The AI Accountability Standard

How Mathematics Solves the Compliance Problem for Artificial Intelligence

A Cross-Sector Policy White Paper

April 2026

Prepared for regulators, enterprise AI leaders, policy researchers, and standards bodies working on the accountability infrastructure of deployed artificial intelligence systems

Executive Summary

Artificial intelligence has moved from research into consequential deployment across nearly every regulated sector of the economy. Credit decisions, medical determinations, employment screening, insurance underwriting, child welfare assessment, content moderation, criminal justice risk scoring, and government benefit eligibility are now routinely shaped by algorithmic systems whose behavior neither the deployers, the subjects, nor the regulators can independently verify.

The governance response to this situation, across jurisdictions and sectors, has converged on a common set of mechanisms: model cards, ethics committees, principles statements, third-party audits, impact assessments, and attestation frameworks. These mechanisms are reasonable. They are also collectively insufficient, and increasingly visibly so. The reason is structural rather than operational: all of these mechanisms rely on the deploying organization to produce evidence about its own behavior, in forms that external parties must accept without independent verification. The evidence is as credible as the organization is.

This paper argues that the AI compliance problem has been misdiagnosed in most policy discussions. The problem is not that organizations lack principles, processes, or good intentions. Most have all three. The problem is that governance built on organizational attestation cannot scale to algorithmic decisioning at population volume, and cannot produce the form of evidence that regulators, courts, and subjects increasingly demand. What is missing is not better principles. What is missing is verifiable evidence.

Cryptographic research has, over the past decade, produced techniques that allow the party performing a computation to generate a small mathematical certificate attesting to specific properties of that computation — model identity, structural independence from prohibited inputs, record integrity, credentialed human review — such that any third party can verify the certificate in milliseconds without access to the underlying model, data, or systems. These techniques are no longer research artifacts. They are practical at the latency and volume that consequential AI decisioning operates at, and they address exactly the gap that current governance cannot close.

This paper proposes a cross-sector AI Accountability Standard built on four verifiable properties that any consequential algorithmic decisioning system can be held to: Model Identity, Structural Independence from Declared Prohibited Inputs, Chained Record Integrity, and Credentialed Human Review. The Standard is sector-agnostic, technology-neutral, and designed to be adoptable as a regulatory requirement, a procurement specification, or a voluntary industry commitment. It does not resolve every question in AI governance — several categories of problems are beyond the reach of mathematical verification and require human judgment — but it closes the specific evidentiary gap that attestation-based governance cannot close, and it does so in a form that scales to the volume at which AI is now being deployed.

The paper is written for a cross-sector audience. Its central claim is that the compliance problem for artificial intelligence is, at its core, a question about what counts as evidence — and that the mathematical techniques now available can provide a form of evidence that organizational attestation structurally cannot. The remainder of the paper develops this claim in detail, names the Standard's specific commitments, catalogs the problems the Standard does and does not address, and walks through sectoral applications across healthcare, financial services, employment, insurance, content moderation, and government services.

1. The Compliance Problem Is Not What Most People Think It Is

1.1 The standard diagnosis and why it is incomplete

Most public and regulatory discussion of AI compliance treats the central problem as one of bias, fairness, or alignment. Algorithmic systems have been shown to encode and amplify racial, gender, and socioeconomic disparities in high-stakes domains; they have been shown to reflect the values of their developers in ways that affect consequential decisions; they have been shown to behave unpredictably in deployment environments that differ from their training data. These are real problems, and substantial research, regulatory, and policy effort is appropriately directed at them.

But bias, fairness, and alignment are substantive problems — questions about whether the system's behavior is good. Compliance, in the ordinary regulatory sense, is a different kind of problem. It is an evidentiary problem: how does any party outside the deploying organization verify what the system actually did on a specific decision, at a specific moment, against a specific standard? If the answer is "trust the deploying organization's account," then compliance has not been established in any rigorous sense. It has been asserted.

The distinction matters because the substantive problems and the evidentiary problem require different solutions. Substantive fairness requires good training data, careful model design, thoughtful deployment, and ongoing monitoring. Evidentiary verification requires something different: a mechanism by which claims about the system's behavior can be confirmed by parties other than the organization making the claims. Current AI governance has developed substantial sophistication on the first question and nearly none on the second. This is the asymmetry the paper addresses.

1.2 Why the evidentiary problem has been invisible

The evidentiary problem has been relatively underdiscussed in AI policy, for several reasons worth naming.

First, in many regulated domains, algorithmic decisioning is a recent overlay on human-decisioning workflows that had their own evidentiary infrastructure. Medical records, financial audit trails, credit decision documentation, employment records — all of these existed before AI was widely deployed, and

they are still maintained. From a superficial view, the evidence appears to be there. What is missed in that view is that the records document what the organization says about the decision, not what the algorithm actually did. The gap between those two things is the evidentiary problem, and it widens as algorithmic sophistication increases.

Second, the AI research community has focused extensively on interpretability and explainability — techniques for rendering model behavior more legible to humans. These are valuable tools, but they do not close the evidentiary gap. An interpretability method run by the organization on its own model, producing an explanation that the organization chooses to share, is not different in evidentiary character from any other organizational attestation. The external party has no way to verify that the explanation corresponds to what the model actually did on the decision in question.

Third, regulatory frameworks have, for understandable reasons, leaned on principles-based approaches that specify what organizations should do rather than what they must prove. Principles-based regulation is flexible and avoids micromanaging rapidly evolving technology, but it presumes that compliance can be assessed through organizational review. When the review relies on the organization's own account of its behavior, the principles-based framework becomes self-referential.

Fourth, the enforcement community — regulators, plaintiff's counsel, oversight agencies, investigative journalists — has historically lacked the technical sophistication to press on the evidentiary gap. This is beginning to change, but the evidentiary sophistication of enforcement is still meaningfully behind the evidentiary sophistication of deployment, and the gap itself has shielded the problem from scrutiny.

1.3 The environment that is making the problem visible

Several forces are now pressing on the evidentiary gap simultaneously, which is why AI accountability has emerged as a policy issue rather than remaining a research concern. Each force operates differently, but they point in the same direction.

Litigation. Courts are increasingly being asked to adjudicate AI-driven decisions, and the existing evidentiary record is proving difficult to work with. Plaintiffs attempt to reconstruct, through years of discovery, what a model actually did on a specific date; defendants produce millions of pages of policy documents and internal communications; neither side can definitively answer the question at issue. The legal system is adapting, but the adaptation is uneven and expensive, and it generates demand for better evidence.

Regulation. Multiple jurisdictions have enacted or proposed AI-specific regulatory frameworks — the EU AI Act, Colorado SB 205, the New York City employment AI rule, California's various AI transparency statutes, state insurance-department AI guidance, federal agency AI enforcement actions under existing authorities. Each framework imposes requirements whose compliance is, under current infrastructure,

attested rather than verified. The frameworks are forcing organizations to make compliance claims that they do not currently have the technical infrastructure to back with evidence.

Procurement. Enterprise buyers of AI — particularly in regulated industries — have begun including AI-governance review in procurement processes. The review, conducted by compliance staff who are themselves evaluating AI vendors' attestations, has become a significant bottleneck in enterprise AI sales cycles. This is a commercial force, not a regulatory one, but it operates on the same underlying issue: attestation-based evidence is being asked to do more work than it can support.

Political salience. Individual AI-decisioning harms — denied coverage, denied credit, improper surveillance, content moderation errors, biased employment screening — have become politically visible in ways that raise the reputational and electoral cost of inadequate accountability. Political salience is not the same as policy effectiveness, but it drives policy attention, and it increases the pressure on governance systems that rely on the public's trust in organizational self-reporting.

Collectively, these forces are producing an environment in which the evidentiary gap is being pressed on in ways it was not pressed on five years ago. Organizations that continue to rely on attestation-based governance will find themselves increasingly unable to meet the evidentiary demands of their regulatory, legal, commercial, and political environments. This is the problem the Standard proposed in this paper is designed to solve.

2. The Four Failure Modes of Current AI Governance

Current AI governance infrastructure is not monolithic. It consists of several distinct mechanisms that have developed somewhat independently, each addressing a legitimate concern. Each has strengths worth preserving. Each also has a specific failure mode that, considered together, produces the evidentiary gap described above.

2.1 Principles and policy documents

Most organizations deploying consequential AI have published AI principles — statements of commitment to fairness, transparency, accountability, and human oversight. Most have developed internal AI policies that translate principles into operational guidelines. These artifacts have value: they create organizational alignment, establish external expectations, and provide a framework for internal decision-making.

The failure mode of principles and policies is that they describe intent rather than behavior. A principles statement cannot be violated by the principles statement itself; it can only be violated by behavior that contradicts it. Verifying whether behavior contradicts a principles statement requires observing the behavior — which returns the governance problem to the evidentiary gap the principles were supposed to help resolve. Principles are necessary but not self-enforcing, and in the absence of a verification

mechanism, they can drift arbitrarily far from operational reality without external parties being able to detect the drift.

2.2 Third-party audits and attestations

More sophisticated organizations engage third-party auditors to review their AI systems against specified standards. SOC 2 Type II reports, ISO certifications, NIST AI Risk Management Framework assessments, and specialized AI audit services all provide forms of third-party opinion about organizational practices. The opinions are more credible than self-attestation because they involve a party with some professional independence.

The failure mode of third-party audits is structural: the auditor is retained by the organization, operates under the organization's disclosure decisions, and produces conclusions constrained by what the organization shows them. The auditor's opinion is honest, but it is an opinion about the organization's processes at a point in time, based on evidence the organization provides. The evidence itself is not independently verifiable by parties outside the audit relationship. This limitation is familiar from financial auditing, where it is well-understood and mitigated by regulatory enforcement backstops and strict liability frameworks. AI auditing has not yet developed equivalent mitigations, and the underlying structural limitation is more acute because AI systems can be reconfigured without leaving physical traces.

2.3 Impact assessments and algorithmic audits

Several jurisdictions now require organizations deploying high-risk AI systems to conduct impact assessments — structured analyses of the system's expected effects on affected populations. Algorithmic audits, a related mechanism, involve independent parties testing the system's behavior across synthetic or sampled inputs to identify disparate impact.

The failure mode of impact assessments and algorithmic audits is that they operate on system behavior in test conditions rather than on specific decisions in deployment. An assessment that a credit model exhibited acceptable disparate impact in audit conditions does not establish that a specific denial the following month was consistent with the audited behavior. Models drift. Thresholds shift. Feature pipelines change. The audit becomes stale, and its findings become unreliable as evidence for any specific decision. This is not a flaw in audit methodology — it is a structural limit on what population-level testing can establish about individual cases.

2.4 Interpretability and explanation

Interpretability research has produced a range of techniques for making model behavior more legible — saliency maps, attention visualizations, feature importance analyses, counterfactual explanations, natural-language rationales. These techniques are valuable for model development, internal auditing, and communication with affected parties.

The failure mode of interpretability-as-governance is that interpretability tools run by the organization, on the organization's models, producing explanations the organization chooses to share, do not provide independent evidence of anything. An explanation the organization generates is an organizational attestation about the model, not a verification of the model. External parties have no mechanism to confirm that the explanation corresponds to the computation the model actually performed. The tools can be run honestly, or they can be run in ways that produce explanations convenient to the organization; the external observer has no way to distinguish.

Each of the four mechanisms above addresses something real and valuable. The failure modes are not arguments for abandoning any of them. They are arguments for recognizing that none of them — alone or together — closes the specific evidentiary gap that consequential AI deployment now faces. Closing that gap requires a fundamentally different kind of mechanism: one in which the evidence is generated by the computation itself, in a form that any third party can verify without relying on the deploying organization's cooperation.

3. What Mathematical Verification Actually Does

3.1 The basic idea, without the mathematics

Cryptographic research has, since the late 2000s, developed a class of techniques known as succinct proof systems — methods by which a party performing a computation can generate a small mathematical certificate attesting that the computation has certain properties, such that any third party can verify the certificate rapidly without access to the underlying computation, data, or infrastructure. The techniques are referred to in the technical literature by acronyms including SNARK, STARK, zk-SNARK, and others, but the acronyms obscure the underlying commercial and evidentiary point.

The point, stated in ordinary language, is this: a mathematical mechanism now exists by which the party running an algorithm can produce a small certificate that proves, to any third party, specific facts about what the algorithm did — and the third party can verify the proof in milliseconds, without trusting the party who produced it. The verification is not a matter of reading the organization's account of its behavior and deciding whether to believe it. The verification is a mathematical check that either succeeds or fails, with no discretion for the verifier and no cooperation required from the party who produced the proof.

What makes these techniques useful for AI accountability is that the certificates can be generated automatically, at decision time, without significant performance cost. For a consequential AI decision — a credit decision, a medical determination, an employment screening outcome, a content moderation action — a certificate attesting to the decision's structural properties can be generated in under one second and verified by any third party in under fifty milliseconds. The certificate itself is small, typically one to two kilobytes, negligible compared to the decision record it accompanies.

3.2 What can be proved

Not every property of an AI decision can be proved by mathematical techniques. The techniques work best on structural properties — facts about what computation occurred, which inputs were used, which model was run — and less well on substantive properties — whether the outcome was fair, whether the training data was representative, whether the decision was wise. The distinction matters and will be developed in Section 5. But the range of structural properties that can be proved is broader than most policy readers realize, and includes several that address the central concerns of current AI governance.

Model identity. A certificate can prove that a specific decision was produced by a specific, cryptographically committed version of a specific model, identified by a hash of the model's weights. Any claim that a different model — an earlier version, a later version, an experimental variant, a manually-adjusted fork — produced the decision is mathematically refuted by the certificate.

Input feature composition. A certificate can prove that a decision was produced using features drawn from a specific, committed feature pipeline, and can prove that the feature vector presented to the model included or excluded specific variables. This addresses the "what did the model actually see" question that is central to disparate-impact analysis.

Structural independence from declared inputs. A certificate can prove that a decision was structurally independent of a declared set of prohibited inputs — that is, the decision would not have changed if those inputs had been zeroed, set to a reference value, or replaced with a redacted version. This is the cryptographic expression of commitments like "race was not considered" or "cost features did not influence the determination."

Record integrity. A certificate can be chained to prior certificates for related decisions, such that no certificate in the chain can be altered or re-ordered without invalidating every subsequent certificate. This addresses allegations of post-hoc record modification that are common in litigation discovery.

Credential binding. For decisions that require human review or approval, a certificate can prove that the reviewing human's credential — professional license, role authorization, organizational seniority — was valid at the time of decision. This addresses claims that decisions attributed to one credential level were produced under a different credential level.

These are the categories of property that mathematical verification addresses well. Each corresponds to a compliance question that current AI governance infrastructure answers through attestation rather than evidence. Each is verifiable in milliseconds by any party holding the certificate and the corresponding public verification key, without the cooperation or consent of the deploying organization.

3.3 The practical performance envelope

The numbers matter for assessing whether the techniques are practical at the volumes AI deployment operates at. Representative performance for a consequential decisioning application with a model of moderate complexity and a feature vector of roughly 100-200 inputs:

Certificate generation: 200 milliseconds to 1 second per decision, depending on the proof composition and the hardware. Specialized hardware (FPGA-based provers) brings the upper end down substantially. At the volumes consequential AI decisioning operates at — millions to hundreds of millions of decisions annually per deploying organization — the total compute cost is significant but not prohibitive, and is a small fraction of the compute cost of the model inference itself.

Certificate size: 1 to 2 kilobytes per decision. At organizational scale, this is an entirely trivial storage cost relative to the decision records the certificates accompany.

Verification time: 10 to 50 milliseconds per certificate on commodity hardware. Verification does not require specialized infrastructure on the verifier's side, which is critical because the verifier is typically not the deploying organization — it is a regulator, auditor, litigant, or subject.

These numbers reflect currently-achievable performance, not projections. They are the reason the techniques have moved from research contexts to deployment contexts over the past three to five years. The mathematical foundations date to the 1980s and 1990s; the engineering that made them practical at production volumes is recent. The transition from research curiosity to deployable infrastructure has occurred, and it is this transition — not any new theoretical breakthrough — that makes cross-sector AI accountability a realistic standard to propose now.

4. The Standard: Four Verifiable Properties

This section defines the AI Accountability Standard proposed by this paper. The Standard consists of four verifiable properties that any consequential algorithmic decisioning system can be designed to attest to. The Standard is sector-agnostic; it can be adopted as a regulatory requirement, a procurement specification, an industry commitment, or an internal governance mandate. Its practical meaning depends on how deploying organizations instantiate it, but the four properties themselves are general.

4.1 Property One: Model Identity

For every consequential decision produced by an algorithmic system, the deploying organization produces a certificate binding the decision to a cryptographically committed version of the specific model that produced it. The commit identifies the model's weights, architecture, and — to the degree organizationally appropriate — training configuration. The commit is registered in a public or semi-public registry that any third party can consult to verify that the committed version matches the version the organization claims was in production at the time of the decision.

Model Identity addresses the drift problem in the clearest form. A decision attributed to "our production model" is only as meaningful as the organization's account of which model that was. A decision bound to a specific committed version by a cryptographic certificate is meaningful independent of the organization's account. Regulators investigating patterns of decisions can verify that the same model was running across the period they care about. Plaintiffs alleging that an unauthorized experimental variant produced their decision can verify the allegation or be refuted by the verification. Internal audit can detect model substitution that might have escaped deployment oversight.

4.2 Property Two: Structural Independence from Declared Prohibited Inputs

The deploying organization publishes, signs, and versions a registry of inputs that the algorithmic system is prohibited from considering in its decisions. The registry is public or semi-public, durable, and versioned such that any certificate generated by the system is bound to the specific version of the registry in force at the time of the decision. For every consequential decision, a certificate is generated proving that the decision was structurally independent of the declared prohibited inputs — that is, the decision would not have changed if those inputs had been zeroed or set to a reference value.

Structural Independence addresses the core concern of current AI fairness regulation in a form that is verifiable rather than attested. A lender committing that race was not considered in a credit decision does not currently produce evidence supporting that commitment; under the Standard, the lender produces a certificate proving structural independence from race-correlated features (specifically enumerated) for every decision. An insurer committing that litigation propensity did not influence a coverage determination produces a certificate demonstrating the same. An employer committing that non-job-related demographic signals did not influence a screening decision produces a certificate.

Critically, the Standard does not specify which inputs must be declared prohibited. That is a sector-specific regulatory, ethical, and commercial question, and deploying organizations will vary in what they commit to. What the Standard specifies is that whatever is committed to must be declared in the public registry and proven at every decision, not merely asserted in policy documents. The content of the registry becomes subject to external scrutiny, criticism, and regulation; the rigor with which the registry reflects genuine prohibitions is the measure of the organization's commitment. Registries that are narrow become publicly narrow. Registries that omit obvious concerns are subject to obvious criticism.

4.3 Property Three: Chained Record Integrity

The certificate for each decision is cryptographically chained to the certificates for prior related decisions — the same subject, the same model version, the same workflow — such that the chain's integrity can be verified by any third party and any alteration in any historical record invalidates the chain going forward. The chain's head is periodically anchored to a public tamper-evident registry that cannot be revised by the deploying organization.

Chained Record Integrity addresses the allegation, common in both litigation and regulatory enforcement, that decision records have been altered after the fact in response to inquiry. Under current infrastructure, such allegations are difficult to refute and difficult to substantiate; they generate lengthy discovery, expert testimony about log file integrity, and inconclusive conclusions. Under the Standard, the chain's cryptographic structure either validates or it does not, with no room for interpretation. Organizations that have not altered records benefit from a mechanism that refutes false allegations; organizations that have altered records lose the ability to do so without detection.

4.4 Property Four: Credentialed Human Review

For decisions that involve human review, approval, or override, the certificate binds the decision to the credential of the reviewing human at the time of the decision. The credential may be a professional license, a role authorization, an organizational approval level, or any other externally-verifiable marker of authority. The credential's validity at the time of decision is itself cryptographically verifiable through a separate attestation chain maintained by the credentialing authority.

Credentialed Human Review addresses the increasingly common scenario in which AI systems produce recommendations that are formally "reviewed" by humans whose actual role in the decision is unclear. A determination attributed to a licensed professional may have been produced by a nurse or junior analyst; a credit review may have been performed by a clerk rather than the underwriter the decision is formally attributed to; a medical decision may have been signed by a physician who did not actually see the case. The Standard does not prohibit any of these practices — that is a substantive policy question beyond the Standard's scope — but it makes the credential-of-actual-reviewer verifiable, so that organizations choosing to use human review as part of their governance posture are held to what they claim.

4.5 Composition and optional extensions

The four properties above constitute the core of the Standard. In most deployment contexts, a decision's certificate will attest to all four simultaneously, composed into a single artifact that a third party can verify in a single operation. In lower-risk contexts, a subset may be appropriate — a content moderation decision may warrant Model Identity and Chained Record Integrity but not Credentialed Human Review, for example, because the workflow does not involve human credentialing. The Standard accommodates these compositions without requiring all four properties to apply in every context.

Several optional extensions warrant mention. Training-time attestations can bind a model's commit to specific training data manifests, training configurations, or fairness evaluations performed at training time. Population-level aggregation can compress many decision-level certificates into a single constant-size proof suitable for regulatory reporting at scale. Compound separation can attest to joint independence from several prohibited-input groups simultaneously rather than each separately. These extensions are not required by the core Standard but are natural companions to it and can be adopted by organizations with higher governance ambitions or in sectors with higher regulatory expectations.

5. What the Standard Does Not Address

No governance mechanism is complete, and the Standard proposed here is not an exception. The paper has argued that mathematical verification closes a specific evidentiary gap that current AI governance cannot close. This section names, as explicitly as possible, the problems the Standard does not close, so that readers can assess what additional governance mechanisms they continue to need and what conversations they still need to have.

5.1 Substantive fairness, accuracy, and quality

The Standard addresses structural properties of AI decisions. It does not address whether the decisions are correct, fair, or wise. A perfectly valid certificate can accompany a decision that is substantively indefensible — a denial that should have been an approval, a classification that is clinically wrong, a risk score that reflects bias encoded in training data. The certificate proves that a specific model ran, that declared prohibited inputs were structurally excluded, that the record is tamper-evident, and that the reviewing credential was valid. It does not prove that the model was well-designed, that the training data was representative, that the output was appropriate, or that the reviewer exercised judgment beyond formal sign-off.

Substantive fairness, accuracy, and quality remain human responsibilities. They are addressed by good training practices, thoughtful model design, ongoing monitoring, diverse evaluation populations, fairness-aware machine learning techniques, and clinical/legal/domain expertise in the deployment context. The Standard does not substitute for any of these; it is designed to be adopted alongside them.

5.2 The completeness of the prohibited-inputs registry

The Standard proves structural independence from declared prohibited inputs. It does not prove that the declaration is complete. If an organization declares prohibition of directly named protected characteristics but omits proxies for those characteristics — ZIP code as a proxy for race, occupation as a proxy for socioeconomic class, purchase history as a proxy for religion — the certificate will validate without catching the proxy influence. The completeness of the registry is a human governance question, not a cryptographic one.

This limitation is, however, more usefully framed as a feature of the architecture than a flaw. The registry is public. Its contents are subject to external evaluation, advocacy pressure, regulatory specification, and academic critique. Under the Standard, the debate about what should be prohibited becomes a debate about specific enumerated items on a public registry — a more productive debate than the one current governance supports, which concerns what organizations have vaguely committed to in principles documents.

5.3 Training-time influence

The core Standard operates at inference time — the moment a decision is produced. Several important forms of bias and inappropriate influence enter a model during training, not at inference. A model trained on biased data will encode the bias in its weights, and the inference-time certificate will validate even though the underlying behavior is problematic. Training-time attestations (mentioned in Section 4.5) address parts of this problem, but they are optional extensions and not yet as mature as inference-time proofs. Organizations and regulators pursuing comprehensive AI accountability should expect to develop training-time mechanisms over the next three to five years.

5.4 Adversarial model and registry updates

The Standard depends on the durability of the model commit and the prohibited-inputs registry. If a deploying organization can silently replace a committed model with a subtly different one, or can modify the registry without public notice, the Standard's guarantees are weakened. Mechanisms for preventing such manipulation — public anchoring of registry versions, mandatory notice periods for registry changes, third-party witnesses to model commits — are part of a full Standard implementation but require governance infrastructure beyond the cryptography itself. A careful adoption will include these mechanisms; a careless one will not.

5.5 The adversarial deployer

The Standard addresses deployers who are willing to make genuine governance commitments and want external verification of those commitments. It offers less against deployers who wish to evade governance entirely. A sufficiently adversarial organization can refuse to produce certificates at all, can commit to minimal registries that do not meaningfully constrain behavior, or can arrange workflows to move consequential decisions outside the Standard's scope. The Standard is a tool for enforcing commitments, not a tool for forcing commitments to exist. Creating the environment in which organizations make meaningful commitments is a regulatory and political task that the Standard supports but does not accomplish on its own.

6. Sectoral Applications

The Standard is general; its practical meaning becomes visible in specific sectors. This section walks through representative applications across six domains where consequential AI decisioning is now widespread. The examples are not exhaustive; they are illustrative of how the Standard's four properties translate into sector-specific evidentiary commitments.

6.1 Healthcare utilization management

A health plan deploying AI-assisted prior authorization produces, for each determination, a certificate attesting that the committed clinical classifier produced the decision, that the decision was structurally independent of cost features and demographic proxies declared in the plan's prohibited-features registry,

that the determination record is chained to prior determinations for the same member, and that the reviewing clinician's credential was valid at decision time. Certificates are available to the member, their provider, state insurance regulators, CMS, and counsel in the event of appeal or litigation. The certificate does not address whether the clinical criteria were appropriate — that remains a clinical and policy question — but it addresses the structural questions that currently dominate utilization-management litigation.

6.2 Financial services and consumer lending

A consumer lender deploying algorithmic underwriting produces, for each decision, a certificate attesting that the committed underwriting model made the decision, that the decision was structurally independent of the registry's prohibited inputs (race, gender, age, and declared proxies), that the decision record is chained to prior decisions for the same applicant, and — for adverse actions requiring human review under Regulation B — that the reviewing underwriter's authorization was valid. The certificate accompanies the adverse action notice required by Regulation B, and CFPB examinations can verify large samples of decisions in computational workflows rather than through document-review audits. The certificate does not address whether the underwriting criteria reflect sound lending practice — that remains a safety-and-soundness question — but it addresses the fair lending structural questions that dominate enforcement.

6.3 Employment screening and hiring

An employer or employment-screening vendor deploying algorithmic candidate evaluation produces, for each screening decision, a certificate attesting to the committed screening model, to structural independence from prohibited inputs (protected characteristics and their declared proxies), and to chained record integrity. The certificate can be provided to candidates receiving adverse screening outcomes and to enforcement agencies (EEOC, state civil rights commissions) investigating patterns. Under New York City Local Law 144 and analogous state statutes, the required bias audit can be augmented by certificate-level evidence rather than relying entirely on the audit's point-in-time findings.

6.4 Insurance underwriting and claims

An insurance carrier deploying AI-assisted underwriting or claims adjudication produces certificates attesting to committed model identity, structural independence from declared prohibited inputs (including the specific factors state insurance departments prohibit in each line of business), chained integrity across related claims, and credentialed adjuster review for claims requiring it. The certificates become part of the carrier's market-conduct examination infrastructure, and state insurance departments can verify compliance at scale rather than through the case-file audits that currently consume most of the examination budget.

6.5 Content moderation and platform governance

A platform deploying algorithmic content moderation produces certificates attesting to the committed moderation model, to structural independence from declared prohibited inputs (such as the user's declared political affiliation, in platforms that commit to such independence), and to chained integrity of moderation records. The certificates are available to affected users in appeals, to regulators enforcing platform transparency laws (EU Digital Services Act, analogous state frameworks), and to independent researchers studying moderation patterns. The certificate does not address whether moderation decisions are substantively correct — that remains a policy and editorial question — but it addresses the "was this treated differently because of who I am" question that drives most moderation controversy.

6.6 Government benefit determinations and public services

A government agency deploying algorithmic tools in benefit eligibility determination, child welfare assessment, unemployment insurance fraud detection, or tax audit selection produces certificates for each determination. The certificates attest to committed model identity, to structural independence from declared prohibited inputs, to chained record integrity, and to credentialed case-worker review where applicable. The certificates are available to affected citizens, to oversight bodies (inspectors general, auditors general, ombudsman offices), and to legislative oversight. In the public sector, the Standard has particular force because government accountability norms are higher than private-sector norms and because the consequences of opaque algorithmic decisioning by government agencies are politically and legally more severe.

Across all six sectors, the pattern is consistent. The Standard does not resolve the substantive policy questions each sector faces — what clinical criteria should be used, what underwriting factors are sound, what employment screening is job-related, what insurance practices are actuarially justified, what moderation is appropriate, what government benefits are deserved. Those questions remain for humans to resolve. What the Standard does is make the structural properties of algorithmic decisioning in each sector verifiable, so that the substantive debates can occur with better evidence and the structural protections can be enforced without relying on the deploying organization's self-report.

7. Adoption and the Path Forward

7.1 Regulatory adoption

The Standard is designed to be adoptable as a regulatory requirement in any sector with existing AI-specific regulatory frameworks and in sectors where such frameworks are being developed. The technical specifications are public, the verification infrastructure is inherently multi-vendor, and the performance characteristics are within the range regulators can reasonably impose. The Standard avoids the common failure mode of regulatory specification — over-prescription of implementation details that becomes obsolete as technology evolves — by specifying properties to be attested to rather than mechanisms by which attestation occurs.

Regulatory adoption requires three things. First, the sector's regulator specifies the minimum prohibited-inputs registry applicable in the sector — the specific inputs that regulated entities must declare and prove structural independence from. This is a policy question that regulators are positioned to answer and that should involve public comment. Second, the regulator specifies which decisions require certificates — the threshold of consequentiality below which certification is not required. This is a cost-benefit judgment appropriate to the sector. Third, the regulator develops or designates a verification capacity — the technical ability to check certificates at the scale of the sector's activity. This requires modest technical investment but not fundamental re-architecting of regulatory agencies.

7.2 Procurement adoption

Even in the absence of regulatory mandate, the Standard is adoptable as a procurement specification. Large enterprise AI buyers — in healthcare, financial services, government, and other regulated industries — can require that vendors supplying consequential decisioning systems produce certificates meeting the Standard's properties. This shifts the incentive landscape for AI vendors toward certificate production regardless of whether any regulator has mandated it, and creates a path to voluntary industry-wide adoption that does not depend on legislative action.

Procurement adoption is likely to develop faster than regulatory adoption in most sectors, for simple commercial reasons. AI vendors will produce certificates because their enterprise customers demand them; enterprise customers will demand them because their own regulatory and legal exposure drives them to; the combination drives de facto standardization without requiring de jure regulation. Regulators can then formalize the de facto standard once it has achieved sufficient adoption.

7.3 Voluntary industry adoption

Several industries have developed voluntary self-regulatory frameworks for AI governance — AHIP's commitments on prior authorization, the financial services industry's various model-risk management guidelines, trade-association frameworks in insurance and employment screening. The Standard can be adopted as a voluntary commitment within these frameworks, and doing so offers the adopting organizations several advantages: differentiation against non-adopting competitors, preemptive positioning against regulatory action, and a credibility signal to enterprise customers and subjects.

Voluntary adoption has the disadvantage of being unevenly distributed — organizations with good governance practices adopt, organizations with poor practices do not, and the non-adopters are exactly the ones governance most needs to reach. But voluntary adoption also accelerates regulatory and procurement adoption by establishing the practical feasibility of the Standard and by generating the technical vendor ecosystem that supports it. The cycle tends to produce industry-wide adoption within five to ten years once voluntary adoption reaches a critical threshold.

7.4 What different actors should do

This paper does not prescribe a universal course of action. It closes with specific observations about what major actors should consider doing, based on the structural analysis above.

Federal and state regulators should begin incorporating certificate-based verification into their AI-specific regulatory frameworks. This does not require displacing existing mechanisms — impact assessments, algorithmic audits, principles-based requirements — but complements them with a verification layer that closes the evidentiary gap those mechanisms cannot close. Regulators should also begin building internal verification capacity, which is a modest technical investment with substantial long-term enforcement leverage.

Enterprise AI buyers should incorporate certificate requirements into procurement specifications for consequential AI systems. This creates commercial pressure for vendor adoption without waiting for regulation, and reduces the buyer's own regulatory and legal exposure. Buyers should also begin evaluating certificates from current vendors where possible, building organizational familiarity with the infrastructure before it becomes mandatory.

AI vendors should evaluate where certificate production fits in their product roadmap. Vendors whose commercial position depends on governance credibility — enterprise AI platforms, vertical AI in regulated industries, tools sold into compliance-sensitive workflows — will find early adoption commercially valuable. Vendors whose current competitive position relies on capabilities that cannot be defended in a public registry should understand that their position is eroding regardless of their own adoption choices.

Policy researchers and standards bodies should engage with the Standard's specifications, propose refinements, and develop the cross-sector infrastructure that makes adoption practical. The specifications in this paper are a starting point, not a finished product. The path to a durable industry standard runs through the standards-body work that will follow early adoptions.

Affected subjects, civil society organizations, and advocacy groups should understand what certificates enable and incorporate them into advocacy, litigation, and research strategies. Certificates shift the balance of evidentiary capacity in favor of subjects and their advocates in ways that current governance does not. Advocacy that understands this shift will be more effective than advocacy that continues to rely on the tools of the attestation-based era.

8. Conclusion

Artificial intelligence has become consequential infrastructure across the regulated economy. The governance mechanisms that have grown up alongside it are reasonable, valuable, and collectively insufficient. Their insufficiency is not a failure of the individual mechanisms — it is a structural limit on what governance built on organizational attestation can accomplish when the activity being governed occurs inside computational systems that the governing parties cannot independently observe.

Mathematics now offers a form of evidence that does not depend on trusting the organization producing it. The techniques are not theoretical — they are deployable at the latency and volume consequential AI decisioning operates at, and they address exactly the evidentiary gap that current governance cannot close. The Standard proposed in this paper specifies four verifiable properties — Model Identity, Structural Independence from Declared Prohibited Inputs, Chained Record Integrity, and Credentialed Human Review — that any consequential algorithmic decisioning system can be designed to produce evidence for.

The Standard does not solve every problem in AI governance. It does not address substantive fairness, accuracy, or quality. It does not substitute for good training practices, thoughtful model design, or ongoing monitoring. It does not close the completeness gap in prohibited-inputs declarations, which remains a human governance question. It does not force organizations to make meaningful commitments; it only makes meaningful commitments verifiable once made. These limits are real and should be named honestly in any conversation about adoption.

What the Standard does is transform the AI accountability conversation from one about organizational self-report to one about verifiable evidence. This is not the whole of AI governance, but it is the part of AI governance that current infrastructure cannot provide and that the mathematical techniques now available can. The shift from governance-by-assertion to governance-by-verification is not a minor improvement in existing mechanisms; it is a change in what counts as compliance, with implications that will reshape regulatory enforcement, enterprise procurement, litigation, and public trust over the next decade.

The shift is occurring whether any individual actor chooses to participate in it. The forces pressing on the evidentiary gap — regulatory, legal, commercial, political — will continue to press on it, and the gap will continue to widen until it is closed. The organizations, agencies, and standards bodies that understand the structural dynamics early will shape the terms on which the transition occurs. The ones that do not will find the terms shaped for them, and will adapt later and at greater cost.

The compliance problem for artificial intelligence is, at its core, a question about what counts as evidence. The answer that served an earlier technological era — organizational attestation, principles-based self-governance, audit opinions constrained by organizational disclosure — is no longer sufficient for the scale and consequentiality at which AI is now deployed. The answer the next era requires is available. The question is who will be the first to adopt it, and who will be the last.

About this paper

This white paper proposes a cross-sector standard for verifiable AI accountability based on mathematical verification techniques that have matured from research artifacts into deployable infrastructure over the past five years. The

paper is vendor-neutral and technology-neutral; the Standard's properties can be implemented using any of several cryptographic techniques (SNARK, STARK, and successors), on any reasonable hardware platform, and by any competent engineering team. The paper's purpose is to make the structural argument accessible to regulators, enterprise buyers, policy researchers, and standards bodies working on AI accountability infrastructure, and to propose a specific framework that adoption and formalization can organize around. The conclusions are the author's and should not be attributed to any organization whose work is referenced or described.