Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

NIST AI 600-1: Confabulation and Information Integrity

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-29 · nist ai 600-1 · nist generative ai profile · confabulation ai risk · information integrity genai nist · genai risk management profile

Original article: https://reliablecontext.com/articles/nist-ai-600-1-generative-ai-profile-confabulation-information-integrity



In this article

1. What the profile adds to the AI RMF
2. The risk list
3. Confabulation: definition and suggested actions
4. Information integrity
5. Value chain and third-party components
6. Provenance and content-credential actions
7. Mapping actions to records you can show
8. Limits: voluntary, general-purpose
9. Sources

What the profile adds to the AI RMF

NIST AI 600-1, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, is dated July 2024 on its cover and was published on July 26, 2024. It calls itself "a cross-sectoral profile of and companion resource for the AI Risk Management Framework (AI RMF 1.0) for Generative AI" (NIST AI 600-1, p. 1). The parent framework, NIST AI 100-1 (January 2023), organizes risk management into four functions (Govern, Map, Measure, Manage) with categories and subcategories.

A *profile* is "an implementation of the AI RMF functions, categories, and subcategories for a specific setting, application, or technology" (p. 1). This one adds two things:

1. **A list of risks "unique to or exacerbated by" generative AI** (GAI), each defined (Section 2).
2. **More than two hundred suggested actions**, each with an Action ID tied to an AI RMF subcategory: "GV-1.1-001 corresponds to the first suggested action for Govern 1.1", with prefixes GV, MP, MS and MG (p. 13). Each action is tagged with the GAI risks it addresses.

The actions are selective: "Not every subcategory of the AI RMF is included", and actions relevant to developers "may not be relevant to GAI deployers" (pp. 12-13). The working group behind the profile focused on Governance, Content Provenance, Pre-deployment Testing and Incident Disclosure (p. 2), which is why provenance recurs throughout the tables.

Status on 28 October 2026. The profile was developed under Executive Order 14110 (p. 1, fn. 2), which Executive Order 14148 revoked on January 20, 2025. NIST still publishes the profile, and its technical reports list shows only the July 2024 version. NIST states that "The AI RMF 1.0 is being revised as part of the White House AI Action Plan" (NIST); that plan recommends revising the AI RMF "to eliminate references to misinformation, Diversity, Equity, and Inclusion, and climate change" (America's AI Action Plan, July 2025, p. 4). The plan names the framework, not the profile. Whether a revised profile will keep the Information Integrity risk, which is defined partly in terms of mis- and disinformation, is unknown.

The risk list

Section 2 defines twelve risks, listed here in NIST's order (pp. 3-5). The second column is this article's reading, not NIST's.

#	Risk (NIST name)	Bearing on factual reliability (interpretation)
1	CBRN Information or Capabilities	Low
2	Confabulation	Direct
3	Dangerous, Violent, or Hateful Content	Low
4	Data Privacy	Indirect: deletions and corrections interact with provenance
5	Environmental Impacts	None
6	Harmful Bias or Homogenization	Indirect
7	Human-AI Configuration	Amplifies confabulation through over-reliance
8	Information Integrity	Direct
9	Information Security	Indirect: poisoning and prompt injection corrupt inputs
10	Intellectual Property	Low
11	Obscene, Degrading, and/or Abusive Content	None
12	Value Chain and Component Integration	Direct

The profile covers risks "for which there is an existing empirical evidence base at the time this profile was written" and excludes speculative ones (p. 3). An optional grouping in a footnote places Confabulation, Data Privacy and Value Chain and Component Integration among "Technical / Model risks" and Information Integrity under "Misuse by humans" (p. 3, fn. 5). The rest of this article covers the four risks that decide whether a generated statement can be traced to evidence.

Confabulation: definition and suggested actions

Confabulation is "The production of confidently stated but erroneous or false content (known colloquially as 'hallucinations' or 'fabrications') by which users may be misled or deceived" (p. 4). Section 2.2 extends it to outputs "that diverge from the prompts or other input or that contradict previously generated statements in the same context" (p. 6). NIST avoids "hallucination" because some commenters said it anthropomorphizes GAI (p. 4, fn. 6); the colloquial term is covered in the glossary entry on hallucination.

Three points in Section 2.2 shape control design (p. 6):

- Confabulations are "a natural result of the way generative models are designed." The risk is managed, not removed.
- They matter most for "open-ended prompts for long-form responses" and in domains that need contextual or expert knowledge.

- Outputs "may also include confabulated logic or citations that purport to justify or explain the system's answer." A citation is not evidence until it is checked.

Confabulation-tagged actions that leave a retainable record include:

- **MS-2.5-003**: "Review and verify sources and citations in GAI system outputs during pre-deployment risk measurement and ongoing monitoring activities."
- **MS-2.5-001**: "Avoid extrapolating GAI system performance or capabilities from narrow, non-systematic, and anecdotal assessments."
- **GV-1.3-002**: "Establish minimum thresholds for performance or assurance criteria and review as part of deployment approval ('go/'no-go') policies, procedures, and processes".
- **MG-4.1-002**: post-deployment monitoring, "particularly for potential confabulation, CBRN, or cyber risks."
- **MG-4.3-002**: "Establish and maintain policies and procedures to record and track GAI system reported errors, near-misses, and negative impacts."

(Quoted from pp. 14, 31, 44-45.) MS-2.5-003 is a citation audit, run before release and again in production. The method auditing an LLM answer against its sources describes one way to perform it.

Information integrity

The **Information Integrity** risk is a "Lowered barrier to entry to generate and support the exchange and consumption of content which may not distinguish fact from opinion or fiction or acknowledge uncertainties, or could be leveraged for large-scale dis- and mis-information campaigns" (p. 4). Section 2.8 defines the property itself, drawing on a 2022 White House research roadmap. High-integrity information "distinguishes fact from fiction, opinion, and inference; acknowledges uncertainties; and is transparent about its level of vetting. This information can be linked to the original source(s) with appropriate evidence. High-integrity information is also accurate and reliable, can be verified and authenticated, has a clear chain of custody, and creates reasonable expectations about when its validity may expire" (p. 9).

Read as a specification (an interpretation), that sentence lists properties a record can carry: a source link, a vetting status, declared uncertainty, a chain of custody and a validity horizon. NIST links the two risks: GAI can spread misinformation at scale, "particularly if the content stems from confabulations" (p. 10).

Relevant Information Integrity actions:

- **MP-2.3-001**: assess output "by comparing it to a set of known ground truth data and by using a variety of evaluation methods (e.g., human oversight and automated evaluation, proven cryptographic techniques, review of content inputs)."
- **MP-2.3-003**: "Deploy and document fact-checking techniques to verify the accuracy and veracity of information generated by GAI systems, especially when the information comes from multiple (or unknown) sources."
- **MS-2.5-005**: "Verify GAI system training data and TEVV data provenance, and that fine-tuning or retrieval-augmented generation data is grounded." TEVV means test, evaluation, validation and verification; see grounding.

- **MS-1.1-009**: "Track and document risks or opportunities related to all GAI risks that cannot be measured quantitatively, including explanations as to why some risks cannot be measured".

(Quoted from pp. 24-25, 29, 31.) MS-1.1-009 asks for a record of what is not measured and why. The method on declaring unknowns applies the same idea to individual facts.

Value chain and third-party components

Value Chain and Component Integration is "Non-transparent or untraceable integration of upstream third-party components, including data that has been improperly obtained or not processed and cleaned due to increased automation from GAI; improper supplier vetting across the AI lifecycle; or other issues that diminish transparency or accountability for downstream users" (p. 5). Section 2.12 says GAI worsens it through data too large to vet, "extensive reuse of limited numbers of models", and integration into other services. With many components, "it may be difficult to attribute issues in a system's behavior to any one of these sources"; even benchmark datasets "can contain label errors" (p. 12).

The actions are mostly documentation:

- **GV-1.6-003**: inventory entries should include "Data provenance information (e.g., source, signatures, versioning, watermarks)" and "Underlying foundation models, versions of underlying models, and access modes."
- **GV-6.1-008**: "Maintain records of changes to content made by third parties to promote content provenance, including sources, timestamps, metadata."
- **GV-6.2-001**: "Document GAI risks associated with system value chain to identify over-reliance on third-party data and to identify fallbacks."
- **MP-2.2-001**: "Identify and document how the system relies on upstream data sources, including for content provenance, and if it serves as an upstream dependency for other systems."
- **MG-3.1-003**: "Re-assess model risks after fine-tuning or retrieval-augmented generation implementation and for any third-party GAI models deployed for applications and/or use cases that were not evaluated in initial testing."

(Quoted from pp. 16, 21, 24, 42.) Appendix A.1.3 adds conventional procurement controls: due diligence, software bills of materials, service level agreements and SSAE attestation reports (p. 48).

Data privacy, briefly. Data Privacy covers "Impacts due to leakage and unauthorized use, disclosure, or de-anonymization of biometric, health, location, or other personally identifiable information or sensitive data" (p. 4). **MG-4.1-006** ties it to provenance: "Track dataset modifications for provenance by monitoring data deletions, rectification requests, and other changes that may impact the verifiability of content origins" (p. 44). A deletion that is not itself recorded leaves a provenance trail pointing at nothing.

Provenance and content-credential actions

Appendix A.1.6 supplies the definitions. "Provenance data tracking records the origin and history for digital content, allowing its authenticity to be determined." "Data provenance refers to tracking the origin and history of input data through metadata and digital watermarking techniques." Provenance metadata "can include

information about GAI model developers or creators of GAI content, date/time of creation, location, modifications, and sources" (p. 51).

Actions in this group:

- **GV-1.5-003**: "Maintain a document retention policy to keep history for test, evaluation, validation, and verification (TEVV), and digital content transparency methods for GAI."
- **MS-1.1-001**: "Employ methods to trace the origin and modifications of digital content."
- **MS-2.8-003**: document "each instance where content is generated, modified, or shared to provide a tamper-proof history of the content", noting that "Robust version control systems can also be applied".
- **MG-2.2-002**: "Document training data sources to trace the origin and provenance of AI-generated content."

(Quoted from pp. 16, 28, 34, 40.) The profile names no standard. One industry specification is C2PA, where a manifest is "The set of information about the provenance of an asset based on the combination of one or more assertions (including content bindings), a single claim, and a claim signature," and "Content Credential" is its non-technical name (C2PA Technical Specification 2.5, §2.3). Content credentials attach to assets such as images. A generated sentence asserting a fact needs provenance at the level of the claim: a link to the dated source record it came from. That distinction is this article's, not NIST's.

Mapping actions to records you can show

The suggested actions say what to do; they rarely say what evidence remains. The "NIST text" column below is abridged from the PDF. The "Record" and "Check" columns are interpretation.

Action ID	NIST text (abridged)	Record you can show (interpretation)	Check (interpretation)
GV-1.3-002	Minimum thresholds in go/no-go approval	Release decision citing measured confabulation rate against the threshold	Threshold predates the measurement
MS-2.5-003	Verify sources and citations in outputs	Citation-audit log: answer ID, source, version retrieved, verdict, reviewer, date	Re-audit of a sample reproduces verdicts
MS-2.5-005	Verify RAG data is grounded	Corpus manifest: source, version, ingestion date per document	Every cited passage resolves to an entry
MS-1.1-009	Document risks not measured, and why	Register of unmeasured risks with reason and owner	Each risk is measured or registered
GV-1.6-003	Inventory with provenance and model versions	Inventory entry: model, version, access mode, data sources	Deployed version equals inventory version
GV-6.1-008	Records of third-party content changes	Change log with source, timestamp, metadata	No content version without a change entry
MG-3.1-003	Re-assess after fine-tuning or RAG	Re-assessment report per adaptation	Report dated after each change
MG-4.1-006	Track deletions and rectification requests	Append-only log of deletions and corrections	Deleted items keep a tombstone entry

Action ID	NIST text (abridged)	Record you can show (interpretation)	Check (interpretation)
MG-4.3-002	Record and track errors and near-misses	Incident register linked to answer IDs	Each confirmed confabulation is registered

Worked example (fictional)

Halden Therapeutics AG and Oberau Fill GmbH are fictional; all details are invented.

Halden runs an internal assistant that answers supplier-qualification questions using retrieval-augmented generation over its quality documents. A buyer asks whether the contract manufacturer Oberau Fill GmbH is currently approved. The assistant answers "Approved; next audit due June 2027" and cites a qualification memo.

Under MS-2.5-003 a reviewer checks the citation. The corpus holds two versions of the memo. The cited one, dated March 2024, gives the June 2027 audit date. A January 2026 version records a suspension pending corrective action. The answer is a confabulation in NIST's sense: confidently stated and false, backed by a real but superseded citation.

The corpus manifest kept under MS-2.5-005 shows both versions were ingested and nothing marked the 2024 memo as superseded. The incident register under MG-4.3-002 gets an entry linked to the answer ID, and MG-3.1-003 triggers a re-assessment once only the current version is retrievable. Without the manifest, Halden could show the answer was wrong but not why. With it, the cause is a missing validity marker, which is exactly NIST's "reasonable expectations about when its validity may expire".

Limits: voluntary, general-purpose

- **Voluntary.** The AI RMF is "intended for voluntary use" (p. 1), the profile's primary considerations are "relevant for voluntary use by any organization" (p. 47), and AI 100-1 calls the framework "voluntary, rights-preserving, non-sector-specific, and use-case agnostic" (NIST AI 100-1). It is United States federal guidance, not a statute or regulation in any jurisdiction.
- **Not a checklist.** The AI RMF Playbook "is neither a checklist nor set of steps to be followed in its entirety"; the profile's actions likewise depend on actor and use.
- **General-purpose.** A cross-sectoral profile sets no sector thresholds. GV-1.3-002 leaves the acceptable confabulation rate to the organization.
- **Dated.** The risks reflect 2024 evidence, the parent framework is under revision, and the profile anticipates "Future revisions" with "additional AI RMF subcategories, risks, and suggested actions" (p. 2).

This article is educational and is not legal or regulatory advice.

Reliable Context is built around the record-level properties this profile points to: one canonical record per real thing; a cited, dated source behind every value; unknowns declared; any past state reproducible, running inside the customer's perimeter.

Sources

1. NIST AI 600-1, AI RMF: Generative Artificial Intelligence Profile, NIST, July 2024 — risks, definitions, Action IDs and text, Appendix A.
2. NIST AI 600-1 publication page, NIST, published July 26, 2024 — publication date.
3. AI Risk Management Framework, NIST, read 28 October 2026 — AI RMF 1.0 under revision.
4. Technical Reports, NIST AIRC, read 28 October 2026 — only the July 2024 profile listed.
5. NIST AI 100-1, AI RMF 1.0, NIST, January 2023 — four functions; voluntary framework.
6. NIST AI RMF Playbook, NIST AIRC — not a checklist.
7. America's AI Action Plan, The White House, July 2025 — recommendation to revise the AI RMF.
8. Executive Order 14148 of January 20, 2025, 90 FR 8237 — revokes Executive Order 14110.
9. C2PA Technical Specification 2.5, October 2026 — C2PA Manifest and Content Credential definitions.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence
2. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
3. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
4. https://www.govinfo.gov/content/pkg/FR-2025-01-28/pdf/2025-01901.pdf
5. https://airc.nist.gov/technical-reports
6. https://www.nist.gov/itl/ai-risk-management-framework
7. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf
8. https://spec.c2pa.org/specifications/specifications/2.5/specs/C2PA_Specification.html
9. https://airc.nist.gov/airmf-resources/playbook

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.