



WHITE PAPER | OCTOBER 2026

Two Audit Trails

Why AI Guardrails Won't Answer the Oversight
Question in Open-Source Intelligence

Michael Nord

Co-founder & CEO, Nordic Road

nordicrd.com

1 Executive Summary

Research increasingly begins inside AI tools rather than with a search box and a page of links. An AI answer arrives already synthesized: the sources have been chosen, weighed and blended before a human sees them.

Much of the public conversation about “auditable AI” is aimed at whether the model is staying inside its guardrails. That is necessary work, but it answers a question about the *system*. It does not answer the question an analyst gets asked about a *judgment*.

For open-source intelligence (OSINT), the narrower, under-served problem is chain of custody for open-source information as it moves into analysis and decisions. ICD 203, ICD 206, ICD 505 and the *IC OSINT Strategy 2024–2026* already point this way.

2 Introduction: The Shift to AI-First Research

For two decades, research started with a search box and a page of links. That habit is changing. In a Pew Research Center survey of U.S. adults conducted in February 2026, 42% said they use AI chatbots to search for information, the most common use reported. Among professional decision-makers the shift is sharper. G2 reported in April 2026 that 51% of the 1,076 B2B software buyers it surveyed now start their research with an AI chatbot more often than with Google, up from 29% a year earlier. There is no comparable public data for intelligence analysts, but there is little reason to think analysts, or the officials they support, are immune.

That changes what the first step of research means. A page of links leaves the analyst to judge each source. An AI answer arrives already synthesized: the sources have been chosen, weighed and blended before a human sees them. If the first step of open-source intelligence (OSINT) now happens inside an AI tool, grading sources and recording how they were used can't be afterthoughts. When an inspector general, a congressional committee or a court challenges a judgment, they ask what you knew, where it came from, and how much you trusted it.

Much of the public conversation about “auditable AI” is aimed at a different problem. Both problems are real, but they are not the same, and OSINT programs that treat them as one will find out during their first serious review.

3 Two Audit Trails

3.1 Audit trail one: is the model staying inside its guardrails?

The first problem is the one most people now recognize. Frontier developers such as OpenAI and Anthropic have published safety frameworks and warned publicly about the risks of increasingly capable models, putting AI safety squarely into public view.

This summer made the stakes concrete. In July 2026, by OpenAI's own account, models in an internal cybersecurity evaluation with reduced safeguards got around controls meant to isolate them from the internet and compromised parts of OpenAI's research infrastructure and Hugging Face's systems. OpenAI called it a "warning shot." Independent investigators at METR and Redwood Research reported that agents tried to disguise cheating, including spoofing tool-call outputs and attempting to edit or delete logs. They found no successful alteration of the authoritative transcripts in the period they examined.

Reconstruction was possible because a source of truth existed outside the agents' reach. For the Intelligence Community (IC), ICD 505, *Artificial Intelligence* (signed January 2025), addresses this class of problem. It calls for IC AI to be auditable for performance and compliance, for AI provenance to be tracked through the AI lifecycle, and for a model registry and a risk management framework.

This is necessary work. But it answers a question about the *system*. It does not answer the question an analyst gets asked about a *judgment*.

3.2 Audit trail two: chain of custody for the information itself

The second problem is narrower, under-served, and central to OSINT: chain of custody for open-source information as it moves into analysis and decisions.

A model can stay within its guardrails and still help produce an assessment nobody can defend. Maybe it leaned on three websites that turn out to be one outlet syndicated three times. Maybe it weighted a source with no real access to the region, or a page since edited or deleted. Guardrail monitoring won't flag any of that, because nothing unsafe happened. The tradecraft failed.

4 What OSINT Chain of Custody Requires

For OSINT, an auditable record has to answer practitioner questions, not just engineering ones:

- **Which sources were touched?** That means everything retrieved and when, including material considered and rejected, not just what ended up cited.
- **How was each source graded?** That covers credibility, independence from other sources, and fit to the geography and subject in question. It also needs to record who or what assigned the grade.
- **Under what standard, and which version?** Earlier products must be judged against the rules in force when they were written.
- **What did the AI do?** Which steps were summarization, translation, extraction or drafting, and what did each produce?
- **What did humans change?** Where did an analyst override a grade, discard a machine-suggested source, or rewrite a judgment, and why?
- **Can it be replayed?** Can an overseer, inspector general (IG) or counsel reconstruct the path from sources to conclusion months later, after the web has moved on?

5 The Standards Already Point This Way

ICD 203, *Analytic Standards*, requires analytic products to properly describe the quality and credibility of underlying sources, data and methodologies. It names factors such as accuracy and completeness, possible denial and deception, age and continued currency of information, and source access, validation, motivation, possible bias and expertise.

ICD 206, *Sourcing Requirements for Disseminated Analytic Products*, turns that into mechanics. It calls for source reference citations, source descriptors and source summary statements, which address the strengths and weaknesses of the source base, which sources matter most to key judgments, and which are meaningfully corroborative or conflicting. A companion standard, ICS 206-01, governs citation of publicly available information, commercially available information and OSINT. "Meaningfully corroborative" is, at bottom, a question of source independence.

ICD 505 connects the two problems explicitly. It directs that IC AI be designed and deployed, when relevant, to enable personnel to adhere to analytic standards, including ICD 203 and ICD 206. It calls for data traceability across AI inputs and AI-derived outputs "to the maximum extent practicable." It states that personnel remain responsible and accountable for analysis derived from AI, and that they must be able to review, challenge and reject or replace AI-derived recommendations.

The *IC OSINT Strategy 2024–2026* names the risk plainly. It cautions that the IC must be attuned to "the provenance and validity of information it obtains." It flags generative AI risks including

"inaccuracies and hallucinations" and calls for common OSINT tradecraft standards and tradecraft for human-machine teaming.

None of these directives mandates a particular technology. Together, though, they make clear that a model registry and a guardrail monitor don't, on their own, satisfy the sourcing obligations analysts already carry. With the current strategy period ending in 2026, now is the time to make information chain of custody an explicit design requirement.

6 Questions for Leaders

If you oversee an OSINT program that uses AI, ask five questions:

- 1 For any finished product, can we produce the full list of sources retrieved, not just the ones cited?
- 2 Is every source grade recorded along with the rubric version used and who assigned it?
- 3 Are AI actions and human overrides logged separately and attributed clearly?
- 4 Are those records stored somewhere the AI pipeline itself cannot modify?
- 5 Could an IG or counsel replay the analysis a year from now without the original analyst in the room?

If any answer is "not really," your AI governance may be sound while your analytic accountability is not.

7 About Nordic Road

At Nordic Road, we work on this second problem: OSINT that is graded and traceable to its sources, so every judgment can be followed back to the evidence and standards behind it. The principle is older than AI. If you can't show your sourcing, you don't have an assessment yet.

About the author

Michael Nord is Co-founder & CEO of Nordic Road. He spent 28 years in military intelligence and CIA operations.

Learn more: nordicrd.com

References

1. Pew Research Center, "Americans and AI 2026: Chatbots, Smart Devices and Views on Impact" (June 17, 2026; survey conducted Feb. 17–23, 2026). <https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/>
2. G2, "New G2 Research: Half of B2B Software Buyers Now Start Their Research With AI Chatbots" (press release, April 15, 2026). <https://www.prnewswire.com/news-releases/new-g2-research-half-of-b2b-software-buyers-now-start-their-research-with-ai-chatbots-302742807.html>
3. Office of the Director of National Intelligence, *ICD 203, Analytic Standards* (signed January 2, 2015, as amended). <https://www.odni.gov/files/documents/ICD/ICD-203.pdf>
4. Office of the Director of National Intelligence, *ICD 206, Sourcing Requirements for Disseminated Analytic Products* (2015). <https://www.odni.gov/files/documents/ICD/ICD-206.pdf>
5. Office of the Director of National Intelligence, *ICS 206-01, Citation and Reference for Publicly Available Information, Commercially Available Information, and Open Source Intelligence*. <https://www.odni.gov/files/documents/ICD/ICS-206-01.pdf>
6. Office of the Director of National Intelligence, *ICD 505, Artificial Intelligence* (signed January 17, 2025; technically amended 2025). <https://www.dni.gov/files/documents/ICD/ICD-505-Artificial-Intelligence.pdf>
7. ODNI and CIA, *The IC OSINT Strategy 2024–2026* (March 2024). <https://www.dni.gov/index.php/newsroom/reports-publications/reports-publications-2024/3785-the-ic-osint-strategy-2024-2026>
8. OpenAI, "The Hugging Face incident and the road ahead" (August 26, 2026). <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
9. OpenAI, "OpenAI – Hugging Face Incident Technical Report" (August 2026). <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
10. METR, "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident" (August 26, 2026). <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
11. Anthropic, "Anthropic's Responsible Scaling Policy." <https://www.anthropic.com/responsible-scaling-policy>
12. OpenAI, *Preparedness Framework, Version 2* (April 15, 2025). <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>