Watermarking Without Standards Is Not AI Governance

Alexander Nemecek [1] **Yuzhou Jiang** [1] **Erman Ayday** [1]

Abstract

Watermarking has emerged as a leading technical proposal for attributing generative AI content and is increasingly cited in global governance frameworks. This paper argues that current implementations risk serving as symbolic compliance rather than delivering effective oversight. We identify a growing gap between regulatory expectations and the technical limitations of existing watermarking schemes. Through analysis of policy proposals and industry practices, we show how incentive structures disincentivize robust, auditable deployments. To realign watermarking with governance goals, we propose a three-layer framework encompassing technical standards, audit infrastructure, and enforcement mechanisms. Without enforceable requirements and independent verification, watermarking will remain inadequate for accountability and ultimately undermine broader efforts in AI safety and regulation.

1. Introduction

"A law without teeth is just a suggestion."

This adage captures a growing concern in AI governance, where policies are advancing more quickly than the technical tools available to enforce them. A prominent example of this mismatch is *watermarking*, a family of techniques designed to embed identifiable signatures into AI-generated content. While policymakers increasingly cite watermarking in governance frameworks, and often include caveats such as "as technically feasible," there remains a structural overreliance on watermarking in the absence of robust alternatives. This reliance persists even with partial awareness of its technical limitations, driven by the urgency of addressing provenance challenges without a mature suite of tools.

Recent advances in generative AI have significantly increased the scale and realism of synthetic content, including text, images, and audio (Spennemann, 2025; Fisher et al., 2024). As such content is deployed in sensitive areas like education, healthcare, and finance, policymakers have identified attribution and provenance as urgent challenges. Misattributed content contributes to the spread of misinformation and to technical failures such as feedback loops in model training, including contamination and eventual model collapse (Shumailov et al., 2024).

In response, watermarking has gained traction in both technical research (Kirchenbauer et al., 2023; Fernandez et al., 2023; Chen et al., 2023) and policymaking. The United States (U.S.) Executive Order (EO) 14110, for example, mandates "state-of-the-art" provenance tools and cites watermarking explicitly (Exe, 2023). The European Union's (EU) AI Act requires machine-readable content markings (EUA, 2024), with similar provisions being proposed in jurisdictions worldwide (Zhao et al., 2024a). Although watermarking is just one of several provenance strategies, it occupies an important role in current governance discourse.

This reliance rests on a flawed foundation. While watermarking techniques vary across modalities, most remain brittle, difficult to audit, and proprietary. Policymakers often assume these methods can be standardized and verified. In practice, industry deployments obscure technical details while asserting compliance, turning watermarking into a box-checking exercise rather than a meaningful tool. Lacking common standards, evaluation infrastructure, or defined threat models, current implementations are unlikely to fulfill their intended governance role.

To address this growing disconnect, **we advance two central positions:**

- **Watermarking schemes must be designed with verifiability and auditability as primary technical requirements, rather than implemented as proprietary black boxes.**
- **Policymakers must establish technical standards and independent testing to ensure that watermarking fulfills governance goals in practice.**

To support these positions, we make three contributions. First, we analyze emerging watermarking mandates and

[1]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH., U.S.A.. Correspondence to: Alexander Nemecek <ajn98@case.edu>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

demonstrate they assume levels of technical feasibility that current systems do not meet. Second, we examine the incentive structures driving industry watermarking deployments, showing how market dynamics often disincentivize robust, auditable implementations. Third, we propose a three-layer framework, spanning technical design, audit infrastructure, and enforcement which realigns the design of watermarking systems with their intended governance functions.

While prior work has analyzed watermarking's technical constraints, our contribution reframes these issues through a governance lens and provides a structured, actionable framework for aligning technical design with regulatory intent. Without intervention at the intersection of policy and design, watermarking is unlikely to deliver meaningful accountability. Instead, it risks becoming a symbolic substitute for the more demanding components of effective AI regulation.

2. Governance Aspirations vs. Technical Realities

As watermarking becomes a feature of AI governance proposals, many mandates, despite explicit references to technical feasibility limits, ultimately rely on assumptions that current systems cannot consistently fulfill at scale. While documents such as NIST AI 100-4 (Chandra et al., 2024) have informed interagency views and illustrate that portions of the policy community recognize watermarking's limitations, this nuanced understanding has not prevented reliance on watermarking in practice.

To concretize our analysis, we focus on a representative subset of prominent governance frameworks (e.g., U.S. EO 14110 (Exe, 2023), EU AI Act (EUA, 2024), California AI Transparency Act (California State Legislature, 2024), China Deep Synthesis Provisions (CAC, 2022)) and widely cited industry deployments (e.g., Google SynthID (Google DeepMind, 2023), OpenAI's classifier (OpenAI, 2023b)) to illustrate the structural mismatch between policy expectations and current watermarking implementations. These were selected due to their explicit references to watermarking and provenance tools, their global relevance, and their documented influence on industry practices.

We identify three assumptions and examine how this gap between qualified expectations and deployment realities creates challenges for effective governance.

2.1. Assumption I: Watermarking Is Technically Robust Against Modification

Many governance proposals expect watermarking to provide substantial robustness against both benign transformations and adversarial tampering, while acknowledging technical constraints. However, the practical interpretation of these qualifiers often remains undefined, creating a gap between policy expectations and measurable technical capabilities. Whether applied to text, images, or audio, watermarking is expected to remain detectable throughout the content lifecycle.

> *"The disclosure is permanent or extraordinarily difficult to remove, to the extent it is technically feasible."* — California AI Transparency Act (California State Legislature, 2024)

> *"Such techniques... should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible..."* — EU AI Act (EUA, 2024)

We acknowledge that policies such as the EU AI Act and U.S. directives often qualify expectations with language like "as technically feasible." However, while these clauses explicitly include broad caveats tied to technical feasibility, they nonetheless function as de facto requirements within governance frameworks that lack alternative provenance mechanisms. Policymakers reference watermarking as a compliance tool without defining measurable thresholds for what constitutes "technically feasible" robustness, leaving room for symbolic adherence without verifiable accountability. These documents reflect the shared policy expectation that watermarking should resist removal or degradation in any environment. None define quantitative thresholds, making compliance and enforcement difficult. Additionally, no shared common metrics exist for evaluating watermark robustness under adversarial conditions.

While it is possible these policies are intended to create incentives for firms to adopt watermarking practices rather than mandating strict compliance, effective governance in safety-critical contexts requires moving beyond aspirational language to testable standards that translate "feasibility" into operational criteria.

In practice, watermarking schemes often fall short of this robustness ideal. Google's SynthID (Google DeepMind, 2023), for example, embeds imperceptible watermarks in AI-generated text. While resilient to certain distortions, its detection can be evaded by simple edits such as character perturbations or short-form text paraphrasing (Dathathri et al., 2024; Nemecek et al., 2024). Across all modalities, there is currently no standardized evaluation protocol for watermark robustness, and few systems offer guarantees under adversarial or worst-case conditions.

While policymakers acknowledge technical feasibility constraints, the absence of shared benchmarks or quantitative thresholds for interpreting these constraints creates two problems: first, it makes meaningful regulatory compliance unverifiable; second, it enables firms to implement weak watermarking schemes while claiming alignment with policy

language that includes appropriate caveats. Without concrete robustness criteria, policy becomes aspirational rather than actionable, with the challenge lying not in policymaker awareness of limitations, but in translating qualified expectations into measurable, enforceable standards.

2.2. Assumption II: Watermarking Enables Independent and Reliable Detection

A second assumption in many governance proposals is that watermarking will enable independent third parties (e.g., regulators, researchers, platform operators) to reliably detect AI-generated content. This assumption seeks to enforce provenance or hold actors accountable for synthetic content.

> *"[Develop] methods for the verification of statements of digital content provenance to ensure authenticity such as watermarking or classifiers."* — S. 3312 (U.S. Senate, 2023)

> *"Deep synthesis service providers shall... establish and complete management structures for algorithmic mechanism... and verification..."* — China's Deep Synthesis Provisions (CAC, 2022)

While these statements do not explicitly mandate external third-party verification, they reflect a policy orientation toward establishing verifiable detection. In practice, effective governance often requires mechanisms that allow independent verification to ensure accountability beyond the claims of the deploying entity. However, most watermarking systems today do not support this expectation due either to limited transparency or configuration dependence, limiting the practical enforceability of provenance requirements.

For instance, OpenAI previously released a classifier for detecting AI-generated text, but withdrew it due to unreliability (OpenAI, 2023b). Google's SynthID offers a more complete approach, with watermarking and detection tools for text, images, and audio. SynthID-Text (Dathathri et al., 2024) supports watermarking and detection in public tools, but detection across modalities like image and video still requires access to internal configurations. Additionally, users need to apply to receive access to the configurations used in those deployments. Unless watermarking keys or models are shared or standardized, third-party detection remains limited to contexts explicitly designed for interoperability.

While the technical foundations for third-party detection are emerging, the practical reality remains constrained. Most watermarking approaches do not currently support universal, auditable detection, and governance frameworks often overlook the infrastructure and standardization needed to make detection viable at scale.

2.3. Assumption III: Industry Will Voluntarily Align with Governance Goals

A final dynamic in governance frameworks, particularly in the U.S. and EU, is a reliance on a combination of voluntary commitments and aspirational regulatory frameworks to encourage generative AI providers to adopt watermarking practices aligned with policy goals.

> *"encouraging... allies and partners to support voluntary commitments similar to those that United States companies have made... and to develop common regulatory and other accountability principles..."* — U.S. EO 14110 (Exe, 2023)

> *"All stakeholders... are encouraged to take into account... ethical principles for the development of voluntary best practices and standards."* — EU AI Act (EUA, 2024)

While policymakers acknowledge that regulatory aspirations, alongside voluntary commitments, are insufficient for robust accountability, a combination of political constraints, perceived trade-offs related to economic competitiveness, and reputational considerations has hindered the adoption of more binding approaches. This has led to a governance-by-consensus model that, while pragmatic under current conditions, remains fragile and difficult to enforce, with enforcement mechanisms needed to transform these commitments into effective oversight still lacking.

The U.S. Biden-Harris Administration secured voluntary commitments from seven leading AI companies explicitly referencing watermarking to lead safe and transparent AI (White House, 2023a). An additional group of eight companies later joined these commitments (White House, 2023b).

However, implementation remains fragmented, with many deployments being firm-specific and lacking interoperability. Few are publicly auditable or developed through shared infrastructure. In the absence of enforceable standards or independent oversight, these commitments risk enabling symbolic compliance, signaling safety while delivering little practical governance capability.

Voluntary alignment also depends on political continuity. In 2025, the U.S. Trump-Vance Administration rescinded EO 14110, directing agencies to "suspend, revise, or rescind" related initiatives (White House, 2025). The administration's stated focus on technological competitiveness leaves the governance status of watermarking ambiguous, casting uncertainty over the future of watermarking as a governance tool. By contrast, China's Deep Synthesis Provisions (CAC, 2022) offer a more directive model, mandating compliance and audits, but even in such regimes, enforcement is uneven

and robustness is not guaranteed due to opaque, regionally inconsistent implementation practices within China's regulatory system (Freedom House, 2024).

While voluntary commitments may promote cooperation, they offer a fragile foundation for governance. When watermarking imposes costs or strategic risks, firms are unlikely to sustain alignment in the absence of legal or institutional compulsion.

3. Why the Gap Persists: Industry Incentives

Despite growing regulatory attention, industry watermarking implementations remain fragmented. This is not simply a technical lag but reflects a deeper misalignment between governance objectives and industry incentives. Without enforceable standards, firms have strong incentives to pursue symbolic or minimal compliance (Aaronson, 2024).

First, watermarking offers reputational value. Companies can signal alignment with governance goals by announcing watermarking initiatives, often without disclosing technical details or enabling independent verification. This satisfies public and regulatory expectations at low cost while sidestepping the challenges of building robust systems.

For example, Google's SynthID (Google DeepMind, 2023) offers multi-modal watermarking capabilities, while Meta's Fundamental AI Research (FAIR) team has announced watermarking techniques across modalities (Fernandez et al., 2024; Sander et al., 2024) with research and announcements continuing over the past several years (Meta, 2023). Similarly, companies such as Microsoft and OpenAI have endorsed initiatives like the Coalition for Content Provenance and Authenticity (C2PA, 2025), which embed cryptographic metadata to support provenance tracking (OpenAI, 2025; Smith, 2024). However, these frameworks face limitations in enforceability and robustness, since metadata can be stripped through common transformations such as screenshotting or re-encoding. Moreover, endorsement of such frameworks does not guarantee durable provenance; for instance, OpenAI's DALL·E (OpenAI, 2023a) watermarking approach has been shown to be easily removable with basic image editing, illustrating the fragility of current watermarking measures even when explicitly deployed as governance tools (Collins, 2024).

Second, robust watermarking entails economic and strategic risks. If detection tools are open source, they can be exploited by modifying outputs to evade detection or crafting content that falsely appears watermarked. Conversely, restricted access to detection capabilities would require companies to develop secure infrastructure to manage watermarking keys, detection thresholds, and access control via cryptographic or trust-based frameworks. These trade-offs make firms hesitant to invest in public, auditable watermarking systems that support third-party verification.

Third, companies face disincentives to move unilaterally. Without coordination, early adopters risk losing users if watermarking is perceived as restrictive, such as for moderation or IP enforcement, especially if competitors offer unmarked alternatives. In a rapidly evolving market, the incentive to retain users outweighs alignment with long-term governance objectives.

Finally, regulatory uncertainty undermines long-term planning, further amplifying industry hesitation. Seen with the revocation of EO 14110 (White House, 2025), shifting political priorities can quickly alter the expected policy landscape. In such an environment, companies are hesitant to invest in watermarking infrastructure that may soon be obsolete.

Together, these forces collectively encourage minimal, firm-specific, and performative watermarking environments, an outcome misaligned with the goals of effective AI oversight.

4. A Three-Layer Approach to Enforceable Watermarking

To bridge the gap between policy ambition and technical feasibility, we propose a three-layer framework: technical requirements, audit infrastructure, and policy enforcement. Each layer targets a key weakness in current watermarking practices and together establishes end-to-end accountability. While this approach faces multiple objections regarding implementation complexity, regulatory fragmentation, and innovation constraints, we argue these challenges are surmountable and necessary for meaningful governance. Figure 1 summarizes the framework, with implementation strategies in Appendix A and discussion of potential practical, political, and philosophical objections in Appendix B.

4.1. Technical Requirements

Effective governance depends on watermarking systems that are technically sound and externally testable. At a minimum, watermarking methods should demonstrate robustness to both benign transformations and adversarial modifications designed to evade detection. Systems must report standardized performance metrics under defined perturbation sets, as specified in a shared technical specification enabling independent third-party evaluation. Crucially, these guarantees do not require full transparency of proprietary model internals but must expose interfaces and artifacts, such as detectors or keys, that enable reproducible evaluation by third parties. Establishing this provides the groundwork for audit protocols, extending technical evaluation practices already emerging in AI governance through initiatives like red-teaming frameworks and standardized benchmark development (Ahmad et al., 2025; Anthropic, 2025), with detailed implementation strategies outlined in Appendix A.1.

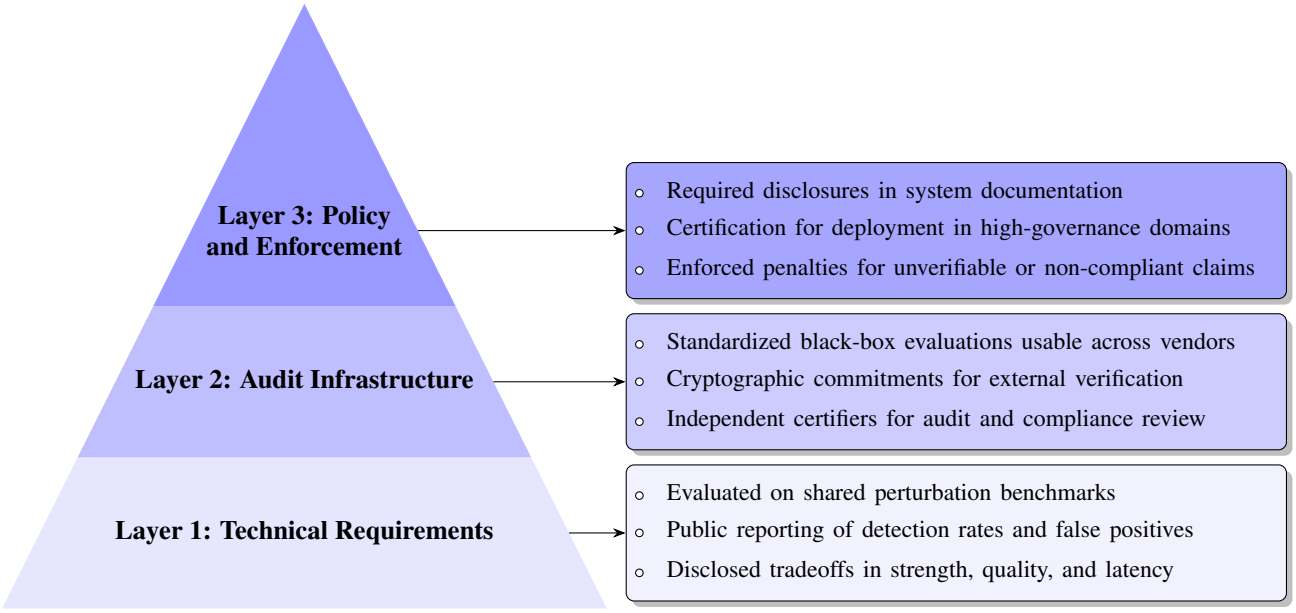

Figure 1. **Three-layer framework for enforceable watermarking.** Each layer represents a distinct governance function: technical guarantees, independent auditability, and regulatory enforcement. Arrows point to concrete mechanisms that instantiate the requirements at each level, linking system design to policy accountability.

4.2. Audit Infrastructure

Robust watermarking requires more than technical performance as it must be verifiable through independent and reproducible evaluation. This layer establishes the infrastructure for third-party audits, including standardized black-box testing protocols, shared testbeds, and recognized certifying bodies. Audits should reflect real-world use conditions rather than idealized lab settings and must work across different vendors without requiring access to proprietary internals. Instead, systems should expose externally testable behaviors or cryptographic commitments that allow verification without reverse engineering. By enabling interoperable, independent testing, this layer transforms watermarking from internal assurance into a publicly accountable mechanism, with specific protocols detailed in Appendix A.2.

This approach builds upon existing audit infrastructure efforts in AI governance, including model cards for standardized reporting (Mitchell et al., 2019), which have demonstrated the value of structured disclosure frameworks. Similarly, emerging practices around data donations for research and auditing purposes (Ohme & Araujo, 2022) and third-party algorithmic auditing initiatives (Costanza-Chock et al., 2022) provide important precedents for the independent verification infrastructure watermarking governance requires.

4.3. Policy and Enforcement

The final layer ensures that technical and audit standards translate into real-world accountability. Without legal mandates and institutional enforcement, even robust watermarking systems may go unused or implemented inconsistently. These mechanisms should be tied to existing regulatory frameworks (e.g., EU AI Act's high-risk application strategies (EUA, 2024)), rather than requiring entirely new institutional apparatus. Public disclosures, building on established frameworks (Mitchell et al., 2019) and extending to watermarking-specific documentation, help standardize expectations and support auditability. Following precedents like SOC 2 cloud compliance (Palo Alto Networks, 2025), certification should be required for deployment in high-risk contexts, with graduated enforcement beginning in government procurement and extending to broader commercial deployment as outlined in Appendix A.3. This layer connects the technical and audit layers to institutional incentives, closing the loop from design to deployment.

5. Conclusion

Watermarking is emerging as a pillar of AI governance, but without enforceable standards and verifiable implementation, it risks becoming symbolic rather than a mechanism of accountability. We argue robustness, verifiability, and auditability must be built into watermarking from the start and not retrofitted for compliance. This vision requires a structural shift for clear technical baselines, independent audit infrastructure, and binding regulatory enforcement. Achieving accountability will depend on sustained collaboration between policymakers, industry actors, and researchers.

References

Executive order 14110 of october 30, 2023: Safe, secure, and trustworthy development and use of artificial intelligence, November 2023. Federal Register, Vol. 88, No. 210, pp. 75191–75208.

Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act), July 2024. Official Journal of the European Union, L 2024/1689, 12 July 2024.

Aaronson, S. Ai safety and watermarking. Talk at Simons Institute, UC Berkeley, 2024.

Ahmad, L., Agarwal, S., Lampe, M., and Mishkin, P. Openai's approach to external red teaming for ai models and systems. *arXiv preprint arXiv:2503.16431*, 2025.

Anthropic. Progress from our frontier red team, 2025. Accessed: May 12, 2025.

C2PA. Coalition for content provenance and authenticity — advancing digital content transparency and authenticity, 2025. Accessed: 2025-07-03.

CAC. Provisions on the administration of deep synthesis internet information services, November 2022. Effective January 10, 2023.

California State Legislature. California ai transparency act, sb 942, chapter 291, September 2024. Approved by the Governor on September 19, 2024; effective January 1, 2026.

Chandra, B., Dunietz, J., and Roberts, K. Reducing risks posed by synthetic content an overview of technical approaches to digital content transparency, 2024-11-20 05:11:00 2024.

Chen, G., Wu, Y., Liu, S., Liu, T., Du, X., and Wei, F. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

Collins, B. The ridiculously easy way to remove chatgpt's image watermarks. *Forbes*, 2024. Accessed: 2025-07-03.

Costanza-Chock, S., Raji, I. D., and Buolamwini, J. Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1571–1583, 2022.

Dathathri, S., See, A., Ghaisas, S., Huang, P.-S., McAdam, R., Welbl, J., Bachani, V., Kaskasoli, A., Stanforth, R., Matejovicova, T., et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035): 818–823, 2024.

Fernandez, P., Couairon, G., Jégou, H., Douze, M., and Furon, T. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.

Fernandez, P., Elsahar, H., Yalniz, I. Z., and Mourachko, A. Video seal: Open and efficient video watermarking. *arXiv preprint arXiv:2412.09492*, 2024.

FIDO Alliance. Fido alliance: Reducing reliance on passwords, 2025. Accessed on 2025-07-06.

Fisher, S. A., Howard, J. W., and Kira, B. Moderating synthetic content: The challenge of generative ai. *Philosophy & Technology*, 37(4):133, 2024.

Freedom House. China: Freedom on the net 2024 country report, 2024. Accessed: 2025-06-28.

Google DeepMind. Synthid, 2023. Accessed: 2025-05-09.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Meta. Stable signature: A new method for watermarking images created by open source generative ai, 2023. Accessed: 2025-07-03.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Nemecek, A., Jiang, Y., and Ayday, E. Topic-based watermarks for llm-generated text. *arXiv preprint arXiv:2404.02138*, 2024.

Ohme, J. and Araujo, T. Digital data donations: A quest for best practices. *Patterns*, 3(4):100467, 2022. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2022.100467.

OpenAI. DALL·E 3 System Card. Technical report, OpenAI, 2023a. Accessed: 2025-07-03.

OpenAI. New ai classifier for indicating ai-written text, January 2023b. Tool discontinued as of July 20, 2023 due to low accuracy.

OpenAI. C2pa in chatgpt images, 2025. Accessed: 2025-07-03.

Palo Alto Networks. What is soc 2 compliance?, 2025.

Pub, F. Secure hash standard (shs). *Fips pub*, 180(4):180–4, 2012.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Sander, T., Fernandez, P., Durmus, A., Furon, T., and Douze, M. Watermark anything with localized messages. *arXiv preprint arXiv:2411.07231*, 2024.

Sasson, E. B., Chiesa, A., Garman, C., Green, M., Miers, I., Tromer, E., and Virza, M. Zerocash: Decentralized anonymous payments from bitcoin. In *2014 IEEE symposium on security and privacy*, pp. 459–474. IEEE, 2014.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Smith, B. Combating abusive ai-generated content: A comprehensive approach. *Microsoft on the Issues*, 2024. Accessed: 2025-07-03.

Spennemann, D. H. Delving into: the quantification of ai-generated content on the internet (synthetic data). *arXiv preprint arXiv:2504.08755*, 2025.

Tabassi, E. Artificial intelligence risk management framework (ai rmf 1.0). 2023.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

U.S. Senate. Artificial intelligence research, innovation, and accountability act of 2023, s. 3312, 118th congress, November 2023. Introduced by Sen. John Thune on November 15, 2023; reported with amendment on December 18, 2024.

White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, July 2023a. Accessed: 2025-05-09.

White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI, September 2023b. Accessed: 2025-05-09.

White House. Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence, January 2025. Accessed: 2025-05-09.

World Wide Web Consortium (W3C). Making the web work, 2025. Accessed on 2025-07-06.

Zhao, X., Gunn, S., Christ, M., Fairoze, J., Fabrega, A., Carlini, N., Garg, S., Hong, S., Nasr, M., Tramer, F., Jha, S., Li, L., Wang, Y.-X., and Song, D. Sok: Watermarking for ai-generated content. *arXiv preprint arXiv:2411.18479*, 2024a.

Zhao, X., Zhang, K., Su, Z., Vasan, S., Grishchenko, I., Kruegel, C., Vigna, G., Wang, Y.-X., and Li, L. Invisible image watermarks are provably removable using generative ai. *Advances in Neural Information Processing Systems*, 37:8643–8672, 2024b.

A. Three-Layer Implementation Examples

A.1. Layer 1: Technical Requirements

Watermarking spans multiple generative modalities, including text, image, and audio (Zhao et al., 2024a). Each modality requires distinct robustness evaluations to assess watermark persistence under realistic conditions. Robustness refers to a watermark's ability to withstand both benign transformations (e.g., compression, summarization) and adversarial modifications (e.g., targeted perturbations designed to evade detection). Layer 1 calls for a shared technical specification to evaluate watermark durability against such transformations in a unified, reproducible manner.

Each modality exhibits specific transformations that risk watermark integrity:

- Text: paraphrasing, spelling alterations, lexical substitutions, summarization, and machine translation (Kirchenbauer et al., 2023).
- Image: lossy compression (e.g., JPEG), geometric manipulations (cropping, rotation, scaling), color shifts, and style transfer (Zhao et al., 2024b).
- Audio: background noise injection, pitch shifts, re-encoding artifacts, and time-stretching (Chen et al., 2023).

A watermarking scheme must be evaluated against these transformations not only for robustness, but also for recoverability without false attribution.

We propose the creation of a shared technical specification for benchmarking watermarking robustness, defining standardized perturbation suites, evaluation protocols, and reporting formats across modalities. This specification would enable both independent researchers and regulatory agencies to implement their own evaluation tools, ensuring independent verification while maintaining consistency across evaluations. Reference implementations should be provided to facilitate adoption and reproducibility, but their use would be optional.

The specification should include, at minimum, performance metrics such as detection accuracy under perturbation, false positive and false negative rates, and area under the ROC curve (AUC), with modality-specific metrics reported where applicable. It should also document tradeoffs introduced by watermarking, including potential reductions in model output quality, inference latency increases, or decreased robustness to noise, enabling users and auditors to understand the operational impact of embedding techniques. To stay ahead of emerging threats, the specification should support evolution through contributions from red-teaming initiatives and adversarial research challenges, ensuring that benchmark suites evolve alongside attack capabilities (Ahmad et al., 2025; Anthropic, 2025). In practice, this would enable both independent researchers and regulatory agencies to evaluate watermarking claims reproducibly across diverse systems.

While Executive Order 14110 (Exe, 2023) identified the National Institute of Standards and Technology (NIST) as a natural candidate for maintaining benchmarking standards, the pace of AI system deployment exceeds NIST's current update cycles. We recommend that NIST serve as the registry and certifier for the technical specification while a more agile, community-driven consortium develops and updates the specification to remain responsive to emerging technical landscapes.

To institutionalize this layer, any system claiming compliance must submit and demonstrate conformance to the shared technical specification, using either their own evaluation implementation or a community reference implementation if desired. Recognizing potential concerns over intellectual property or security, alternative compliance paths, such as exposing a detection API or providing zero-knowledge proofs, could also be considered, provided they allow equivalent third-party evaluation. Outputs will be scored on standardized perturbation suites, with results logged in a public registry. This shifts evaluation from self-reported metrics to verifiable compliance with shared robustness expectations. Such a system ensures that claims of watermark durability are not merely assertions, but demonstrable properties grounded in reproducible tests. These reproducible evaluations form the empirical basis for audit infrastructure in Layer 2, enabling certifiers to assess compliance without needing privileged access.

A.2. Layer 2: Audit Infrastructure

To verify that deployed watermarking schemes meet the minimal technical requirements outlined in Layer 1, robust audit systems must be established. However, watermarking must remain resilient to perturbations while concealing key detection parameters. Full public access to a detection system would enable adversaries to remove or forge watermarks, undermining

the scheme's integrity. Consequently, audit systems must operate in a black-box setting where evaluators do not access model internals or watermarking code, but instead submit test content and observe binary or probabilistic detection outcomes. This approach introduces challenges since black-box detectors are vulnerable to repeated querying attacks, where adversaries iteratively probe a system to learn its boundaries (Sadasivan et al., 2023). To mitigate this, detection interfaces must implement access controls, either via restrained queries or containerized deployments (Tramèr et al., 2016). These include vendor-specific configurations such as closed-source APIs or on-premise audit tools, depending on the operational context or vendor policy.

Watermarking deployments should also support cryptographic commitments that allow external verification without revealing sensitive internals. This can involve traditional hash-based attestations (e.g., SHA-256 commitments to watermark parameters) (Pub, 2012) or emerging cryptographic techniques such as zero-knowledge proofs (Sasson et al., 2014). For example, a system might prove to an auditor that it conforms to a certified detection threshold, without disclosing the underlying keys or watermark design.

A central question concerns auditor identity and governance. Potential certifiers include NIST-accredited laboratories, third-party security firms, or consortia established for compliance auditing (Tabassi, 2023). Certification levels may include baseline compliance for systems meeting minimum robustness, interoperability certification for systems supporting cross-vendor detection, or adversarial resilience certification for schemes robust under threat-model testing. Audit cadence is equally critical. Evaluations should occur: (i) pre-deployment, to certify governance readiness, (ii) periodically post-deployment, to catch degradation or drift, and (iii) post-incident, in response to adversarial exploits or system failures. As adversaries evolve, the audit layer must ensure that claims of durability and detectability remain valid over time. Auditors must be able to reproduce robustness metrics defined in Layer 1 using only externally exposed interfaces. This includes performance on benchmark perturbation sets, detection behavior under adversarial conditions, and public system documentation. Certified systems should be logged in a public registry indicating audit status and any revocations or compliance failures. While audit processes introduce operational overhead, they establish trust by linking technical claims to external validation. Layer 2 builds directly on the technical guarantees of Layer 1, ensuring they are externally verified, reproducible, and trustworthy in deployment contexts.

A.3. Layer 3: Policy and Enforcement

Layer 3 implementation follows established regulatory precedents rather than requiring novel institutional frameworks. Technical standards development mirrors successful multi-stakeholder initiatives such as the FIDO Alliance for authentication (FIDO Alliance, 2025) or W3C for web standards (World Wide Web Consortium (W3C), 2025). Audit infrastructure builds on existing frameworks such as SOC 2 compliance (Palo Alto Networks, 2025), where independent auditors verify controls without accessing proprietary systems. Enforcement leverages existing regulatory mechanisms: the EU AI Act's classification system (EUA, 2024), U.S. agency authorities (i.e., Federal Trade Commission (FTC), Cybersecurity and Infrastructure Security Agency (CISA)), and sectoral frameworks in healthcare (Food and Drug Administration (FDA)) and finance (Securities and Exchange Commission(SEC)).

To ensure that technical requirements and audit mechanisms translate into industry commitment, enforceable legal and regulatory consequences must be in place. Without binding consequences, the outputs of Layers 1 and 2 risk being symbolic, allowing firms to claim compliance without delivering accountability. Layer 3 operationalizes these technical and audit layers by anchoring them in policy frameworks that mandate compliance and impose consequences for deviation.

Certification based on audit results should be required for deployment in high-governance contexts (e.g., elections, education, public interfaces). Systems that fail certification would face deployment restrictions or public disclosure of non-compliance. In tandem, mandated disclosures in system documentation must include audit status, robustness guarantees, and known failure modes, extending documentation frameworks to watermarking-specific requirements. These disclosures provide transparency while enabling downstream accountability. Non-compliant systems failing to meet audit standards or refusing audit participation should face graduated enforcement actions. These may include fines, removal of deployment licenses for regulated sectors, or public listing in non-compliance registries. For firms, such penalties also carry reputational risk, further incentivizing alignment with certification pathways.

Enforcement authority must be clearly defined. Governance bodies such as NIST (U.S.) (Exe, 2023), designated regulators under the EU AI Act (EUA, 2024), or China's Cyberspace Administration (CAC, 2022) can serve as enforcement agents within their jurisdictions. However, because AI systems operate across borders, enforcement must also address international fragmentation. A model deployed in one jurisdiction, producing outputs accessible in another, would make alignment across

regimes critical for consistent governance.

Finally, policy enforcement must evolve alongside the underlying technologies. Just as audit protocols adapt to new adversarial techniques, enforcement mechanisms must include provisions for periodic policy review and revision. Feedback loops from auditors, researchers, and affected stakeholders can help keep regulatory frameworks responsive and legitimate. By linking technical performance and audit compliance to real-world consequences, Layer 3 closes the loop from design to deployment. It ensures that watermarking systems are not merely well-engineered, but meaningfully accountable in practice.

B. Opposing Views

While we argue for enforceable watermarking as a critical step toward meaningful AI governance to transform it from a symbolic gesture into a mandatory mechanism, it is important to acknowledge opposing perspectives that highlight real implementation challenges across practical, political, and philosophical dimensions.

B.1. Practical Objections

Our proposed three-layer framework, though necessary for governance, poses substantial practical hurdles. The technical complexity alone requires the development of robust watermarking schemes across modalities, supported by standardized benchmarks and test protocols. Beyond technical design, the framework calls for the creation of new institutions to oversee audit infrastructure and certification processes. These demands would foreseeably slow the current rapid pace of generative AI innovation, particularly for smaller firms and open-source communities.

The computational overhead and implementation costs of robust watermarking could create significant barriers to entry, potentially consolidating market power among larger technology firms capable of absorbing these costs. Open-source AI development, which has been crucial for democratizing access to generative models, faces particular challenges under mandatory watermarking regimes. Unlike proprietary systems, open-source models cannot rely on centralized key management or restricted detection APIs, making them inherently more vulnerable to watermark removal while facing higher compliance burdens.

Additionally, the standardization process itself presents practical challenges. Achieving consensus on technical specifications across diverse stakeholders, including competing technology firms, academic researchers, and international regulatory bodies, would require substantial negotiation periods. The rapid evolution of generative AI creates a fundamental mismatch with the typically slower pace of regulatory processes.

B.2. Political Objections

In addition to technical challenges, the fragmented landscape of both AI development and global regulation limits the interoperability and enforceability of watermarking systems. Proprietary models often adopt incompatible watermarking implementations, while open-source forks can bypass governance requirements entirely. On the regulatory side, jurisdictions such as the U.S., EU, and China are developing divergent standards, ranging from voluntary commitments to strict mandates, resulting in conflicting requirements. A model compliant in one region may not meet the standards of another, undermining global accountability efforts.

Political opposition to mandatory watermarking often centers on competitiveness concerns. Industry advocates argue that unilateral implementation of strict watermarking requirements could disadvantage domestic firms relative to international competitors operating under more permissive regimes. Furthermore, the enforcement mechanisms required for effective watermarking governance could conflict with existing regulatory frameworks and jurisdictional boundaries. Cross-border AI services complicate enforcement, as content generated in one jurisdiction may be consumed in another with different regulatory requirements.

Amid this fragmentation, some would argue that voluntary compliance and soft norms offer a more flexible and innovation-friendly alternative to rigid enforcement. As noted in Section 2.3, shifts in U.S. policy have demonstrated this volatility while one administration secured voluntary commitments from fifteen major AI firms, a subsequent administration rescinded those governance efforts. For some, this approach appears more feasible and politically viable than establishing binding standards.

B.3. Philosophical Objections

At a deeper level, watermarking itself could be questioned on whether it aligns with principles of digital rights and technological development. In terms of privacy, the argument is mandatory watermarking creates infrastructure for content tracking that could enable broader surveillance capabilities, even if initially designed for provenance verification. From a technological philosophy perspective, others contend that watermarking attempts to impose artificial scarcity and control on digital information that is inherently copyable and modifiable, viewing it as incompatible with the open, transformative nature of digital content creation and sharing.

B.4. Response to Opposing Views

However, we contend that voluntary commitments alone are insufficient. In the absence of enforceable requirements and independent audits, companies are more likely to implement minimal or symbolic watermarking measures that fall short of supporting meaningful oversight. While the practical, political, and philosophical objections reflect legitimate concerns, they do not negate the governance challenges that watermarking aims to address. The practical challenges of implementation, while significant, are not insurmountable given sufficient coordination and investment, and political fragmentation makes coordinated standards even more necessary to prevent inconsistent governance practices. In contexts where AI-generated content poses demonstrable risks to public welfare, some degree of technical accountability infrastructure is necessary to preserve the broader benefits of AI systems while mitigating potential harms.