Annotated Research Bibliography

Malevolent Use of AI in Safeguards Implementation

23 external references | Relevance and limitations | Verified 19 October 2026

This bibliography develops the research question in Goffredo Pierbattisti's proposal on malevolent AI use in safeguards implementation. It contains 23 external references, selected to connect safeguards verification with AI threat assessment, evidence integrity and governance. Each annotation explains how to use the source and where its evidence stops.

Recommended reading sequence

Begin with the IAEA workshop report (03) and glossary (01), then develop the threat taxonomy with NIST (08). Use the risk-assessment guide (06) to document assumptions and uncertainty. The transparency report (10) and C2PA specification (15) support control evaluation, while the AI RMF (19) supports governance.

Sources by research objective

Research objective	Priority references
Define safeguards evidence and conclusions	01-03
Map adversarial capabilities and AI threats	07-09
Assess likelihood and impact	06, 08, 19-20
Evaluate synthetic media and provenance	10-16
Design human review and information handling	04-05, 17-18
Compare governance approaches	19-23

Questions for extending the proposal

The literature can support a documented taxonomy and evaluation protocol, but it cannot determine the risk ranking for an unspecified safeguards workflow. Research should establish which evidence channels and judgements are in scope, whether sources are independent, how control errors affect verification, and how uncertainty changes the priorities. These are suggested research questions rather than results in the source abstract.

Selection and access

References were checked on 19 October 2026 against official institutional records, original research manuscripts or publisher records. Final publications take precedence over drafts. Nuclear security guidance is identified separately from safeguards authority. Access limits appear beside the relevant entries, including public ISO descriptions and a partially accessible EU record.

Starting 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. The source contains no bibliography or empirical findings. The accompanying BibTeX file contains the 23 external references.

Safeguards foundations

01 IAEA Safeguards Glossary

International Atomic Energy Agency (2022). IAEA Safeguards Glossary: 2022 Edition. Reference glossary; International Nuclear Verification Series No. 3 (Rev. 1); STI/PUB/2003.

Primary source

Research use. Establish a consistent safeguards vocabulary before constructing the threat taxonomy. Use its definitions of safeguards conclusions, information evaluation and data authentication to specify which verification functions a threat could compromise. Limitation. The glossary has no legal status. It explains terminology and does not create verification powers or replace applicable safeguards agreements and protocols.

02 Safeguards techniques and equipment

International Atomic Energy Agency (2011). Safeguards Techniques and Equipment: 2011 Edition. Technical reference; International Nuclear Verification Series No. 1 (Rev. 2).

Primary source

Research use. Ground the analysis in verification methods and evidence channels, including material accountancy, surveillance and remote monitoring. Chapter 9 on data security, beginning on printed page 123, is relevant to integrity, authenticity and confidentiality. Limitation. This is a 2011 technical baseline. Equipment descriptions and cryptographic implementations require current validation; the publication does not evaluate modern generative AI, prompt injection or contemporary model supply chains.

03 AI for nuclear verification

International Atomic Energy Agency, Department of Safeguards (2025). Emerging Technologies Workshop: Artificial Intelligence for Nuclear Verification. Workshop Report - Summary. Official safeguards workshop summary; STR-406; workshop held in Vienna, 27-29 January 2025.

Primary source

Research use. The most direct contextual source for the Department of Safeguards AI agenda. Use its discussion of nuclear verification opportunities and risks, human validation and strategic planning to identify relevant research and stakeholder questions. Limitation. A workshop summary records emerging priorities and expert discussion. It does not establish binding policy, quantify adversarial attack likelihood or validate the proposed countermeasures.

Information security and assessment

04 Computer security for nuclear facilities

International Atomic Energy Agency (2021). Computer Security Techniques for Nuclear Facilities. Nuclear security technical guidance; IAEA Nuclear Security Series No. 17-T (Rev. 1); STI/PUB/1921.

Primary source

Research use. Use its treatment of digital assets, defence in depth, security levels and zones, and lifecycle protection to structure the analysis of AI-related vulnerabilities and controls. Limitation. Nuclear security guidance for facilities. Its control principles need explicit adaptation to safeguards verification and do not represent Departmental internal AI policy.

05 Security of nuclear information

International Atomic Energy Agency (2015). Security of Nuclear Information. Nuclear security implementing guide; IAEA Nuclear Security Series No. 23-G; STI/PUB/1677.

Primary source

Research use. Inform the information-governance workstream through its treatment of sensitive information, classification, authorised sharing and staff awareness. Useful when considering access controls and information-handling procedures. Limitation. Its nuclear security scope differs from safeguards authority and confidentiality arrangements. Apply its principles by analogy; it predates generative AI.

06 Risk assessment method

Joint Task Force Transformation Initiative (2012). Guide for Conducting Risk Assessments. NIST Special Publication 800-30 Revision 1. DOI 10.6028/NIST.SP.800-30r1.

Primary source DOI

Research use. Provides a documented process for preparing, conducting and maintaining risk assessments. Use it to structure assumptions, evidence and uncertainty behind the proposal's semi-quantitative judgements. Limitation. Written for US federal information systems. Safeguards impact and detectability require context-specific definitions, and the report does not supply a validated safeguards scoring model.

Threat models and AI systems

07 Malicious use of AI

Brundage, M., et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Multi-author research report; arXiv:1802.07228, original 2018 version. DOI 10.48550/arXiv.1802.07228.

Primary source DOI

Research use. Provides a foundational framework for considering how AI may change adversarial capabilities and defensive needs. Useful for the cross-cutting capability-uplift dimension in the proposal. Limitation. A forward-looking report, rather than a safeguards-specific empirical assessment. Its 2018 framing needs updating against current models and verification workflows.

08 Adversarial machine learning taxonomy

Vassilev, Apostol; Oprea, Alina; Fordyce, Alie; Anderson, Hyrum; Davies, Xander; Hamin, Maia (2025). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. Technical report; NIST AI 100-2e2025 (final). DOI 10.6028/NIST.AI.100-2e2025.

Primary source DOI

Research use. Provides a common taxonomy for predictive and generative AI threats across the model lifecycle. Adapt its attacker capability, knowledge and objective dimensions to the proposal's safeguards risk register. Limitation. Generic voluntary guidance; it neither estimates safeguards-specific likelihood nor defines acceptable risk. Scenario scoring and control effectiveness require contextual evidence.

09 Indirect prompt injection

Greshake, Kai; Abdelnabi, Sahar; Mishra, Shailesh; Endres, Christoph; Holz, Thorsten; Fritz, Mario (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. Peer-reviewed conference paper; Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, pp. 79-90; open author manuscript. DOI 10.1145/3605764.3623985.

Primary source DOI

Research use. Establishes why retrieved documents must remain untrusted data in LLM-assisted analysis. Informs defensive testing of document ingestion, tool permissions and human review before safeguards-relevant actions. Limitation. Proof-of-concept demonstrations on 2023 systems do not quantify present-day attack prevalence or IAEA exposure. Current deployments require controlled, application-specific assessment.

Synthetic content and imagery

10 Synthetic content transparency

Chandra, Bilva; Dunietz, Jesse; Roberts, Kathleen; Lee, Yooyoung; Fontana, Peter; Awad, George (2024). Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. Technical report; NIST AI 100-4 (final November 2024). DOI 10.6028/NIST.AI.100-4.

Primary source DOI

Research use. Compares provenance, watermarking, synthetic-content detection and evaluation. Supports layered protection of declarations, imagery and communications, with explicit attention to false positives and human interpretation. Limitation. No individual technique provides comprehensive assurance. Provenance concerns origin and history; it cannot establish that a declaration is factually true or complete.

11 Synthetic geospatial imagery

Bo Zhao; Shaozeng Zhang; Chunxue Xu; Yifan Sun; Chengbin Deng (2021). Deep fake geography? When geospatial data encounter Artificial Intelligence. Cartography and Geographic Information Science, 48(4), 338-352. DOI 10.1080/15230406.2021.1910075.

Primary source DOI

Research use. Connects geospatial evidence credibility to AI-generated imagery. A domain-specific starting point for the proposal's satellite-imagery workstream and for research on provenance and corroboration. Limitation. The research concerns geospatial image manipulation. It does not establish the vulnerability of an IAEA workflow or validate a safeguards detection method.

Access: author and publisher metadata; full article not retrieved.

12 Facial image manipulation benchmark

Rössler, Andreas; Cozzolino, Davide; Verdoliva, Luisa; Riess, Christian; Thies, Justus; Nießner, Matthias (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. Peer-reviewed conference paper; 2019 IEEE/CVF International Conference on Computer Vision (ICCV); open author manuscript. DOI 10.1109/ICCV.2019.00009.

Primary source DOI

Research use. Provides a foundational facial-forgery benchmark with human and automated baselines under compression. Useful for designing staff-awareness exercises and evaluating detection claims against realistic media degradation. Limitation. Covers four historical facial-manipulation methods, not all synthetic media, voice cloning or safeguards imagery. Benchmark performance cannot establish operational authentication reliability.

13 Cross generator image detection

Wang, Sheng-Yu; Wang, Oliver; Zhang, Richard; Owens, Andrew; Efros, Alexei A. (2020). CNN-Generated Images Are Surprisingly Easy to Spot... for Now. Peer-reviewed conference paper; 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); open author manuscript. DOI 10.1109/CVPR42600.2020.00872.

Primary source DOI

Research use. Introduces cross-generator evaluation and shows how augmentation affects transfer beyond one training generator. Supplies a historical baseline for the proposal's synthetic-imagery detection workstream. Limitation. Evidence concerns the evaluated CNN generator families. It does not establish universal detection of later diffusion models or domain-specific satellite and surveillance imagery.

Generalisation and provenance

14 Detection across generative models

Ojha, Utkarsh; Li, Yuheng; Lee, Yong Jae (2023). Towards Universal Fake Image Detectors that Generalize Across Generative Models. Peer-reviewed conference paper; 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24480-24489; open author manuscript. DOI 10.1109/CVPR52729.2023.02345.

Primary source DOI

Research use. Shows failures when GAN-trained detectors encounter other generator families and evaluates pretrained-feature alternatives. Supports testing safeguards-relevant detectors on unseen generators, rather than relying on familiar benchmark accuracy. Limitation. Improved generalization on the studied datasets is not universal assurance. Performance must be revalidated for new generators, image transformations and safeguards-specific imagery.

15 C2PA content credentials

Coalition for Content Provenance and Authenticity (C2PA) (2026). Content Credentials: C2PA Technical Specification, Version 2.4. Published technical specification; version 2.4, April 2026.

Primary source

Research use. Specify testable provenance and signature-validation controls for documents and imagery, including manifests, content bindings and trust relationships. Version 2.4 is the current published edition identified by the official index at verification. Limitation. Cryptographic provenance does not establish factual truth; coverage, signer trust and workflow implementation require separate evaluation.

16 Distributed ledger integrity

Dylan Yaga; Peter Mell; Nik Roby; Karen Scarfone (2018). Blockchain Technology Overview. NISTIR 8202. DOI 10.6028/NIST.IR.8202.

Primary source DOI

Research use. Introduces distributed ledgers, tamper evidence and tamper resistance. Use it to formulate trust assumptions and compare ledger-based integrity with simpler signed-record approaches. Limitation. A general technical overview, with substantial cryptocurrency discussion. It does not establish that a ledger verifies the truth of declarations or improves safeguards implementation.

Human review and risk governance

17 Human use and misuse of automation

Raja Parasuraman; Victor Riley (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, 39(2), 230-253. DOI 10.1518/001872097778543886.

Primary source DOI

Research use. Provides a foundation for analysing over-reliance, monitoring failures and disuse associated with false alarms. Use it to formulate human review requirements and distinguish different failures of automation use. Limitation. Predates modern generative AI and concerns automation across domains. It supplies conceptual grounding rather than evidence about IAEA inspectors or current safeguards tools.

18 Interventions for over reliance

Zana Buçinca; Maja Barbara Malaya; Krzysztof Z. Gajos (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), Article 188, 1-21. DOI 10.1145/3449287.

Primary source DOI

Research use. Reports an experiment with 199 participants on interventions that reduced over-reliance compared with simple explanation approaches. Useful for designing independent reviewer judgement and measuring performance alongside subjective trust. Limitation. The experimental task and participants differ from safeguards practice. Benefits varied with cognitive motivation, and more effective interventions received less favourable subjective ratings.

19 AI risk management framework

Elham Tabassi (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Voluntary risk management framework; NIST AI 100-1. DOI 10.6028/NIST.AI.100-1.

Primary source DOI

Research use. Structure the research protocol around governance, contextual mapping, measurement and management. Useful for documenting assumptions, responsibilities, uncertainty and residual risk across safeguards AI lifecycles. Limitation. Voluntary, sector-agnostic guidance; it provides neither a safeguards-specific scoring formula nor evidence of control effectiveness.

Governance benchmarks and standards

20 Generative AI risk profile

Chloe Autio; Reva Schwartz; Jesse Dunietz; Shomik Jain; Martin Stanley; Elham Tabassi; Patrick Hall; Kamie Roberts (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Cross-sectoral generative-AI risk profile; NIST AI 600-1. DOI 10.6028/NIST.AI.600-1.

Primary source DOI

Research use. Develop the generative-AI component of the taxonomy and evaluation plan, particularly content provenance, information integrity, human-AI interaction, adversarial testing and incident disclosure. Suggested actions include assessing authentication errors. Limitation. A cross-sectoral profile; its suggested actions require adaptation and testing in authorised safeguards workflows.

21 EU Artificial Intelligence Act

European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act). European Union regulation; original Official Journal text, OJ L, 2024/1689, 12 July 2024.

Primary source

Research use. Use the original adopted regulation as a comparative governance reference, including its concepts of risk management, technical documentation and human oversight. Limitation. No IAEA legal applicability is inferred. This entry cites the original 2024 text; any current legal assessment requires the applicable consolidated text and separate analysis.

Access: official metadata and indexed original text; live retrieval limited.

22 AI risk management standard

International Organization for Standardization and International Electrotechnical Commission (2023). ISO/IEC 23894:2023 - Information technology - Artificial intelligence - Guidance on risk management. International Standard; first edition, February 2023.

Primary source

Research use. Use as a reference when integrating AI risk assessment into organisational risk management and tailoring the approach to safeguards workflows. It complements the proposal's attention to technical, procedural and governance measures. Limitation. Only the public abstract and catalogue metadata were consulted. The complete standard requires licensed access; no clause-level interpretation or conformity claim is made.

Access: public ISO abstract and catalogue only; licensed standard not read.

23 AI management system standard

International Organization for Standardization and International Electrotechnical Commission (2023). ISO/IEC 42001:2023 - Information technology - Artificial intelligence - Management system. International Standard; first edition, December 2023.

Primary source

Research use. Inform the organisational governance workstream: an AI management system provides a reference for policies, objectives, responsibilities and continual improvement. Useful for distinguishing institutional management arrangements from the performance of individual technical controls. Limitation. Only the public ISO description and catalogue metadata were consulted. Full requirements need licensed access; management-system alignment alone does not demonstrate effective safeguards verification.

Access: public ISO description and catalogue only; licensed standard not read.

The AI Act is listed as a regulatory benchmark, as in the proposal. Standards and management frameworks require adaptation and do not by themselves demonstrate the effectiveness of safeguards verification.