VCDNP). (https://vcdnp.org/wp-content/uploads/2025/10/Background-AI-Models-and-Systems.pdf)

2 For more information, see The Role of AI in Reducing the Risk of Weapons of Mass Destruction, Johns Hopkins University, 1 August 2025 (available online at: https://washingtondc.jhu.edu/news/ai-wmd-risk-reduction)

3 See, for example OpenAI's Preparedness Framework (available online at: https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf), the International AI Safety Report, updated October 2025 (https://internationalaisafetyreport.org/sites/default/files/2025-10/first-key-update_0.pdf/)

AI models are increasingly able to rapidly and autonomously gather information, process it, identify useful patterns, and generate novel content, making them a valuable tool for a wide range of criminals.[4] Recently, there has been an increase in discussion on how to prevent AI models – particularly frontier or cutting-edge AI models[5] – from being used to uplift, or provide new capabilities to, a criminal or malicious actor bent on creating weapons of mass destruction, whether nuclear, chemical, or biological.[6,7]

The frontier AI community has produced valuable advice focused on preventing these models from being of material assistance in the construction of biological weapons,[8] and there has been increasing discussion of how to reduce the potential of frontier AI models to assist in the development of chemical weapons. However, there has been less discussion in the AI community of the nexus of AI models with the potential development and construction of nuclear weapons. This shortfall may be related to the perceived difficulty in making progress on this path without extensive expertise in nuclear science and technology and access to classified and highly sensitive information.[9]

However, significant progress in preventing frontier AI models from assisting a malicious actor in constructing a nuclear weapon or using a nuclear facility as a radiological weapon can be made without the need to discuss the technical details of the devices themselves. In short, this can be accomplished by focusing on the most urgent concern: the security of nuclear materials and facilities.

Constructing a modern nuclear weapon is a difficult and complex undertaking, likely to be beyond the capabilities of a non-State actor. Nonetheless, there is a possibility that a non-State actor would be able to construct a crude nuclear device using information that is already publicly available.[10] Thus, access to nuclear materials is generally considered to be the main bottleneck to the construction of a nuclear weapon by such an actor. These materials are largely kept in highly secured facilities around the world, protected by armed guards, fortified barriers, and other measures (including cyber security measures). Preventing a non-State actor from gaining access to nuclear material has traditionally focused on ensuring the physical security of the facility housing the materials using measures such as “guns, gates and guards”, and more recently, incorporating cyber security measures where necessary.

In addition to protecting the facilities housing weapons usable materials, other high-consequence nuclear facilities also need to be protected from attack. Notably, an adversary might seek to sabotage a nuclear power plant, due to its potential to release a massive amount of radiation outside the facility perimeter if successfully sabotaged.[11] This makes protecting such facilities a high priority for States and for operators. Typical security measures involve preventing physical access (and cyber access, if digitised) to key safety systems that stop such a radiation release from occurring.

Thus, when considering the risk of AI models enabling a malicious actor to create either a crude nuclear device or cause a massive radiological release from a facility, a key first step would be to understand whether or not they

4 Joe Burton, Ardi Janjeva, Simon Moseley and Alice, AI and Serious Online Crime, Alan Turing Institute Center for Emerging Technology and Security, March 2025. (https://cetas.turing.ac.uk/sites/default/files/2025-03/cetas_research_report_-_ai_and_serious_online_crime_0.pdf).

5 These are, according to the Frontier Model Forum, “those general purpose AI models that constitute the state of the art, a collection which will shift over time as the field progresses.”See: https://www.frontiermodelforum.org/about-us/.

6 Contemporary AI foundation models increase biological weapons risk, Brent, R. an McKelvey, G., Submitted to arXiv on12 June 2025 (https://arxiv.org/abs/2506.13798/).

7 Mitigating Risks at the Intersection of Artificial Intelligence and Chemical and Biological Weapons, RAND, 28 January 2025 (https://www.rand.org/pubs/research_reports/RRA2990-1.html).

8 See, for example, A Preliminary Taxonomy of AI-Bio Misuse Mitigations, Frontier Model Forum, 30 July 2025 (https://www.frontiermodelforum.org/issue-briefs/preliminary-taxonomy-of-ai-bio-misuse-mitigations/).

9 See Issue Brief: Frontier AI and Nuclear Security, Frontier Model Forum (in publication).

10 Bunn, M and Weir, A, Terrorist Nuclear Weapon Construction: How Difficult? The Annals of the American Academy of Political and Social Science, Vol. 607, Confronting the Specter of Nuclear Terrorism (Sep., 2006), pp. 133-149 (https://scholar.harvard.edu/files/matthew_bunn/files/bunn_wier_terrorist_nuclear_weapon_construction-_how_difficult.pdf).

11 The possible outcome of such an event can be seen in the safety event that occurred at the Chernobyl nuclear facility in April of 1986, where the radiation effects are still significant nearly 40 years later.

can assist that actor in overcoming the security measures that protect nuclear materials from theft and prevent the sabotage of nuclear facilities.[12]

While assessing and mitigating this risk will need to be a priority for both AI developers and international nuclear security experts,[13,14] there has so far been little interaction between the two communities. To address this lack of interaction, over the second half of 2025, the Vienna Center for Disarmament and Non-Proliferation (VCDNP) collaborated with the Frontier Model Forum (FMF) to bring together nuclear security experts and frontier AI developers to:

- Identify key risks and challenges posed by frontier AI models to the security of nuclear materials and facilities, and
- Consider potential mitigations and next steps for the frontier AI community and the nuclear security community to jointly address these risks and challenges.

To enable fruitful discussion on these topics, the VCDNP and FMF conducted two virtual meetings between representatives of the AI industry and international nuclear security experts in July and October 2025, as well as a number of consultations with various experts from both communities. These discussions, as well as the individual and group consultations, focused on the following questions:

1. Which frontier AI capabilities are considered to pose the most risk for uplift of a malicious actor planning a theft of nuclear material or high-consequence sabotage of a nuclear facility?

2. What is needed for and unique to nuclear security risk assessment of frontier AI, as compared to overlapping fields like infrastructure security and cybersecurity?

3. What mitigation measures (model-level, system-level or nuclear security approaches and measures) would most effectively address these risks? Which communities have the ability to implement them?

The scope of the project was limited to capabilities that frontier AI models could provide to a malicious actor when seeking to carry out a very high-consequence attack. In particular, it focuses on misuse of generative, general-purpose AI models, potentially with reasoning and multimodal capabilities, as well as agentic systems, although other types may also pose risks.

From the nuclear security perspective, theft of Category I nuclear material[15] and sabotage of facilities resulting in unacceptable radiological consequences[16] were the primary focus. Further, the exploitation of potential vulnerabilities of AI models and systems integrated into nuclear facilities was explicitly not addressed, to maintain a clear focus for this initial project.

This report summarises and builds on the results of these discussions facilitated by the VCDNP and FMF. It is intended to provide useful information as well as initial considerations and recommendations for nuclear security professionals on managing the coming risks that frontier AI models may pose to security at nuclear facilities around the globe.

12 This is also true for facilities housing high-activity radioactive materials, however, to maintain a well-defined scope, the current report focused only on nuclear materials.
13 Hewes, M., How Artificial Intelligence will Change Information and Computer Security in the Nuclear World, IAEA Bulletin (2023) (https://www.iaea.org/bulletin/how-artificial-intelligence-will-change-information-and-computer-security-in-the-nuclear-world).

14 Case Lackner, S. and Zarka, M, Nuclear Security and the Nuclear Supply Chain in the Age of AI, VCDNP, April 2025 (https://vcdnp.org/wp-content/uploads/2025/04/VCDNP_AI-and-Security-of-the-Nuclear-Supply-Chain_web.pdf).

15 This is generally considered to be the category of material that could lead most directly to the development of an improvised nuclear weapon. According to Annex II of the Amendment to the Convention on the Physical Protection of Nuclear Material (INFCIRC/274/Rev.1/Mod.1 (Corrected)), Category I nuclear material 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. (https://www.iaea.org/sites/default/files/publications/documents/infcircs/1979/infcirc274r1m1c.pdf).

16 As defined in the IAEA Nuclear Safety and Security Glossary (2022), unacceptable radiological consequences are “a level of radiological consequences, established by the State, above which the implementation of nuclear security measures is warranted.” (https://www.iaea.org/publications/15236/iaea-nuclear-safety-and-security-glossary) This definition varies by State, but is generally considered a level of consequences that is high enough to cause harm to populations outside the perimeter of the facility.

Unit 1-2 of Korea Shin-Kori Nuclear Power Plant. Credit: Korea Shin-Kori NPP.

Frontier AI's Challenges for Global Nuclear Security

As noted in the previous section, the key bottleneck for a malicious actor seeking to develop a crude nuclear device is limited access to nuclear material, which is typically held in highly secured nuclear facilities.[17] At the same time, preventing sabotage of a nuclear facility resulting in unacceptable radiological consequences depends on maintaining security at that facility. Thus, a key challenge for a malicious actor seeking to develop a nuclear weapon or turn a nuclear facility into a radiological weapon is to find methods to overcome security (and potentially safety) measures at an identified nuclear facility.

A useful framework for considering an adversary's attack on a nuclear facility or another target involves three phases:

- **Target selection**, in which the adversary determines which facility to attack, based on information about the facility, such as the types and quantities of nuclear material and the safety, security, and other systems operating in it;
- **Attack planning and skill building**, in which the adversary develops a detailed attack plan and acquires[18] the skills needed to conduct the attack; and
- **Execution**, in which the attack is put into action.

17 The necessary skills may be self-developed, automated, purchased or obtained through other means.

18 A nuclear facility is "a facility (including associated buildings and equipment) in which nuclear material is produced, processed, used, handled, stored or disposed of." IAEA Nuclear Safety and Security Glossary (2022) (https://www-pub.iaea.org/MTCD/Publications/PDF/IAEA-NSS-GLOweb.pdf).

When considering how frontier AI models could provide new adversary capabilities or enhance existing capabilities, there are three key issues: (1) how can frontier AI models assist an adversary with target selection; (2) how can they assist with attack planning and skill building; and (3) how can they assist with the actual execution of an attack.

Risk-Relevant Capabilities of Frontier AI

In our discussions with both AI and nuclear security experts, the primary risk was generally considered to be the ability of frontier AI models to draw useful conclusions by rapidly aggregating and processing vast quantities of detailed information. In the civilian nuclear sector, much relevant information is already publicly available regarding nuclear facilities and their safety, security, and operational systems, from sources such as regulatory applications, environmental and siting evaluations, or legal proceedings. While this information is too vast in quantity, detailed, and technical for most human adversaries to draw useful conclusions from it without machine assistance, the aid of a frontier model in gathering information, analysing potential weaknesses and "filling in" redactions might significantly increase the capabilities of even a less capable actor.

It was also noted that the code generation capabilities of frontier AI models could be of value in facilitating cyber-attacks on nuclear facilities, whether isolated or as support for physical attacks, for example, by disabling perimeter detection systems or gathering sensitive facility information.

It was generally agreed that the main uplift by frontier AI models is likely to be in the target selection and attack planning phases, where information gathering and processing would be most useful. These capabilities are most likely to enhance and accelerate the ability of an attacker to choose an appropriately weak target to improve the chances of success, rather than present entirely new modes of attack.

Some experts noted, however, that the ability to rapidly sift through large amounts of data is not unique to the most recent generation of frontier AI. Some uncertainty therefore remains regarding how much unique and novel uplift the most recent generations might provide to potential attackers. However, the very definition of frontier AI models is dynamic and depends on new developments. Thus, when frontier AI models develop new and powerful capabilities, novel risks may emerge.

The following sub-sections set out some specific considerations relevant to risks from frontier AI models in each of the three attack phases set out previously. These considerations are intended to provide an initial basis for nuclear security professionals to research and improve their understanding of novel and evolutionary risks from frontier AI to the facilities for which they are responsible.

Target Selection

The primary concern with respect to target selection is the ability of a frontier model to aggregate disparate public data to identify vulnerable nuclear targets, as well as to identify specific system vulnerabilities that would prioritise one target over another. For example, data available publicly might enable frontier models to draw conclusions regarding:

- **Facility Layout and Security Inferences**: A model may be able to infer sensitive information by combining, for example, publicity information about the facility, regulatory applications and environmental impact statements, and satellite imagery, among other information.
- **System Identification**: Specific types, models, or manufacturers of safety, security, or operational systems used at the facility might be identified, as well as potential vulnerabilities or exploits.
- **Analysis Gap Identification**: A model may be able to review a facility's own safety and security analyses to identify logical gaps, unexamined assumptions, or overlooked failure modes that create an exploitable vulnerability.

The information made publicly available and accessible to the model may not need to be specific to the nuclear facility being assessed, as information on other facilities known to be of the same design or layout may be able to provide key information on the facility being assessed.

Attack Planning and Support

The primary concern with respect to attack planning and support is the generation of specialised knowledge and plans that would be required to carry out an attack. For example, the following paths could pose concerns:

- **Historical Attack Analysis**: A model may be able to use information from past physical and cyber-attacks to identify security gaps and adversary techniques.
- **Physical Security Disablement**: A model may be able to generate a plan for disabling or circumventing physical security systems, based on manufacturer specifications and known vulnerabilities.
- **System Weakness Exploitation**: A model may be able to generate credible scenarios for sabotaging safety-critical systems or identify common cause failures.
- **Supply Chain Vulnerability Mapping**: A model might be able to map the digital and/or physical supply chain for a critical component of a nuclear facility and/or identify vulnerable targets in the supply chain.
- **Social Engineering**: A model may be able to generate highly realistic and convincing tailored phishing emails or social media campaigns designed to elicit sensitive operational information or recruit an insider from staff known to occupy key roles in the facility.

Attack Execution Support

While a frontier AI model may be able to provide assistance in the execution of an attack on a nuclear facility, it is important to keep in mind that in the case of the highest-risk facilities, physical access is likely to be necessary to complete the attack.

However, nuclear security professionals should account for the potential assistance that a frontier AI model might provide, particularly to enhance an adversary's capabilities. For example, such models could be used for:

- **Enhancing Capabilities for Cyber Attacks**, for example, by writing functional exploit code for a known vulnerability.
- **Falsification or manipulation of data**, with the goal of misleading staff, concealing illicit activity, or otherwise seeking to manipulate facility procedures and processes.

The most effective path forward will be to establish ongoing informal channels of communication between frontier AI developers and the nuclear security community.

Conclusions and Recommendations

Conclusions

Several key conclusions were drawn over the course of the project that can be of value for nuclear security professionals in determining how to best address the nuclear security challenges posed by adversary misuse of frontier AI models.

1. Considering realistic risk scenarios and mitigations will need the involvement of both frontier AI developers and nuclear security experts. While nuclear security operators and regulators can assess some of the risks to their own facilities, input from the frontier AI community will be essential to account for rapidly shifting capabilities of frontier AI. In particular, nuclear security experts need to understand the realistic current capabilities of frontier AI in order to develop their own security measures and mitigations. In parallel, nuclear security experts can help AI developers to understand the highest risks in the nuclear sector and to develop effective AI safeguards to prevent model misuse where possible and necessary. If risk scenarios and mitigations are developed solely by either of the two communities, they run the risk of either not accounting accurately for the technical aspects of the nuclear sector or of incorrectly estimating the capabilities of frontier AI, both now and into the future.

2. Responsibility for mitigating these risks will need to be shared. Frontier AI developers are motivated to take responsibility for ensuring the safety and security of their models.[19] However, the diversity of nuclear facilities around the world and their often stringent secrecy measures, alongside the range of deployment contexts and use cases of frontier AI, makes it unrealistic to design AI safeguards that prevent all possible pathways for misuse of frontier models to aid in attacking nuclear facilities. Thus, some responsibility for mitigating the risks unique to nuclear facilities will also need to fall on the nuclear security community. For example, AI developers should implement AI safeguards to prevent their models from reconstructing non-public information related to sensitive areas such as facility layout or failure modes that create an exploitable vulnerability. However, nuclear security regulators and operators will need to ensure that adequate security measures are in place to account for the information gathering and processing capabilities of frontier AI models. This is an important area for future collaboration across the two communities.

3. The responsibility for mitigating the risks will fall disproportionately on the international nuclear security community. Given the rapid pace of change of the capabilities of frontier AI models, it will be increasingly important to maintain current awareness and analyse how a malicious actor may be able to misuse frontier AI models to assist with an attack on a specific nuclear facility. To be as comprehensive as possible, these analyses should use a systems approach, incorporating the risk to safety, security, and operational systems. In particular, and in light of the evolving challenges posed by frontier AI models, discussion may need to be initiated within States on the appropriate balance between cyber and physical security measures, enhancing resilience of facility operations against attack, and the need for transparency via the release of information on nuclear facilities to the public.

Further, the increasing prevalence of open-source and open-weight frontier AI models highlights the need for the nuclear security community to take the lead in establishing security measures that account for frontier AI capabilities. While for proprietary models it may be possible to work with the developers and establish some broad AI safeguards against misuse, open-source and open-weight models are more vulnerable, as AI safeguards can easily be disabled by a malicious actor. Further, there will always remain a risk that some AI developers – whether for expediency, lack of awareness of the risks, or even malicious reasons – may not include adequate AI guardrails in their models.

Recommendations

1. Establish ongoing informal channels of communication between frontier AI developers and the nuclear security community. Developing robust risk assessments and measures to reduce this risk, whether nuclear security measures or AI model mitigations, will require technical knowledge of both nuclear security and frontier AI models. Few experts in both fields currently exist. Thus, the most effective path forward would be the establishment of an ongoing conversation between the two communities to build technical understanding on both sides.

Via standing informal channels of communication, international nuclear security regulators and policymakers could communicate with frontier AI developers about nuclear security opportunities and challenges related to these models as well as improve their understanding of model capabilities. These channels of communication could also strengthen connections between the two communities and serve to support the development of a core group of experts with knowledge of both nuclear materials and facility protection and frontier AI models. Notably, these channels could serve as a path not only for communication on risks, but also on opportunities through which frontier AI models could improve security of nuclear materials and facilities.

19 See Frontier AI Safety Commitments, AI Seoul Summit 2024. (https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024).

Such communication channels could, for example, take the form of regularly scheduled, informal, non-attributable convenings, providing the opportunity for sensitive information to be shared on both sides. These convenings could be organised by relevant and interested civil society organisations with close links to each community.

2. Develop nuclear security systems and measures that account for the dynamic, systemic, and changing nature of the risks to nuclear security posed by frontier AI models. It is essential that future nuclear security measures manage the new and evolving risk environment associated with frontier AI models, among other new and emerging risks. The capacity and need for these measures will vary from State to State and will need to account for a graded approach. However, the capabilities of frontier models are rapidly changing and their dynamic nature cannot be responded to effectively by a static, checklist approach to nuclear security.

Five elements can serve as a starting point for national considerations as regulators, policymakers, and operators discuss and prepare enhanced and novel strategies to manage these risks:

- **Prioritising data security and management strategies.** The data processing capabilities of frontier AI models mean that developing and implementing strategies for data security and management that account for frontier AI model capabilities will need to be prioritised by the nuclear sector.
- **Taking a systems approach to thinking about nuclear security.** A systems approach incorporating physical protection and cyber security considerations into hazards analyses is increasingly essential in the nuclear sector, given the potential capabilities of frontier AI models to identify common cause failures among multiple types of system and generate falsified or manipulated information.
- **Strengthening and adapting cyber security measures.** Cyber security measures will need to account for the possibility of an AI-enhanced adversary, including with code generation and increasingly autonomous goal-seeking capabilities. As the digital transformation increasingly affects the nuclear sector, this will become only more critical.
- **Adapting the approach to supply chain security.** Measures to establish security in the nuclear supply chain, both digital and physical, will need to account for the evolving capabilities of AI models and systems, including their ability to assist an adversary to masquerade as a trusted supplier.
- **Implementing a strategy that accounts for increasingly sophisticated social engineering attacks.** Social engineering attacks such as spear phishing and AI-driven recruitment of insider threats will require equally sophisticated measures in response, including AI-enhanced detection of phishing emails and stronger insider threat training that accounts for the risks of AI-powered chatbots, among other risks.

VCDNP

Vienna Center for Disarmament and Non-Proliferation

The VCDNP is an international non-governmental organisation that promotes peace and security by conducting research, facilitating dialogue, and building capacity on nuclear non-proliferation and disarmament.

vcdnp.org

@VCDNP

info@vcdnp.org

VCDNP