DUAL-USE ATTESTATIONS: A VERIFIABLE DISCLOSURE PRIMITIVE FOR MILITARIZATION-RISK GOVERNANCE IN ML RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Dual-use repurposing of machine learning (ML) artifacts (papers, code, datasets, weights) into military and surveillance contexts is widely discussed in policy and civil society, yet there is no standardized, verifiable disclosure mechanism embedded in mainstream ML research dissemination. We propose *Dual-Use Attestations*: a governance primitive that binds structured, minimum-elements militarization-risk disclosures to ML artifacts via signed supply-chain attestations and append-only transparency logs. Our approach is intentionally non-operational: it does not optimize weapons, nor adjudicate legal compliance. Instead, it makes author risk-positioning and release-surface choices *machine-readable*, *tamper-evident*, and *auditable at scale*, leveraging emerging AI supply-chain standards (CycloneDX/ECMA-424 ML-BOM; OWASP AIBOM; SPDX AI-BOM) and established provenance tooling (in-toto/SLSA, Sigstore/Rekor, Certificate Transparency). We provide a meticulous documentary analysis crosswalking standards, conference governance scaffolding, and dual-use oversight precedents from other fields.

1 INTRODUCTION

Research artifacts in ML increasingly function as *components* in downstream systems: reused through open repositories, model registries, and weight-sharing ecosystems. The same research objects can plausibly support beneficial civilian applications and also be repurposed into military or coercive surveillance pipelines. Policy debate has sharpened around military AI governance (e.g., state-level norms and directives on autonomy and AI in weapons contexts) (U.S. Department of Defense, 2023; U.S. Department of State, 2023; United Nations General Assembly, 2023; International Committee of the Red Cross, 2025), and around the risks and benefits of widely available model weights (National Telecommunications and Information Administration, 2024). Yet within mainstream ML research dissemination, dual-use positioning tends to remain informal, inconsistent, and difficult to audit.

Core gap. Existing documentation practices (model cards, datasheets, data statements, factsheets) were designed to improve transparency and accountability, but they largely operate as *static narratives* rather than *verifiable, persistent* commitments (Mitchell et al., 2019; Gebru et al., 2021; Bender & Friedman, 2018; Hind et al., 2019). Separately, AI supply-chain security work has emphasized provenance and integrity for models and datasets (Chaudhuri et al., 2024; Ecma International, 2025; CycloneDX Project, 2026; Linux Foundation Research, 2024b). These two threads—ethical disclosure and supply-chain provenance—have not been systematically fused into a conference-and-procurement-ready mechanism for militarization-risk governance.

Thesis. We treat militarization-risk disclosure as a *supply-chain traceability* problem: disclosures should be structured (machine-readable), bound to artifact digests (cryptographically linked), and revision-auditable (tamper-evident). This is directly analogous to why SBOMs emerged for software: to turn qualitative risk into auditable component metadata (Executive Office of the President, 2021; National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025).

Contributions. We provide:

1. A *Dual-Use Card* (DUC) specification: a minimum-elements disclosure object oriented to ML artifact release surfaces (paper/code/data/weights), designed to be safe (non-operational) and comparable across submissions.
2. A *Dual-Use Annex—Escrow* (DUA-E) specification: a deeper governance layer for cases requiring additional scrutiny (e.g., procurement, auditors, conference ethics escalation), without publishing operational detail.
3. A *verifiability design* (DUA-ATT): binding DUC/DUA-E to artifact digests using in-toto/SLSA-style attestations and recording them in an append-only transparency log (Sigstore Rekor; CT-inspired) (in-toto Project, 2024; SLSA, 2026; Sigstore Project, 2026; Laurie et al., 2013).
4. A *documentary analysis* linking this primitive to (a) ML documentation literature (Mitchell et al., 2019; Gebru et al., 2021; Hind et al., 2019), (b) AI BOM standards (CycloneDX/ECMA-424; OWASP AIBOM; SPDX AI BOM) (Ecma International, 2025; OWASP Foundation, 2026; SPDX Project, 2024; Linux Foundation Research, 2024a), and (c) governance adoption pathways (conference ethics scaffolding; EU procurement clauses; EU AI Act documentation duties) (International Conference on Learning Representations, 2026; Neural Information Processing Systems, 2026; International Conference on Machine Learning, 2025; Public Buyers Community (European Commission), 2025; European Union, 2024).

2 RELATED WORK: DOCUMENTATION, PROVENANCE, AND AI SUPPLY CHAINS

Documentation artifacts in ML. Model Cards propose standardized reporting for trained models, emphasizing intended use, performance, and evaluation context (Mitchell et al., 2019). Datasheets for Datasets and Data Statements similarly operationalize structured disclosure for training data and its context (Gebru et al., 2021; Bender & Friedman, 2018). FactSheets extend the analogy to supplier declarations of conformity, aiming to increase trust in AI services (Hind et al., 2019). These contributions motivate our emphasis on standardization, but do not by themselves provide *cryptographic binding* to artifacts or *tamper-evident revision history*.

SBOM to AI BOM / ML-BOM. The cybersecurity-driven rise of SBOM practice (EO 14028; NIST and NTIA minimum-elements framing; CISA updates) illustrates that minimum structured disclosure can be adopted broadly when it is operationally feasible and contractible (Executive Office of the President, 2021; National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025). CycloneDX standardized as ECMA-424 explicitly includes ML models and supports a Machine Learning BOM capability emphasizing model and dataset transparency (Ecma International, 2025; CycloneDX Project, 2026). SPDX has also expanded toward AI/dataset profiles, with the Linux Foundation publishing an AI BOM guide using SPDX 3.0 (SPDX Project (Linux Foundation), 2026; Linux Foundation Research, 2024a; SPDX Project, 2024).

Provenance and transparency logs. in-toto provides a general attestation framework, and SLSA defines predicate types for provenance within that ecosystem (in-toto Project, 2024; SLSA, 2026). Sigstore Rekor provides a transparency log for signed metadata, while Certificate Transparency formalizes append-only log properties and proofs (Sigstore Project, 2026; Laurie et al., 2013). In AI supply-chain security, Google's Secure AI Framework materials highlight the need to adapt supply-chain security and provenance to AI artifacts (datasets, models, evaluation) (Chaudhuri et al., 2024). Recent work also proposes end-to-end attestable ML pipeline frameworks combining provenance specs and transparency logs (Spoczynski et al., 2025).

3 DUAL-USE ATTESTATIONS: CONCEPTUAL DESIGN

3.1 DESIGN GOALS AND THREAT MODEL (GOVERNANCE-CENTRIC)

We target *governance failures* in open research dissemination (traceability, accountability, comparability), not operational adversaries. Our framing is informed by minimum-elements disclosure in software supply-chain governance (SBOM practice) (Executive Office of the President, 2021; National Institute of Standards and Technology, 2021; National Telecommunications and Information

Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025) and by the emergence of AI BOM standards that treat models/datasets as inventory objects (CycloneDX/ECMA-424 ML-BOM; SPDX AI BOM; OWASP AIBOM) (Ecma International, 2025; CycloneDX Project, 2026; SPDX Project, 2024; Linux Foundation Research, 2024a; OWASP Foundation, 2026).

Concretely, we address:

1. **Non-repudiation gap:** after copying/mirroring, authors cannot prove what they disclosed at release time.
2. **Revision ambiguity:** disclosures can be silently revised post hoc without a durable audit trail.
3. **Non-comparability:** organizers/reviewers cannot systematically compare dual-use positioning across submissions at scale.
4. **Procurement friction:** buyers cannot contract for militarization-risk disclosures without standardized deliverables (Public Buyers Community (European Commission), 2025; European Union, 2024).

Design requirements. Disclosures should be (i) *minimum-elements structured* (adoptable baseline), (ii) *digest-bound* to released artifacts (portable under copying), (iii) optionally *tamper-evident* via append-only logs, and (iv) strictly *non-enabling* (governance metadata only) (National Telecommunications and Information Administration, 2021; in-toto Project, 2024; Sigstore Project, 2026; Laurie et al., 2013).

3.2 Two-layer disclosure: DUC and DUA-E

We propose a layered mechanism aligned with minimum-elements adoption logic (National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025) and motivated by policy attention to *release surface* (especially widely available weights) (National Telecommunications and Information Administration, 2024):

- DUC (public, minimum-elements): a short, machine-readable disclosure safe for public release and conference review.
- DUA-E (escrowable depth): additional governance metadata available under controlled access (e.g., auditors, procurement, ethics escalation), to reduce disclosure-avoidance due to IP/safety concerns.

What we standardize (high specificity, non-operational). A DUC minimally specifies: (1) **release surface** (paper/code/data/weights/recipe/API), (2) **immutable digests** for each released artifact, (3) **categorical foreseeable high-risk contexts** (e.g., coercive surveillance, targeting-support) stated non-operationally, and (4) **mitigations** (e.g., staged release, gating, documentation, redactions). The DUA-E may add provenance/evaluation coverage summaries and monitoring/incident-response contacts, optionally expressed in a risk-management vocabulary compatible with organizational frameworks (e.g., AI RMF) (National Institute of Standards and Technology, 2023). Both layers intentionally exclude deployment instructions.

3.3 Verifiable binding: attestations + transparency logs

Dual-Use Attestations bind DUC/DUA-E disclosures to artifact digests using signed in-toto statements and optionally record them in a transparency log (in-toto Project, 2024; SLSA, 2026; Sigstore Project, 2026). When logged, CT-style append-only properties make silent replacement detectable (Laurie et al., 2013). These properties are valuable for governance because they allow third parties to verify *what was disclosed*, *by whom*, *for which artifact version*, and (when logged) *when* (Sigstore Project, 2026; Laurie et al., 2013). This is consistent with the broader movement to adapt supply-chain provenance mechanisms to AI artifacts (Chaudhuri et al., 2024; Spoczynski et al., 2025).

4 ADOPTION PATHWAYS: CONFERENCES AND PROCUREMENT

4.1 CONFERENCE INTEGRATION

ICLR and NeurIPS maintain ethics-oriented policies (International Conference on Learning Representations, 2026; Neural Information Processing Systems, 2026), and ICML demonstrates scalability of structured reflection via impact statements (International Conference on Machine Learning, 2025). We propose:

1. **DUC requirement:** if artifacts beyond the PDF are released (code/data/weights/recipes/APIs), attach a DUC at submission (or provisional if release is post-acceptance).
2. **DUA-E by trigger:** require DUA-E under explicit high-risk triggers (e.g., weight release + high-risk context categories; explicit military framing), via controlled-access review.
3. **Optional integrity upgrade:** for accepted artifact releases, include attestation and (optionally) log reference to support verifiability (in-toto Project, 2024; Sigstore Project, 2026).

4.2 PROCUREMENT INTEGRATION

EU model contractual clauses and the EU AI Act provide a concrete implementation surface for standardized documentation deliverables (Public Buyers Community (European Commission), 2025; European Union, 2024). We propose that procurement request: (i) an AI BOM inventory (CycloneDX ML-BOM or SPDX AI BOM), plus (ii) DUC for each delivered model/version, plus (iii) DUA-E (escrow) when high-risk triggers apply (Ecma International, 2025; CycloneDX Project, 2026; Linux Foundation Research, 2024a; SPDX Project, 2024). We do not claim legal sufficiency; rather, we provide an auditable documentation primitive that can be contractually required and versioned.

5 LIMITATIONS AND SAFETY BOUNDARY

Dual-Use Attestations cannot prevent classified adoption, compel honest disclosure, or enforce downstream behavior. They can, however, reduce ambiguity by making disclosures comparable, artifact-bound, and (optionally) tamper-evident—an adoption pattern mirrored in SBOM minimum-elements practice (National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021). We maintain a strict non-enablement boundary: disclosures are governance metadata and MUST NOT include operational deployment guidance (Appendix L).

6 CONCLUSION

We propose Dual-Use Attestations as a verifiable disclosure primitive linking ML documentation norms with supply-chain provenance and transparency logs. The result is a minimum-elements DUC plus escrowable DUA-E that is digest-bound, optionally log-auditable, and adoptable by conferences and procurement as a lightweight mechanism for dual-use governance.

REFERENCES

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 2018. URL https://aclanthology.org/Q18-1041/. ACL Anthology entry.

Nick Bostrom. Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, 10:44–79, 2011. URL https://nickbostrom.com/information-hazards.pdf. Author-hosted PDF.

Glenn A. Bowen. Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2):27–40, 2009. doi: 10.3316/QRJ0902027.

Scott Bradner. Rfc 2119: Key words for use in rfcs to indicate requirement levels, 1997. URL https://www.rfc-editor.org/rfc/rfc2119.

Miles Brundage et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018. URL https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf. Multi-institution report.

Shamik Chaudhuri, Kingshuk Dasgupta, Isaac Hepworth, Michael Le, Mark Lodato, Mihai Maruseac, Sarah Meiklejohn, Tehila Minkus, and Kara Olive. Securing the ai software supply chain (secure ai framework), 2024. URL https://storage.googleapis.com/gweb-research2023-media/pubtools/pdf/26731199bc024241177e212ec9a0183690ddc07f.pdf. Google white paper (PDF).

Scott A. Crosby and Dan S. Wallach. Efficient data structures for tamper-evident logging. In *Proceedings of the 18th USENIX Security Symposium*, pp. 317–334. USENIX Association, 2009.

Cybersecurity and Infrastructure Security Agency. 2025 minimum elements for a software bill of materials (sbom), 2025. URL https://www.cisa.gov/resources-tools/resources/2025-minimum-elements-software-bill-materials-sbom.

CycloneDX Project. Machine learning bill of materials (ml-bom), 2026. URL https://cyclonedx.org/capabilities/mlbom/. Capability overview; accessed Jan 2026.

Stephanie O. M. Dyke, Anthony A. Philippakis, Jordi Rambla De Argila, Dina N. Paltoo, Elizabeth S. Luetkemeier, Bartha M. Knoppers, et al. Consent codes: Upholding standard data use conditions. *PLOS Genetics*, 12(1):e1005772, 2016. doi: 10.1371/journal.pgen.1005772. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005772.

Ecma International. Ecma-424: Cyclonedx bill of materials specification, 2025. URL https://ecma-international.org/publications-and-standards/standards/ecma-424/. Standardization of CycloneDX; includes ML models in scope (see ECMA-424 page).

European Union. Regulation (eu) 2024/1689 (artificial intelligence act), jun 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Official Journal text via EUR-Lex.

Executive Office of the President. Executive order 14028: Improving the nation's cybersecurity, 2021. URL https://www.federalregister.gov/documents/2021/05/17/2021-10460/improving-the-nations-cybersecurity. Federal Register publication.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021. URL https://arxiv.org/abs/1803.09010. Originally circulated as arXiv:1803.09010.

Shafi Goldwasser, Silvio Micali, and Ronald L. Rivest. A digital signature scheme secure against adaptive chosen-message attacks. *SIAM Journal on Computing*, 17(2):281–308, 1988. doi: 10.1137/0217017.

Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004. URL https://www.jstor.org/stable/25148625.

Michael Hind, Matthew Arnold, et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity, 2019. URL https://arxiv.org/abs/1808.07261. arXiv:1808.07261 (v2).

Allen D. Householder, Garret Wassermann, Art Manion, and Chris King. The cert guide to coordinated vulnerability disclosure, August 2017. URL https://www.sei.cmu.edu/library/the-cert-guide-to-coordinated-vulnerability-disclosure-2/. Special Report CMU/SEI-2017-SR-022, Software Engineering Institute, Carnegie Mellon University.

in-toto Project. in-toto attestation framework, 2024. URL https://github.com/in-toto/attestation. Specification repository; accessed Jan 2026.

International Committee of the Red Cross. Autonomous weapon systems and international humanitarian law: Selected issues, oct 2025. URL https://www.icrc.org/sites/default/files/media_file/2025-10/ICRC-Position_Paper-Autonomous_Weapon_Systems_and_IHL-Selected_issues_Oct2025.pdf. Position paper PDF.

International Conference on Learning Representations. Iclr code of ethics, 2026. URL https://iclr.cc/public/CodeOfEthics. Conference policy page; accessed Jan 2026.

International Conference on Machine Learning. Icml 2025 call for papers, 2025. URL https://icml.cc/Conferences/2025/CallForPapers. Includes impact statement requirement; accessed Jan 2026.

International Organization for Standardization and International Electrotechnical Commission. Iso/iec 29147:2018 information technology — security techniques — vulnerability disclosure, 2018. URL https://www.iso.org/standard/72311.html. International Standard; reference page.

International Organization for Standardization and International Electrotechnical Commission. Iso/iec 30111:2019 information technology — security techniques — vulnerability handling processes, 2019. URL https://www.iso.org/standard/69725.html. International Standard; reference page.

International Organization for Standardization and International Electrotechnical Commission. Iso/iec 23894:2023 information technology — artificial intelligence — guidance on risk management, 2023a. URL https://www.iso.org/standard/77304.html. International Standard; reference page.

International Organization for Standardization and International Electrotechnical Commission. Iso/iec 42001:2023 information technology — artificial intelligence — management system, December 2023b. URL https://www.iso.org/standard/81230.html. International Standard; reference page.

Jonathan Katz and Yehuda Lindell. *Introduction to Modern Cryptography*. CRC Press, 2 edition, 2014.

Barbara Kitchenham. Procedures for performing systematic literature reviews. Technical report, Keele University and NICTA, 2004. URL https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf. Technical Report.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, 4 edition, 2018. ISBN 9781506395661.

Ben Laurie, Adam Langley, and Emilia Kasper. Rfc 6962: Certificate transparency, 2013. URL https://www.rfc-editor.org/rfc/rfc6962.

Jonathan Lawson, Moran N. Cabili, Giselle Kerry, Tiffany Boughtwood, Adrian Thorogood, Pinar Alper, Sarion R. Bowers, Rebecca R. Boyles, Anthony J. Brookes, et al. The data use ontology to streamline responsible access to human biomedical datasets. *Cell Genomics*, 1(2):100028, 2021. doi: 10.1016/j.xgen.2021.100028. URL https://www.sciencedirect.com/science/article/pii/S2666979X21000355. Open access.

Jonathan Lawson, Elena M. Ghanaim, Jinyoung Baek, Harin Lee, and Heidi L. Rehm. Aligning nih's existing data use restrictions to the ga4gh duo standard. *Cell Genomics*, 3(9):100381, 2023. doi: 10.1016/j.xgen.2023.100381. URL https://www.sciencedirect.com/science/article/pii/S2666979X23001787. Open access.

Linux Foundation Research. Implementing ai bill of materials (ai bom) with spdx 3.0, 2024a. URL https://www.linuxfoundation.org/research/ai-bom. AI BOM report; DOI listed on LF page.

Linux Foundation Research. Implementing ai bill of materials (ai bom) with spdx 3.0, 2024b. URL https://www.linuxfoundation.org/research/ai-bom. Alias key for LF_AI_BOM_Report_2024; same resource.

Ralph C. Merkle. A digital signature based on a conventional encryption function. In *Advances in Cryptology—CRYPTO '87*, volume 293 of *Lecture Notes in Computer Science*, pp. 369–378. Springer, 1988.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019. URL https://arxiv.org/abs/1810.03993. Also available as arXiv:1810.03993.

National Institute of Standards and Technology. Fips pub 180-4: Secure hash standard (shs), August 2015. URL https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf. Federal Information Processing Standards Publication.

National Institute of Standards and Technology. Software security in supply chains: Software bill of materials (sbom), 2021. URL https://www.nist.gov/itl/executive-order-14028-improving-nations-cybersecurity/software-security-supply-chains-software-1. NIST EO 14028 SBOM resource page.

National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0), 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. NIST.AI.100-1 PDF.

National Research Council. Biotechnology research in an age of terrorism, 2004. URL https://www.bureaubiosecurity.nl/sites/default/files/2020-03/2004%20-%20Fink%20report%20-%20Biotechnology%20Research%20in%20an%20Age%20of%20Terrorism.pdf. National Academies report; PDF mirror.

National Science Advisory Board for Biosecurity. Proposed framework for the oversight of dual use life sciences research: Strategies for minimizing the potential misuse of research information, June 2007. URL https://osp.od.nih.gov/wp-content/uploads/Proposed-Oversight-Framework-for-Dual-Use-Research.pdf. NSABB report.

National Telecommunications and Information Administration. The minimum elements for a software bill of materials (sbom), 2021. URL https://www.ntia.gov/report/2021/minimum-elements-software-bill-materials-sbom.

National Telecommunications and Information Administration. Dual-use foundation models with widely available model weights, 2024. URL https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf. NTIA report PDF.

Neural Information Processing Systems. Neurips ethics guidelines, 2026. URL https://neurips.cc/public/EthicsGuidelines. Ethics guidelines page; accessed Jan 2026.

Chitu Okoli. A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, 37, 2015. doi: 10.17705/1CAIS.03743. URL https://aisel.aisnet.org/cais/vol37/iss1/43/.

OWASP Foundation. Owasp ai bill of materials (aibom) project, 2026. URL https://owasp.org/www-project-aibom/. Project page; accessed Jan 2026.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. doi: 10.1136/bmj.n71. URL https://www.bmj.com/content/372/bmj.n71.

Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007. doi: 10.2753/MIS0742-1222240302.

Public Buyers Community (European Commission). Updated eu ai model contractual clauses, 2025. URL https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/updated-eu-ai-model-contractual-clauses. Procurement resource; accessed Jan 2026.

Heidi L. Rehm, Angela J. H. Page, Lindsay Smith, Jeremy B. Adams, et al. Ga4gh: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics*, 1 (2):100029, 2021. doi: 10.1016/j.xgen.2021.100029. URL https://www.sciencedirect.com/science/article/pii/S2666979X21000367. Perspective; open access.

Phillip Rogaway and Thomas Shrimpton. Cryptographic hash-function basics: Definitions, implications, and separations for preimage resistance, second-preimage resistance, and collision resistance. In *Fast Software Encryption*, volume 3017 of *Lecture Notes in Computer Science*, pp. 371–388. Springer, 2004.

Sigstore Project. Rekor transparency log overview, 2026. URL https://docs.sigstore.dev/logging/overview/. Documentation; accessed Jan 2026.

SLSA. Slsa provenance v1.0 specification, 2026. URL https://slsa.dev/spec/v1.0/provenance. Provenance predicate within in-toto ecosystem; accessed Jan 2026.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release strategies and the social impacts of language models, 2019. URL https://arxiv.org/abs/1908.09203. arXiv:1908.09203.

SPDX Project. Implementing an ai bom, 2024. URL https://spdx.dev/implementing-an-ai-bom/. SPDX post on AI BOM with SPDX 3.0 AI/dataset profiles.

SPDX Project (Linux Foundation). Spdx specifications, 2026. URL https://spdx.dev/use/specifications/. Spec hub including SPDX 3.0 formats; accessed Jan 2026.

Marcin Spoczynski, Marcela S. Melara, and Sebastian Szyller. Atlas: A framework for ml lifecycle provenance & transparency, 2025. URL https://systex-workshop.github.io/2025/papers/systex25-final68.pdf. Workshop/technical paper PDF.

United Nations General Assembly. United nations general assembly resolution a/res/78/241: Lethal autonomous weapons systems, 2023. URL https://documents.un.org/doc/undoc/gen/n23/431/11/pdf/n2343111.pdf. UN document PDF.

United States Government. United states government policy for oversight of life sciences dual use research of concern, mar 2012. URL https://aspr.hhs.gov/S3/Documents/us-policy-durc-032812.pdf. Policy PDF.

U.S. Department of Defense. Dod directive 3000.09: Autonomy in weapon systems, jan 2023. URL https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf. Directive PDF.

U.S. Department of State. Political declaration on responsible military use of artificial intelligence and autonomy, 2023. URL https://www.state.gov/wp-content/uploads/2023/10/Latest-Version-Political-Declaration-on-Responsible-Military-Use-of-AI-and-Autonomy.pdf. PDF text.

Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pp. 1–10, 2014. doi: 10.1145/2601248.2601268.

World Health Organization. Global guidance framework for the responsible use of the life sciences: mitigating biorisks and governing dual-use research, 2022. URL https://www.who.int/publications/i/item/9789240056107. WHO publication page.

A Appendix A: Documentary method and source corpus

This paper is a *design-and-synthesis* contribution grounded in documentary analysis. We construct and analyze a corpus spanning (i) ML documentation literature, (ii) software and AI supply-chain standards and "minimum elements" guidance, (iii) provenance/attestation and transparency-log

specifications, and (iv) governance and procurement texts relevant to dual-use militarization risk. Our objective is to derive *explicit, auditable requirements* for a disclosure primitive (Dual-Use Attestations) and to demonstrate *interoperability* with emerging AI BOM ecosystems, while maintaining a strict *non-enablement* boundary.

A.1 A.1 Study design and research questions

We follow a document-analysis approach that treats policy/standards texts as analyzable artifacts with normative and operational implications (Bowen, 2009). Our synthesis is also consistent with design-science principles in information systems research: we propose an artifact (a disclosure primitive) and justify its design by tracing requirements to authoritative sources and by formalizing properties elsewhere in the appendix (Hevner et al., 2004; Peffers et al., 2007).

We structure the analysis around four research questions (RQs), each directly tied to claims in the main body:

- **RQ1 (Disclosure gap):** What are the limitations of existing ML documentation artifacts (e.g., model cards, datasheets) for dual-use militarization-risk governance, particularly regarding *verifiability* and *revision auditability*? (Mitchell et al., 2019; Gebru et al., 2021; Bender & Friedman, 2018; Hind et al., 2019)
- **RQ2 (Minimum-elements doctrine):** What constitutes a feasible "minimum elements" disclosure baseline in complex ecosystems, and how does this translate from software SBOM practice to AI/ML artifacts? (Executive Office of the President, 2021; National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025)
- **RQ3 (Verifiability primitives):** Which standardized technical mechanisms exist for (a) binding claims to immutable artifact identifiers and (b) providing tamper-evident history, and what security properties can they provide under standard assumptions? (in-toto Project, 2024; SLSA, 2026; Sigstore Project, 2026; Laurie et al., 2013)
- **RQ4 (Adoption surfaces):** Where can such disclosures be operationalized as *contractible* deliverables (conference submission workflows; procurement clauses; regulatory documentation regimes), and what evidence supports feasibility? (International Conference on Learning Representations, 2026; Neural Information Processing Systems, 2026; International Conference on Machine Learning, 2025; Public Buyers Community (European Commission), 2025; European Union, 2024)

A.2 A.2 Corpus construction: source classes and primary anchors

We assemble a corpus across eight source classes, prioritizing *primary* and *normative* documents (standards, laws, official guidance) where possible. To avoid conflating normative authority with commentary, we treat journalistic and vendor materials as *contextual triangulation* only unless they reference primary texts.

Search strategy and "structured snowballing". We do not claim exhaustive coverage of all relevant documents. Instead, we adopt a targeted retrieval strategy for canonical primary texts, complemented by backward/forward snowballing from anchor documents to identify standards adjacency and governance linkages (Wohlin, 2014; Okoli, 2015). The approach is informed by systematic review best practices (explicit inclusion criteria, traceable selection decisions), without asserting PRISMA completeness (Page et al., 2021; Kitchenham, 2004).

A.3 A.3 Inclusion/exclusion criteria and evidence weighting

We apply explicit selection rules to maintain evidentiary discipline:

Inclusion criteria. A document is included if it satisfies at least one of:

1. **Normative authority:** it is a law, regulation, directive, treaty text, or official policy statement relevant to AI governance, documentation, or procurement (European Union, 2024; U.S. Department of Defense, 2023).

Table 1: Source classes and representative anchors used in the documentary corpus. "Role" indicates how the class constrains the design of DUC/DUA-E and DUA-ATT.

Source class	Representative anchors (non-exhaustive)	Role in design
ML documentation literature	Model Cards (Mitchell et al., 2019); Datasheets (Gebru et al., 2021); Data Statements (Bender & Friedman, 2018); FactSheets (Hind et al., 2019)	Field norms for transparency; highlights gaps (binding, auditability)
Software supply-chain transparency	EO 14028 (Executive Office of the President, 2021); NIST SBOM resource (National Institute of Standards and Technology, 2021); NTIA minimum elements (National Telecommunications and Information Administration, 2021); CISA minimum elements update (Cybersecurity and Infrastructure Security Agency, 2025)	Minimum-elements adoption doctrine; contractibility logic
AI BOM / ML-BOM standards	ECMA-424 CycloneDX (Ecma International, 2025); CycloneDX ML-BOM (CycloneDX Project, 2026); OWASP AIBOM (OWASP Foundation, 2026); SPDX AI BOM materials (SPDX Project, 2024; Linux Foundation Research, 2024a; SPDX Project (Linux Foundation), 2026)	Interoperability substrate; inventory representation for models/datasets
Provenance / attestations	in-toto Attestation Framework (in-toto Project, 2024); SLSA provenance (SLSA, 2026)	Digest binding; issuer attribution; structured predicates
Transparency logs	Sigstore Rekor (Sigstore Project, 2026); Certificate Transparency RFC (Laurie et al., 2013)	Tamper-evidence; inclusion/consistency proofs; revision auditability
AI supply-chain security analyses	Google Secure AI supply chain (Chaudhuri et al., 2024); Attestable ML lifecycle proposals (Spoczynski et al., 2025)	Feasibility arguments for adapting supply-chain approaches to AI artifacts
Dual-use / military AI governance	DoD autonomy directive (U.S. Department of Defense, 2023); Political Declaration (U.S. Department of State, 2023); UNGA LAWS resolution (United Nations General Assembly, 2023); ICRC AWS/IHL position (International Committee of the Red Cross, 2025)	Stakes framing; legal/ethical constraint salience; non-enablement boundary
Procurement + regulatory regimes	EU model contractual clauses (Public Buyers Community (European Commission), 2025); EU AI Act (European Union, 2024); AI RMF (National Institute of Standards and Technology, 2023)	Contractible deliverables; documentation duties; governance vocabulary alignment

2. **Standards authority:** it is a formal standard or a standards-body specification relevant to inventory, provenance, attestations, or transparency logs (Ecma International, 2025; SPDX Project (Linux Foundation), 2026; Laurie et al., 2013).
3. **Field-defining scholarship:** it is a widely cited peer-reviewed or archival research artifact proposing documentation norms or accountability mechanisms for ML (Mitchell et al., 2019; Gebru et al., 2021; Hind et al., 2019).
4. **Implementation-relevant governance guidance:** it provides widely used risk/governance vocabulary and process framing applicable to disclosure fields (e.g., AI RMF) (National Institute of Standards and Technology, 2023).

Exclusion criteria. We exclude:

1. documents that are primarily marketing material without stable technical or normative content;
2. operational guidance that could increase harmful deployment capability (non-enablement boundary);
3. sources that cannot be stably referenced (no persistent identifiers, no stable publication record), unless used only as triangulation.

Evidence weighting. To avoid "citation laundering," we weight evidence in descending order:

Table 2: Extraction codebook (summary).

Field	How used in synthesis
Document authority (law/standard/guidance/scholarship)	Evidence weighting; adoption feasibility
Document version and "living" status	Forces access-date recording; version pinning
Scope (software/AI/ML; procurement; military; transparency)	Determines which design claims it can support
Normative requirements (MUST/SHOULD language)	Directly translated into DUC/DUA-E requirements
Definitions (e.g., component, inventory, provenance, log)	Unifies terminology across ecosystems
Implementation surface (conference/procurement/regulation)	Determines adoption pathway section placement
Risk governance vocabulary (e.g., AI RMF functions)	Provides interoperable language for DUA-E
Non-enablement sensitivity	Determines redaction and exclusion boundaries

1. **Primary normative texts** (laws, regulations, official directives, treaty texts) (European Union, 2024; U.S. Department of Defense, 2023; United Nations General Assembly, 2023).
2. **Formal standards/specifications** (ECMA, IETF RFCs, SPDX specs) (Ecma International, 2025; Laurie et al., 2013; SPDX Project (Linux Foundation), 2026).
3. **Official guidance** (NIST/NTIA/CISA minimum-elements documents) (National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025).
4. **Peer-reviewed/archival scholarship** (documentation frameworks; governance studies) (Mitchell et al., 2019; Gebru et al., 2021; Bender & Friedman, 2018).
5. **Technical whitepapers and research prototypes** for feasibility triangulation (Chaudhuri et al., 2024; Spoczynski et al., 2025).

A.4 A.4 Extraction protocol and analytic codebook

We apply a structured extraction protocol inspired by qualitative document analysis and content analysis methods (Bowen, 2009; Krippendorff, 2018). Each document is coded along two axes: (i) *metadata* (authority, scope, revision status) and (ii) *design-relevant propositions* (what requirement it implies for disclosure, binding, auditability, or adoption).

Unit of analysis. The primary unit is a *normative or design-relevant proposition* (e.g., "minimum elements" lists; definitional statements; required documentation elements; formal properties like append-only logs). Propositions are extracted conservatively, with preference for explicit statements over inferred intent.

Extraction codebook (high-level). Table 2 lists the principal fields used during extraction and synthesis.

A.5 A.5 Synthesis procedure: from propositions to design requirements

Synthesis proceeds in three stages:

Stage 1: Requirement derivation. We translate extracted propositions into *design requirements* using a conservative mapping rule: a requirement is included only if supported by at least one high-weight source class (law/standard/guidance) or by convergence of multiple independent sources. For example, SBOM minimum-elements documents motivate minimum-elements structure and adoption pragmatics (National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025), while in-toto/SLSA motivate digest binding and issuer attribution (in-toto Project, 2024; SLSA, 2026).

Stage 2: Crosswalk construction. We create a crosswalk aligning DUC/DUA-E fields to AI BOM ecosystems (CycloneDX/ECMA-424 ML-BOM; SPDX AI BOM; OWASP AIBOM) to ensure interoperability and contractibility (Ecma International, 2025; CycloneDX Project, 2026; OWASP Foundation, 2026; SPDX Project, 2024; Linux Foundation Research, 2024a). This crosswalk is reported in Appendix I.

Stage 3: Formal property justification. Where the main text claims verifiability properties (binding, authenticity, tamper-evidence), we formalize them as theorems under standard cryptographic assumptions and log models in Appendix H. The underlying primitives are anchored in established specifications (in-toto/SLSA; Rekor; CT RFC) (in-toto Project, 2024; SLSA, 2026; Sigstore Project, 2026; Laurie et al., 2013).

A.6 A.6 Validity, triangulation, and limitations

Triangulation. We triangulate requirements across at least two independent source classes where feasible (e.g., SBOM minimum-elements guidance + AI BOM standards; provenance specs + transparency-log RFCs). This reduces reliance on any single institutional perspective and avoids overfitting the design to one ecosystem.

Living documents and version pinning. Many relevant artifacts are living standards or evolving governance pages. To mitigate citation drift, we (i) cite formally versioned documents when available (e.g., ECMA-424 edition/version) (Ecma International, 2025), and (ii) record access dates for web-hosted materials in BibTeX notes (e.g., OWASP AIBOM; Sigstore docs) (OWASP Foundation, 2026; Sigstore Project, 2026).

Limitations. This corpus does not fully capture non-public procurement practices or classified adoption contexts, and it may under-represent non-English governance documents. We therefore treat the proposal as a *portable disclosure primitive* for open research ecosystems rather than a complete solution to downstream militarization.

A.7 A.7 Reproducibility artifacts (recommended)

To make the documentary analysis auditable, we recommend packaging the following artifacts alongside the paper (where venue policy permits):

1. A **source inventory** enumerating all corpus items with version/access-date fields.
2. An **extraction sheet** (CSV/JSON) containing the codebook fields in Table 2.
3. A **claim-to-evidence matrix** (Appendix M) mapping main-body claims to supporting sources.

These artifacts operationalize the same auditability principles the paper advocates: standardized metadata, versioning, and traceability.

B Appendix B: Definitions, notation, and scope boundaries

This appendix makes the core objects and claims in the main body precise. We (i) define the artifacts and disclosure objects manipulated by Dual-Use Attestations, (ii) establish cryptographic and log-theoretic notation used in later appendices (formal properties and proofs), and (iii) state scope boundaries that enforce non-enablement.

B.1 B.1 Core objects and terminology

ML artifact. An *ML artifact* is any research output that can be reused as a component in downstream systems, including: (a) a paper/preprint, (b) source code, (c) a dataset (or dataset manifest), (d) pretrained model weights, (e) a training recipe/config, (f) an evaluation harness, or (g) an inference API endpoint. This definition is motivated by ML documentation and transparency frameworks that treat models/datasets/services as auditable objects (Mitchell et al., 2019; Gebru et al., 2021; Bender & Friedman, 2018; Hind et al., 2019) and by AI BOM initiatives that expand inventory concepts

beyond software to include ML models and datasets (Ecma International, 2025; CycloneDX Project, 2026; OWASP Foundation, 2026; Linux Foundation Research, 2024a; SPDX Project, 2024).

Release surface. The *release surface* of a submission is the set of artifact types actually released (beyond the PDF), e.g., {`code`, `dataset`, `weights`}. Release surface is a governance-relevant dimension because it changes reuse friction (e.g., releasing weights vs only code), and is foregrounded in policy debate on widely available weights (National Telecommunications and Information Administration, 2024) as well as in supply-chain disclosure logic (enumerating what is shipped) (National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025).

Artifact identifier and digest. An *identifier* is a stable reference such as DOI/arXiv/OpenReview ID, repository URL, and commit hash. A *digest* is a cryptographic hash computed over a canonical byte representation of an artifact:

$$d \leftarrow H(\mathsf{canon}(a)),$$

where $\mathsf{canon}(\cdot)$ deterministically maps the artifact a (e.g., a weight file, a tarball of code, or a dataset manifest) to bytes. Digest algorithms should be standardized and algorithm-agile; SHA-256/SHA-512 are canonical examples standardized by NIST (National Institute of Standards and Technology, 2015). The point of digests in this paper is *binding*: they allow verifiers to establish that a disclosure refers to a particular artifact version, even under copying.

Dual-Use Card (DUC). A DUC is a *public, minimum-elements* disclosure object intended to be feasible at scale and comparable across submissions. It includes: release surface enumeration, digests for released artifacts, categorical foreseeable high-risk contexts, and mitigations (e.g., gating, staged release, redactions). Minimum-elements framing is adopted from SBOM practice, which operationalizes transparency through a small set of required fields that can be contractually requested and audited (National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025).

Dual-Use Annex—Escrow (DUA-E). A DUA-E is an optional controlled-access annex intended for contexts requiring deeper governance scrutiny (procurement, auditors, ethics escalation). It contains *non-operational* depth (e.g., evaluation coverage summaries, monitoring plan summary, incident response contacts). The annex can be expressed using risk-governance vocabulary compatible with organizational frameworks, e.g., NIST AI RMF functions (National Institute of Standards and Technology, 2023).

Attestation (DUA-ATT). A Dual-Use Attestation is a signed statement that binds a DUC or DUA-E payload to a set of artifact digests. We treat in-toto statements as the canonical container, with SLSA-style provenance as a precedent for structured predicates (in-toto Project, 2024; SLSA, 2026). Formally, an attestation is a tuple:

$$\text{DUA-ATT} := (S, \sigma, pk),$$

where S is the statement including a subject digest set and a disclosure payload, σ is a signature over S, and pk is the public verification key associated with an issuer identity.

Transparency log entry. A transparency log is an append-only authenticated data structure that supports later verification that a statement was logged, and that the log has only grown by appending entries. We treat Certificate Transparency as the canonical log model and use Rekor as a practical instance (Laurie et al., 2013; Sigstore Project, 2026). A log entry commits to (at least) the signed statement bytes or its digest.

Categorical "foreseeable high-risk contexts". When we refer to high-risk contexts (e.g., coercive surveillance, mass identification, targeting-support), these are *categorical governance labels* used to support review/audit comparability. They are not operational descriptions and MUST NOT include deployment guidance. This boundary is consistent with the paper's non-enablement stance and with the existence of ongoing international deliberations about military AI governance (U.S. Department of Defense, 2023; United Nations General Assembly, 2023; International Committee of the Red Cross, 2025).

B.2 B.2 Cryptographic notation and security definitions

We use standard cryptographic notation and game-based definitions as in modern cryptography references (Katz & Lindell, 2014). All adversaries are probabilistic polynomial-time (PPT) in a security parameter λ. A function $\mathsf{negl}(\lambda)$ is negligible if it vanishes faster than any inverse polynomial.

Hash functions and collision resistance. Let $H : \{0,1\}^* \to \{0,1\}^k$ be a cryptographic hash function. Collision resistance is defined by the adversary's advantage in finding distinct inputs hashing to the same digest:

$$\mathsf{Adv}^{\mathsf{cr}}_H(A) := \Pr\left[(x, x') \leftarrow A(1^\lambda) : x \neq x' \wedge H(x) = H(x')\right].$$

We assume $\mathsf{Adv}^{\mathsf{cr}}_H(A) \leq \mathsf{negl}(\lambda)$. For a detailed discussion separating collision, preimage, and second-preimage resistance, see Rogaway & Shrimpton (2004). This assumption underpins our binding claims: if an attestation references a digest $d = H(\mathsf{canon}(a))$, then (absent collisions) a different artifact $a' \neq a$ will not match d.

Digital signatures and EUF-CMA. A signature scheme is $\Sigma = (\mathsf{KeyGen}, \mathsf{Sign}, \mathsf{Verify})$. We assume existential unforgeability under chosen-message attack (EUF-CMA) (Goldwasser et al., 1988; Katz & Lindell, 2014): after adaptively querying a signing oracle on messages of its choice, an adversary cannot output a new message-signature pair that verifies. This assumption underpins issuer attribution: if $\mathsf{Verify}(pk, S, \sigma) = 1$, then (except with negligible probability) the issuer holding the signing key produced σ.

B.3 B.3 Transparency log notation (Merkle trees, inclusion, consistency)

We use an append-only Merkle tree log model aligned with Certificate Transparency (Laurie et al., 2013). Merkle trees originate in classic work by Merkle (Merkle, 1988) and are widely used for tamper-evident logging (e.g., Crosby & Wallach (2009)).

Merkle tree construction (abstract). Let leaf entries be byte strings $e_1, \ldots, e_n$. Define leaf hashes (with domain separation) as

$$\ell_i := H(0x00 \parallel e_i),$$

and internal node hashes as

$$v := H(0x01 \parallel v_L \parallel v_R),$$

where v_L, v_R are child hashes. The *tree root* at size n is denoted r_n.

Inclusion proof. An *inclusion proof* $\pi^{\mathsf{inc}}(e_i, n)$ is a sequence of sibling hashes that allows a verifier to recompute r_n from e_i. In CT-style logs, proof size is $O(\log n)$ and verification requires $O(\log n)$ hash computations (Laurie et al., 2013).

Consistency proof. A *consistency proof* $\pi^{\mathsf{con}}(n, m)$ (for $n < m$) allows a verifier to check that the tree of size n is a prefix of the tree of size m (append-only growth). This is the formal mechanism by which "tamper-evidence" is operationalized in CT-style logs (Laurie et al., 2013). Later appendices use this model to formalize what we mean by revision-auditable disclosures.

Log operator and monitors. We distinguish: (i) *submitters* (authors or repositories) who submit attestations, (ii) the *log operator* who maintains the append-only structure, and (iii) *monitors* who store observed signed tree heads and verify inclusion/consistency over time. These roles mirror CT monitoring logic (Laurie et al., 2013) and the practical logging/verification workflow described for Rekor (Sigstore Project, 2026).

B.4 B.4 Scope boundaries and non-enablement constraints

Non-enablement boundary (hard constraint). Dual-Use Attestations are *governance metadata*. A DUC or DUA-E MUST NOT contain:

- step-by-step operational deployment instructions (e.g., how to integrate into targeting or coercive surveillance workflows);

- tactics, field procedures, sensor-placement guidance, or engagement-selection rules;
- optimization guidance that would improve weapons-related or coercive surveillance capability.

This boundary is motivated by humanitarian/legal concerns surrounding military AI governance and autonomy, and by the workshop/paper aim to support harm-preventive governance rather than capability uplift (International Committee of the Red Cross, 2025; United Nations General Assembly, 2023).

What our claims do and do not assert. Our "verifiable" claim is limited to integrity and attribution properties: a verifier can check that (i) a disclosure binds to an artifact version via digests, (ii) a disclosure is attributable to an issuer via signatures, and (iii) when logged, disclosure history is tamper-evident in the CT sense. We do *not* claim that disclosures are truthful, complete, or legally compliant. Formal theorems are stated in Appendix H.

Relationship to later appendices. Appendix C elaborates release-surface taxonomy and categorical risk semantics. Appendix H formalizes binding/authenticity/tamper-evidence claims under the assumptions introduced here. Appendix I provides standards crosswalks (CycloneDX/ECMA-424; SPDX AI BOM; OWASP AIBOM), and Appendix J/K provide conference/procurement adoption packages.

C Appendix C: Release-surface taxonomy and dual-use semantics

This appendix operationalizes a central claim of the main body: *release surface* (what is actually released beyond the PDF) is a primary governance lever for dual-use risk in ML research dissemination. The argument is deliberately *non-operational*: we analyze how different release modalities change *reuse friction*, *copyability*, and *governability* (monitoring, revocation, contracting), not how to deploy systems in high-risk settings.

Our taxonomy is informed by (i) ML transparency frameworks that treat models/datasets/services as documentation targets (Mitchell et al., 2019; Gebru et al., 2021; Bender & Friedman, 2018; Hind et al., 2019), (ii) minimum-elements supply-chain disclosure doctrine that emphasizes enumerating what is shipped (National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025), and (iii) the rapid emergence of AI BOM standards that represent ML models/datasets as first-class inventory components (CycloneDX/ECMA-424 ML-BOM; OWASP AIBOM; SPDX AI BOM) (Ecma International, 2025; CycloneDX Project, 2026; OWASP Foundation, 2026; Linux Foundation Research, 2024a; SPDX Project, 2024; SPDX Project (Linux Foundation), 2026). It is additionally motivated by policy analysis explicitly framing widely available model weights as a focal point of dual-use risk-benefit governance (National Telecommunications and Information Administration, 2024) and by technical supply-chain security analyses that treat AI artifacts (models/datasets/evaluations) as requiring integrity and provenance controls (Chaudhuri et al., 2024; Spoczynski et al., 2025).

C.1 C.1 Release-surface typology

We define a submission's release surface as a structured description of what is released and how. Let

$$\mathcal{A} := \{\texttt{paper}, \texttt{code}, \texttt{dataset}, \texttt{weights}, \texttt{recipe}, \texttt{eval}, \texttt{api}\}$$

denote the set of artifact types considered here. A *release surface* is a tuple

$$\mathsf{RS} := (S, \mathsf{Access}, \mathsf{DigestMap}),$$

where (i) $S \subseteq \mathcal{A}$ enumerates released artifact types, (ii) Access describes the access modality for each type (open/gated/contractual/escrow), and (iii) DigestMap assigns cryptographic digests to each released artifact instance (Appendix B).

We distinguish the following artifact types:

`paper`. The PDF/preprint and accompanying narrative disclosures. Papers can carry conceptual knowledge, evaluation claims, and design rationales, but typically do not provide executable capability without additional artifacts. Model Cards and related documentation efforts primarily attach to this modality of disclosure (Mitchell et al., 2019).

`code`. Source code, scripts, build files, or containers enabling reproduction or reuse. Code reduces *implementation friction* and often enables direct integration into downstream systems. In SBOM terms, code constitutes a deliverable component whose dependencies and versions can be inventoried (National Institute of Standards and Technology, 2021).

`dataset`. Training/evaluation datasets, dataset manifests, or pointers to stable dataset releases. Datasets can reduce *data acquisition friction* and may encode sensitive properties (composition, labeling). Datasheets and Data Statements highlight why provenance, composition, and intended-use constraints are essential (Gebru et al., 2021; Bender & Friedman, 2018).

`weights`. Pretrained model parameters or checkpoints. Weights are uniquely salient because they can materially lower compute and expertise barriers to deploying capabilities, and are often trivially redistributable once released. The governance salience of widely available weights is a primary focus of NTIA's dual-use analysis (National Telecommunications and Information Administration, 2024).

`recipe`. Training recipes/configurations: hyperparameters, pre-processing steps, training scripts, fine-tuning instructions, or reproducible pipelines. Recipes reduce *retraining and adaptation friction* by lowering experimentation overhead and clarifying how to reproduce capabilities. In AI supply-chain security terms, recipes are part of the "how built" provenance story (Chaudhuri et al., 2024).

`eval`. Evaluation harnesses: benchmark code, test protocols, metric implementations, and (where appropriate) evaluation datasets. Eval artifacts can reduce *selection and validation friction*: they make it easier for downstream actors to assess whether a capability meets desired performance criteria. AI BOMs increasingly treat evaluation artifacts as relevant lifecycle metadata (e.g., provenance and assessment coverage) (CycloneDX Project, 2026; Linux Foundation Research, 2024a).

`api`. Inference APIs or hosted model endpoints. APIs provide capability access without necessarily distributing weights. They can be more governable than weights (rate limits, user gating, revocation), but can also enable scalable access. FactSheets and supplier declarations are particularly aligned with service/API contexts (Hind et al., 2019).

Access modalities (cross-cutting). For each artifact type, the access modality matters: open release, gated release (e.g., click-through terms), contractual access (procurement), or escrow access (controlled review). Minimum-elements doctrine emphasizes that what is disclosed must be *contractible* and *auditable* in real organizational workflows (National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025); in our setting, Access is therefore a first-class disclosure dimension rather than an afterthought.

C.2 C.2 Why release surface changes governance risk

We use "risk" in a governance sense: *the extent to which a release materially lowers barriers for repurposing an artifact into high-risk contexts, relative to the mitigations and controls declared*. We do not claim to measure this empirically in this paper; instead, we provide a precise conceptual model that supports comparability across disclosures.

Reuse friction as a governance-relevant latent variable. Define *reuse friction* as an abstract cost that must be paid to instantiate a capability from the released artifacts:

$$\mathsf{Friction}(\mathsf{RS}) := \mathsf{F}_{\text{impl}} + \mathsf{F}_{\text{data}} + \mathsf{F}_{\text{compute}} + \mathsf{F}_{\text{ops}},$$

where the terms represent implementation, data acquisition, compute/training, and operationalization costs. Each artifact type plausibly reduces at least one component:

- `paper` primarily reduces conceptual uncertainty but not necessarily implementation/compute costs;
- `code` reduces F_{impl};
- `dataset` reduces F_{data};
- `weights` reduces $\mathsf{F}_{\text{compute}}$ (and often F_{impl} via reference implementations);

- `recipe` reduces $\mathsf{F}_{\text{compute}}$ and adaptation overhead;
- `eval` reduces F_{impl} and F_{ops} by clarifying acceptable performance;
- `api` reduces F_{impl} and F_{ops} for consumers (capability access), but increases *provider governability*.

This framing aligns with the NTIA report's emphasis that the availability of weights changes the ease and breadth of downstream use (National Telecommunications and Information Administration, 2024), and with AI supply-chain security work emphasizing that AI systems have complex lifecycle dependencies whose transparency matters for governance (Chaudhuri et al., 2024; Spoczynski et al., 2025).

Governability: copyability vs control. Release surface changes not only friction but also *governability*. We highlight two orthogonal governance dimensions:

- **Copyability:** how easily the artifact can be replicated and redistributed without the issuer's control (weights and datasets are typically highly copyable once released; APIs are less copyable).
- **Control surface:** the degree to which the issuer can impose or enforce access constraints, monitoring, revocation, or update policies (APIs typically provide more control; open weights provide less).

Minimum-elements SBOM guidance underscores that effective governance depends on being able to *enumerate components*, *track versions*, and *support audits* (National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025). AI BOM standards extend this to ML artifacts, making release surface and artifact identity natural "inventory" units for governance (Ecma International, 2025; CycloneDX Project, 2026; SPDX Project, 2024; Linux Foundation Research, 2024a; OWASP Foundation, 2026).

A monotonicity lemma (conceptual). We formalize a minimal governance intuition used in our risk semantics.

Axiom C.1 (Friction monotonicity). If two release surfaces have identical access modality and digests are available, and $S \subseteq S'$, then

$$\mathsf{Friction}(S') \leq \mathsf{Friction}(S),$$

i.e., releasing additional artifacts does not increase the minimum effort required to instantiate the capability.

Axiom C.2 (Misuse opportunity monotonicity). Let $\mathsf{Opp}(\cdot)$ be an abstract "opportunity for repurposing" measure (not observed here). Assume Opp is non-increasing in friction:

$$\mathsf{Friction}(S') \leq \mathsf{Friction}(S) \implies \mathsf{Opp}(S') \geq \mathsf{Opp}(S).$$

Lemma C.1 (Release-surface monotonicity). Under Axioms C.1–C.2, for $S \subseteq S'$ (with comparable access modality), the governance-relevant repurposing opportunity is non-decreasing:

$$\mathsf{Opp}(S') \geq \mathsf{Opp}(S).$$

Proof. Immediate from Axiom C.1 and Axiom C.2. □

This lemma is not an empirical claim; it is a conservative structural assumption used to justify why release surface should be explicitly disclosed and treated as a primary axis in DUC/DUA-E semantics. It is consistent with minimum-elements supply-chain practice: shipping more components increases the set of downstream compositions that become feasible (National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021) and with policy analysis emphasizing the distinct governance implications of widely available weights (National Telecommunications and Information Administration, 2024).

C.3 C.3 Categorical high-risk context list (non-operational)

To make disclosures comparable at scale, DUC uses a fixed categorical list of *foreseeable high-risk contexts*. These categories are descriptive and non-operational: they do not specify tactics, procedures,

Table 3: Foreseeable high-risk context categories used in DUC. Definitions are categorical (governance labels), not operational guidance.

Category	One-sentence definition
coercive_surveillance	Use in monitoring/analysis pipelines that enable coercion, suppression, intimidation, or repression of individuals/communities.
mass_identification	Use in identification/verification at population scale (e.g., biometric or identity matching) without case-specific, individualized constraints.
border_enforcement	Use in migration/border-control decision systems where errors and asymmetries can produce severe rights impacts.
detention_or_custody_decisions	Use to inform detention, custody, or deprivation-of-liberty decisions (risk scoring, triage, eligibility) with high stakes.
targeting_support	Use as a component in systems that support target identification/selection or related engagement decision workflows (categorical label only).
battlefield_intelligence_fusion	Use to fuse or prioritize multi-source sensing/ISR information into actionable intelligence products in conflict settings.
autonomous_platform_control	Use in control loops for physical platforms (air/land/sea) where autonomy may affect safety or rights outcomes.
cyber_operations_support	Use supporting cyber operations (e.g., automated analysis/prioritization) where misuse can amplify harm.
other	Any additional foreseeable high-risk context, described at a categorical level without operational details.

or deployment configurations. They are motivated by the paper's emphasis on militarization and coercive surveillance contexts, and are consistent with the broader salience of military AI governance debates (U.S. Department of Defense, 2023; United Nations General Assembly, 2023; International Committee of the Red Cross, 2025).

Interpretation rule. Selecting a category does *not* assert intent or knowledge of downstream deployment; it indicates that, given the capability and release surface, the authors consider repurposing into that category plausible enough to warrant governance visibility and mitigations.

C.4 C.4 Risk-rating semantics (governance meaning of low/medium/high)

We define DUC risk ratings as *governance labels* that summarize (i) release surface, (ii) foreseeable high-risk contexts, and (iii) mitigations. They are not claims about legality or compliance, and do not attempt to predict specific downstream deployments. The goal is comparability and auditable reasoning, consistent with minimum-elements doctrine (National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025) and risk-management vocabulary alignment (e.g., AI RMF framing of mapping and managing risks) (National Institute of Standards and Technology, 2023).

Risk rating factors. We treat the following factors as the minimum semantic ingredients:

- **R1: Release surface enabling power.** Weights/datasets/recipes typically lower friction more than paper-only release; APIs may increase control but still provide capability access (National Telecommunications and Information Administration, 2024).
- **R2: Copyability vs control.** Open weights are highly copyable and hard to revoke; APIs support stronger centralized monitoring/revocation but may scale access (Hind et al., 2019).
- **R3: Foreseeable high-risk contexts selected.** Selection of categories in Table 3 indicates which governance concerns must be addressed.
- **R4: Mitigations and access modality.** Staged release, gating, redactions, and documentation constraints reduce governance risk by increasing friction or control. Layered disclosure echoes dual-use governance precedents that emphasize oversight processes and controlled dissemination

rather than purely voluntary norms (National Research Council, 2004; United States Government, 2012; World Health Organization, 2022).

Operational meaning for governance (not deployment). We define the rating levels as follows:

- **Low.** The release surface is limited (e.g., `paper`-only, or `paper+code` without weights/datasets/recipes), and the disclosed capability is unlikely to materially lower barriers to repurposing in any category in Table 3. Mitigations are consistent with the limited surface and disclosures are complete enough to support review (Mitchell et al., 2019; National Institute of Standards and Technology, 2021).
- **Medium.** The release surface includes enabling artifacts (e.g., code+eval, or datasets without weights; or APIs providing capability access), but mitigations and access controls plausibly preserve meaningful governance leverage (e.g., gating, terms, monitoring summaries, redactions). Foreseeable high-risk contexts may be present but are addressed with explicit mitigation statements (National Telecommunications and Information Administration, 2024; National Institute of Standards and Technology, 2023).
- **High.** The release surface includes artifacts that materially lower reuse friction for broad repurposing (e.g., open weights and/or datasets and/or detailed recipes), especially when foreseeable high-risk contexts include coercive surveillance or targeting-support categories. "High" indicates that the disclosure should trigger additional governance attention (e.g., DUA-E escrow; enhanced review), not that deployment is intended (National Telecommunications and Information Administration, 2024; Cybersecurity and Infrastructure Security Agency, 2025).

Why a categorical rating is still useful. Minimum-elements programs succeed when they enable *triage*: a low-friction baseline that identifies when deeper review is warranted. This mirrors SBOM practice (baseline inventories enabling escalation for critical components) (National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025) and dual-use oversight traditions in other fields (committee-based escalation and controlled dissemination) (National Research Council, 2004; World Health Organization, 2022).

C.5 C.5 Deliverable: release surface × governance risk mechanisms

Table 4 summarizes how each release element changes governance risk mechanisms. The table is descriptive and non-operational; it is intended to guide consistent DUC completion and reviewer interpretation.

Connection to later appendices. Appendix D uses this taxonomy to define required DUC fields (release surface + digests + contexts + mitigations). Appendix J uses the high-risk categories and rating semantics to define trigger-based escalation to DUA-E escrow. Appendix I maps these elements onto CycloneDX/ECMA-424 ML-BOM and SPDX AI BOM representations.

Table 4: Release surface × governance risk mechanisms. "Friction reduced" describes which barriers are lowered; "Governance concerns" describes why this matters for dual-use governance; "Governance levers" are non-operational mechanisms (documentation, gating, auditability) consistent with minimum-elements doctrine and AI BOM traceability.

Release element	Primary friction reduced	Key governance concerns (non-operational)	Typical governance levers (non-operational)
`paper`	Conceptual uncertainty; reproducibility cues	Hard to audit what downstream code/weights correspond to the described system; narrative-only disclosures are not artifact-bound	Model-card style documentation; explicit release-surface statement; identifiers (Mitchell et al., 2019)
`code`	Implementation friction	Increases reusability and composability; makes downstream integration easier; dependency supply-chain becomes relevant	Component inventory (BOM); commit pinning; dependency disclosure; provenance claims (National Institute of Standards and Technology, 2021; Ecma International, 2025)
`dataset`	Data acquisition/labeling friction	Potential sensitivity in composition/labels; enables reproduction and capability transfer; downstream use may diverge from intended-use norms	Datasheets/data statements; dataset digests/manifests; licensing constraints; access gating when appropriate (Gebru et al., 2021; Bender & Friedman, 2018)
`weights`	Compute/training friction; time-to-deploy	High copyability; difficult revocation; can materially lower barriers for repurposing; central in open-weights governance debate	Weight digests; staged release; gating/terms; explicit foreseeable-context disclosure and mitigations (National Telecommunications and Information Administration, 2024)
`recipe`	Adaptation and retraining friction	Encodes "how to reproduce" details; can increase portability across domains; increases the set of feasible downstream variants	Provenance summaries; controlled dissemination when warranted; clear intended/non-intended use; escalation to DUA-E under triggers (Chaudhuri et al., 2024)
`eval`	Selection/validation friction	Benchmarking artifacts can enable rapid capability selection and integration; evaluation coverage becomes a governance question	Evaluation coverage summary; limitations disclosure; lifecycle metadata capture in AI BOM/provenance (CycloneDX Project, 2026; Spoczynski et al., 2025)
`api`	Consumer integration/ops friction	Provides scalable capability access; offers stronger provider control but can broaden access; shifts accountability toward service governance	Service factsheets; monitoring summaries; revocation policies; contractual constraints; versioned endpoint documentation (Hind et al., 2019)

D Appendix D: Normative DUC specification (minimum elements)

This appendix specifies the *Dual-Use Card* (DUC) as a *minimum-elements* disclosure object intended to be: (i) feasible at scale for conference submissions, (ii) comparable across papers, and (iii) auditable/contractible in downstream governance settings. The design is inspired by the minimum-elements doctrine in SBOM guidance (National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025) and by ML transparency artifacts

(Model Cards, Datasheets, Data Statements, FactSheets) (Mitchell et al., 2019; Gebru et al., 2021; Bender & Friedman, 2018; Hind et al., 2019). We additionally ensure interoperability with AI BOM ecosystems that treat ML models/datasets as inventory objects (Ecma International, 2025; CycloneDX Project, 2026; OWASP Foundation, 2026; SPDX Project (Linux Foundation), 2026; Linux Foundation Research, 2024a; SPDX Project, 2024).

D.1 D.1 Normative language, conformance, and safety boundary

Normative keywords. The key words **MUST**, **MUST NOT**, **REQUIRED**, **SHALL**, **SHALL NOT**, **SHOULD**, **SHOULD NOT**, **RECOMMENDED**, **MAY**, and **OPTIONAL** are to be interpreted as described in RFC 2119 (Bradner, 1997).

Conformance classes. We define two conformance classes for adoption flexibility:

- **DUC-Core** (baseline): includes all *required* fields in Table 5.
- **DUC-Plus** (enhanced): includes all required fields plus the *recommended* fields, and provides explicit links to an AI BOM inventory (CycloneDX/ECMA-424 ML-BOM or SPDX AI BOM) when applicable (Ecma International, 2025; SPDX Project (Linux Foundation), 2026; Linux Foundation Research, 2024a).

Non-enablement boundary (hard constraint). A DUC is governance metadata. A DUC **MUST NOT** include operational instructions that would increase harmful capability (e.g., deployment playbooks for targeting or coercive surveillance). This constraint is consistent with the paper's scope and the salience of humanitarian/legal concerns in military AI governance discourse (United Nations General Assembly, 2023; International Committee of the Red Cross, 2025).

What DUC does and does not claim. A DUC is *not* a legal compliance determination. It is a standardized disclosure of release surface, foreseeable high-risk contexts (categorical), and mitigations. Verifiability properties (digest binding; issuer attribution; optional log-auditability) are addressed by attestations and logs in other appendices, but DUC is structured to support those bindings (Appendix B).

D.2 D.2 DUC data model overview

A DUC is a record with five logical blocks:

1. **Header and identity:** versioning, issuer identity, and stable paper identifiers.
2. **Artifact inventory and release surface:** an explicit enumeration of released artifacts and their immutable identifiers (digests).
3. **Use positioning:** intended use and non-intended use statements.
4. **Foreseeable high-risk contexts and governance risk rating:** categorical labels (Appendix C) and a governance-only rating.
5. **Mitigations and governance commitments:** what constraints/controls are applied to what artifacts, and how.

Design rationale (minimum-elements doctrine). Minimum structured disclosure succeeds when it is: (i) *small enough* to be completed reliably, and (ii) *precise enough* to support auditing and contracting (National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025). Accordingly, DUC-Core is deliberately compact, and deeper governance detail is deferred to DUA-E (Appendix on escrow), rather than forcing maximal disclosure for every submission.

D.3 D.3 Required and recommended fields

Table 5 lists normative DUC fields. "Type" indicates expected structure but is not tied to any single serialization format. Fields are written as monospace names for clarity; a venue may implement them as an OpenReview form, a PDF template, or a structured document.

D.4 D.4 Field-level requirements and interpretation rules

This section clarifies how fields should be interpreted so DUCs remain comparable across submissions.

D.4.1 D.4.1 Artifact inventory and digests

Digest requirement (binding intent). Every released artifact beyond the PDF **MUST** have a digest entry. This is the minimum condition needed to later bind disclosures to immutable artifact versions (Appendix B). Digest algorithms **SHOULD** follow standardized secure hash functions such as SHA-256 (National Institute of Standards and Technology, 2015). If an artifact is too large or access-restricted (e.g., gated dataset), the digest **MUST** cover a stable manifest or signed release package description, and `digest.scope` must state what is covered.

Canonicalization disclosure. `digest.scope` **MUST** describe the canonical bytes being hashed. This prevents ambiguity for artifacts that can be represented in multiple equivalent forms (e.g., repository snapshots, container images, dataset shards). Canonicalization is treated as part of the disclosure, not a hidden implementation detail.

Third-party dependencies. If an artifact depends on third-party packages, datasets, or pretrained components, the DUC **SHOULD** include an AI BOM reference (`interop.aibom_ref`) that inventories those dependencies where feasible (Ecma International, 2025; CycloneDX Project, 2026; SPDX Project, 2024).

D.4.2 D.4.2 Intended and non-intended use

Boundedness. `use.intended` and `use.non_intended` **MUST** be concise and bounded: they describe plausible use domains and explicitly excluded domains without attempting to enumerate all possibilities. This aligns with the spirit of Model Cards and Datasheets (intended use and limitations) while keeping the DUC minimum-elements feasible (Mitchell et al., 2019; Gebru et al., 2021).

Non-operational constraint. Non-intended use statements **MUST NOT** include operational guidance; they should remain categorical (e.g., "not intended for coercive surveillance"), consistent with the non-enablement boundary (Appendix B.4).

D.4.3 D.4.3 Foreseeable high-risk contexts

Taxonomy requirement. `risk.contexts[]` **MUST** be selected from the fixed category list in Appendix C.3 (Table 3), optionally with a categorical `other` label. This supports cross-submission comparability for reviewers and conference governance workflows.

Interpretation. Selecting a context category **MUST** be interpreted as a plausibility disclosure, not an intent claim. It signals that the authors believe repurposing into that context is foreseeable enough to warrant mitigation discussion.

D.4.4 D.4.4 Risk rating

Governance semantics only. `risk.rating` **MUST** follow the governance definitions in Appendix C.4. "High" indicates that the combination of release surface and foreseeable contexts plausibly warrants *additional governance attention* (e.g., DUA-E escrow triggers), not that harm is intended.

Rationale requirement. `risk.rationale` **MUST** explicitly reference: (i) what was released, (ii) which context categories were selected, and (iii) which mitigation statements apply. This enforces auditable reasoning rather than ungrounded labels.

D.4.5 D.4.5 MITIGATIONS

Artifact linkage. Each mitigation **MUST** specify the artifacts it applies to via `mitigations[j].applies_to`. This requirement makes mitigations auditable and prevents purely aspirational mitigation lists disconnected from the actual release surface.

Mitigation types. The mitigation type vocabulary is intentionally small to preserve feasibility. For additional depth (e.g., monitoring plans, evaluation coverage summaries), issuers should use DUA-E (escrow) rather than bloating DUC.

D.5 D.5 COMPLETENESS SCORING AND REVIEW UTILITY

To support reviewer triage and conference workflows, we define a DUC completeness score:

$$\mathsf{Score}(\mathrm{DUC}) := \sum_{k=1}^{K} \mathbf{1}\{\text{required field } k \text{ is present and non-empty}\},$$

where K is the number of required fields in Table 5. The score ranges from 0 to K and is *not* a quality judgment; it is a missingness indicator. Conferences may use this score to enforce minimum completeness thresholds without interpreting the substantive content.

D.6 D.6 WORKED EXAMPLE DUC (ILLUSTRATIVE; NON-OPERATIONAL)

The following example demonstrates DUC-Core structure. Values are illustrative; venues may replace these with an OpenReview form.

Header and identity

- `duc.version`: 1.0
- `duc.created_at`: 2026-01-19
- `paper.identifier`: OpenReview: `<ID>` ; arXiv: `<ID>`
- `issuer.name`: Authors of submission `<ID>`
- `issuer.contact`: `<governance-contact@domain>`

Release surface and artifact inventory

- `release.surface`: {`paper, code, weights, eval`}
- `artifacts[0]` (`code`):
  - `locator`: `https://repo/#commit=<hash>`
  - `access`: `open`
  - `license`: `<SPDX-id or URL>`
  - `digest.alg`: SHA-256 (National Institute of Standards and Technology, 2015)
  - `digest.value`: `<sha256-of-tarball>`
  - `digest.scope`: tarball snapshot of repository at pinned commit
- `artifacts[1]` (`weights`):
  - `locator`: `https://host/model-weights-v1.bin`
  - `access`: `gated`
  - `license`: `<terms URL>`
  - `digest.alg`: SHA-256
  - `digest.value`: `<sha256-of-weight-file>`
  - `digest.scope`: exact binary weight file as hosted
- `artifacts[2]` (`eval`):
  - `locator`: `https://repo/eval/#commit=<hash>`

  - `access`: `open`
  - `license`: `<URL>`
  - `digest.alg`: SHA-256
  - `digest.value`: `<sha256-of-eval-package>`
  - `digest.scope`: evaluation harness package (code + metric implementation)

Use positioning

- `use.intended`: Intended for research on `<benign domain>`; evaluation and educational use.
- `use.non_intended`: Not intended for coercive surveillance, mass identification, or targeting-support contexts.

Foreseeable contexts and rating

- `risk.contexts`: {`coercive_surveillance`, `mass_identification`, `other: <categorical>`}
- `risk.rating`: `high`
- `risk.rationale`: Weights are released (gated) alongside code/eval; foreseeable repurposing exists for selected contexts; mitigations include gated access and documentation constraints.

Mitigations

- `mitigations[0].type`: `gating`; `applies_to`: {`weights`}; `description`: access request + terms; provides governance leverage relative to fully open weights.
- `mitigations[1].type`: `documentation`; `applies_to`: {`paper`, `code`, `weights`}; `description`: explicit non-intended use and limitations; categorical context disclosures.

D.7 D.7 Security, privacy, and interoperability considerations

Security and auditability. The DUC alone is a disclosure object; verifiability properties are realized when DUC is bound to artifacts via attestations and optionally logged (in-toto Project, 2024; SLSA, 2026; Sigstore Project, 2026; Laurie et al., 2013). Nonetheless, requiring digests in DUC is essential to enable binding.

Privacy and data minimization. DUC fields are designed to be publishable. Sensitive details (e.g., private dataset composition, proprietary pipeline details) should be deferred to DUA-E under controlled access, rather than omitted entirely.

Interoperability with AI BOMs. DUC is designed to link to an AI BOM inventory rather than replace it. When an AI BOM is provided, `interop.aibom_ref` should point to the BOM, and the DUC focuses on dual-use positioning and mitigations. This aligns with the separation of concerns in supply-chain practice: inventory (BOM) vs governance/disclosure profiles (Ecma International, 2025; SPDX Project (Linux Foundation), 2026; OWASP Foundation, 2026).

Table 5: DUC fields (minimum elements). "Req." indicates whether the field is required for DUC-Core, recommended for DUC-Plus, or optional. All `artifact.digest` values SHOULD use standardized secure hash functions (e.g., SHA-256) (National Institute of Standards and Technology, 2015).

Field	Req.	Type	Semantics
`duc.version`	Required	string	DUC spec version (for parsing/interpretation stability).
`duc.created_at`	Required	date	Creation date (UTC preferred) for traceability.
`paper.identifier`	Required	string/list	Stable identifier(s): OpenReview ID, DOI, arXiv ID.
`issuer.name`	Required	string	Issuer entity for the disclosure (e.g., "Authors of submission X").
`issuer.contact`	Required	string	Governance contact email or URL for inquiries/incident reports.
`release.surface`	Required	set(enum)	Subset of {`paper`, `code`, `dataset`, `weights`, `recipe`, `eval`, `api`} (Appendix C).
`artifacts[]`	Required	list	List of released artifacts with fields below.
`artifacts[i].type`	Required	enum	One of the artifact types in `release.surface`.
`artifacts[i].locator`	Required	string	Stable location pointer (URL + version/commit/tag).
`artifacts[i].access`	Required	enum	{`open`, `gated`, `contractual`, `escrow`}.
`artifacts[i].license`	Required	string	License/terms reference (URL or identifier).
`artifacts[i].digest.alg`	Required	string	Hash algorithm identifier (e.g., SHA-256).
`artifacts[i].digest.value`	Required	string	Digest of canonical bytes ($\mathsf{canon}(\cdot)$) (Appendix B).
`artifacts[i].digest.scope`	Required	string	What bytes were hashed (e.g., "tarball at commit", "weight file", "dataset manifest").
`use.intended`	Required	text	Bounded intended use statement (what the authors expect/support).
`use.non_intended`	Required	text	Explicit non-intended uses (high-level, non-operational).
`risk.contexts[]`	Required	list(enum+text)	Selected high-risk context categories (Table 3) + optional categorical `other`.
`risk.rating`	Required	enum	{`low`, `medium`, `high`} with governance semantics (Appendix C.4).
`risk.rationale`	Required	text	Short justification referencing release surface + contexts + mitigations.
`mitigations[]`	Required	list	List of mitigation statements with fields below.
`mitigations[j].type`	Required	enum	{`staged_release`, `gating`, `redaction`, `documentation`, `monitoring`, `legal/terms`, `other`}.
`mitigations[j].applies_to`	Required	list	Which `artifacts[i]` the mitigation covers.
`mitigations[j].description`	Required	text	Non-operational mitigation description and intended governance effect.
`interop.aibom_ref`	Recommended	string	Link/pointer to AI BOM (CycloneDX/ECMA-424 ML-BOM or SPDX AI BOM) if provided.
`funding.disclosure`	Recommended	text/list	Funding disclosure summary (aligned to venue requirements where applicable).
`assurance.attestation_ref`	Optional	string	Pointer to signed attestation/log entry (if published post-acceptance) (in-toto Project, 2024; Sigstore Project, 2026).

E DUAL-USE ANNEX—ESCROW (DUA-E): NORMATIVE SPECIFICATION AND CONTROLLED-ACCESS WORKFLOW

Status and requirement language. This appendix is *normative*. Requirement keywords (MUST, SHOULD, MAY, etc.) are interpreted as in Bradner (1997). DUA-E is designed as the *escrowable-depth* companion to the public DUC described in the main text (§3), enabling *controlled access* to additional governance metadata without expanding the public release surface.

E.1 WHY AN ESCROW LAYER IS GOVERNANCE-RELEVANT

Dual-use oversight precedents: controlled dissemination is not exceptional. In biosecurity governance, dual-use oversight frameworks emphasize that certain *classes* of information warrant structured scrutiny and, in some cases, modified dissemination practices rather than unconditional public release (National Research Council, 2004; United States Government, 2012; World Health Organization, 2022; National Science Advisory Board for Biosecurity, 2007). In security engineering, coordinated vulnerability disclosure (CVD) practices similarly recognize that *timing, audience, and content* of disclosure can materially affect harm outcomes; standards and guidance formalize staged, role-based information sharing to reduce downstream risk while preserving accountability (International Organization for Standardization & International Electrotechnical Commission, 2018; 2019; Householder et al., 2017).

Information hazards framing. The DUA-E layer operationalizes a conservative principle from the information hazards literature: some information can be socially valuable to evaluate and govern, yet harmful if broadcast without controls (Bostrom, 2011). DUA-E therefore functions as an *audit substrate*: it preserves evidence and structure for legitimate reviewers (conference ethics escalation; procurement auditors; institutional oversight) while keeping the public-facing DUC strictly non-operational.

Responsible release discourse in AI. AI governance discussions around widely available model weights have elevated *release surface* as a central lever (National Telecommunications and Information Administration, 2024). Research on release strategies for foundation-model artifacts likewise treats staged and conditioned releases as a governance instrument rather than a technical optimization objective (Solaiman et al., 2019; Brundage et al., 2018). DUA-E is our proposal to *standardize* this instrument in conference/procurement pipelines using machine-readable, attestable metadata rather than informal narratives.

E.2 SCOPE, SAFETY BOUNDARY, AND NON-ENABLEMENT CONSTRAINTS

Non-enablement boundary (hard constraint). DUA-E MUST NOT contain operational deployment instructions, tactical guidance, or technical details whose primary utility would be to enable military/surveillance deployment (or any other intentional harm). This restriction is consistent with the workshop scope and the paper's stated non-operational design goal (§5). DUA-E is governance metadata only: release-surface characterization, categorical risk semantics, process commitments, and audit-relevant provenance summaries.

Escrow is for *depth*, not for *secrets-as-default*. DUA-E SHOULD be used to store *additional* governance detail that is either (i) not suitable for public dissemination because it increases misuse risk, or (ii) not suitable because it would unduly reveal proprietary or sensitive institutional information (e.g., internal incident response contacts). It MUST NOT be used to avoid making the DUC disclosures required for comparability and baseline transparency.

E.3 ACTORS, ROLES, AND CONTROLLED-ACCESS MODEL

Role model. We adopt a role-based access framing aligned with governance practice in controlled disclosure (biosecurity oversight; CVD; data access committees) (International Organization for Standardization & International Electrotechnical Commission, 2018; 2019; Householder et al., 2017; National Science Advisory Board for Biosecurity, 2007). The intent is to make *who can access what, and why* explicit and auditable.

Table 6: Recommended DUA-E access roles and obligations (governance-centric).

Role	Access level	Primary legitimate purpose	Mandatory obligations
Issuer (authors / org)	Write + read own DUA-E	Provide escrow depth; respond to queries; maintain contact for governance follow-up	Accuracy attestation; maintain update log; designate responsible contact; follow redaction policy (§E.9)
Conference ethics body (authorized reviewers)	Read (case-by-case)	Ethics escalation; evaluate high-risk triggers; ensure DUC completeness and non-enablement	Confidentiality; access-logging; minimize retention; produce a public decision rationale where feasible
Procurement / compliance auditor (authorized)	Read (contractual / statutory)	Verify deliverable documentation; check versioned disclosures; assess governance controls	Confidentiality; use limitation; record-of-access; avoid re-disclosure
Independent oversight (e.g., IRB/ethics committee equivalent)	Read (case-by-case)	Institutional review and accountability; post-release incident handling	Confidentiality; conflict-of-interest controls; written determinations; retention limits
General public	No access (default)	N/A	N/A

Access decision record. Any access to DUA-E MUST generate an access decision record containing: requester identity/affiliation, purpose, scope of access (which fields), timestamp, and retention period. This mirrors the harm-minimization logic in controlled disclosure systems (International Organization for Standardization & International Electrotechnical Commission, 2018; Householder et al., 2017). The record itself SHOULD be non-sensitive and MAY be summarized publicly when appropriate.

E.4 DUA-E CONFORMANCE CLASSES

To enable incremental adoption (minimum-elements logic (National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025)), we define two conformance classes:

1. **DUA-E-Core** (minimum escrow): MUST include identity linkage to DUC, artifact digests, categorical risk semantics, and governance process commitments.
2. **DUA-E-Plus** (audit-ready): MUST include DUA-E-Core plus structured crosswalks to risk management frameworks and a minimal incident-response/monitoring plan (still non-operational), supporting procurement and organizational governance (National Institute of Standards and Technology, 2023; International Organization for Standardization & International Electrotechnical Commission, 2023a;b).

E.5 DUA-E DATA MODEL: REQUIRED AND RECOMMENDED FIELDS

Design principle: DUA-E extends DUC without changing the public surface. DUA-E MUST explicitly reference the DUC it extends (by digest and version) so that escrow material cannot be detached from the public minimum-elements disclosure (non-repudiation and revision-audit goals).

E.6 TRIGGER MATRIX: WHEN DUA-E IS REQUIRED

Trigger logic. Triggers are defined over *release surface* and *categorical contexts* (Appendix C). They are not claims about intent and do not require operational capability thresholds. The design goal is to make escalation *legible and consistent* (reducing ad hoc decision-making).

Table 7: DUA-E-Core / DUA-E-Plus field specification (governance metadata only). "Req." indicates requirement level for the given conformance class.

Field	Req.	Type	Meaning and constraints (non-operational)	Interoperability hooks
`duae.version`	Core: MUST	string	Schema version for parsing and comparison.	Align w/ DUC versioning
`duae.extends_duc_digest`	Core: MUST	digest	Cryptographic digest of the DUC instance this annex extends.	Attestation binding (§3.3)
`duae.extends_artifact_digests`	Core: MUST	list[digest]	Digest list for all relevant released artifacts (code/data/weights/etc.).	AI BOM link: ML-BOM/SPDX
`duae.confidentiality_tier`	Core: MUST	enum	e.g., `committee-only`, `procurement-audit`, `institutional-review`.	Access policy enforcement
`duae.access_policy_ref`	Core: MUST	uri/string	Pointer to escrow access policy (who, why, retention, logging).	CVD/VDP analog (International Organization for Standardization & International Electrotechnical Commission, 2018)
`risk.high_risk_contexts_expanded`	Core: MUST	list[enum]	The categorical high-risk contexts selected in DUC with expanded rationale. MUST remain categorical and non-operational.	Uses DUC taxonomy
`risk.release_surface_mechanisms`	Core: MUST	list[string]	Explain which release elements reduce reuse friction (e.g., weights+recipe+eval harness) and why, at a governance level.	Crosswalk to Appendix C
`risk.uncertainty_notes`	Core: SHOULD	string	What is unknown/assumed; data gaps; contested interpretations.	Supports auditability
`mitigations.controls_detail`	Core: SHOULD; Plus: MUST	structured text/list	Governance controls: gating, staged release, licensing intent, documentation redactions, contact for concerns. MUST NOT include operational misuse guidance.	Maps to AI RMF "Manage" (National Institute of Standards and Technology, 2023)
`mitigations.release_conditions`	Core: MAY; Plus: SHOULD	list[string]	Conditions under which additional artifacts may be released (e.g., post-acceptance, after third-party review).	Procurement clauses (Public Buyers Community (European Commission), 2025)
`provenance.coverage_summary`	Core: SHOULD; Plus: MUST	structured text	Summary of provenance captured (lineage granularity, build steps). Reference in-toto/SLSA artifacts if available.	in-toto/SLSA (in-toto Project, 2024; SLSA, 2026)
`evaluation.coverage_summary`	Core: SHOULD; Plus: MUST	structured text	High-level evaluation coverage relevant to foreseeable misuse contexts (e.g., bias/robustness categories). Must not provide operational targeting guidance.	Model cards/datasheets (Mitchell et al., 2019; Gebru et al., 2021)
`incident.contact_channels`	Core: MAY; Plus: SHOULD	string	Designated reporting channels for concerns (abuse reports, misuse signals). Keep minimal; do not publish if it increases risk.	CVD analog (Householder et al., 2017)
`incident.response_commitments`	Core: MAY; Plus: SHOULD	structured text	Governance commitments: triage policy, revision cadence, notification procedures (non-operational).	ISO/IEC 42001 governance (International Organization for Standardization & International Electrotechnical Commission, 2023b)
`framework.crosswalks`	Core: MAY; Plus: MUST	table/list	Crosswalk to recognized governance frameworks (AI RMF; ISO risk guidance).	(National Institute of Standards and Technology, 2023; International Organization for Standardization & International Electrotechnical Commission, 2023a)
`redaction.public_projection_digest`	Core: SHOULD	digest	Digest of the public projection (the DUC) to enforce stable coupling.	Formal relation (§E.8)

Table 8: Illustrative DUA-E trigger matrix (governance escalation).

Trigger condition (non-operational)	**Rationale**	**Minimum DUA-E scope required**	**Default action**
Weights released (or committed to be released) and DUC includes any high-risk context category	Weight availability materially lowers reuse friction; emphasized in policy discourse (National Telecommunications and Information Administration, 2024)	`risk.*` + `mitigations.*` + `provenance.coverage_summary`	DUA-E-Core REQUIRED
Training recipe + code + evaluation harness released (even without weights) and high-risk context selected	Recipe/harness can reduce reproduction friction and facilitate downstream systemization	`risk.release_surface_mechanisms` + mitigations and staged-release conditions	DUA-E-Core REQUIRED
Submission framed around security/military/surveillance applicability (self-described)	Higher likelihood of contested governance interpretation; requires clearer documentation trail	Expanded categorical rationale + controls detail; crosswalk recommended	DUA-E-Core REQUIRED
Procurement/deployment contract explicitly references dual-use governance deliverables	Contractibility requires audit-ready, versioned documentation (Public Buyers Community (European Commission), 2025; European Union, 2024)	DUA-E-Plus (crosswalks + incident commitments)	DUA-E-Plus REQUIRED

E.7 Crosswalk: DUA-E fields to governance frameworks

Why crosswalks matter. A key adoption barrier is that conferences and organizations already operate within established governance vocabularies (risk management, management systems, procurement controls). DUA-E-Plus therefore standardizes a minimal crosswalk to reduce translation overhead and make DUA-E contractible and auditable (National Institute of Standards and Technology, 2023; International Organization for Standardization & International Electrotechnical Commission, 2023a;b; Public Buyers Community (European Commission), 2025).

E.8 Formal relation: DUA-E to DUC as a public projection

Why formalize the relation. To prevent "escrow-only" disclosures that undermine baseline transparency, we model DUC as a *public projection* of DUA-E. This is a governance constraint: it ensures that the public minimum-elements disclosure is not optional and stays coupled to the deeper annex.

Objects. Let $\mathcal{A}$ be the set of valid DUA-E documents and $\mathcal{C}$ the set of valid DUC documents under their respective schemas. Let $d(\cdot)$ be a cryptographic digest function satisfying standard preimage/second-preimage resistance assumptions (Rogaway & Shrimpton, 2004; National Institute of Standards and Technology, 2015).

Projection function. Define a (deterministic) projection $\pi : \mathcal{A} \to \mathcal{C}$ that drops escrow-only fields and preserves minimum-elements fields:

$$\text{DUC} \;=\; \pi(\text{DUA-E})\,. \tag{1}$$

DUA-E MUST include `duae.extends_duc_digest` such that

$$\texttt{duae.extends\_duc\_digest} \;=\; d(\pi(\text{DUA-E}))\,. \tag{2}$$

Audit invariants (informal). Under standard cryptographic assumptions, this provides:

1. **Coupling invariant:** a DUA-E instance commits to exactly one public DUC projection (by digest).

Table 9: Minimal suggested crosswalk (illustrative): DUA-E-Plus to AI RMF and ISO guidance.

DUA-E component	AI RMF function	ISO alignment (high level)	Governance output
`risk.high_risk_contexts_expanded` + `risk.release_surface_mechanisms`	MAP (National Institute of Standards and Technology, 2023)	Risk framing guidance (International Organization for Standardization & International Electrotechnical Commission, 2023a)	Comparable risk semantics
`mitigations.controls_detail` + `mitigations.release_conditions`	MANAGE (National Institute of Standards and Technology, 2023)	Management system controls / PDCA-style governance (International Organization for Standardization & International Electrotechnical Commission, 2023b)	Documented governance commitments
`provenance.coverage_summary` + `evaluation.coverage_summary`	MEASURE (National Institute of Standards and Technology, 2023)	Evidence expectations for organizational risk management (International Organization for Standardization & International Electrotechnical Commission, 2023a)	Audit-ready evidence summary
`incident.*` fields	GOVERN + MANAGE (National Institute of Standards and Technology, 2023)	Management system incident processes (International Organization for Standardization & International Electrotechnical Commission, 2023b)	Escalation pathway clarity

2. **Non-repudiation support:** a later-presented DUC that does not match the committed digest is detectably inconsistent.
3. **Redaction discipline:** the existence of π forces authors to articulate what remains public (governance minimum elements) and what remains escrowed (safety/IP sensitive), consistent with information-hazard reasoning (Bostrom, 2011).

We emphasize this is a governance property: it does not prevent nondisclosure, but it makes *silent substitution* and *escrow-only evasion* easier to detect.

E.9 Redaction and safety patterns for DUA-E

Redaction principle: minimize marginal misuse value. Redactions SHOULD remove information whose marginal value is primarily enabling misuse, while retaining information that supports governance decisions (risk category selection, release-surface reasoning, mitigations and oversight commitments). This mirrors CVD guidance cautioning that partial disclosure can sometimes increase harm if it materially lowers attacker effort (Householder et al., 2017).

Recommended redaction motifs (non-exhaustive).

1. **Mechanism-at-a-distance:** keep rationale at the level of *which released components lower reuse friction* (weights/recipe/harness), not *how* to operationalize them.
2. **Categorical pathways only:** describe foreseeable harmful application classes (Appendix C) without specifying target selection, operational settings, or deployment integration.
3. **Contact minimization:** include incident channels only if necessary; otherwise provide an organizational intake alias and avoid personal identifiers.
4. **Time-bounded sensitivity:** if certain details become less sensitive over time (e.g., after mitigations deployed), authors MAY submit an updated DUA-E (revision-auditable) and update DUC accordingly.

Analogy: machine-readable access constraints in sensitive data governance. Work on machine-readable data use restrictions (e.g., Consent Codes; DUO) demonstrates that restrictions can be standardized and interoperable without exposing the underlying sensitive data itself (Dyke et al.,

2016; Lawson et al., 2021; 2023; Rehm et al., 2021). We treat DUA-E similarly: standardize the *governance semantics* without disclosing harmful operational detail.

E.10 Implementation note: relationship to attestations and logs

DUA-E MAY be bound to artifact digests and recorded as an attestation reference (with the DUA-E payload itself kept in escrow), consistent with in-toto/SLSA patterns and transparency-log auditability (in-toto Project, 2024; SLSA, 2026; Sigstore Project, 2026; Laurie et al., 2013). The transparency log SHOULD record only non-sensitive metadata (e.g., digest pointers), not escrow content.

F Appendix F: Attestation construction (in-toto/SLSA alignment)

This appendix specifies how Dual-Use Attestations (DUA-Att) bind DUC/DUA-E disclosures to ML artifacts using supply-chain attestations. The design follows the in-toto Attestation Framework statement model (in-toto Project, 2024) and is compatible with the SLSA provenance ecosystem (as an existence proof that structured predicates can be adopted at scale) (SLSA, 2026). Requirement keywords are interpreted as in Bradner (1997). This appendix is non-operational and concerns governance metadata integrity only.

F.1 F.1 Object model

Artifacts and digests. Let $\{a_i\}_{i=1}^{n}$ be the set of released artifacts (code, dataset manifests, weights, recipes, eval harness, API spec) associated with a submission. Each artifact has a canonical byte representation $\mathsf{canon}(a_i)$ and a digest

$$d_i := H(\mathsf{canon}(a_i)),$$

where H is a standardized secure hash (e.g., SHA-256) (National Institute of Standards and Technology, 2015). Canonicalization scope MUST be described (cf. `digest.scope` in Appendix D).

Disclosure payloads. Let C denote a DUC instance (public minimum-elements disclosure), and let E denote a DUA-E instance (escrow annex). In the escrow case, the attestation SHOULD avoid embedding sensitive escrow text; instead it binds to an escrow pointer and digest (Appendix E).

F.2 F.2 Statement format (in-toto compatible)

Statement structure. A Dual-Use Attestation is an in-toto statement S with:

- **Subject:** a list of subject digests $\{d_i\}$ corresponding to released artifacts.
- **Predicate type:** a stable identifier indicating whether the predicate encodes a DUC or a DUA-E binding.
- **Predicate:** the disclosure payload (either embedded DUC, or a DUA-E reference structure).

We write this abstractly as:

$$S := \big(\mathsf{subject} = \{d_i\}_{i=1}^{n},\ \mathsf{predicateType} = t,\ \mathsf{predicate} = P\big),$$

consistent with in-toto statement semantics (in-toto Project, 2024).

DUC predicate. For public DUC attestations, P SHOULD include the DUC content (or a canonical serialization digest of the DUC), and MUST include: (i) DUC version, (ii) DUC digest, (iii) reference to the paper identifier(s), and (iv) a mapping from DUC artifacts to subject digests. This makes the DUC disclosure *artifact-bound*.

DUA-E predicate (escrow-safe). For DUA-E attestations, P MUST include: (i) digest of the DUC projection $\pi(E)$ (Appendix E.8), (ii) digest of the escrow annex blob (or escrow package manifest), (iii) an access policy reference, and (iv) the artifact digest set $\{d_i\}$. The predicate MUST NOT embed operational details. Its function is to preserve auditability (existence, binding, issuer attribution) without expanding public release.

Algorithm agility. The statement MUST indicate digest algorithm identifiers and SHOULD support algorithm agility (ability to migrate hashes/signatures while preserving audit trails) (National Institute of Standards and Technology, 2015; Rogaway & Shrimpton, 2004).

F.3 F.3 Signing and issuer identity

Signature binding. An issuer signs the statement bytes to produce a signature σ verifiable under a public key pk:

$$\sigma \leftarrow \mathsf{Sign}(sk, S), \qquad \mathsf{Verify}(pk, S, \sigma) = 1.$$

Security relies on standard EUF-CMA assumptions for the signature scheme (Appendix B.2) (Katz & Lindell, 2014; Goldwasser et al., 1988).

Issuer granularity. The issuer identity MAY be: (i) *individual* (author-held key), (ii) *organizational* (lab/company key), or (iii) *delegated* (conference-managed signing service for post-acceptance releases). The choice affects governance accountability but not the cryptographic verification procedure.

Multi-signer support. If multiple entities must endorse a disclosure (e.g., authors + institution), the system MAY attach multiple signatures $\{\sigma_j\}$ over the same statement S, or issue multiple attestations referencing the same subject digests. This is a governance choice (threshold or consensus) and does not change the non-enablement boundary.

F.4 F.4 Versioning, updates, and anti-overwrite rules

Append-only disclosure evolution. Updates MUST be issued as new attestations; prior attestations MUST NOT be overwritten. Each new attestation SHOULD reference the prior attestation digest (or log entry reference) to support revision traceability.

Change triggers (non-exhaustive). A new attestation SHOULD be issued when: (i) any subject artifact digest changes (new weights/code/dataset manifest), (ii) the DUC content changes (e.g., revised release surface, updated mitigations), (iii) DUA-E content changes (new escrow package), or (iv) access modality changes (open $\leftrightarrow$ gated/contractual/escrow).

F.5 F.5 Optional transparency logging

Log commitment. To make attestations publicly auditable in time ("what existed when"), issuers MAY submit the signed statement (or its digest) to an append-only transparency log such as Rekor (Sigstore Project, 2026) and rely on CT-style inclusion/consistency properties for tamper-evident history (Laurie et al., 2013). For escrowed DUA-E, only non-sensitive commitments SHOULD be logged (e.g., digests and references), not escrow content.

What logging adds. Logging supports third-party verification that a given attestation existed at or before a point in time and was not silently replaced, assuming monitors and CT-style consistency checking (Laurie et al., 2013).

F.6 F.6 Verification procedure (what a verifier checks)

Given artifacts $\{a_i\}$, a DUC/DUA-E payload reference, and an attestation (S, σ, pk), a verifier SHOULD:

1. **Recompute digests:** compute $H(\mathsf{canon}(a_i))$ and check equality with $\{d_i\}$ in the statement subject.
2. **Verify signature:** check $\mathsf{Verify}(pk, S, \sigma) = 1$.
3. **Check predicate type and schema:** ensure predicateType corresponds to the expected DUC or DUA-E schema version.
4. **If logged:** verify transparency-log inclusion (and optionally consistency) proofs for the attestation/log entry (Laurie et al., 2013; Sigstore Project, 2026).

Failure at any step indicates that the disclosure is not verifiably bound to the artifact version, not attributable to the issuer, or not time-auditable (if logging was claimed). These checks do not establish truthfulness of the disclosure; they establish integrity and auditability (formalized in Appendix H).

G TRANSPARENCY LOG MODEL AND VERIFICATION COMPLEXITY

This appendix specifies the transparency-log model assumed when we claim "tamper-evident" disclosure history. We use the Certificate Transparency (CT) append-only Merkle log abstraction (Laurie et al., 2013) and treat Sigstore Rekor as a practical implementation surface (Sigstore Project, 2026). This is a governance primitive: it supports auditability of *existence, timing, and revision history* of attestations, not truthfulness of disclosures.

G.1 G.1 LOG OBJECTS

Logged item. The log entry MUST commit to one of: (i) the attestation statement bytes S (Appendix F), (ii) the signature bundle (S, σ, pk), or (iii) a digest of S and a stable pointer to where the bundle is stored. For escrowed DUA-E, the log entry SHOULD NOT include escrow content; it should include only non-sensitive commitments (digests and references).

Merkle tree construction. Let log leaf entries be $e_1, \ldots, e_n$ (byte strings). A standard CT-style Merkle tree (originating with Merkle (Merkle, 1988)) defines:

$$\ell_i := H(0x00 \parallel e_i), \qquad v := H(0x01 \parallel v_L \parallel v_R),$$

where H is a secure hash (e.g., SHA-256) (National Institute of Standards and Technology, 2015). The tree root for size n is r_n.

G.2 G.2 PROOF TYPES AND WHAT THEY GUARANTEE

Inclusion proof. An inclusion proof $\pi^{\mathsf{inc}}(e_i, n)$ allows a verifier to recompute r_n from e_i and a path of sibling hashes. If the verifier accepts, then (under standard hash assumptions) e_i was included in the log of size n (Laurie et al., 2013).

Consistency proof (append-only growth). A consistency proof $\pi^{\mathsf{con}}(n, m)$ for $n < m$ allows a verifier to check that the log of size m extends the log of size n without removing or rewriting earlier entries (Laurie et al., 2013). This is the formal basis for "append-only" auditability.

Tamper-evidence. Tamper-evidence means: (i) entries are hard to silently remove without breaking consistency, and (ii) equivocation (showing different histories to different observers) can be detected by monitors who gossip or compare observed roots (CT monitoring model) (Laurie et al., 2013). The data-structure intuition aligns with tamper-evident logging literature (Crosby & Wallach, 2009).

G.3 G.3 VERIFICATION PROCEDURE (LOG LAYER)

A verifier that receives (a) an attestation bundle (S, σ, pk), (b) a log entry reference, and (c) proofs, SHOULD:

1. **Verify signature:** check $\mathsf{Verify}(pk, S, \sigma) = 1$ (Appendix B.2).
2. **Verify inclusion:** using π^{inc}, check that the committed log root r_n includes the entry (or its digest) (Laurie et al., 2013).
3. **Verify consistency (optional but recommended):** if the verifier has previously observed $r_{n'}$, validate $\pi^{\mathsf{con}}(n', n)$ to ensure append-only growth (Laurie et al., 2013).

G.4 G.4 COMPLEXITY (WHAT IT COSTS TO VERIFY)

Let n be the current log size.

- **Inclusion proof size:** $O(\log n)$ hashes; verification time $O(\log n)$ hash computations (Laurie et al., 2013).
- **Consistency proof size:** $O(\log n)$; verification time $O(\log n)$ (Laurie et al., 2013).
- **Signature verification:** treated as $O(1)$ w.r.t. log size (depends on signature scheme; Appendix B.2) (Katz & Lindell, 2014).

The intent is to keep verification lightweight enough for conference/procurement workflows: check a small number of hashes and a signature, not a full artifact rebuild.

H FORMAL PROPERTIES AND PROOFS

This appendix formalizes what our mechanism can guarantee. We prove integrity/auditability properties for (i) binding disclosures to artifacts, (ii) attributing disclosures to an issuer, and (iii) making disclosure history tamper-evident when logged. We do *not* prove truthfulness or legal compliance.

H.1 H.1 ASSUMPTIONS

We assume:

1. **Collision resistance:** H is collision resistant (Rogaway & Shrimpton, 2004; National Institute of Standards and Technology, 2015).
2. **Signature unforgeability:** the signature scheme is EUF-CMA secure (Goldwasser et al., 1988; Katz & Lindell, 2014).
3. **CT-style log correctness:** inclusion/consistency proofs behave as specified in Laurie et al. (2013).
4. **Deterministic canonicalization:** $\mathsf{canon}(\cdot)$ is deterministic for each artifact type (Appendix B).

H.2 H.2 THEOREM: ARTIFACT–DISCLOSURE BINDING

Theorem H.1 (Digest binding). Let an attestation statement S include a subject digest $d = H(\mathsf{canon}(a))$ for an artifact a. If a verifier recomputes $H(\mathsf{canon}(a))$ from the obtained artifact bytes and matches d, then any different artifact $a' \neq a$ that also matches d implies a collision in H.

Proof (sketch). Assume there exists $a' \neq a$ with $H(\mathsf{canon}(a')) = H(\mathsf{canon}(a)) = d$. Then $x = \mathsf{canon}(a)$ and $x' = \mathsf{canon}(a')$ form a collision for H, contradicting collision resistance. □

H.3 H.3 THEOREM: ISSUER AUTHENTICITY

Theorem H.2 (Issuer attribution). Let (S, σ, pk) be an attestation bundle such that $\mathsf{Verify}(pk, S, \sigma) = 1$. Under EUF-CMA security, producing such a valid signature on a new statement without the issuer's signing key has negligible probability.

Proof (sketch). Standard reduction: an adversary that outputs a fresh valid (S, σ) pair with respect to pk can be used to break EUF-CMA unforgeability (Goldwasser et al., 1988; Katz & Lindell, 2014). □

H.4 H.4 THEOREM: TAMPER-EVIDENT HISTORY UNDER TRANSPARENCY LOGS

Theorem H.3 (Inclusion implies prior publication). If a verifier accepts a valid inclusion proof $\pi^{\mathsf{inc}}(e, n)$ for an entry e and root r_n, then (absent hash collisions) the entry e was included in the log at size n (Laurie et al., 2013).

Proof (sketch). The verifier recomputes the Merkle root from e and the proof path; acceptance means the recomputed root equals r_n. If e were not in the committed tree, acceptance would require constructing a second preimage/collision in the Merkle hashing structure, which reduces to collision resistance of H. □

Theorem H.4 (Consistency implies append-only). If a verifier accepts a consistency proof $\pi^{\mathsf{con}}(n, m)$ between roots r_n and r_m with $n < m$, then (absent hash collisions) the first n entries of the log at size m match those of the log at size n (no deletion/rewriting) (Laurie et al., 2013).

Proof (sketch). CT consistency proofs certify that the size-n tree is a prefix of the size-m tree. A log operator attempting to remove or rewrite an earlier entry while still producing a valid consistency proof would need to create conflicting Merkle roots with the same proof structure, implying collisions. □

H.5 H.5 Corollary: "non-repudiation" as a governance property

Corollary H.5 (Governance non-repudiation). If an issuer signs S (Theorem H.2) and S (or its digest) is included in a transparency log (Theorem H.3), then a third party can later verify that a disclosure statement attributable to the issuer existed by (or before) the logged time, subject to the log's timestamp semantics (Laurie et al., 2013; Sigstore Project, 2026).

What is not proven. These results do not prove that the disclosure content is accurate, complete, or legally compliant. They prove integrity (binding), attribution (authenticity), and, when logged, tamper-evident history.

I Standards crosswalk (ECMA-424, SPDX AI BOM, OWASP AIBOM)

This appendix provides an *easy* interoperability map: how DUC/DUA-E can coexist with AI BOM inventories rather than replacing them. We treat AI BOMs as *inventory* (what artifacts exist, how they relate), and DUC/DUA-E as *governance positioning* (foreseeable contexts, mitigations). Relevant standards/projects include ECMA-424 (CycloneDX) and its ML-BOM capability (Ecma International, 2025; CycloneDX Project, 2026), SPDX specifications and AI BOM guidance (SPDX Project (Linux Foundation), 2026; SPDX Project, 2024; Linux Foundation Research, 2024a), and OWASP AIBOM (OWASP Foundation, 2026).

I.1 I.1 High-level mapping strategy

Strategy.

- Use **CycloneDX/SPDX** to represent the *artifact inventory*: components (models, datasets, code packages) and relationships.
- Attach DUC as a **linked governance record**: either embedded as a property/annotation or referenced as an external document with a digest.
- Treat DUA-E as **escrow-only depth**: referenced by digest/pointer, not embedded in public BOMs.

I.2 I.2 Practical crosswalk (conceptual)

Key point. This crosswalk does not require BOM schemas to natively "understand" militarization risk. It only requires that BOMs can link to external governance records (the DUC) and carry stable identifiers/digests, which is consistent with the goals of AI BOM initiatives (Ecma International, 2025; SPDX Project (Linux Foundation), 2026; OWASP Foundation, 2026).

Table 10: Conceptual crosswalk: DUC elements to AI BOM representations. (Kept high-level for robustness across evolving schemas.)

DUC element	CycloneDX / ECMA-424 (ML-BOM)	SPDX (AI BOM guidance)	OWASP AIBOM
Artifact list + digests	Represent each released artifact as a component; record cryptographic hashes; capture version/locator	Represent artifacts as SPDX elements; record checksums/IDs; express relationships (model–dataset, etc.)	Inventory items emphasizing model/dataset/service identity
Release surface (`paper/code/dataset/model`)	Component types + scope + external refs; treat non-file items (API/spec) as referenced artifacts	Element types + relationships; link external documents for non-file artifacts	Sections for model, data, training, evaluation, deployment surface
Foreseeable high-risk contexts (categorical)	Attach as governance properties/annotations or external reference to DUC	Attach as annotations/external refs to DUC; keep categories machine-readable	Risk/governance metadata fields (categorical labels)
Mitigations (gating, staged release, redactions)	Record access modality and terms as metadata; link mitigation statements via DUC reference	Record access/licensing; link mitigation narrative as external document + digest	Governance section commitments; avoid operational detail
Attestation reference (optional)	External reference to in-toto/SLSA attestation bundle (in-toto Project, 2024; SLSA, 2026)	External reference to attestation; bind via digests	Treat attestations as verifiable provenance links

J Conference adoption package (OpenReview + reviewer rubric)

This appendix gives a minimal, easy-to-adopt conference workflow. It is designed to fit within existing ethics/scoping mechanisms used by major ML venues (International Conference on Learning Representations, 2026; Neural Information Processing Systems, 2026) and the precedent of standardized impact reflections at scale (International Conference on Machine Learning, 2025).

J.1 J.1 Submission-time requirements

1. **Release-surface declaration (always).** Every submission must declare whether it releases artifacts beyond the PDF (Appendix C).
2. **DUC required if artifacts are released.** If any non-PDF artifacts are released at submission time, attach a DUC (Appendix D). If artifacts will be released only post-acceptance, submit a provisional DUC (without final digests) and finalize later.
3. **Funding disclosure alignment.** Authors include the venue-required funding disclosure pointer in the DUC (if applicable).

J.2 J.2 Reviewer rubric (easy checklist)

Reviewers (or an ethics subcommittee) SHOULD use the following checklist:

1. **Completeness:** required DUC fields present (Appendix D).
2. **Artifact binding readiness:** artifacts listed with stable locators; digests present when available.
3. **Categorical clarity:** high-risk contexts are categorical (Appendix C.3), not operational.
4. **Mitigation linkage:** mitigations specify which artifacts they apply to.
5. **Non-enablement:** no operational deployment guidance (Appendix L).

J.3 J.3 Escalation triggers for DUA-E (controlled access)

A venue MAY require DUA-E (Appendix E) under trigger-based escalation, e.g.:

- weights released (or planned) *and* any high-risk context selected (National Telecommunications and Information Administration, 2024);
- explicit military/security framing in the submission;
- reviewer flags indicating missing mitigations for a high-risk context.

J.4 J.4 Post-acceptance artifact integrity (optional but recommended)

For accepted papers that release artifacts, conferences MAY request:

1. final artifact digests in the DUC;
2. an in-toto/SLSA-style attestation binding DUC to artifacts (in-toto Project, 2024; SLSA, 2026);
3. (optional) transparency-log entry reference (Sigstore Project, 2026; Laurie et al., 2013).

K Procurement adoption package (clauses + verification)

This appendix provides a simple procurement integration pattern compatible with EU contractual clause templates and documentation-oriented regulation, without claiming legal sufficiency (Public Buyers Community (European Commission), 2025; European Union, 2024).

K.1 K.1 Contract deliverables (recommended minimum)

A buyer SHOULD request three deliverables per model/version delivered:

1. **AI BOM inventory:** CycloneDX ML-BOM (ECMA-424) or SPDX-based AI BOM inventory (Ecma International, 2025; CycloneDX Project, 2026; SPDX Project (Linux Foundation), 2026; Linux Foundation Research, 2024a).
2. **DUC per delivered version:** minimum-elements dual-use positioning bound to the delivered artifacts (Appendix D).
3. **DUA-E when triggered:** escrow annex for high-risk triggers (weights released, high-risk categories selected, or contractual requirement) (National Telecommunications and Information Administration, 2024).

K.2 K.2 Model clause text (easy template)

Clause (documentation deliverables). *Supplier shall provide (a) an AI Bill of Materials inventory for each delivered model version, (b) a Dual-Use Card (*DUC*) disclosure for each delivered model version, and (c) a Dual-Use Annex—Escrow (*DUA-E*) upon trigger conditions defined by the buyer (including, at minimum, release of model weights and/or selection of high-risk context categories). Each* DUC/DUA-E *must reference immutable artifact digests for delivered weights/code/data. Supplier shall provide verifiable attestations binding disclosures to artifacts and, where applicable, evidence of transparency-log inclusion.*

(Organizations can adapt this clause style to procurement templates such as EU model contractual clauses (Public Buyers Community (European Commission), 2025).)

K.3 K.3 Verification checklist for buyers/auditors

Given delivered artifacts and documentation, the buyer/auditor SHOULD:

1. **Match inventory:** confirm the AI BOM lists the delivered artifacts and versions (Ecma International, 2025; SPDX Project (Linux Foundation), 2026).
2. **Match digests:** recompute digests and compare with DUC/DUA-E digests (Appendix B).

3. **Verify issuer signatures:** verify in-toto/SLSA-style attestation signatures (in-toto Project, 2024; SLSA, 2026).
4. **If logged:** validate inclusion/consistency proofs for transparency-log references (Laurie et al., 2013; Sigstore Project, 2026).

K.4 K.4 Regulatory alignment note (non-legal claim)

Documentation-oriented regimes (e.g., EU AI Act) emphasize traceability and documentation for certain systems (European Union, 2024). DUC/DUA-E and AI BOM deliverables can support an organization's ability to demonstrate structured documentation and versioned traceability; however, this paper does not claim they satisfy any particular legal requirement.

L Safety and non-enablement safeguards

This appendix states practical safeguards to keep DUC/DUA-E non-operational and aligned with harm-preventive governance goals.

L.1 L.1 Allowed vs disallowed content

Allowed (governance metadata).

- release surface, artifact digests, version pointers;
- categorical foreseeable high-risk contexts (Appendix C.3);
- high-level mitigations (gating, staged release, redactions, monitoring commitments);
- governance contacts and escalation pathways (non-sensitive).

Disallowed (capability uplift). DUC/DUA-E MUST NOT include:

- deployment playbooks for targeting or coercive surveillance;
- tactics, procedures, or operational integration guidance;
- optimization steps whose primary effect is improving harmful deployment capability.

L.2 L.2 Redaction and escalation

1. **Default to DUC minimality:** keep public disclosures categorical and short (Appendix D).
2. **Escrow for sensitive depth:** use DUA-E when deeper governance detail is needed (Appendix E), but do not place operational detail there either.
3. **Conference escalation:** if reviewers identify operationally enabling content, escalate to an ethics process consistent with venue norms (International Conference on Learning Representations, 2026; Neural Information Processing Systems, 2026).

L.3 L.3 Why these safeguards are justified

Dual-use oversight traditions in other fields highlight the importance of structured governance without broadcasting misuse-enabling details (National Research Council, 2004; United States Government, 2012; World Health Organization, 2022). In the military AI context, humanitarian and international deliberations underscore the need for careful governance framing (United Nations General Assembly, 2023; International Committee of the Red Cross, 2025). Our safeguards implement this orientation in the narrow context of conference/procurement documentation.

M Claim-to-evidence matrix

This matrix makes the paper easy to audit: each main claim is linked to primary sources and to the appendix section where it is elaborated.

Table 11: Claim-to-evidence matrix (non-exhaustive).

Main-body claim	**Key evidence (citations)**	**Where elaborated**
Military AI governance and autonomy are active policy/legal topics.	(U.S. Department of Defense, 2023; U.S. Department of State, 2023; United Nations General Assembly, 2023; International Committee of the Red Cross, 2025)	App. A, App. L
Widely available model weights are a focal dual-use governance issue.	(National Telecommunications and Information Administration, 2024)	App. C
ML documentation artifacts exist but are largely narrative and not artifact-bound.	(Mitchell et al., 2019; Gebru et al., 2021; Bender & Friedman, 2018; Hind et al., 2019)	App. B, App. D
Minimum-elements SBOM doctrine supports feasible, contractible disclosure baselines.	(Executive Office of the President, 2021; National Institute of Standards and Technology, 2021; National Telecommunications and Information Administration, 2021; Cybersecurity and Infrastructure Security Agency, 2025)	App. A, App. D
AI BOM standards/projects treat models and datasets as inventory objects.	(Ecma International, 2025; CycloneDX Project, 2026; SPDX Project (Linux Foundation), 2026; SPDX Project, 2024; Linux Foundation Research, 2024a; OWASP Foundation, 2026)	App. I
Attestations can bind claims to artifact digests; logs support auditability over time.	(in-toto Project, 2024; SLSA, 2026; Sigstore Project, 2026; Laurie et al., 2013)	App. F, App. G
Verification is lightweight (log proofs $O(\log n)$).	(Laurie et al., 2013; Crosby & Wallach, 2009)	App. G
Formal guarantees are integrity/attribution/tamper-evidence (not truthfulness).	(Rogaway & Shrimpton, 2004; Goldwasser et al., 1988; Katz & Lindell, 2014; Laurie et al., 2013)	App. H
Conferences already have ethics scaffolding that can host a DUC/DUA-E workflow.	(International Conference on Learning Representations, 2026; Neural Information Processing Systems, 2026; International Conference on Machine Learning, 2025)	App. J
Procurement frameworks provide a surface to contract for documentation deliverables.	(Public Buyers Community (European Commission), 2025; European Union, 2024)	App. K
Life-sciences dual-use oversight precedents motivate controlled, accountable disclosure.	(National Research Council, 2004; United States Government, 2012; World Health Organization, 2022)	App. E, App. L