Law Reform Institute

lawreforminstitute.org

April 2, 2026

Multiple Award Schedule Program Management Office
Federal Acquisition Service
U.S. General Services Administration

Via email: maspmo@gsa.gov

Re: Comments on 552.239-7001, Basic Safeguarding of Artificial Intelligence Systems

To the MAS Program Management Office:

The Law Reform Institute ("LRI") appreciates the opportunity to comment on the proposed clause 552.239-7001, Basic Safeguarding of Artificial Intelligence Systems.

I. About the Law Reform Institute

LRI is a non-profit organization, staffed by former State Department attorneys, that is focused on federal AI policy. LRI's work on AI security concerns has addressed topics such as (1) the ability of publicly available frontier AI models to output export-controlled information,[1] (2) the executive branch's options for responding to distillation attacks on U.S. AI companies,[2] (3) the need for a narrow antitrust exemption to ensure that AI companies can adequately collaborate on security risks posed by frontier models,[3] and (4) the need for adjustments to the existing export control framework to facilitate evaluations of AI models' biosecurity capabilities.[4]

LRI strongly supports GSA's intent to include subclauses addressing human oversight and incident reporting. LRI submits these comments to suggest several ways in which those subclauses could be strengthened to better achieve GSA's goals, particularly in the context of procurement of frontier AI models. LRI is limiting its comments to subclauses (e)(3), (e)(4), and (h) and is not commenting on the remainder of draft 552.239-7001.

---

[1] Tim Schnabel and Joe Khawam, *AI Outputs and National Security Controls*, Law Reform Institute (Oct. 14, 2025), *available at* https://lawreforminstitute.org/report101425.pdf; *see also* Joe Khawam and Tim Schnabel, "AI Model Outputs Demand the Attention of Export Control Agencies," Just Security (Dec. 12, 2025), https://www.justsecurity.org/126643/ai-model-outputs-export-control/.

[2] Joe Khawam and Tim Schnabel, *Sanctions and Export Control Responses to Adversarial Distillation*, Law Reform Institute (Mar. 13, 2026), *available at* https://lawreforminstitute.org/distillation031326.pdf.

[3] Law Reform Institute, *Antitrust Exemption for AI Frontier Model Risks* (Aug. 12, 2025), *available at* https://lawreforminstitute.org/antitrust081225.pdf.

[4] Doni Bloomfield, Joe Khawam, and Tim Schnabel, *Application of U.S. Export Controls to AI Biosecurity Evaluations* (Oct. 14, 2025), *available at* https://lawreforminstitute.org/bio101425.pdf; *see also* Doni Bloomfield, Joe Khawam, and Tim Schnabel, "How U.S. Export Controls Risk Undermining Biosecurity," Lawfare (Dec. 2, 2025), https://www.lawfaremedia.org/article/how-u.s.-export-controls-risk-undermining-biosecurity.

II. Risks Related to Government Procurement of AI

In the context of litigation, the U.S. Government has recently expressed concerns regarding complications that arise in the procurement of frontier AI models. The Department of War has recently highlighted the "unique, opaque nature" of AI.[5] In particular, DoW stated that AI models may "'drift' or degrade as new data is introduced," that they "require constant tuning," and that they "are acutely vulnerable to manipulation" (such as through "subtl[e] poison[ing of] the training data").[6] DoW also noted that contractors providing AI models to the Government may have the "ability to unilaterally alter system guardrails and model weights without [the Government's] consent," which could "fundamentally change the system's function."[7] DoW is thus concerned about the possibility that the Government may have to "operate a black box controlled by" a contractor.[8] The issues raised by DoW arose in a national security context. However, to the extent that GSA believes that similar concerns could arise in the Government's procurement of AI models (particularly frontier AI models) outside the context of national security systems, parts of draft 552.239-7001 are relevant to addressing the concerns.

III. Strengthening Draft 552.239-7001 to Address the Risks

Draft 552.239-7001 would take important steps toward addressing these risks, but the draft language could be strengthened to accomplish this goal more effectively.

A. Incident Reporting

First, the incident reporting requirement in subclause (e)(4) may require some clarification in the context of frontier AI procurement. Inclusion of a robust reporting requirement, including a short deadline for reporting, is important. However, including only the standard language, without further clarification of how this requirement applies in the context of frontier AI procurement, creates the risk that (e)(4) may be ambiguous in its scope of application and potentially applied in an underinclusive manner.

The framework established under 44 U.S.C. § 3552 defines "incident" to encompass a range of "occurrence[s]," including those that jeopardize the "integrity, confidentiality, or availability of information on an information system" and those that constitute a violation of "security policies" or "security procedures," among others. In the frontier AI procurement context, the Government needs contractors to report incidents that fall within this definition but that would not plausibly arise in other areas of procurement—both in terms of the information systems that are relevant and the types of incidents potentially occurring on those networks.

To avoid under-reporting of incidents, GSA should ensure that (e)(4) is sufficiently clear that the full range of relevant incidents, including those unique to the AI context, would be covered.

---

[5] Under Secretary of War Emil Michael, Memorandum for the Record, "Urgent Supply Chain Risk Analysis: Anthropic's Refusal to Permit Lawful AI Use" (March 2026), *available at* https://storage.courtlistener.com/recap/gov.uscourts.cand.465515/gov.uscourts.cand.465515.96.2_2.pdf.
[6] *Id.*
[7] *Id.*
[8] *Id.*

For example, with respect to information systems, GSA should ensure that contractors are expressly required to report incidents occurring not only on the information systems hosting a model currently used by the Government, but also information systems that are used to develop a possible "successor model" covered by subclause (h) (Change Management). Depending on how the contractor's networks are structured, training of a potential "successor model" may or may not occur on the same "information system" involved in making the original model available to the Government. Yet if an incident occurs during training of a "successor model," the Government needs to be made aware promptly: the right to evaluate a "successor model" is hollow if the Government does not have the ability to promptly assess security incidents during that model's training, as subsequent disclosure and review after the completion of training may be inadequate given the black-box nature of AI models.

Moreover, even before the Government gets access to a "successor model," the contractor might use that new model to adjust the model already being used by the Government. For example, if the Government is using Model 1 and the contractor is developing Model 2 (potentially on a different network), the contractor may use the outputs from Model 2 to make changes to Model 1 (e.g., fine-tuning using synthetic data generated by Model 2).[9] Incidents affecting Model 2's integrity may therefore be highly relevant to Model 1's integrity even absent a formal succession event under (h).[10] Additionally, even where infrastructure used for Model 1 and Model 2 is nominally separate, a compromise in one environment may be probative of risk to the other due to shared personnel, shared security credentials, shared data pipelines, and shared organizational security posture.

With respect to the types of incidents, GSA should ensure that (e)(4) unambiguously requires reporting of security-relevant occurrences related to the model weights (e.g., theft or compromise). GSA should also ensure that (e)(4) requires reporting of relevant occurrences triggered by an AI model itself (not only more traditional incidents caused by external hostile actors or insider threats from employees). In particular, GSA should ensure that a contractor has to report incidents—including in the course of training a possible successor model—relating to unauthorized attempts by a model itself to access higher-privilege credentials, evade security controls (such as monitors, logging, or firewalls), or exfiltrate data.[11] As models develop more agentic capabilities, the Government needs to be made aware of incidents suggesting that the models might exert those capabilities in ways that could undermine the security of Government information systems. GSA should thus ensure that (e)(4) unambiguously covers such incidents,

---

[9] *See generally* Nathan Lambert, "Synthetic Data & Distillation," https://rlhfbook.com/c/12-synthetic-data.

[10] *See, e.g.,* NIST, Adversarial Machine Learning § 2.3 (Poisoning Attacks and Mitigations) (2025), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf; Giovanni De Muri et al., *Pay Attention to the Triggers: Constructing Backdoors That Survive Distillation* (arXiv:2510.18541, 2025), https://arxiv.org/abs/2510.18541; Alex Cloud et al., *Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data* (arXiv:2507.14805, 2025), https://arxiv.org/abs/2507.14805.

[11] *See, e.g.*, OpenAI, "How We Monitor Internal Coding Agents for Misalignment" (Mar. 19, 2026), https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/ (noting deception by agents as "[c]ommon" and "[u]nauthorized data transfer" as "[r]are but high severity"); Alexander Meinke et al., *Frontier Models are Capable of In-Context Scheming* (arXiv:2412.04984, 2024), https://arxiv.org/abs/2412.04984; Apollo Research, *More Capable Models Are Better At In-Context Scheming* (2025), https://www.apolloresearch.ai/blog/more-capable-models-are-better-at-in-context-scheming/.

whether related to models currently supplied to the Government or potential successor models under (h).[12]

B. Human Oversight

Second, the "human oversight" provision in (e)(3) is a valuable inclusion but needs to be strengthened. Currently, (e)(3) would require contractors to provide a minimal level of visibility into "intermediate processing" by the AI, such as by providing summaries. For current frontier models, such visibility would be possible via chain-of-thought (CoT) monitoring, which—even if imperfect—is a valuable tool for gaining insights into how frontier AI models reason and for blocking undesirable actions.[13] CoT monitoring has been used to identify risks such as "intentional hallucinations, reward hacking, alignment faking, and scheming. In addition, monitoring may also protect against attackers that attempt to take advantage of CoT to produce harmful outputs, either by leveraging prompt injections or other adversarial techniques."[14]

The draft clause rightly recognizes that access to reasoning models' intermediate processing (such as CoT) is valuable for oversight purposes. To ensure that oversight is not undermined, GSA should reconsider whether access to mere summaries of intermediate reasoning adequately protects the Government's interest in ensuring oversight of the model's reasoning. A subsequently developed summary of intermediate processing may be nothing more than a post hoc rationalization—potentially incomplete and inaccurate—rather than a clear window into a model's genuine reasoning process.[15] Making the actual CoT (or other intermediate processing) available could be significantly more helpful in reducing this risk.

Additionally, CoT monitorability is fragile, and not all contractors will equally prioritize ensuring that their models' CoT remain monitorable and that they faithfully represent the models' actual reasoning processes. Moreover, "there are potential architectures that would enable latent reasoning in an opaque, black-box manner, which would harm monitorability."[16] To provide the Government with greater confidence in its ability to oversee models, (e)(3) should also require the contractor to do the following:

---

[12] LRI believes that the standard incident reporting requirements, if properly understood and explicitly clarified in the text of the draft provision, would require reporting of all these types of incidents. However, as an alternative approach, GSA could also choose to include these AI-specific incidents in a supplemental reporting requirement for frontier AI procurement.

[13] Frontier Model Forum, *Issue Brief: Chain of Thought Monitorability*, Jan. 27, 2026, https://www.frontiermodelforum.org/issue-briefs/chain-of-thought-monitorability/.

[14] *Id. See also* Tomek Korbak et al., *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety* (arXiv:2507.11473, 2025), https://arxiv.org/abs/2507.11473 ("When models act in a misaligned way, for example, exploiting flaws in their reward functions during training, manipulating data to achieve an outcome, or falling prey to prompt injection attacks, they often explicitly say so in their reasoning traces.").

[15] *See generally* Melody Guan et al., *Monitoring Monitorability* (arXiv:2512.18311, 2025), https://arxiv.org/abs/2512.18311 (finding that "monitors with access to the CoT alone perform surprisingly well compared to monitors with access to only actions and final outputs"); Ivan Arcuschin et al., *Chain-of-Thought Reasoning in the Wild Is Not Always Faithful* (arXiv:2503.08679, 2025), https://arxiv.org/abs/2503.08679 (finding that post-hoc rationalization is a risk even with a full CoT).

[16] Frontier Model Forum, *supra* n. 13.

(1) certify that the model's intermediate processing has not been subject to *direct* optimization pressure from monitors or reward signals in a manner that could induce obfuscation or otherwise degrade monitorability, and fully disclose any use of such direct mechanisms regardless of effect on monitorability;[17]

(2) fully disclose any *indirect* mechanisms, such as process supervision, that could have risked making the model's intermediate processing less faithful to the model's actual decision process;[18]

(3) evaluate the model's performance on benchmarks relevant to the monitorability[19] and controllability[20] of its intermediate processing, and disclose the evaluation results;

(4) ensure that any subsequent changes to the model during the contract term do not degrade the monitorability and controllability of its intermediate processing; and

(5) ensure that any proposed "successor model" has at least equivalent scores on the monitorability and controllability evaluations as the model being replaced.

At least until further architectural breakthroughs alter the types of risks faced in procuring and deploying frontier AI models, the more detailed requirements proposed above could significantly bolster the Government's ability to ensure human oversight of AI models. Consultation with the Department of Commerce's Center for AI Standards and Innovation on this element may also be prudent.

Conclusion

LRI believes that these minor adjustments discussed above would strengthen the laudable goals of those subclauses. We would be happy to discuss these issues further.

Respectfully submitted,

Tim Schnabel
President
Law Reform Institute

---

[17] *See, e.g.*, Bowen Baker et al., *Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation* (arXiv:2503.11926, 2025), https://arxiv.org/abs/2503.11926; Max Kaufmann et al., *Aligned, Orthogonal or In-conflict: When can we safely optimize Chain-of-Thought?* (arXiv:2603.30036, 2026), https://arxiv.org/abs/2603.30036.

[18] *See, e.g.*, Joey Skaf et al., *Large Language Models Can Learn and Generalize Steganographic Chain-of-Thought under Process Supervision* (arXiv:2506.01926, 2025), https://arxiv.org/abs/2506.01926.

[19] *See, e.g.,* Guan, *supra* n.15.

[20] *See, e.g.*, Chen Yueh-Han et al., *Reasoning Models Struggle to Control their Chains of Thought* (arXiv:2603.05706, 2026), https://arxiv.org/abs/2603.05706.