The Collection Fallacy: AI Foundation Models and Executive Order 12333

Tal Feldman*

ABSTRACT

The legal framework governing U.S. intelligence activities was built for a world of filing cabinets and databases. Executive Order 12333 and its implementing Attorney General Guidelines assume that information moves in discrete, retrievable units that can be tagged, queried, and deleted. Artificial intelligence foundation models do not work this way. These models transform training data into statistical weights, learning patterns rather than storing records.

In November 2024, the Office of the Director of National Intelligence released interim guidance applying this framework to AI for the first time. While the guidance sensibly avoids treating the mere acquisition of a commercial model as "collection" of its training data, it provides that agencies which fine-tune or modify models using covered data must comply with the full suite of traditional retention, querying, and dissemination controls. This Comment argues that those obligations, designed for searchable databases, have no coherent application to AI models that do not store or retrieve individual records. The practical result is to foreclose in-house AI development, pushing the intelligence agencies toward commercial black-box systems that are harder to inspect, adapt, or secure. Because the guidance is scheduled for review every six months, there is a narrow window to correct course. This Comment proposes two reforms: safe harbors for models built with verifiable privacy protections and targeted amendments to the Attorney General Guidelines that distinguish between storing information and learning from it.

INTRODUCTION

For over forty years, the rules governing U.S. intelligence have operated on a consistent, record-based logic. Executive Order 12333, first issued in 1981, and each agency's Attorney General-approved guidelines ("AG Guidelines") that implement it are the primary authorities governing how intelligence agencies collect, retain, and disseminate information, particularly about U.S. persons.[1] These rules reflect a

* J.D. Candidate, Yale Law School. I am grateful to Josh Geltzer for conversations that inspired this Comment, and to Liv Hale and the editorial team of the Journal of National Security Law & Policy for their careful edits. All errors are my own.

1. Exec. Order No. 12,333, 46 Fed. Reg. 59941, 59950 (Dec. 4, 1981), *as amended by* Exec. Orders No. 13,284 (2003), 13,355 (2004) and 13,470 (2008), https://perma.cc/W97T-NDEX (§ 2.3 states, "Agencies within the Intelligence Community are authorized to collect, retain or disseminate information concerning United States persons only in accordance with procedures established by the head of the agency concerned and approved by the Attorney General . . . .") [hereinafter E.O. 12333].

database-centric paradigm: information is "collected" when it enters a government system, "retained" when it is stored, and "disseminated" when it is shared.[2] The architecture assumes that information moves in discrete, traceable units subject to tagging, oversight, and deletion.

That assumption no longer holds. New classes of AI systems are entering national security workflows, from intelligence analysis and threat detection to operational planning and decision support.[3] In particular, intelligence analysts are increasingly using large language models and other generative AI systems in classified settings to process raw collection, draft analytic products, and identify patterns across massive datasets.[4] These systems do not store information as traditional records. Instead, they rely on statistical generalization across vast datasets, making it difficult, and often impossible, to retrieve or isolate individual data points.[5] This creates a mismatch: how do rules designed for databases apply to systems that do not function like databases?

In November 2024, the Office of the Director of National Intelligence ("ODNI") took an important first step toward addressing this gap by issuing the *Interim Guidance Regarding the Acquisition and Use of Artificial Intelligence Foundation Models* ("Interim Guidance") that maps the compliance requirements of Executive Order 12333 and the AG Guidelines onto AI foundation models for the first time.[6] This guidance was urgently needed: absent clarity on how existing intelligence-law frameworks apply to AI systems, agencies faced significant uncertainty about the legality of acquiring, developing, and deploying such tools. Without such guidance, for example, an agency acquiring a commercially trained foundation model could arguably be deemed to have "collected" the vast troves of personal data embedded in that model's training dataset, triggering retention, minimization, and oversight obligations that were not designed to apply to AI models.

---

2. See *infra* Parts I and IV for a discussion of this legal architecture generally and a specific example, respectively.

3. Frank Bajak, *Takeaways: How Intelligence Agencies' Are Cautiously Embracing Generative AI*, ASSOC. PRESS (May 23, 2024), https://perma.cc/58JA-UZSB.

4. Derek B. Johnson, *The US Intelligence Community Is Embracing Generative AI*, NEXTGOV FCW (July 3, 2024), https://perma.cc/PW6J-KQY6; Carten Cordell, *How the CIA Is Using Generative AI — Now and into the Future*, FEDSCOOP (June 27, 2024), https://perma.cc/9VBZ-2CMY. *See also* Jim Kelly, *New Google Public Sector Research Shows that Nearly 90% of Federal Agencies Are Already Using AI*, GOOGLE CLOUD (Jan. 13, 2026), https://perma.cc/JL2F-LD93; Cole Stryker, *What are LLMs?*, IBM, https://perma.cc/WPE3-LVQM.

5. Kai Riemer & Sandra Peter, *Conceptualizing Generative AI as Style Engines: Application Archetypes and Implications*, 79 INT'L J. INFO. MGMT., art. no. 102824, Dec. 2024, at 1, 1, https://perma.cc/FVB2-SAVG ("As probabilistic technologies, generative AIs do not store, in any traditional sense, any data or content. Rather, essential features of training data become encoded in deep neural networks as patterns . . . .").

6. OFF. OF THE DIR. OF NAT'L INTEL., COMMON INTELLIGENCE COMMUNITY INTERIM GUIDANCE REGARDING THE ACQUISITION AND USE OF FOUNDATION AI MODELS 2 (2024), https://perma.cc/8RDV-2V3Y [hereinafter INTERIM GUIDANCE]. *See also* Charlie Savage, *Spy Agency Memo Sets Rules for Artificial Intelligence and Americans' Private Data*, N.Y. TIMES (Nov. 14, 2024), https://perma.cc/RSK4-UB9W.

The Interim Guidance goes a long way toward avoiding such categorical outcomes and is largely sound. It recognizes that foundation models involve "technologies that differ substantially from those considered" when the existing legal framework was developed, avoids sweeping all commercial AI purchases into the compliance regime, and commits to revisiting its own conclusions every six months.[7] To date, no such revision has been publicly issued.

However, the Interim Guidance also provides that when an agency adapts a model by training it on data that would usually be subject to its AG Guidelines, such as U.S. person information ("USPI"), the agency must comply with the applicable AG Guidelines, including rules on "use, querying, retention, and dissemination of the covered information."[8] The difficulty, as this Comment explains, is that the AG Guidelines were designed for systems that store information as discrete records, not for AI models that encode statistical patterns learned from data. The two are fundamentally different: a database preserves information that can be queried, shared, or deleted; an AI model transforms data into weights and parameters that cannot be meaningfully subjected to those same operations.[9] Still, intelligence agencies are legally required to comply with their AG Guidelines.[10] If those guidelines cannot coherently apply to in-house AI development, agencies may be restricted from doing any of their own AI development even as the operational demand for tailored AI tools continues to grow.

This Comment proceeds in four parts. Part I outlines the legal architecture that governs intelligence data handling, tracing the database-centric assumptions embedded in Executive Order 12333 and the AG Guidelines. Part II explains how foundation models differ fundamentally from traditional databases—a distinction the Interim Guidance itself acknowledges—and why the compliance obligations designed for databases do not translate to systems that learn from data rather than store it. Part III analyzes the ODNI's Interim Guidance, arguing that while it is an important first step in bringing government AI use within the fold of existing regulations, its provision on model modification imports compliance obligations that lack coherent application in this context. Part IV examines the practical consequences: risk-averse agencies in the Intelligence Community ("IC")[11] will reasonably read the modification provision as foreclosing in-house AI development, pushing them toward general-purpose commercial models they cannot inspect or adapt for intelligence work. The conclusion proposes reforms, including safe harbors for models developed with verifiable privacy protections and targeted amendments to the AG Guidelines, to fill this doctrinal gap.

---

7. INTERIM GUIDANCE, *supra* note 6, at 1–2.

8. *Id.* at 2.

9. *See infra* Part II for a discussion of memorization risks in AI systems and available technical safeguards.

10. E.O. 12333, *supra* note 1, § 2.3.

11. Comprising eighteen agencies—including the NSA, CIA, and military intelligence branches—the Intelligence Community is a sprawling, decentralized bureaucracy overseen by the Office of the Director of National Intelligence. *Members of the IC*, OFF. OF THE DIR. OF NAT'L INTEL., https://perma.cc/MKC4-5G4X.

I. OVERVIEW OF THE LEGAL FRAMEWORKS GOVERNING INTELLIGENCE

A. Executive Order 12333 and the Attorney General Guidelines

Executive Order ("E.O.") 12333 remains the central authority governing how intelligence agencies collect, retain, and share information, particularly about "United States persons."[12] The Order sets out the structure, responsibilities, and limitations of the IC's activities, including rules for signals intelligence, covert action, and coordination among agencies.[13] Critically, it requires that any collection, retention, or dissemination of USPI be conducted under procedures approved by the Attorney General.[14] Each agency must develop its own set of implementing AG Guidelines which translate the Order's high-level directives into operational rules tailored to that agency's mission and authorities.[15] Once information is "collected," the Guidelines impose a set of privacy-preserving obligations, wherein agencies must minimize identifying details, retain data only as long as necessary, and tightly control dissemination to others.[16]

Accordingly, much turns on when information is considered "collected." While E.O. 12333 does not define the term, the AG Guidelines adopted by different intelligence agencies do, and their definitions are broadly consistent.[17] Across the IC, information is generally treated as "collected" when it is received by an agency component.[18] These definitions also clarify what is not collection: data that merely passes through a system, is viewed without being saved or used, or is maintained on behalf of another agency without access for intelligence purposes.[19] This threshold permits agencies to filter irrelevant or incidental material, such as information swept up in signals intelligence, before compliance obligations attach. But once information is deemed collected, the full suite of retention, minimization, and dissemination obligations attaches.[20]

---

12. E.O. 12333 defines a "United States person" as "a United States citizen, an alien known by the intelligence agency concerned to be a permanent resident alien, an unincorporated association substantially composed of United States citizens or permanent resident aliens, or a corporation incorporated in the United States, except for a corporation directed and controlled by a foreign government or governments." E.O. 12333, *supra* note 1, § 3.4(i).

13. E.O. 12333, *supra* note 1, at pmbl., §§ 1.1, 1.2.

14. E.O. 12333, *supra* note 1, § 2.3.

15. *Id.*

16. *See infra* Part IV.A.

17. *See, e.g.*, DEP'T OF DEF., DOD MANUAL 5240.01: PROCEDURES GOVERNING THE CONDUCT OF DOD INTELLIGENCE ACTIVITIES 45 (2016) [hereinafter DOD MANUAL 5240.01]; DEP'T OF THE TREASURY, PROCEDURES FOR INTELLIGENCE ACTIVITIES 35 (2022); OFF. OF THE DIR. OF NAT'L INTEL., INTELLIGENCE ACTIVITIES PROCEDURES APPROVED BY THE ATTORNEY GENERAL PURSUANT TO EXECUTIVE ORDER 12333 29 (2020). To access the full list of agency Attorney General Guidelines, see generally OFF. OF THE DIR. OF NAT'L INTEL., ATTORNEY GENERAL APPROVED U.S. PERSON PROCEDURES UNDER E.O. 12333 (2024), https://perma.cc/LS2R-KM5C.

18. *See, e.g.*, DOD MANUAL 5240.01, *supra* note 17, at 45 (defining collection).

19. *Id.*

20. *Id.* at 15–23.

B. Other Legal Authorities

While this Comment focuses on E.O. 12333 and its implementing AG Guidelines, other legal frameworks, including the Foreign Intelligence Surveillance Act ("FISA") and the Privacy Act of 1974 ("Privacy Act"), also impose restrictions on the handling of U.S. person information. These statutes are noted in the Interim Guidance as potentially applicable to AI systems.[21] Unlike E.O. 12333, both FISA and the Privacy Act establish private rights of action against agencies.[22] This introduces an additional layer of legal risk, transforming compliance failures from internal oversight issues into potential judicial and financial consequences.

FISA governs electronic surveillance and physical searches involving U.S. persons and domestic communications, subject to court approval and strict minimization procedures.[23] The Privacy Act regulates federal systems of records retrievable by name or other identifiers and imposes disclosure and correction obligations.[24] AI systems complicate compliance with both FISA and the Privacy Act. These authorities' conceptual models mirror that of E.O. 12333 and presume that information can be stored and queried as discrete units.

II. AI FOUNDATION MODELS ARE NOT DATABASES

Before evaluating how the existing compliance obligations outlined in Part I might apply, it is necessary to understand what AI foundation models do with data, and how that differs from traditional storage or retrieval. Foundation models are large AI systems trained on massive datasets to learn general patterns in language.[25] They are not built to store or retrieve records but are instead optimized to predict likely outputs based on prior examples.[26] During training, the model's internal parameters—numerical values known as "weights"—are repeatedly adjusted to capture patterns in the data.[27] The result is a mathematical representation of relationships within the training data, not a copy of the data itself.[28]

To use the Interim Guidance's language, intelligence agencies may also "modify" a foundation model through further training or fine-tuning on new data.[29] This

21. INTERIM GUIDANCE, *supra* note 6, at 2.

22. *See* Foreign Intelligence Surveillance Act of 1978, 50 U.S.C. § 1810; 18 U.S.C. § 2712; Privacy Act of 1974, 5 U.S.C. § 552a(g).

23. *See* 50 U.S.C. §§ 1801–1813, 1821–1829 (governing electronic surveillance and physical searches, respectively).

24. *See* 5 U.S.C. § 552a(a)(5), (d).

25. Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models* 3 (Stan. Ctr. for Rsch. on Found. Models, 2021), https://perma.cc/NQJ4-ULEN.

26. *Id.* at 78–79 (explaining that foundation models encode generalized knowledge implicitly in learned parameters, while explicit factual storage and retrieval are usually external to the model).

27. Xiaoguang Tu, Zhi He, Yi Huang, Zhi-Hao Zhang, Ming Yang & Jian Zhao, *An Overview of Large AI Models and Their Applications*, 2 VISUAL INTEL., art. no. 34, Dec. 2024, at 1, 15, https://perma.cc/8NVE-KUQU; AARON COURVILLE, IAN GOODFELLOW & YOSHUA BENGIO, DEEP LEARNING 149–50 (2016) (discussing stochastic gradient descent).

28. Adam Buick, *Copyright and AI Training Data—Transparency to the Rescue?*, 20 J. INTELL. PROP. L. & PRAC. 182, 186 (2024), https://perma.cc/W8MT-EMMC.

29. INTERIM GUIDANCE, *supra* note 6, at 4.

is already happening. Scale AI's "Defense Llama," for instance—Meta's Llama model fine-tuned on military doctrine, international humanitarian law, and IC guides—is already in experimental and operational use on classified U.S. government networks.[30] Fine-tuning is how a general-purpose AI system becomes useful for a specific mission.[31] A model trained on the open internet knows very little about particular national security objectives, but further training on an agency's own data can teach it to perform those tasks.[32] The process works by updating the model's weights so that the resulting model reflects features of the new training data in its responses.[33] After training, there is no stored copy of the input documents inside of the model—only a changed set of weights that shape how the model responds to prompts.[34]

Agencies may also "augment" a model by attaching an external search tool or database.[35] This allows the model to pull in outside data while generating outputs without altering the model's internal parameters.[36] One common approach, Retrieval-Augmented Generation ("RAG"), lets the model query external sources while generating responses.[37] Unlike modification, augmentation leaves the model's weights unchanged.[38]

In some cases, AI models have been shown to memorize and regurgitate passages from their training data, particularly when specific sequences appear frequently or contain unusual patterns that make them statistically distinctive.[39] This raises the fundamental risk that the Interim Guidance is trying to address: that an AI model might inadvertently output sensitive information about U.S. persons

30. Brandi Vincent, *Scale AI Unveils "Defense Llama" Large Language Model for National Security Users*, DEFENSESCOOP (Nov. 4, 2024), https://perma.cc/QHQ6-6MKY; The Scale Team, *Defense Llama: The LLM Purpose-Built for American National Security*, SCALE AI BLOG (Nov. 4, 2024), https://perma.cc/5CJ9-LH2U.

31. Bommasani et al., *supra* note 25, at 3, 85–89.

32. *Id.*

33. *Id.*; INTERIM GUIDANCE, *supra* note 6, at 4.

34. Bommasani et al., supra note 25, at 3, 85–89.

35. INTERIM GUIDANCE, *supra* note 6, at 4.

36. *Id.*

37. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel & Douwe Kiela, Conference Paper, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 9459, 9459 (2020), https://perma.cc/TD8W-49BR; Kim Martineau, *What Is Retrieval-Augmented Generation?*, IBM RSCH. BLOG (Aug. 22, 2023), https://perma.cc/T63V-N6DQ.

38. Lewis et al., *supra* note 37, at 9459; Martineau, *supra* note 37; INTERIM GUIDANCE, *supra* note 6, at 4.

39. *See, e.g.*, Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr & Chiyuan Zhang, Conference Paper, *Quantifying Memorization Across Neural Language Models*, INT'L CONF. ON LEARNING REPRESENTATIONS 1, 1 (2023), https://perma.cc/JWE3-ZJ4U; Michael Aerni, Javier Rando, Adoardo Debenedetti, Nicholas Carlini, Daphne Ippolito & Florian Tramèr, Conference Paper, *Measuring Non-Adversarial Reproduction of Training Data in Large Language Models*, INT'L CONF. ON LEARNING REPRESENTATIONS 1, 1 (2025), https://perma.cc/3B7N-CBDL; Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch & Nicholas Carlini, Conference Paper, *Deduplicating Training Data Makes Language Models Better*, 1 60TH ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS: PROC. CONF. (LONG PAPERS) 8424, 8428–29 (2022), https://perma.cc/89TL-CG2L.

present in its training data.[40] However, this risk is increasingly manageable through known technical safeguards: differential privacy techniques inject noise during training to prevent the model from encoding identifiable information; output filtering can block responses that contain specific categories of sensitive data; red-teaming exercises systematically probe models for memorization vulnerabilities before deployment; audit logging tracks model outputs to detect and respond to any inadvertent disclosures after the fact.[41] Most importantly for intelligence law, these techniques can be targeted specifically at personally identifiable information, preventing models from outputting certain categories of data regardless of what appears in the training data.[42] None of these safeguards is perfect, but in combination they can reduce memorization risk to a level far below what the Interim Guidance's compliance framework appears to assume.

These technical capabilities are directly relevant to the doctrinal question this Comment raises: if a model can be designed to prevent the disclosure of USPI, the case for subjecting it to compliance obligations designed for queryable databases is considerably weakened. Yet, as the next Part shows, the ODNI's Interim Guidance does not draw this distinction.

III. THE NOVEMBER 2024 GUIDANCE AS A WATERSHED MOMENT

The November 2024 Interim Guidance represents a watershed moment in how the IC interprets and applies legacy surveillance law to AI systems. Formally, it is merely guidance.[43] However, in the intelligence and national security space,

40. INTERIM GUIDANCE, *supra* note 6, at 1 ("Whether the acquisition of a specific FM [foundation model] constitutes the 'collection' of any training data will depend on the specific facts and circumstances of the FM, and in particular whether the acquiring IC element has the capability to access the training data (either directly or by reverse-engineering the model and/or de-anonymizing the data) in its original form as well as the authorization and intent to do so.").

41. *See* Jian Du, Song Li, Xiangyi Chen, Siheng Chen & Mingyi Hong, Dynamic Differential-Privacy Preserving SGD 1 (Oct. 30, 2021) (unpublished article) (on file on arXiv, https://perma.cc/6LAT-2ZWE); NAT'L INST. OF STANDARDS & TECH., ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK: GENERATIVE ARTIFICIAL INTELLIGENCE PROFILE (NIST AI 600-1) 35 (2024) (calling for AI red-teaming to detect output of training data); NAT'L INST. OF STANDARDS & TECH., MANAGING MISUSE RISK FOR DUAL-USE FOUNDATION MODELS (NIST AI 800-1) 11–13 (2024) (recommending structured red-team testing to mitigate leakage risks); *Security Planning for LLM-Based Applications*, MICROSOFT, https://perma.cc/R3QG-6TA9 (advising continuous audit-logging of LLM access); NAT'L INST. OF STANDARDS & TECH., GUIDE TO COMPUTER SECURITY LOG MANAGEMENT (NIST SP 800-92) 1–2 (2006) (explaining how audit logs deter and detect unauthorized disclosure of sensitive data).

42. *See, e.g.*, Shubhi Asthana, Ruchi Mahindru, Bing Zhang & Jorge Sanz, *Adaptive PII Mitigation Framework for Large Language Models* (IBM Rsch., 2025), https://perma.cc/KX34-4R4H (introducing a system for redacting personally-identifiable information from AI outputs); *Presidio: Data Protection & De-identification SDK*, MICROSOFT PRESIDIO, https://perma.cc/59AC-2KM8 (demonstrating an open-source toolkit that automatically detects and redacts personally identifiable information before model output reaches the user); Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang & Xuebing Zhou, *PrivacyScalpel: Enhancing LLM Privacy via Interpretable Feature Intervention with Sparse Autoencoders*, PROC. 8TH BLACKBOXNLP WORKSHOP: ANALYZING & INTERPRETING NEURAL NETWORKS FOR NLP 226 (2025), https://perma.cc/PF7M-YSQW (demonstrating a technique that drives personally-identifiable information-leakage rates to 0 percent while preserving more than 99 percent of model utility).

43. INTERIM GUIDANCE, *supra* note 6, at 1 ("This Interim Guidance is supplemental to and does not supersede any applicable law, including E.O. 12333 and the AG Guidelines.").

ODNI guidance carries near-mandatory weight.[44] The document is thus far more than a procedural memo. It signals how the IC intends to categorize and constrain AI activity going forward—and its interpretations are likely to shape not just compliance practice, but also future policy, oversight, and internal culture.

A. Acquisition of AI Models

The main concern driving the Interim Guidance is preventing intelligence agencies from gaining unauthorized access to covered information, particularly USPI.[45] This is evident in its treatment of AI models acquired from vendors. On its own, AI model acquisition does not constitute "collection" of the model's training data, including USPI, so long as the agency does not have the ability to access that data "in its original form," either directly or through reverse engineering.[46] The Interim Guidance thus adopts an access-based compliance model: if an agency cannot inspect or retrieve the underlying data, then it is not treated as having "collected" it for legal purposes.

This approach is both sensible and consistent with how modern foundation models typically function. Most commercial models acquired by the IC are black-box systems, meaning the agency receives only the final model with no access to the training dataset.[47] In such cases, the model is not functionally equivalent to a trove of documents.

B. Modification of AI Models

The Interim Guidance's treatment of model modification appears to diverge from its access-based approach to acquisition. When an agency modifies or augments a foundation model using "covered information collected by the IC element," the Interim Guidance requires compliance with the full suite of requirements imposed by its AG Guidelines, "including where applicable the relevant rules on the use, querying, retention, and dissemination of the covered information."[48] Furthermore,

---

44. OFF. OF THE DIR. OF NAT'L INTEL., INTELLIGENCE COMMUNITY POLICY GUIDANCE 101.1 1 (2009), https://perma.cc/CJV2-2WXP (noting that Intelligence Community Policy Guidance, along with Intelligence Community Directives, "establish policy and guidance, and provide formal and definitive direction to the IC for the purposes of achieving a unified, integrated and effective IC"). For discussions of Intelligence Community Directives, which are a higher-priority level of guidance, see PRIVACY OFF., OFF. FOR C.R. & C.L. & DEP'T OF HOMELAND SEC., EXEC. ORDER 13636 PRIVACY & CIVIL LIBERTIES ASSESSMENT REPORT 3 n.13 (2014), https://perma.cc/V4Q6-MNVN (suggesting that Intelligence Community Directives are "binding on the entire IC"); *Nomination of Robert S. Litt & Stephen W. Preston: Hearing Before the S. Select Comm. on Intelligence*, 111th Cong. 10 (2009) (statements of Robert S. Litt & Stephen W. Preston) (affirming that Intelligence Community Directives are binding on agencies).

45. "Covered information refers to the category of data to which an individual agency's specific AG guidelines apply, including U.S. person information." INTERIM GUIDANCE, *supra* note 6, at 4.

46. INTERIM GUIDANCE, *supra* note 6, at 2.

47. *See, e.g.*, OPENAI, GPT-4 TECHNICAL REPORT (2023), https://perma.cc/B56Z-NMJB; ANTHROPIC, CLAUDE 3 MODEL CARD (2024), https://perma.cc/VU4T-Y8E2; *Our Next-Generation Model: Gemini 1.5*, GOOGLE (Feb. 15, 2024), https://perma.cc/8NVJ-FWPH; The Scale Team, *supra* note 34.

48. INTERIM GUIDANCE, *supra* note 6, at 2.

the Interim Guidance says that these activities "may also implicate" rules under FISA and the Privacy Act.[49]

This framework suggests that these rules attach not only to the training data, but to the model itself. A reasonable reading of these provisions indicates that once a model has been trained or fine-tuned on covered data, it becomes subject to the same legal restrictions that would govern a database containing that information.[50] The modified model, in effect, inherits the compliance status of its training data.

The Interim Guidance also imposes substantial procedural requirements: agencies must "take all necessary steps ahead of any modification or augmentation" to ensure compliance, including consulting with the General Counsel's office, Chief AI Officer, and others.[51] "Covered information cannot be used to modify or augment a [foundation model] absent an authorized purpose."[52] Agencies must coordinate with the Department of Justice ("DOJ") and ODNI before modifying models using FISA-obtained or FISA-derived information.[53]

The guidance goes further still. It provides that querying rules contained in the AG Guidelines "may also apply to prompts if such rules apply to data used to train or modify a [foundation model]." By way of example, the guidance states that if an agency fine-tunes a model on data collected under Executive Order 12333, "the applicable querying rules under E.O. 12333 would apply" to prompts of that model. This effectively treats every prompt of a fine-tuned model as a query of the data used to train it. The provision illustrates the broader doctrinal problem: compliance obligations designed for searchable databases are being extended to systems that do not function as such.

However, a properly designed model cannot reproduce specific training records on demand, and modern safeguards can further reduce memorization risk.[54] Yet the Interim Guidance's explicit invocation of "querying" and "retention" rules, designed for systems where analysts search identifiable documents and can delete them when they must, suggests it treats modified models as if they function like databases.

---

49. *Id.*

50. *See infra* Part IV. An alternate reading of the Interim Guidance would treat the phrase "where applicable" as a limiting modifier, giving agencies discretion to determine which AG Guideline obligations attach to AI systems and which do not. But this reading is difficult to sustain. The overarching command is mandatory: agencies "must comply" with their AG Guidelines. The "where applicable" clause modifies only the illustrative list that follows—use, querying, retention, and dissemination—and those terms are themselves database-native operations that map poorly onto foundation models. A risk-averse General Counsel's office is far more likely to read the obligation broadly and the exception narrowly, applying the full compliance framework rather than relying on an ambiguous modifier to carve out AI-specific flexibility. As Part IV demonstrates, this practical reading creates significant operational and conceptual difficulties.

51. INTERIM GUIDANCE, *supra* note 6, at 2.

52. *Id.*

53. *Id.*

54. *See supra* Part II.

IV. THE EFFECTS OF THE GUIDANCE'S INTERPRETATION

A. Limiting In-House AI Development

The previous Part established that the Interim Guidance extends the full suite of AG Guidelines obligations to any model modified using covered information. This Part examines what that compliance looks like in practice—and why, when applied to foundation models, these obligations do not merely create heavy compliance burdens but cease to function coherently.

The Department of Defense's AG Guidelines ("DoD Guidelines") illustrate the problem.[55] Though other agencies have their own rules, the DoD Guidelines reflect the broader approach across the Intelligence Community.[56] These Guidelines were written for documents, not for statistical systems like AI models.

Retention. Under the DoD Guidelines, unevaluated USPI may be retained for no more than five years, even if collected lawfully.[57] During that period, agencies must document the basis for retention, conduct periodic reviews, and limit access only to those who have clearance and mission requirements to view the data.[58] But when applied to AI foundation models, this retention logic breaks down. If a model has been fine-tuned on a dataset that includes a single document later subject to deletion, the agency may be forced to discard the entire model. There is no reliable way to isolate or remove the influence of individual documents embedded in billions of parameters, short of full retraining.[59]

Dissemination. Dissemination rules fare no better. Under the DoD Guidelines, USPI may generally only be disseminated to individuals with a need for it and appropriate training, with varying levels of further procedural requirements depending on the person receiving the information.[60] If dissemination involves a large volume of unevaluated USPI, heightened approvals are required, including certification by senior leadership that no lesser amount would suffice.[61] Applying this to foundation models, a person accessing a certain model would need to have clearance to view the entire training dataset, often millions of documents, even if the model never returned any of the information to the user. Furthermore, there is

55. *See generally* DOD MANUAL 5240.01, supra note 17.

56. For the complete list of agency AG Guidelines, see OFF. OF THE DIR. OF NAT'L INTEL., ATTORNEY GENERAL APPROVED U.S. PERSON PROCEDURES UNDER E.O. 12333 (2024), https://perma.cc/LS2R-KM5C.

57. DOD MANUAL 5240.01, *supra* note 17, at 15–16.

58. *Id*. at 18.

59. Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo & Yang Liu, *Rethinking Machine Unlearning for Large Language Models*, 7 NATURE MACH. INTELL. 181, 183 (2025) ("In the literature, 'exact' unlearning, which involves retraining the model from scratch after removing specific training data points, is often considered the gold standard." and discussing the "infeasibility of pinpointing and attributing training data points designated for unlearning"); Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie & Nicolas Papernot, *Machine Unlearning*, 2021 IEEE SYMP. ON SEC. & PRIV. 141, 143 (2021).

60. DOD MANUAL 5240.01, *supra* note 17, at 20–21.

61. *Id*. at 22.

no meaningful way to disseminate "less" of a model. These rules, built for memo distribution and cable sharing, are structurally incompatible with model deployment.

Querying. Under the DoD Guidelines, electronic queries of unevaluated USPI must be mission-relevant, narrowly tailored, and supported by written justifications.[62] In a database context, this is manageable: queries are logged, scoped, and reviewed. But foundation models do not retrieve records. A query presupposes the existence of stored, retrievable information; a model prompt triggers statistical generation. The compliance framework assumes a fundamentally different operation.

The Interim Guidance suggests that other legal authorities, such as FISA and the Privacy Act, apply, but these, too, collapse in application to foundation models. Under FISA, the agency must follow "minimization procedures" for acquisition and retention of data, prohibit some of its dissemination, and more.[63] Under the Privacy Act, assuming a model qualifies as a "system of records," the agency must let any covered individual learn whether the system contains a record about them, obtain an intelligible copy, seek amendment, and more.[64] These obligations are incredibly difficult, if not infeasible, to apply to AI models.

The result is that agencies face a choice between avoiding in-house model development or navigating a compliance framework designed for an entirely different technology. Even where privacy risk is low—due to red-teaming, output filtering, and audits—the compliance uncertainty is high. The safest path becomes reliance on commercial black-box systems, which may undermine oversight and create dependencies on vendors outside government control.

B. The Interpretation Favors Acquisition of AI Models

The Interim Guidance draws a stark distinction between externally acquired and internally developed AI models. By contrast to modification, the acquisition of a model from a third party is not treated as implicating the full suite of E.O. 12333 and AG Guidelines obligations, so long as the agency cannot access the underlying training data in its original form.[65] The practical effect of this distinction is to create a significant regulatory asymmetry between externally acquired and internally developed models. While this distinction is grounded in a defensible principle—agencies should face fewer obligations when they cannot access training data—its practical consequence may be significant.

Indeed, the guidance pushes agencies to acquire models from external vendors rather than develop them in-house using privacy-protective techniques like differential privacy, red-teaming, and output filtering. The security implications may be serious. Vendor-provided models are often opaque: agencies usually cannot inspect their training data, audit their internal reasoning, or verify that they meet government security standards. They cannot be adapted to mission-specific

62. *Id*. at 18.
63. 50 U.S.C. § 1801(h).
64. 5 U.S.C. § 552a(b)-(e).
65. *See supra* Section III.A.

requirements without triggering the very compliance obligations agencies sought to avoid. Further, they may introduce supply-chain dependencies on private companies whose incentives may not align with national security priorities. The framework, in other words, trades a manageable privacy risk for a far less manageable security one.

C. The Interpretation Favors Augmentation over Modification

Alongside its permissive stance on model acquisition, the Interim Guidance implicitly encourages another workaround: augmentation. Techniques like RAG allow agencies to pair foundation models with external databases without modifying the model itself.[66] When prompted, the system retrieves relevant documents, feeds them to the model as context, and discards them after generating a response.[67] Even though these systems still interact with covered information, and thus trigger compliance obligations, the key architectural choice is that the model and database remain technically disassociated.[68] This separation allows agencies to map legacy legal frameworks onto the system more easily, since the sensitive data is never embedded in the model's weights.

This approach offers a legally safer architecture, but it is no substitute for full model adaptation. Augmentation has limitations: it adds latency, requires well-maintained databases, and may not match the fluency or performance of a model fine-tuned on mission-specific data.[69] It appears that the legal architecture now rewards this constrained design over more capable, secure, and auditable in-house development. Agencies are left choosing between suboptimal functionality and the heavy compliance burdens associated with training and fine-tuning.

CONCLUSION

Modern AI systems are already shaping how wars are fought, how threats are detected, and how intelligence is produced. If the United States cannot build and adapt its own models because of well-intentioned but ill-fitting rules, then national security itself is at risk. The November 2024 Interim Guidance was an admirable first attempt to translate legacy surveillance law into the AI era, but even ODNI recognized it was only a starting point, committing to update the document every six months.[70] That promise should now be honored.

The Guidance's built-in six-month review cycle provides the natural vehicle for these reforms. In the short term, ODNI and DOJ should carve out a safe harbor for truly privacy-protective models: systems that demonstrate non-memorization, undergo rigorous red-teaming, and filter outputs for personally identifiable

66. Lewis et al., *supra* note 37, at 9459; Martineau, *supra* note 37.

67. Lewis et al., *supra* note 37, at 9459; Martineau, *supra* note 37.

68. Tal Feldman, *The Law Already Supports AI in Government — RAG Shows the Way*, JUST SEC. (May 16, 2025), https://perma.cc/5GPH-4GRN.

69. Jignesh Patel, *The Limitations of Model Fine-Tuning and RAG*, INFOWORLD (May 14, 2024), https://perma.cc/28QT-C5JE.

70. INTERIM GUIDANCE, *supra* note 6, at 1.

information. Compliance should attach to what the user sees, not to what affected the statistical weights hidden deep in an AI system. Agencies that can demonstrate their models do not retain or disclose covered information in identifiable form should not face the same compliance burden as agencies operating traditional data repositories.

The longer project is reforming the underlying rules. The AG Guidelines, and the E.O. 12333 ecosystem they implement, need a clear distinction between storing a document and learning from it. "Collection," "retention," and "dissemination" make sense for files; they break down for AI models. Updating these definitions to focus on access and observable outputs would preserve civil-liberties protections while freeing technologists to build mission-specific tools.

Until the legal framework is updated, agencies will be unable to train or fine-tune models internally. Instead, they will be limited to acquiring general-purpose systems developed by external vendors—tools that often cannot be meaningfully inspected, adapted, or aligned with mission needs. Meanwhile, adversaries are already building and deploying models purpose-built for intelligence, warfare, and strategic influence. Without reform, the United States risks falling behind not because it lacks the talent or resources, but because its own legal constraints prevent it from using them.

***