Executive Summary

Malevolent Use of Artificial Intelligence in Safeguards Implementation

Research proposal by Goffredo Pierbattisti | Specializing Master in Nuclear Safeguards, Politecnico di Milano | 19 October 2026

Artificial intelligence creates a risk to the credibility of nuclear safeguards when an adversary uses it to falsify information, manipulate verification personnel or compromise analytical tools. The proposal examines this risk across safeguards implementation and asks which technical, procedural and governance measures could reduce it. Its intended contribution is a structured threat taxonomy, a semi-quantitative assessment of risks and recommendations for the Department of Safeguards' AI guidelines.

Status of the research

The source is a one-page master's thesis proposal abstract. It states a research question and four objectives, but reports no completed assessment, empirical dataset, quantitative findings or validated mitigation performance. The contributions described here are therefore planned outputs. The abstract envisages work within the IAEA Department of Safeguards in Vienna, subject to arrangements to be agreed with Politecnico di Milano.

Safeguards significance

Safeguards conclusions depend on the credibility of State-declared information, independent verification by inspectors and analysts, and the integrity of information systems. The abstract argues that generative AI can reduce the cost of creating convincing forged documents, synthetic imagery and impersonations. Adversarial machine learning can also degrade the tools used to analyse safeguards information. These threats could affect the evidence and judgements underlying safeguards conclusions.

Research question and scope

The central question is how malevolent uses of AI by State or non-State actors could undermine the integrity of safeguards implementation, and which measures could most effectively minimise that risk. The proposed taxonomy covers three attack surfaces.

- Information used in safeguards conclusions. State declarations, nuclear material accountancy reports, design information, supporting documentation, satellite and surveillance imagery, and open-source information may contain manipulated or synthetic content.
- People performing verification. Inspectors and analysts may face deepfakes, voice cloning and AI-assisted social engineering, with consequences for the confidentiality and integrity of safeguards information.
- The Department's own AI tools. Data poisoning, evasion, prompt injection and over-reliance may affect the use of these tools across their lifecycle.

A cross-cutting concern is the additional capability AI may give an adversary to plan diversion or concealment, including internally consistent falsification across several facilities. The proposal identifies this as a dimension for analysis, without demonstrating a particular scenario or estimating its feasibility.

Proposed assessment approach

The abstract proposes a semi-quantitative evaluation of likelihood, impact on safeguards conclusions and detectability. It does not specify scales, weights, aggregation rules, evidence thresholds or a final ranking. Those choices will determine how reproducible and useful the prioritisation becomes. A research protocol should define the dimensions and record the evidence and uncertainty behind each judgement before scores are compared.

Candidate mitigation measures

At the technical level, the proposal considers data provenance and authentication, digital signatures and content credentials, distributed-ledger integrity for declarations and inspection data, corroboration across multiple sources, and detection tools. These are candidate controls. The abstract does not establish that any one of them can authenticate the truth of a declaration or reliably identify every AI-generated artefact.

At the procedural level, it considers verification protocols, human review, red teaming, staff awareness and incident reporting. Their relevance is to the way personnel handle evidence, challenge suspicious material and respond to incidents. At the governance level, it considers Departmental AI policy and guidelines, information security, cooperation with State and regional authorities, and alignment with international standards and regulatory benchmarks such as the EU AI Act. The reference to the Act does not establish its legal applicability to the IAEA.

Expected contribution

The planned outputs are a safeguards-specific taxonomy, a prioritised account of the assessed risks, and a set of mitigation recommendations that could inform Departmental AI guidelines. The source supports these as research objectives. It does not yet support a conclusion about the most likely threat, the most consequential vulnerability or the effectiveness of a preferred control.

Priorities for developing the study

The following are analytical recommendations for taking the proposal forward. Define the safeguards workflows and trust assumptions within scope, establish a documented scoring method, and test candidate controls using authorised, non-sensitive material. Evaluation should distinguish content authenticity from factual truth and measure both missed manipulation and false alarms. Independent review of the scenarios and scoring would help assess whether the resulting priorities are defensible.

For a safeguards audience, the main proposition is that AI misuse deserves assessment at the level of verification evidence and safeguards conclusions. The research will need to show how specific controls preserve those functions, what residual uncertainty remains and where additional human verification is necessary.

Source: Goffredo Pierbattisti, Minimizing the Risk of Malevolent Use of Artificial Intelligence in Safeguards Implementation: Threat landscape, risk assessment and mitigation measures, master's thesis proposal abstract, one page, undated. Local source: Papers/Minimizing the Risk of Malevolent Use of AI in SG Implementation Abstract - Goffredo Pierbattisti.pages. The recommendations in the final section extend the proposal and are not reported findings.