Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Versioned Governance for Frontier AI: A Software-Engineering-Inspired Approach

Zhixin Xue

franka328162810@gmail.com

Independent Researcher

Research Article

Keywords: AI governance, version management, software engineering, frontier AI, transparency, regulatory documentation

Posted Date: October 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11082416/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background FrontierAI laboratories govern deployed systems via model specifications, usage policies, and safety frameworks. These governance documents undergo repeated revisions, yet current practices rarely distinguish substantive normative policy changes from mere editorial clarifications. This produces silent drift: continuous normative modification without clear documentation of changes, reasons for modification, and the institutional authority under which changes were made.

The following are examples of an AI API that identifies new drugs and vaccines for a given disease, scaling from a simple “Name that drug!” function to much more advanced models.

This paper reviews the revision histories for five publicly available AI governance instruments that are slated for release between 2019 and 2025; OpenAI usage policies and Model Specific, Anthropic Responsible Scaling Policy, the EU Artificial Intelligence Act and the NIST AI Risk Management Framework. This paper identifies the areas where current practices seemed aligned with formal versioning discipline, and it pinpointed the areas where there are persistent implementation gaps.

Version number tracking vocabulary is becoming more frequent at the whole document level, while the specific term level is still not being used. The type of AI disruptive technology that incites a substantial shift from a normative route is an exogenous shock. Current changelogs are descriptive, but they are not change classified. Explicit structural rollback mechanisms are not found in any of the examined instruments. Through this analysis, we realize that there are genuine limits of the software-as-a-policy analogy, taking these limitations as hard design constraints for possible AI regulation instead of unconsidered counter-arguments.

Implications: AI labs can “self-engineer” a lightweight globally adaptable, versioning approach to formally defined meta-changes, clause granular history logs, design rollback aware implementations, and misclassification preserved as independently audit-bounded events that can be independently audited by regulators/independent auditors and formally accepted by regulators as an independent practice. Does not imply specific compelling AI system substantive policy content.

1. Introduction

1.1 The problem of silent drift

A frontier AI laboratory in 2025 maintains, at minimum, four kinds of public-facing governance document. There is a model specification or its equivalent, stating what the model is supposed to do and what it is supposed to refuse. There is a usage policy, stating what users may do with the system. There is a safety framework, often labeled with a phrase such as responsible scaling or preparedness, stating which capability thresholds trigger which mitigations. And there are model cards or system cards documenting specific releases.

Each of these documents is revised over time. The revisions are sometimes announced, sometimes not. They are occasionally accompanied by a dated changelog, more often not. They are almost never versioned in a way that distinguishes substantive normative change from editorial clarification.

The costs of this arrangement are easy to underestimate because they fall on people outside the laboratory. A compliance officer at an enterprise customer needs to know whether the policy her firm certified against last quarter still says what it said. A court asked to evaluate conduct that occurred in March needs the March text, not the current one, and needs some assurance about how the two differ. A journalist who notices a revised sentence cannot tell, from the outside, whether the sentence was rewritten for elegance or because a permission boundary moved. In each case the question is the same: what changed, at what level of significance, and on whose authority? Current practice gives no structured answer. I call the resulting condition silent drift: continuous change in the normative layer of a deployed system, without a legible record of the change itself.

Silent drift is not deception. Nothing in it requires bad faith. It is a missing grammar of change—an institutional artifact produced when documents built for static publication are used to govern systems that are anything but static. Exhortations toward greater transparency do not repair a missing grammar. They produce more text.

1.2 Why software engineering provides a template

Software engineering has spent several decades developing exactly the discipline that AI governance documentation now needs. The discipline did not arise from philosophical reflection. It arose from the practical pain of managing long-lived systems used by populations that depended on stable behavior across upgrades. Semantic versioning, structured changelogs, deprecation policy, and rollback eligibility are answers to questions that any system of sufficient longevity and consequentiality eventually has to answer, and frontier AI governance has reached that point. Deployed models persist across years. Their governance documents are consulted by regulators, courts, journalists, enterprise customers, and competing laboratories.

The analogy has limits, and this paper does not romanticize it. Policy is not code. Still, policy that governs adaptive technical systems cannot remain pre-versional. Section 2.3 states the limits honestly, including the uncomfortable fact that versioning discipline is enforced imperfectly even in software itself. Section 5.2 returns to the deeper disanalogies and takes them seriously one by one.

1.3 Contribution of this paper

This paper has three contributions focused on practical AI governance implementation. The first is diagnostic: this paper identifies silent drift as a structural failure in current frontier AI governance practice, which is outside the scope of the research undertaken in certain labs. The conventional transparency framing means that the temporal governance gap does not come under the purview of static rigid disclosures. The second is constructive: this paper presents a practical idea of a four-tiered

versioning model, with the name vM.m.p.b, for the AI governance documents, which creates a difference between the major items, minor ones, patches, and updates. Each of these update levels has explicit changelog requirements, rollback eligibility prerequisites, and public audit requirements that are usable in real world AI deployments and compliance practices. The third is empirical: through the systematic study of the version history of five real-world governance tools that were implemented between 2019 and 2025 (with planned future retrospective), this paper shows that the versioning discipline features are already observable in the fragmented real-world deployments; the gaps in the versioning implementation are huge in the areas with extreme rapid normative change; and the inherent limitations of the software policy analogy are observable in the empirical public record itself.

2. Background

2.1 The transparency literature and its

diachronic blind spot

Calls for transparency in algorithmic systems have a long history. Pasquale (2015) belled the black box society and called for consequential systems to be opened up. (Turilli and Floridi 2009) already had a robust ethical reason for transparency in transparency as a pro tanto condition of good governance. The later literature encumbered the ideal, Burrell (2016) identified three types of opacity, Kroll (2018) revealed that inscrutability is typically a policy choice rather than a technical limitation and Selbst and Barocas (2018) warned that intuitive explanations that legitimise, but do not constrain, systems. Ananny and Crawford (2018) pressed home the point; transparency as seeing-without-knowing can provide merely the semblance of accountability.

Along with the critique, the field generated proofs. Datasheets for datasets (Gebru et al. 2021) and model cards (Mitchell et al. 2019) introduced the format of disclosure that should be considered during the process of releasing a new dataset. (Raji et al. 2020) developed integrated internal auditing procedures. (Bommasani et al. 2023) created a Transparency Index for Foundation Models. This index evaluates developers using hundreds of indicators.

For all their differences, these instruments share one property: they are synchronic. A model card reports what its authors believed on release day. An index scores a year. An audit certifies a period. None of them is built to represent change—which document superseded which, at what significance, and with what consequences for parties who relied on the earlier text. Yet the object that matters most in frontier AI governance is precisely the object that changes: the policy text that mediates between a deployed system and everyone affected by it. The transparency literature has, in short, a diachronic blind spot. This paper's entry point is the missing layer, which I call diachronic transparency—transparency of change itself. Versioning is its native instrument.

2.2 The governance artifacts

The documents at play are now of a subtype. Labs have published usage provisions, model cards (some consulted with techniques like Constitutional AI), safety and scaling guarantees, and system cards for larger versions (OpenAI 2023), ((OpenAI 2025a), (OpenAI 2025b) ; (Anthropic 2023a); for updates on the Responsible Scaling policy, see (Anthropic 2024), (Anthropic 2025). Public-minded entities offered their alternatives: the EU AI Act (Regulation (EU) 2024/1689), initially supported by the European Commission (for a detailed assessment thereof, see (Veale and Zuiderveen Borgesius 2023), the NIST AI RMF and the Generative AI Profile, along with NIST's wider Cybersecurity Framework(NIST 2023), Qualification criteria / principles from the UK AI Safety Institute (2024), and the report of the UN High-Level Advisory Body (United Nations 2024). International standards, such as ISO/IEC 42001 or OECD AI Principles, join in the governance efforts. The governance research disabuses principles, regulation proposals, and pacing opinions (Calo 2017;Dafoe 2018;Marchant et al. 2011;Anderljung et al. 2023); see also (Cihon 2019) and standards, and (Hadfield & Clark 2023) and regulatory dynamic pricing.

What the literature addressing this gap does not discuss is revision discipline as such. We have arguments on what rules should be spelled out. Rarely do we ask how the norms are versioned as they undergo modification—whether a reader will know that the authority is revoked, or whether the permission is restored, or whether the current text even acknowledges its predecessors. That is the gap this article works on.

2.3 The software precedent, without romance

Semantic versioning is a straightforward specification: a version number with the MAJOR.MINOR.PATCH structure, where the MAJOR digit increase means an incompatible update, the MINOR digit means a backward-compatible addition, and the PATCH means a backward-compatible bug fix (Preston-Werner 2023). From around this contract, there has grown an ecosystem of version control, changelogs, deprecation rules, as well as rollback conventions (Spinellis 2005). The template is, in a way, export-ready. Realistic idea is that the first question to be asked is just if it works inside home.

The empirical answer is: partially, and often badly. Studying the Maven Central repository, (Raemaekers et al. 2017) found breaking changes are common, that deprecation is rarely used as intended, and that version numbers frequently fail to signal the impact of upgrades. Across seven packaging ecosystems, (Decan et al. 2019) found semantic versioning adoption uneven and its constraints widely ignored in practice. Versioning vocabulary, in other words, is cheap; versioning discipline is expensive, and at home it is sustained only where tooling, registries, and downstream dependency pressure make misclassification costly.

This discovery becomes the cornerstone for everything else. If discipline in versioning is neglected at the home front, which lacks enforcement mechanisms, then importing related vocabulary into the AI governance lexicon without including those things that would enforce the law would result into mere decoration instead of discipline. In recognition of this limitation, the design of the framework in Section 3 has been tailored toward verification of commitments and visibility of mischaracterization as falsifiable, and misclassification as visible and auditable events.

3. The versioned governance framework

3.1 The four-level scheme

Governance documents for frontier AI systems should be versioned according to a four-level scheme, denoted vM.m.p.b. The levels are major, minor, patch, and build. Each carries a distinct meaning, a distinct changelog requirement, and a distinct relationship to rollback and audit.

Major (M): normative revision. A change is major when it alters red lines, capability boundaries, the allocation of responsibility, or the conditions under which deployment is permitted. Major revisions require a full changelog entry per clause, a stated rationale, a deprecation or transition window before enforcement where reliance interests exist, and explicit rollback analysis.

Minor (m): policy extension. A change is minor when it adds scenarios, exceptions, or review procedures without overturning superordinate principles. Minor revisions require clause-level changelog entries and rationale, but not transition windows by default.

Patch (p): clarification. A change is a patch when it rewords, disambiguates, corrects references, or harmonizes terminology without altering normative force. Patches require a changelog entry but may be batched.

Build (b): branch adaptation. A build is a derived edition of a document for a jurisdiction, sector, model family, or research environment, carrying an explicit pointer to the base version it branches from and a record of every divergence.

Three mechanical rules keep the scheme honest. Numbers are monotonic: a document never reuses an identifier. Lower levels reset on any higher-level increment. And every published version receives an immutable identifier—identifier, date, and a hash of the text—so that citation to a version is citation to a fixed object. What the scheme is not is a claim about model weights or code releases; it applies to governance texts, the layer where normative commitments live.

3.2 Classification rules and examples

Classification cannot rest on intuition, because intuition is exactly what silent drift exploits. I propose three tests, applied in order. Normative force: does the change alter what is required, permitted, or prohibited? Reliance: have identifiable parties organized conduct around the prior text such that the change undermines their position? Permission boundary: does the change move the conditions under which the system may be deployed at all? A yes to the third test makes the change major; a yes to the first or second makes it at least minor; a no to all three makes it a patch. Builds are classified by structure rather than content: any derived edition is a build.

Four examples fix the idea. Elevating political persuasion from a general behavioral restriction to a high-sensitivity domain with its own controls changes what is prohibited and who may attempt it: major.

Adding a human-review gate for a newly supported class of multimodal agent actions extends procedure without touching superordinate principles: minor. Harmonizing the definitions of "user" and "customer" across a policy suite without changing obligations: patch. Issuing a medical-research sandbox edition of a usage policy, with relaxed provisions for approved institutions and an explicit pointer to the base version: build.

The classification, plus a one-paragraph rationale, must be declared by the author at release. The declaration is a claim, not a proof—and that is the point. A public claim can be wrong in a way that an unclassified edit cannot. What happens when it is wrong is the subject of Section 3.4.

3.3 Changelogs and the governance toolchain

A version number without a changelog is a label without content. The framework therefore specifies a minimal toolchain, described here at the level of public interfaces rather than internal implementation.[2] Six components suffice.

A policy registry lists every governance document an organization maintains, with current version and status. A version ledger records every release of every document—identifier, date, classification, rationale, per-clause changelog, and hash—append-only and timestamped, so that history cannot be silently rewritten. A review queue holds proposed changes with their proposed classifications before release, visible at least to internal reviewers and external auditors, so that classification happens prospectively rather than as post-hoc rationalization. Rollback triggers are pre-registered conditions under which a deployment-relevant change is reverted or suspended, stated at the time of the change rather than improvised during an incident. An audit API exposes the registry and ledger in machine-readable form to accredited auditors. Finally, a cross-reference map links each policy version to the evaluations, red-team protocols, and deployment states that were governed by it, so that "under which rules was this system operating in March?" is a query, not an investigation.

None of this is exotic. Configuration management standards have specified analogous artifacts for decades (IEEE 2012); continuous-delivery practice treats reversible deployment and release ledgers as routine (Humble and Farley 2011); community conventions like Keep a Changelog already define the difference between a changelog and a marketing post (Lacan 2023). Internal audit frameworks for AI have asked for precisely these hooks (Raji et al. 2020); broader audit methodologies (Mökander et al. 2023; Schuett 2023) and critical perspectives on audit independence (Laux et al. 2021) reinforce the need for verifiable records. Regulatory initiatives such as Canada’s Directive on Automated Decision-Making similarly demand transparent decision trails. And the verifiable-claims literature has shown how such commitments can be made checkable rather than rhetorical (Brundage et al. 2020). The design principle is: interfaces open, implementations tiered, critical deployments controlled. A fuller engineering specification exists as a companion preprint (Xue, Z. 2026c); what matters here is that the toolchain is a repackaging of known practice, not an invention.

3.4 Rollback and misclassification as auditable events

Why versioning rather than archiving? Because an archive answers "what did the text say?" while governance needs answers to "what may change back, under what conditions, and who decided?" The reason is the Collingridge dilemma: early in a technology's life, change is easy but its consequences are poorly understood; later, the consequences are understood but change is expensive and contested (Collingridge 1980). Frontier AI lives inside that dilemma. The only way to keep early reversibility available late is to design for it in advance—which is what rollback eligibility means.

Two asymmetries govern rollback design. Safety-relevant tightenings should take effect immediately and require no transition window; loosenings should be subject to deprecation windows, because reliance interests form around permissions. And rollback of a loosening should restore prior restrictions only through review, never automatically, since downstream conduct may have organized around the permission. These rules are developed at length in a companion analysis (Xue, Z. 2026b); the essential point is that governance rollback is asymmetric in a way code rollback is not, and the framework encodes that asymmetry rather than ignoring it.

The enforcement heart of the framework is a single rule: misclassification is itself an auditable event. If a change labeled patch is later judged—by an auditor, a regulator, a court, or the laboratory's own review—to have been major, that judgment enters the ledger as an event with its own record. No sanction needs to attach at first. The record disciplines, because a pattern of misclassifications is visible in a way that a pattern of silent edits never is. This is the direct answer to the lesson of Section 2.3: versioning fails at home when classification is free, so the framework makes classification expensive—not procedurally, but reputationally and legally. In frontier AI, undocumented normative drift is itself a governance failure. The framework's job is to make that failure visible, dated, and attributable.

4. Empirical analysis: five instruments, 2019–2025

4.1 Method

I selected five governance instruments against four criteria: they are public; they govern or constrain frontier AI systems directly; they were revised at least once between 2019 and 2025; and their revision history is at least partially documented. The five are: the usage policies of OpenAI; the Model Spec of OpenAI; Anthropic's Responsible Scaling Policy; the EU Artificial Intelligence Act; and the NIST AI Risk Management Framework. For context, Section 4.2 first examines four adjacent instruments with established edition cultures.

Sources are official publication pages, dated announcements and changelog posts, legislative records, and archived captures from public web archives. The unit of analysis is the document revision; where clause-level diffs are available, the analysis descends to the clause. Each instrument was coded on six dimensions: distinct public versions; dated changelog; clause-level diff; declared change classification; rollback provision; and scheduled review. Table 1 summarizes the coding.

The analysis is limited to the public record. I have no access to internal repositories, and my coding is that of an external reader. I treat this limitation as partly constitutive rather than merely unfortunate: the external reader is exactly the position of the auditor, the journalist, the enterprise customer, and the court. If a versioning discipline cannot be perceived from outside, it does not exist for governance purposes.

4.2 Versioned precedents around AI

Edition culture is not foreign to technology governance; it surrounds frontier AI on all sides. The NIST Cybersecurity Framework moved from version 1.0 (2014) to 1.1 (2018) to 2.0 (2024), each transition labeled, consulted on, and treated by its community at the significance the numbering implied—the 2.0 transition, which broadened scope beyond critical infrastructure, consumed years of drafts precisely because everyone understood it to be major (NIST 2024b). The OECD AI Principles, adopted in 2019, were formally revised in May 2024 to address general-purpose and generative systems, with the revision process and its outputs published as such (OECD 2024). Canada's Directive on Automated Decision-Making has been amended repeatedly since 2019, each amendment published with an effective date and summary, and its companion Algorithmic Impact Assessment is itself a versioned instrument, so that obligations pin to dated editions of a tool (Treasury Board of Canada Secretariat 2019). ISO/IEC 42001:2023, the AI management-system standard, inherits the standards world's deepest convention: certification pins to a year-identified edition, and when editions change, transition windows are announced in advance—deprecation policy in governance form (ISO/IEC 2023; Cihon 2019).

The pattern matters for what follows. Statutes, standards, and directives version; frontier laboratory policies, which change faster than any of them, largely do not.

4.3 The five instruments

OpenAI usage policies. The usage policies were revised publicly several times between 2019 and 2025, including dated updates in March 2023 and January 2024, and a consolidation into a simplified universal structure in January 2025 (OpenAI 2025b). One episode illustrates the framework's diagnostic value. The January 2024 revision, among other edits, removed an explicit blanket prohibition on military use. The change was discovered not because it was announced at its significance but because journalists compared versions by hand; the laboratory subsequently characterized the edit as a clarification. Whether one accepts that characterization is beside the point here. The episode shows a change whose classification was contested, in an environment where the author was never required to declare a classification at all. Under the framework, the declaration "patch" would have been a falsifiable public claim; the ensuing dispute would have been an auditable misclassification event rather than a news cycle. Since 2025, a dated changelog page exists—vocabulary arriving, as predicted, at the document level first.

OpenAI Model Spec. First published as a draft in May 2024 and substantially revised in February 2025, the Model Spec is the closest thing to versioned governance among laboratory documents (OpenAI 2025a). The February 2025 revision shipped with a summary of changes and a CC0 dedication inviting

external reuse. Yet its version identifiers are dates, not semantics: a reader cannot tell from "2025-02" whether the revision is major, and the changelog summarizes rather than classifies. The raw material of versioning discipline is present; the grammar is not.

Anthropic Responsible Scaling Policy. The RSP was published as version 1.0 in September 2023, revised to version 2.0 in October 2024, and updated again in May 2025 (Anthropic 2023a), (Anthropic 2024),(Anthropic 2025). It is the only laboratory instrument in the set with explicit semantic-style numbering, and the 2024 revision was accompanied by a public account of what changed and why. Its capability-threshold architecture—AI Safety Levels that trigger defined mitigations—gives the policy a quasi-executable core, the nearest analogue to a compiler that normative text has produced; this reflects the layered governance instinct visible across Anthropic's public architecture (Anthropic 2023b; Bai et al. 2022). Still absent: clause-level diffs, and any rollback provision. The thresholds say when mitigations must engage; nothing says what happens if a policy revision itself must be reverted (Xue, Z. 2026b).

EU Artificial Intelligence Act. The Act traveled from the Commission proposal of April 2021 (European Commission 2021) through the Council's general approach of December 2022 and the Parliament's position of June 2023 to Regulation (EU) 2024/1689, with staged application running to 2027 (Veale and Zuiderveen Borgesius 2021; Regulation (EU) 2024/1689). Two features stand out. First, its most consequential mid-process change—the insertion of obligations for general-purpose AI models into the Parliament's 2023 position—was a direct response to an exogenous capability shock, the public arrival of ChatGPT in late 2022. Second, the legislative process carried that change with full article-level traceability: trilogue records make the evolution of every provision reconstructable. The Act also contains the closest thing to rollback in the set—periodic evaluation and review obligations (Article 112) and staged entry into force—though these are forward-looking review, not reversal. The statute versions better than the laboratories, at the price of being slowest where the laboratories are fastest.

NIST AI Risk Management Framework. The RMF was released explicitly as version 1.0 in January 2023, with a public roadmap and a commitment to periodic review (NIST 2023). When generative AI demanded treatment, NIST issued a Generative AI Profile as a separately numbered companion rather than a silent expansion of the base text (NIST 2024a). The standards-institution habits hold: numbering signals significance, scope changes arrive as numbered artifacts. What the RMF lacks is the clause-level machinery—unsurprising, since it is a framework rather than an operative policy.

Table 1

Versioning features of five frontier-AI governance instruments, 2019–2025

Instrument	Distinct public versions	Dated changelog	Clause-level diff	Declared classification	Rollback / scheduled review
OpenAI Usage Policies	Multiple; dated updates 2023-03, 2024-01, 2025-01	Yes (from 2025)	No	No	None
OpenAI Model Spec	2024-05 draft; 2025-02; later 2025 revisions	Yes	Partial	No	None
Anthropic RSP	v1.0 (2023-09); v2.0 (2024-10); update (2025-05)	Yes	Partial	Partial (threshold logic)	None
EU AI Act	Proposal 2021; Council 2022; Parliament 2023; Regulation 2024/1689	Full legislative record	Yes (article-level)	Formal procedural categories	Staged application; Art. 112 review; no rollback
NIST AI RMF	1.0 (2023-01); GAI Profile (2024-07)	Numbered editions	Yes	Conveyed by numbering	Periodic review commitment; no rollback

4.4 Findings

Five findings emerge.

F1. Versioning vocabulary is spreading at the document level, not the clause level. Numbered versions, dated changelogs, and edition identifiers are now common. Per-clause classification of what changed is absent everywhere outside the legislative record.

F2. Breaking changes are driven by exogenous capability shocks. The largest revisions in the set—general-purpose model obligations in the AI Act, the RSP's 2024 restructuring, the 2024–2025 laboratory policy revisions—track capability events, not review schedules. Governance change is event-driven; the documentation of change is not.

F3. Changelogs summarize rather than classify. Where changelogs exist, they narrate. They do not declare significance levels, and the burden of classification falls on the reader—precisely the burden silent drift exploits, as the January 2024 usage-policy episode shows.

F4. Rollback is absent everywhere. No instrument in the set specifies conditions under which a governance revision can be reverted. Forward review exists (Article 112, NIST's review commitment); reversal does not.

F5. Discipline correlates with institutional form, not technical capacity. The statute and the standards version best; the laboratories, which employ the world's deepest pool of software-engineering talent, version least. The constraint is incentive, not capability—which is why the framework targets the cost of misclassification rather than the availability of tools.

4.5 An operational scenario

Suppose a frontier model gains agentic coding and browser-use capabilities in a single release. Four governance documents must move: the usage policy (new permission boundaries for autonomous action—major); the evaluation standard (a new review gate for agentic trajectories—minor); the red-team protocol (harmonized terminology with the evaluation standard—patch); and the medical-research sandbox edition (a build diverging from the new base). Under current practice these changes would arrive as silent edits across four pages, at different times, discoverable only by hand. Under the framework they arrive as one versioned event: declared classifications, per-clause changelogs, a cross-reference map linking the policy versions to the deployment they govern, and pre-registered rollback triggers for the agentic permissions. The difference is not cosmetic. It is the difference between a governance system that can be audited and one that can only be observed.

5. Discussion

5.1 Diachronic transparency

The empirical record reframes the transparency debate of Section 2.1. Snapshot instruments—model cards, indices, audits—measure states; the failures documented in Section 4 are failures of transitions. A transparency index can tell you that a laboratory published a changelog in 2025; it cannot tell you whether the changelog classified anything (Bommasani et al. 2023). Ananny and Crawford (2018) warned that transparency fails when it produces seeing without knowing; a changelog without classification is nearly the pure case—everything is visible and nothing is legible. Diachronic transparency does not replace the synchronic kind. It supplies the temporal skeleton that makes snapshots comparable, and it gives auditors, regulators, and publics a shared object to argue about: the declared classification of a dated change.

5.2 Six objections, taken seriously

The template fails at home. Semantic versioning is honored as much in the breach as in the observance (Raemaekers et al. 2017; Decan et al. 2019). This is the strongest objection, and the framework is built around it rather than around it notwithstanding. The home-domain evidence shows that discipline follows enforcement infrastructure, not vocabulary—registries, tooling, and downstream cost. The framework therefore does not ask laboratories to promise to version well; it makes each classification a falsifiable public claim and misclassification an auditable event. Where the analogue to dependency pressure exists—enterprise contracts, regulatory recognition, procurement—the incentive to classify honestly follows.

Normative text has no compiler. Code can be tested for breakage; policy cannot. True, and irrelevant to the design choice that matters. The ASL architecture shows that bright-line capability thresholds can give policy a quasi-executable core; declared behavioral tests and deprecation windows supply the rest. Classification is not a proof of significance. It is a claim about significance, published so that it can be challenged. An unclassified edit cannot even be wrong.

Version theater. Laboratories may adopt the vocabulary as decoration. They may; every transparency instrument is vulnerable to theater, and the critique of transparency-as-performance applies here in full (Ananny and Crawford 2018; Laux et al. 2021). The countermeasure is structural: theater requires that performance be costless, and the ledger makes a pattern of empty classifications a permanent, queryable record. Audit frameworks can read the record (Raji et al. 2020; Mökander et al. 2023). Theater that leaves evidence is a failing strategy.

Classification manipulation. The worse case is not decoration but gaming—systematically downgrading majors to patches. Downgrade incentives are real, which is why the framework distributes classification review: internal review queues prospectively, external auditors through the audit API, regulators where version classes are recognized in law, and the public record as backstop. Manipulation must then survive four audiences with different interests.

Rollback is normatively dangerous. Reverting a safety policy is not like reverting a library release; people organize conduct around permissions, and yanking permissions back can itself harm. The framework concedes the asymmetry and encodes it: tightenings take effect immediately, loosenings face deprecation windows, and rollback of a loosening proceeds only through review. Long-horizon normative commitments—what a governance system owes future reliance interests—are analyzed separately at greater length (Xue, Z. 2026a).

Audience asymmetry. A clause-level diff is legible to engineers and opaque to most publics. The answer is layered disclosure: a plain-language summary for every release, a clause diff for professionals, a machine-readable ledger for machines. Release notes and commit logs have coexisted in software for decades; governance can sustain the same division of labor.

5.3 Limitations and future work

The framework governs organizational documentation, not statutes; extending version grammar to inter-institutional and national instruments is future work. Classification still requires an interpretive community, and builds carry fragmentation risk that coordination mechanisms must answer. Empirically, the analysis rests on the public record; internal revision histories, if ever opened, would test F1–F5 harder. The natural integrations are with audit practice (Mökander et al. 2023), three-lines-of-defense architectures (Schuett 2023), regulatory-market proposals (Hadfield and Clark 2023), and frontier regulation more broadly (Anderljung et al. 2023)—each of which needs, and currently lacks, a grammar of change to anchor to.

6. Conclusion

Frontier AI laboratories publish governance documents that regulators, courts, customers, and competitors rely on, and they revise those documents under a discipline that would embarrass a mid-sized open-source project. The result is silent drift: continuous change without a legible record, concentrated exactly where change is fastest and stakes are highest. This paper has proposed a versioned governance framework—four levels of change, classified changelogs, rollback-aware design, misclassification treated as an auditable event—and tested it against the documented revision histories of five real instruments. The record shows pieces of the discipline already in place: numbered safety policies, dated changelogs, edition-pinned standards, statutory review. It also shows the gaps: no clause-level classification, no rollback anywhere, and an inverse relationship between speed of change and discipline of change. The software analogy, examined honestly, survives its own critique: versioning works when classifications are public, falsifiable, and costly to fake, and fails when it is vocabulary without enforcement. The framework imports the enforcement conditions, not just the vocabulary. None of it requires new law, new institutions, or agreement about what the rules should say. A laboratory could adopt it unilaterally tomorrow; an auditor could require it next quarter; a regulator could recognize it in the next guidance cycle. Frontier AI will be governed in versions whether institutions admit it or not. The real question is whether those versions remain opaque, ad hoc, and weakly accountable, or whether they become explicit, reviewable, and reversible.

Declarations

Competing interests. The author declares no competing interests.

Funding. The author received no funding for this research.

Ethics approval. Not applicable; the study analyzes publicly available documents only.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. All instruments analyzed are public. A versioned source list with archival captures is available from the author on request.

Code availability. Not applicable; the article does not produce or rely on new software code.

Preprints. Pre-publication versions of the author's book chapters cited in this article are openly available on the non-commercial academic repository Zenodo (Xue, Z. 2026a), (Xue, Z. 2026b), ((Xue, Z. 2026c), as indicated by the DOIs in the reference list. The present manuscript has not been published elsewhere and is not under consideration by any other journal.

Author contributions. The author is the sole author and is responsible for all aspects of the work.

Use of AI tools. The author used AI-based tools for language editing and proofreading only. After using these tools, the author reviewed, revised, and approved the content as needed and takes full responsibility for the final manuscript.

Footnotes

[1] This paper is one of a series of standalone papers on versioned governance for advanced AI systems. Each paper in the series is independently argued and independently readable.

[2] The companion preprints cited below (Xue, Z. 2026a),(Xue, Z. 2026b),(Xue, Z. 2026c) are chapters of a larger book project in preparation. They are cited here as preprints for readers who want the extended treatment; no claim in this paper depends on them as core evidence.

References

1. AI Safety Institute (2024) Advanced AI evaluations at AISI: May update. UK Government, London. https://www.aisi.gov.uk/blog/advanced-ai-evaluations-may-update. Accessed 1 September 2026
2. Ananny M, Crawford K (2018) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc* 20(3):973-989. https://doi.org/10.1177/1461444816676645
3. Anderljung M, Barnhart J, Korinek A, Leung J, O'Keefe C, Whittlestone J, et al (2023) Frontier AI regulation: managing emerging risks to public safety. arXiv preprint. https://arxiv.org/abs/2307.03718
4. Anthropic (2023a) Anthropic's Responsible Scaling Policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy. Accessed 1 September 2026
5. Anthropic (2023b) Core views on AI safety: when, why, what, and how. https://www.anthropic.com/news/core-views-on-ai-safety. Accessed 1 September 2026
6. Anthropic (2024) Updating our Responsible Scaling Policy. https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-policy. Accessed 1

September 2026

7. Anthropic (2025) Responsible Scaling Policy, version 2.2 (effective May 14, 2025). https://www.anthropic.com/rsp-updates. Accessed 1 September 2026
8. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al (2022) Constitutional AI: harmlessness from AI feedback. arXiv preprint. https://arxiv.org/abs/2212.08073
9. Bommasani R, Klyman K, Longpre S, Kapoor S, Maslej N, Xiong B, Zhang D, Liang P (2023) The Foundation Model Transparency Index. arXiv preprint. https://arxiv.org/abs/2310.12941
10. Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, et al (2020) Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint. https://arxiv.org/abs/2004.07213
11. Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc* 3(1):2053951715622512. https://doi.org/10.1177/2053951715622512
12. Calo R (2017) Artificial intelligence policy: a primer and roadmap. *UC Davis Law Rev* 51(2):399-435
13. Cihon P (2019) Standards for AI governance: international standards to enable global coordination in AI research & development. Future of Humanity Institute, University of Oxford, Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf. Accessed 1 September 2026
14. Collingridge D (1980) *The social control of technology*. St. Martin's Press, New York
15. Dafoe A (2018) AI governance: a research agenda. Future of Humanity Institute, University of Oxford, Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/AI-Governance_-A-Research-Agenda.pdf. Accessed 1 September 2026
16. Decan A, Mens T, Grosjean P (2019) An empirical comparison of dependency network evolution in seven software packaging ecosystems. *Empir Softw Eng* 24(1):381-416. https://doi.org/10.1007/s10664-017-9589-y
17. European Commission (2021) Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final, European Commission, Brussels. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. Accessed 1 September 2026
18. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, Crawford K (2021) Datasheets for datasets. *Commun ACM* 64(12):86-92. https://doi.org/10.1145/3458723
19. Hadfield GK, Clark J (2023) Regulatory markets: the future of AI governance. arXiv preprint. https://arxiv.org/abs/2304.04914
20. Humble J, Farley D (2011) *Continuous delivery: reliable software releases through build, test, and deployment automation*. Addison-Wesley, Upper Saddle River
21. IEEE (2012) IEEE Standard for configuration management in systems and software engineering (IEEE Std 828-2012). Institute of Electrical and Electronics Engineers, New York. https://doi.org/10.1109/IEEESTD.2012.6170935

22. ISO/IEC (2023) Information technology — Artificial intelligence — Management system (ISO/IEC 42001:2023). International Organization for Standardization, Geneva. https://www.iso.org/standard/81230.html
23. Kroll JA (2018) The fallacy of inscrutability. *Philos Trans R Soc A Math Phys Eng Sci* 376(2133):20180084. https://doi.org/10.1098/rsta.2018.0084
24. Lacan O (2023) Keep a Changelog (v1.1.0). https://keepachangelog.com. Accessed 1 September 2026
25. Laux J, Wachter S, Mittelstadt B (2021) Taming the few: platform regulation, independent audits, and the risks of capture created by the DMA and DSA. *Comput Law Secur Rev* 43:105613. https://doi.org/10.1016/j.clsr.2021.105613
26. Marchant GE, Allenby BR, Herkert JR (eds) (2011) *The growing gap between emerging technologies and legal-ethical oversight: the pacing problem*. Springer, Dordrecht
27. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, pp 220-229. https://doi.org/10.1145/3287560.3287596
28. Mökander J, Schuett J, Kirk HR, Floridi L (2024) Auditing large language models: a three-layered approach. *AI Ethics* 4(4):1085-1115. https://doi.org/10.1007/s43681-023-00289-2
29. NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology, Gaithersburg. https://doi.org/10.6028/NIST.AI.100-1. Accessed 1 September 2026
30. NIST (2024a) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology, Gaithersburg. https://doi.org/10.6028/NIST.AI.600-1. Accessed 1 September 2026
31. NIST (2024b) The NIST Cybersecurity Framework (CSF) 2.0 (NIST CSWP 29). National Institute of Standards and Technology, Gaithersburg. https://doi.org/10.6028/NIST.CSWP.29. Accessed 1 September 2026
32. OECD (2024) Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449, as amended 2024). Organisation for Economic Co-operation and Development, Paris. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Accessed 1 September 2026
33. OpenAI (2023) GPT-4 system card. https://cdn.openai.com/papers/gpt-4-system-card.pdf. Accessed 1 September 2026
34. OpenAI (2025a) Model Spec (May 2024 draft; February 2025 revision and later updates). https://model-spec.openai.com. Accessed 1 September 2026
35. OpenAI (2025b) Usage policies (version 2025-10-29). https://openai.com/policies/usage-policies/. Accessed 1 September 2026
36. Pasquale F (2015) *The black box society: the secret algorithms that control money and information*. Harvard University Press, Cambridge, MA

37. Preston-Werner T (2013) Semantic Versioning 2.0.0. https://semver.org/spec/v2.0.0.html. Accessed 1 September 2026
38. Raemaekers S, van Deursen A, Visser J (2017) Semantic versioning and impact of breaking changes in the Maven repository. *J Syst Softw* 129:140-158. https://doi.org/10.1016/j.jss.2016.04.008
39. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. ACM, New York, pp 33-44. https://doi.org/10.1145/3351095.3372873
40. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union* L, 12 July 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Accessed 1 September 2026
41. Schuett J (2025) Three lines of defense against risks from AI. *AI Soc* 40(2):493-507. https://doi.org/10.1007/s00146-023-01811-0
42. Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. *Fordham Law Rev* 87(3):1085-1139
43. Spinellis D (2005) Version control systems. *IEEE Softw* 22(5):108-109. https://doi.org/10.1109/MS.2005.140
44. Treasury Board of Canada Secretariat (2019) Directive on automated decision-making (as amended). Government of Canada, Ottawa. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592. Accessed 1 September 2026
45. Turilli M, Floridi L (2009) The ethics of information transparency. *Ethics Inf Technol* 11(2):105-112. https://doi.org/10.1007/s10676-009-9187-9
46. United Nations (2024) Governing AI for humanity: final report of the High-level Advisory Body on Artificial Intelligence. United Nations, New York. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf. Accessed 1 September 2026
47. Veale M, Zuiderveen Borgesius F (2021) Demystifying the draft EU Artificial Intelligence Act: analysing the good, the bad, and the unclear elements of the proposed approach. *Comput Law Rev Int* 22(4):97-112. https://doi.org/10.9785/cri-2021-220402
48. Xue, Z. (2026a) The axiomatic formulation of cosmic responsibility ethics. In: *The normative reconstruction of interstellar civilization: a new tripartite outlook and governance under deep uncertainty*, 2nd edn, vol 1, chap 10. Zenodo. https://doi.org/10.5281/zenodo.22044834
49. Xue, Z. (2026b) Version-based governance and rollback mechanisms. In: *The normative reconstruction of interstellar civilization: a new tripartite outlook and governance under deep uncertainty*, 2nd edn, vol 3, chap 19. Zenodo. https://doi.org/10.5281/zenodo.22046229
50. Xue, Z. (2026c) Versioned governance toolchain. In: *The normative reconstruction of interstellar civilization: a new tripartite outlook and governance under deep uncertainty*, 2nd edn, vol 4, chap 9. Zenodo. https://doi.org/10.5281/zenodo.22046536