Research One Pager

Malevolent Use of AI in Safeguards Implementation

Goffredo Pierbattisti | Specializing Master in Nuclear Safeguards, Politecnico di Milano

AI-assisted falsification, impersonation and attacks on analytical tools could compromise the evidence underlying safeguards conclusions. This research proposal examines how to assess those threats and select measures that protect the integrity of verification.

Research question

How can malevolent AI use by State or non-State actors undermine safeguards implementation, and which technical, procedural and governance measures could most effectively minimise the risk?

Scope of the proposed taxonomy

- Information: declarations, nuclear material accountancy reports, design information, supporting documents, satellite and surveillance imagery, and open-source information affected by synthetic content.
- People: inspectors and analysts exposed to deepfakes, voice cloning and AI-assisted social engineering.
- AI tools: data poisoning, evasion, prompt injection and over-reliance across the lifecycle of the Department's analytical systems.

Across these areas, the proposal considers AI's potential to support adversarial planning of diversion or concealment, including consistent falsification across multiple facilities.

Assessment and candidate controls

The proposed semi-quantitative assessment uses likelihood, impact on safeguards conclusions and detectability. Scales, weights and rankings remain to be defined. Candidate technical controls include provenance and authentication, digital signatures, content credentials, distributed ledgers, multi-source corroboration and detection tools. Procedural measures include verification protocols, human review, red teaming, awareness and incident reporting. Governance covers AI guidelines, information security and cooperation with State and regional authorities.

Expected contribution and evidence status

The intended outputs are a safeguards-specific threat taxonomy, prioritised risks and mitigation recommendations that could inform Departmental AI guidelines. The source is a one-page thesis proposal abstract and reports no empirical results, risk scores or validated control effectiveness. Its envisaged placement within the IAEA Department of Safeguards remains subject to institutional arrangements.

Next research priority

A proposed extension is to define a reproducible scoring protocol and evaluate controls with authorised, non-sensitive material, recording false alarms, missed manipulation and residual uncertainty. This would test whether the proposed measures preserve the credibility of verification evidence.

Source: 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, undated. The final section is an analytical recommendation. Prepared 19 October 2026.