AGS-1

Auburn Governance Stack Architecture Specification

A Layered Infrastructure for Verifiable AI Compliance

Cryptographic Model Attestation from Hardware Trust Roots
to Sector-Specific Regulatory Evidence

Ryan Fields

Principal Researcher, Governance Infrastructure & AI Safety

UncleBroFields@proton.me
fieldsryanchristopher@gmail.com

March 2026

Auburn Patent Family Fields

Version 1.0

Intellectual Property (IP) Declaration

Auburn Patent Family Fields

Ownership. The methods, logic structures, architectural designs, layer definitions, composition rules, dependency structures, and governance infrastructure specifications contained in this work are the sole property of Ryan Fields.

Public License (Non-Commercial)

- **Academic Use:** Researchers may share and use this framework for non-commercial academic purposes, provided full attribution is given to Ryan Fields.
- **No Derivatives:** No modifications or adaptations of the architectural specifications, layer definitions, or composition rules are permitted without express written consent.

Commercial Prohibition

Commercial use of this framework is strictly prohibited. This includes, but is not limited to:

- Integration into commercial AI governance, risk management, or compliance platforms.
- Use within proprietary conformity assessment, attestation verification, or audit infrastructure.
- Incorporation into consulting methodologies, regulatory advisory services, or certification programs.
- Use by standards bodies, industry consortia, or technology vendors in commercial product development without license.

Relationship to Other Auburn Documents

AGS-1 is the architectural meta-document of the Auburn Governance Stack. It names the full stack, defines every layer, declares the dependency structure, and specifies the composition rules through which all other Auburn Patent Family documents interoperate. Individual Auburn clauses (AI-1 through AI-9 and beyond) define specific invariants, protocols, and compliance profiles; AGS-1 defines the system within which they operate.

Contents

List of Tables

List of Figures

1 Executive Summary

The Auburn Governance Stack (AGS) is a layered architecture specification for verifiable AI compliance. It defines the complete infrastructure—from hardware trust roots to sector-specific regulatory evidence—required to produce, verify, and enforce cryptographic proof that an AI system was operating in a certified internal state at the moment of inference or training.

AGS-1 is the architectural meta-document. It names the stack, defines the seven layers plus a cross-cutting bridge category, declares the dependency structure among all constituent documents, specifies the composition rules through which lower-layer evidence flows to upper-layer applications, and maps the architecture against the global standards landscape. Every other document in the Auburn Governance Stack either feeds evidence into the composition waist defined here or consumes evidence from it. No document operates independently of the architecture AGS-1 describes.

1.1 The Problem

The global AI governance ecosystem has produced an extensive body of frameworks, regulations, and standards. The NIST AI Risk Management Framework defines organizational risk processes. ISO/IEC 42001 defines management system requirements. The EU AI Act defines outcome-based obligations for high-risk systems. The IETF Remote ATtestation procedureS (RATS) working group defines cryptographic attestation primitives. Supply chain integrity frameworks (SCITT, SLSA, in-toto) define provenance mechanisms for software artifacts.

None of these efforts defines a verification architecture that connects governance requirements to measurable model properties to cryptographically verifiable attestation anchored in hardware trust roots. Each operates at a single layer—governance process, measurement science, or cryptographic primitive—with no specification binding them into an end-to-end system. The result is an ecosystem that can tell an organization what risks to manage but cannot tell it how to *technically verify* that those risks are being managed at inference time.

1.2 The Architecture

The AGS follows the same structural logic as the TCP/IP protocol stack. An hourglass architecture channels all lower-layer evidence through a narrow composition waist—the Model Attestation Interface (MAI-1) and this architectural specification (AGS-1)—ensuring that every component integrates through a single, standardized interface. Lower layers produce evidence: hardware platform attestation (Layer 1), model state health invariants (Layer 2), and supply chain provenance (Layer 3). Upper layers consume evidence: conformance testing (Enforcement Layer) and sector-specific compliance profiles (Application Layer). A foundational theory layer (Layer 0) establishes the intellectual authority of the stack, and a cross-cutting bridge category connects the AGS to the broader standards ecosystem.

The stack comprises documents organized across these layers, of which a core set are published on Figshare under CC BY-NC-ND 4.0 with Auburn Patent Family IP declarations. The architecture is designed to be *compositional*: new invariants, new sector profiles, and new enforcement mechanisms can be added without modifying the composition waist, following the principle that protocol infrastructure should standardize slowly while scientific content evolves rapidly.

1.3 Design Commitments

Four principles govern the architecture:

1. **Standardize the protocol, let the content evolve.** The cryptographic primitives, attestation token format, and inter-layer binding mechanism are infrastructure—they change

slowly. The specific health invariants, threshold values, and measurement algorithms are science—they change rapidly and **MUST** not be prematurely frozen.

2. **The honest framing.** The stack provides probabilistic risk reduction and accountability infrastructure, not behavioral safety guarantees. This is analogous to financial auditing, which certifies process compliance without guaranteeing future solvency.

3. **Binary compliance.** Conformance is binary: a system either passes or fails. There is no "partial compliance" and no interpretive flexibility. This is the design decision that creates enforcement pressure.

4. **Self-authorizing documents.** Each document is designed to be forwarded without explanation, read as final, and referenced without the author's involvement.

1.4 Intended Audience

This specification is intended for AI governance architects, chief information security officers, conformity assessment bodies, standards body participants (CEN/CENELEC JTC 21, ISO/IEC JTC 1/SC 42, IETF RATS), regulatory staff drafting implementing acts, insurance underwriters building AI risk models, procurement officers writing AI system requirements, and researchers studying governance infrastructure design.

2 The Governance Infrastructure Gap

Every major AI governance effort published as of March 2026 operates at one of three layers. No published specification spans all three with defined interfaces.

- **Governance/Process Layer.** Defines what risks to manage and what organizational structures to maintain. Includes NIST AI RMF, ISO/IEC 42001, EU AI Act requirements, OMB guidance, and corporate governance policies. Prescribes no technical verification mechanisms.
- **Measurement/Evaluation Layer.** Defines what properties to test and how to evaluate them. Includes NIST ARIA/CoRIx, MLCommons AILuminate, FDA PCCP requirements, ISO/IEC TS 42119, and the OWASP AI Testing Guide. Does not specify how to cryptographically bind evaluation results to deployment-time enforcement.
- **Cryptographic/Attestation Layer.** Defines signing, verification, and transparency mechanisms. Includes IETF RATS (RFC 9334, RFC 9711), SCITT, C2PA, in-toto/SLSA, and OpenSSF Model Signing. Has no AI-specific profiles, no binding to governance requirements, and no runtime enforcement protocols.

The gap between these layers is not a deficiency in any individual framework. Each was designed with appropriate scope. The deficiency is *architectural*: no specification defines the interfaces, composition rules, and end-to-end verification flows that would connect governance requirements to measurable properties to cryptographically verifiable evidence anchored in hardware trust roots.

2.1 Framework-by-Framework Analysis

2.1.1 NIST AI Risk Management Framework

The NIST AI RMF 1.0 (AI 100-1, January 2023) defines four core functions—Govern, Map, Measure, Manage—with seven trustworthiness characteristics. It is explicitly voluntary, non-certifiable, use-case agnostic, and technology-agnostic. The companion Generative AI Profile (AI 600-1, finalized July 2024) defines twelve risk categories and over 200 suggested actions mapped to RMF subcategories.

The NIST AI ecosystem has continued to expand. AI 100-2e2025 (Adversarial Machine Learning taxonomy, March 2025) provides attack and mitigation terminology. NIST IR 8596 (Cyber AI Profile, preliminary draft December 2025) maps AI risks to the Cybersecurity Framework 2.0. The COSAiS project (SP 800-53 control overlays for AI systems) published a concept paper in August 2025 and an annotated outline in January 2026. The ARIA evaluation program produced operational measurement tools including the Contextual Robustness Index (CoRIx).

Gap: The entire NIST AI ecosystem specifies zero cryptographic verification mechanisms, zero runtime monitoring architectures, and zero hardware trust root dependencies. AI RMF's Measure function defines *what* to measure but not *how to cryptographically attest* that measurement occurred on certified hardware with certified software. The COSAiS overlays, when finalized, may reference existing SP 800-53 controls that address hardware integrity, but no current draft articulates this connection for AI systems.

2.1.2 ISO/IEC 42001 and the SC 42 Portfolio

ISO/IEC 42001:2023 (AI Management System) follows the Annex SL / PDCA pattern structurally identical to ISO 27001. It contains ten clauses and 38–39 Annex A controls covering AI risk assessment, data management, and system lifecycle. Certification is growing: Microsoft (365

Copilot), Cognizant, Miro, and Cornerstone Galaxy have achieved accredited certification through bodies including BSI, SGS, and Schellman.

The SC 42 portfolio has expanded with ISO/IEC 42005:2025 (AI system impact assessment), ISO/IEC 42006:2025 (certification body requirements), and ISO/IEC TS 42119-2:2025 (testing of AI systems). Under active development are ISO/IEC 42007 (conformity assessment schemes), ISO/IEC 42105 (human oversight), and ISO/IEC 42119-3 (verification and validation analysis).

Gap: No published SC 42 standard specifies cryptographic attestation, runtime invariant monitoring, or hardware trust root requirements. ISO 42001 requires "monitoring" at a governance/process level (Clause 9) but prescribes no monitoring architectures, drift detection thresholds, or real-time inference verification systems. It makes zero references to Trusted Platform Modules, Trusted Execution Environments, Hardware Security Modules, secure enclaves, or remote attestation protocols. CEN/CENELEC has independently determined that ISO 42001 does not cover all quality management requirements of the EU AI Act, leading to the development of the separate European standard prEN 18286.

2.1.3 EU AI Act and Harmonized Standards

The EU AI Act (Regulation 2024/1689) defines outcome-based obligations for high-risk AI systems across Articles 8–15, with conformity assessment procedures under Article 43. The European Commission issued standardization requests M/593 (May 2023) and M/613 (June 2025) to CEN/CENELEC for harmonized standards development.

The first harmonized standard to enter public enquiry was prEN 18286 (Quality Management System, October 2025). Additional drafts under development include prEN 18228 (Risk Management, Article 9), prEN 18229-1 (Logging, Transparency, Human Oversight, Articles 12–14), prEN 18229-2 (Accuracy and Robustness, Article 15), prEN 18282 (Cybersecurity, Article 15(5)), prEN 18284 (Dataset Quality, Article 10), and prEN 18283 (Bias Management). CEN/CENELEC adopted acceleration measures in October 2025 targeting completion by Q4 2026. No harmonized standards have been formally referenced in the Official Journal as of March 2026.

The enforcement timeline is subject to revision. Under the original Act, high-risk obligations for Annex III systems apply August 2, 2026. The Digital Omnibus proposal (COM(2025) 836, November 2025) proposes extending this to December 2, 2027 for Annex III systems and August 2, 2028 for Annex I systems, conditional on standards readiness. The proposal is under legislative negotiation.

Gap: The EU AI Act is deliberately technology-neutral. It defines no cryptographic attestation mechanisms, no runtime monitoring protocol, no hardware trust root dependency, and no verification protocol architecture. Article 12 requires automatic event logging but specifies no log format. Article 15 requires cybersecurity resilience but mandates no specific mechanism. Article 50(2) requires AI-generated content to be detectable by "technical means" and marked in "machine-readable format" but delegates specifics to codes of practice. The conformity assessment infrastructure itself is not yet operational—very few Member States have designated notified bodies for AI Act purposes.

2.1.4 IETF RATS and Related Attestation Work

The IETF RATS working group has produced the most architecturally mature attestation framework available. RFC 9334 (RATS Architecture, January 2023) defines the roles (Attester, Verifier, Relying Party, Endorser, Reference Value Provider), conceptual messages (Evidence, Endorsements, Reference Values, Attestation Results), and topological patterns (Passport model, Background-check model). RFC 9711 (Entity Attestation Token, April 2025) defines EAT as CWT/JWT with attestation-oriented claims using COSE/JOSE security envelopes, including a profile mechanism for use-case-specific token definitions. CoRIM (draft-ietf-rats-corim-09,

October 2025) defines CBOR-based reference integrity manifests. AR4SI (draft-ietf-rats-ar4si-09, August 2025) defines a Trustworthiness Vector with eight appraisal dimensions.

Individual Internet-Drafts have begun exploring AI applications: draft-messous-eat-ai-00 (February 2026) defines an EAT profile for autonomous AI agents; draft-huang-rats-agentic-eat-cap-attest-00 (June 2025) proposes capability attestation extensions for agentic AI; draft-aylward-aiga-1-00 (November 2025) proposes an AI Governance and Accountability Protocol. None have achieved working group adoption.

Gap: RATS provides the composable attestation substrate—token formats, signing envelopes, reference value distribution, and appraisal architecture—but no working-group-adopted document defines an AI-specific attestation profile. There is no standard mapping of model health invariants to EAT claims, no RATS profile for runtime inference attestation, and no composition algebra for multi-model attestation. The individual drafts are early-stage and address narrow use cases (agentic capability, agent identity) rather than comprehensive model state governance.

2.1.5 Supply Chain Integrity Frameworks

The supply chain integrity ecosystem is converging toward AI model provenance through three efforts. OpenSSF Model Signing (OMS, library v1.0 April 2025, specification June 2025) provides Sigstore-based cryptographic signing of ML model artifacts, adopted by NVIDIA for all NGC Catalog models. The Atlas framework (arXiv, February 2025) composes C2PA manifests, in-toto attestation layouts, Intel TDX hardware attestation, and SCITT transparency logs into an end-to-end model provenance architecture. SLSA v1.2 (November 2025) adds a Source Track to the existing Build Track.

Gap: These frameworks address model artifact signing—verifying that a model binary has not been tampered with since publication. They do not address runtime model state—verifying that the model is *currently healthy* (entropy above floor, gradients stable, no distribution drift, structural coherence maintained) at inference time. The gap between signing a model artifact and continuously verifying model behavior in production is the central architectural challenge that no supply chain framework addresses.

2.2 The Gap in Tabular Form

Table 1 summarizes the analysis. For each major framework, it records whether the framework specifies governance process requirements, technical measurement methods, cryptographic verification mechanisms, runtime monitoring architectures, or hardware trust root dependencies.

Table 1: AI Governance Standards Landscape: Capability Matrix (March 2026)

Framework	Governance Process	Measurement Methods	Crypto Verification	Runtime Monitoring	Hardware Trust Root	End-to-End Architecture
NIST AI RMF 1.0	**Yes**	Partial[a]	**No**	**No**	**No**	**No**
NIST AI 600-1	**Yes**	Partial[a]	**No**	**No**	**No**	**No**
ISO/IEC 42001	**Yes**	**No**	**No**	**No**	**No**	**No**
EU AI Act	**Yes**	Outcome-based[b]	**No**	**No**	**No**	**No**
CEN/CENELEC prEN Series	In development	In development	**No**[c]	**No**	**No**	**No**
IETF RATS (RFC 9334/9711)	**No**	**No**	**Yes**	**No**[d]	Partial[e]	**No**
SCITT	**No**	**No**	**Yes**	**No**	**No**	**No**
C2PA v2.3	**No**	**No**	**Yes**	**No**	**No**	Content only[f]
OpenSSF Model Signing	**No**	**No**	**Yes**	**No**	**No**	**No**
SLSA v1.2 / in-toto	**No**	**No**	**Yes**	**No**	**No**	**No**
Google SAIF 2.0	**Yes**	Partial	**No**	Vendor-specific	**No**	**No**
Anthropic RSP v3.0	**Yes**	**Yes**	**No**	Internal only	**No**	**No**
Auburn Governance Stack	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

[a] ARIA/CoRIx provides measurement science; AI RMF defines measurement as an organizational function but prescribes no specific technical method.
[b] Articles 9, 10, 15 define outcome requirements (accuracy, robustness, data quality) but delegate technical specifics to harmonized standards not yet published.
[c] Draft harmonized standards (prEN 18229-2 for accuracy/robustness, prEN 18282 for cybersecurity) are under development but none published as of March 2026.
[d] RATS defines attestation at discrete points; no RATS document specifies continuous runtime monitoring architecture.
[e] RFC 9711 EAT includes hardware-oriented claims (UEID, security level, boot state) but RATS mandates no specific hardware trust root.
[f] C2PA provides end-to-end architecture for content provenance (media authenticity) but is scoped to output labeling, not AI system governance.

2.3 Implications

The empty cells in Table 1 are not criticisms of the frameworks listed. NIST AI RMF was correctly designed as a voluntary risk management process. ISO 42001 was correctly designed as a management system standard. IETF RATS was correctly designed as a general-purpose attestation architecture. The gap exists *between* them—in the absence of a specification that composes governance requirements, measurement science, and cryptographic attestation into a layered architecture with defined interfaces, composition rules, and end-to-end verification flows.

The Auburn Governance Stack is that specification. The remainder of this document defines its architecture.

3 Architecture Overview

The Auburn Governance Stack follows an hourglass architecture. Lower layers produce cryptographic evidence about hardware integrity, model health, and supply chain provenance. Upper layers consume that evidence for conformance testing and sector-specific regulatory compliance. All evidence flows through a narrow composition waist—MAI-1 (the Model Attestation Interface) and AGS-1 (this architectural specification)—ensuring that every component integrates through a single, standardized channel.

This design is not novel in structure. The TCP/IP protocol stack, the PCI-DSS security standard, the IEC 62443 industrial security architecture, and the AUTOSAR automotive software platform all follow the same principle: a layered architecture with a narrow composition waist that decouples lower-layer implementation from upper-layer consumption. What is novel is the *application* of this principle to AI governance—the recognition that verifiable AI compliance requires infrastructure, not merely policy.

3.1 The Hourglass Model

Figure 1 presents the full hourglass architecture. The vertical axis represents the evidence flow direction: evidence is produced at the bottom (hardware trust roots) and consumed at the top (sector-specific compliance profiles). The narrow waist at the center—MAI-1 and AGS-1—is the universal composition point through which all evidence passes.

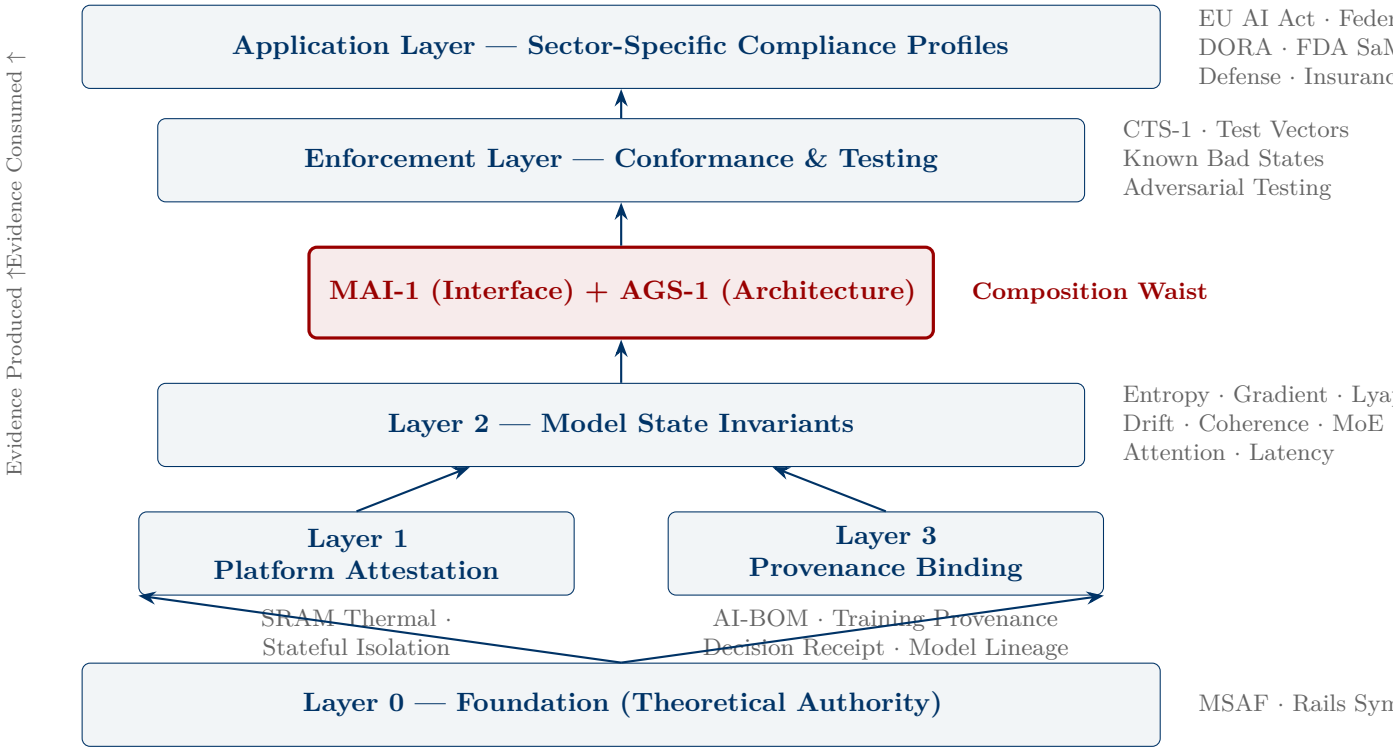

Figure 1: Auburn Governance Stack: Hourglass Architecture. All lower-layer evidence flows upward through the MAI-1/AGS-1 composition waist. All upper-layer applications consume evidence through standardized interfaces. Layers 1 and 3 operate in parallel, both feeding Layer 2.

3.2 Evidence Flow Direction

The hourglass enforces a strict evidence flow:

1. **Layer 0 (Foundation)** establishes the theoretical justification for why cryptographic AI attestation is necessary, the three-tier architecture design, the impossibility bounds, and the accessible tutorial. Foundation documents do not produce attestation evidence; they produce *intellectual authority*.

2. **Layer 1 (Platform Attestation)** produces hardware-level evidence: TEE attestation reports, firmware measurements, SRAM thermal integrity readings, and multi-tenant isolation proofs. This evidence answers the question: *Is the hardware platform trustworthy?*

3. **Layer 3 (Provenance Binding)** produces supply chain evidence: AI bills of materials, training provenance chains, decision receipts, contamination detection results, and model lineage records. This evidence answers the question: *Can you trace this output back to a certified origin?*

4. **Layer 2 (Model State Invariants)** consumes platform evidence (to anchor measurements in trusted hardware) and provenance evidence (to bind measurements to certified model versions), then produces health metrics: entropy floor compliance, gradient stability, distribution drift bounds, structural coherence, and extended measurements (MoE routing, attention thermodynamics, inference latency). This evidence answers the question: *Is the model healthy right now?*

5. **Composition Waist (MAI-1 + AGS-1)** receives all Layer 1–3 evidence and packages it into a single, cryptographically signed attestation artifact with a canonical token format, mandatory fields, and normative encoding rules. MAI-1 defines the interface; AGS-1 defines the architecture within which that interface operates.

6. **Enforcement Layer** consumes attestation artifacts from the composition waist and applies binary pass/fail rules, reference test vectors, known bad state detection, freshness requirements, and adversarial testing. This layer answers the question: *Does this system pass or fail?*

7. **Application Layer** consumes enforcement results and maps them to sector-specific regulatory requirements, insurance underwriting criteria, and procurement eligibility language. This layer answers the question: *Is this system compliant with my regulatory regime?*

Layers 1 and 3 operate in parallel—neither depends on the other. Both feed Layer 2. This parallelism is intentional: hardware attestation and supply chain provenance are independent evidence streams that converge at the model state measurement layer.

3.3 The TCP/IP Analogy

The AGS follows the structural logic of the TCP/IP protocol stack. Table 2 maps the analogy.

Table 2: TCP/IP Analogy Mapping

TCP/IP Component	Auburn Equivalent	Role
Cerf & Kahn (1974)	AGS-1 (this document)	Names the stack, defines the layers, declares the dependency structure
RFC 791 (IP)	MAI-1 (Clause AI-5)	The composition waist—the universal interface through which all evidence flows
RFC 793 (TCP)	CTS-1	Conformance testing and reliability guarantee
RFC 1180 (TCP/IP Tutorial)	Rails Symposium	Accessible capstone tutorial that makes the architecture legible to non-specialists
Physical Layer	Layer 1 (Platform)	Hardware trust root—the silicon upon which everything depends
Network Layer	Layers 2 + 3 (Invariants + Provenance)	Health metrics and supply chain evidence—the packet-level guarantees
Transport Layer	Enforcement Layer	Reliability and conformance—ensuring evidence delivery is correct
Application Layer	Sector Profiles	Regulatory, procurement, and insurance applications that deliver value to end users

The analogy is structural, not metaphorical. The AGS composition waist serves the identical architectural function as the IP layer in TCP/IP: it is the narrow point through which all lower-layer guarantees are delivered and all upper-layer applications consume services. Just as IP decouples physical network technologies (Ethernet, Wi-Fi, fiber) from application protocols (HTTP, SMTP, DNS), MAI-1 decouples hardware attestation technologies (Intel TDX, ARM CCA, AMD SEV-SNP) and model measurement methods (entropy estimation, gradient norm tracking, KL divergence monitoring) from regulatory compliance applications (EU AI Act profiles, DORA reports, FDA submissions).

This decoupling is the architectural property that enables the stack to evolve without breaking. When a new TEE technology emerges, only Layer 1 documents change. When a new health invariant is discovered, only Layer 2 documents change. When a new regulation is enacted, only an Application Layer profile is added. The composition waist remains stable.

3.4 Design Principles

Four principles govern the architecture. These are not aspirational statements; they are structural constraints enforced by the document design.

Principle 1: Standardize the Protocol

The cryptographic primitives (COSE signing, Merkle tree construction, EAT token format), the attestation endpoint specification (MAI-1 canonical endpoint), the inter-layer binding mechanism, and the conformance test structure are **infrastructure**. They change slowly and benefit from interoperability. They **SHALL** be versioned conservatively, with backward compatibility requirements.

The specific health invariants (entropy floor values, gradient stability thresholds, drift detection algorithms), measurement frequencies, and threshold calibrations are **science**. They change rapidly as measurement techniques improve. They **SHALL** be versioned independently of the protocol infrastructure and **MUST NOT** be frozen prematurely.

This separation ensures that the stack can incorporate new scientific understanding without destabilizing the protocol infrastructure, and that protocol stability does not prevent scientific progress.

Principle 2: The Honest Framing

The Auburn Governance Stack provides probabilistic risk reduction and accountability infrastructure. It does **not** provide behavioral safety guarantees.

A system that passes all AGS conformance tests is a system whose hardware platform has been attested, whose model state invariants are within certified bounds, whose provenance chain is intact, and whose conformance has been verified against binary pass/fail criteria. It is **not** a system that is guaranteed to behave correctly, produce accurate outputs, avoid harmful content, or satisfy any specific ethical standard.

This is analogous to financial auditing: a clean audit certifies that financial statements conform to accounting standards and that internal controls are functioning. It does not guarantee that the company will remain solvent, that management decisions will be wise, or that fraud will never occur.

Every document in the Auburn Governance Stack **SHALL** maintain this honest framing. Documents that claim behavioral guarantees beyond what cryptographic attestation can provide are non-conformant with the architecture.

Principle 3: Binary Compliance

Conformance to the Auburn Governance Stack is binary. A system either passes or fails. There is no "partial compliance," no "compliance score," and no interpretive flexibility in the pass/fail determination.

This principle is enforced structurally: CTS-1 defines test assertions with or verdicts. Each assertion evaluates a single, unambiguous condition. The aggregation rules for determining overall conformance level (MAI-C0, MAI-C1, MAI-C2) are deterministic.

The rationale is enforcement economics. If compliance can be argued, it cannot be enforced. Binary pass/fail eliminates the interpretive space in which non-compliant systems claim partial credit. This is the design decision that creates institutional pressure: organizations cannot claim "directional alignment" with the AGS; they either conform or they do not.

Principle 4: Self-Authorizing Documents

Each document in the Auburn Governance Stack is designed to be forwarded without explanation, read as final, and referenced without the author's involvement. Once published, the document's influence operates independently of its author.

This means every document **SHALL** contain sufficient context, definitions, and scope statements to be understood by a qualified reader who has not read any other Auburn document. Cross-references to other Auburn documents **SHALL** include sufficient summary that the referencing document remains self-contained.

This principle is a deliberate architectural choice. Governance infrastructure that requires an intermediary to explain or advocate for it is not infrastructure—it is consulting. The AGS is designed as infrastructure.

4 Layer Specifications

This section provides the normative definition of each layer in the Auburn Governance Stack. For each layer, the specification defines: what the layer produces, what it consumes, what it guarantees, what it cannot guarantee, which existing standards it composes with, and which Auburn documents populate it.

The key words **MUST**, **MUST NOT**, **SHALL**, **SHOULD NOT**, **SHOULD**, and **MAY** in this section are to be interpreted as described in RFC 2119 and RFC 8174.

4.1 Layer 0: Foundation (Theoretical Authority)

Layer 0 Definition

Layer 0 establishes the intellectual authority of the entire stack. It contains the theoretical justification for why cryptographic AI attestation is necessary, the three-tier architecture design that structures the attestation approach, the impossibility bounds that define what attestation cannot guarantee, and the accessible capstone tutorial that makes the architecture legible to non-specialist audiences.

4.1.1 Produces

Layer 0 does not produce attestation evidence in the cryptographic sense. It produces *intellectual authority*: the published arguments, formal analyses, and pedagogical materials that establish why the remaining layers are necessary and why they are designed the way they are. Layer 0 documents are the answer to the question: *Why should anyone believe this architecture is sound?*

4.1.2 Consumes

Layer 0 consumes the academic literature on AI safety, cryptographic attestation, hardware security, regulatory governance, and formal verification. It synthesizes these into a coherent architectural argument.

4.1.3 Guarantees

Layer 0 guarantees that the Auburn Governance Stack rests on published, reviewable theoretical foundations rather than unsupported assertions. It guarantees that the impossibility bounds—what the stack *cannot* achieve—are stated explicitly alongside the capabilities.

4.1.4 Cannot Guarantee

Layer 0 cannot guarantee that the theoretical analysis is correct in all respects, that the impossibility bounds are tight, or that the architectural design is optimal. Theoretical authority is subject to revision as understanding advances.

Honest Framing

Layer 0 is the most conventional academic contribution in the stack. Its value is not operational but epistemic: it demonstrates that the author understands the problem space at a depth sufficient to design the remaining layers. Readers who find errors in Layer 0 documents are invited to contact the author directly.

4.1.5 Standards Composition

Layer 0 composes with no external standards directly. It references and synthesizes the academic literature, IETF architectural documents (RFC 9334), and regulatory texts that inform the stack's design.

4.1.6 Auburn Documents

Table 3: Layer 0: Foundation Documents

Designation	Title	Status	Role
MSAF	The Model State Attestation Framework	Published	Three-tier architecture design, theoretical justification, impossibility bounds, regulatory landscape survey
—	Rails Symposium	Published	Accessible capstone tutorial making the architecture legible to policy audiences and non-specialists

4.2 Layer 1: Platform Attestation (Hardware Root of Trust)

Layer 1 Definition

Layer 1 proves the silicon is real and uncompromised. It contains specifications for Trusted Execution Environment attestation, firmware measurement chains, SRAM thermal integrity monitoring, multi-tenant stateful isolation, and honest disclosure of TEE physical limitations. Layer 1 evidence anchors all subsequent layers in hardware—without Layer 1, all higher-layer measurements could be fabricated by compromised software.

4.2.1 Produces

Layer 1 produces the following evidence categories:

1. **TEE Attestation Reports.** Cryptographically signed evidence that inference or training is executing within a verified Trusted Execution Environment (Intel TDX, ARM CCA, AMD SEV-SNP, or equivalent). The report **SHALL** include the TEE vendor identity, firmware version, security patch level, and a measurement of the loaded runtime.

2. **Firmware Measurement Chains.** Hash chains from the hardware root of trust through each firmware component to the AI runtime, following the DICE (Device Identifier Composition Engine) model or equivalent measured boot architecture. Each link in the chain **SHALL** be independently verifiable against reference values.

3. **SRAM Thermal Integrity Evidence.** Continuous monitoring of on-chip thermal conditions that could compromise computation integrity. The SRAM Thermal Integrity Bound (Auburn Clause AI-4) defines the temperature envelope within which silicon computations remain reliable and the attestation requirements when thermal conditions approach or exceed certified bounds.

4. **Stateful Isolation Proofs.** Evidence that multi-tenant GPU environments maintain state isolation between workloads. The Stateful Isolation Law (Auburn Clause AI-1) defines the contamination functional, six isolation clauses, composability theorems, and compliance tiers that characterize acceptable isolation boundaries.

5. **Side-Channel Disclosure.** Honest reporting of known side-channel vulnerabilities in the deployed hardware platform, following the principle that attestation of known limitations is more valuable than silence about them.

4.2.2 Consumes

Layer 1 consumes hardware vendor endorsements (Intel, AMD, ARM reference values), firmware signing keys, and TEE configuration policies. In RATS terminology, Layer 1 implements the Attester role for the hardware platform, consuming Endorsements and Reference Values from hardware vendors acting as Endorsers and Reference Value Providers.

4.2.3 Guarantees

Layer 1 guarantees that the hardware platform on which AI computations execute has been identified, measured, and attested through a cryptographic chain rooted in silicon. It guarantees that the TEE attestation report is bound to the specific firmware and runtime configuration present at attestation time. It guarantees that thermal conditions and isolation boundaries are monitored and reported.

4.2.4 Cannot Guarantee

Honest Framing

Layer 1 cannot guarantee the absence of undisclosed hardware vulnerabilities (e.g., future Spectre/Meltdown-class attacks), the correctness of hardware vendor endorsements, or the physical security of the deployment environment. TEE attestation proves that a specific firmware configuration was loaded; it does not prove that the firmware is free of bugs. Side-channel disclosures are limited to known vulnerabilities; unknown vulnerabilities remain, by definition, undisclosed.

Hardware attestation is necessary but not sufficient. A system with perfect hardware attestation and a compromised model is still dangerous. Layer 1 provides the *anchor*; Layers 2 and 3 provide the *content* that makes the anchor meaningful.

4.2.5 Standards Composition

Layer 1 composes with the following external standards:

- **IETF RATS** (RFC 9334, RFC 9711): Layer 1 evidence **SHALL** be encodable as EAT claims within the RATS Passport or Background-check model. TEE attestation reports map to EAT's hardware-oriented claims (UEID, OEM identification, security level, boot state).
- **TCG DICE** (Device Identifier Composition Engine): Firmware measurement chains **SHOULD** follow the DICE layering model for hardware-rooted identity and attestation.
- **Confidential Computing Consortium**: TEE profiles **SHOULD** align with CCC attestation architecture recommendations where applicable.
- **FIPS 140-3**: Cryptographic modules used in Layer 1 attestation **SHOULD** meet FIPS 140-3 Level 2 or higher where deployment context requires federal compliance.

4.2.6 Auburn Documents

Table 4: Layer 1: Platform Attestation Documents

Clause	Title	Status	Role
AI-4	SRAM Thermal Integrity Bound	Published	Defines temperature envelope for reliable silicon computation and attestation requirements at thermal boundaries
AI-1	Stateful Isolation Law	Published	Defines contamination functional, six isolation clauses, composability theorems, and compliance tiers for multi-tenant GPU environments
—	GPU TEE Profile	In Development	Vendor-specific attestation profiles for Intel TDX, ARM CCA, AMD SEV-SNP
—	Firmware Integrity Specification	In Development	Measured boot chain requirements and reference value distribution
—	Side-Channel Disclosure Protocol	In Development	Structured reporting of known hardware vulnerabilities

4.3 Layer 2: Model State Invariants (Continuous Health)

Layer 2 Definition

Layer 2 defines the health metrics measured at inference time and training time. These are the continuous signals that indicate whether a model is operating within certified bounds. Layer 2 answers the question: *Is the model healthy right now?* It contains the five mandatory invariants defined in MAI-1 (entropy floor, gradient stability, distribution drift, structural coherence, thermal integrity) plus extended specifications for Mixture-of-Experts routing, attention thermodynamics, inference latency, and quantization integrity.

4.3.1 Produces

Layer 2 produces health metric evidence for each monitored invariant. MAI-1 defines five mandatory invariants that every conformant system **MUST** report:

Table 5: Layer 2: Mandatory Model State Invariants

Clause	Invariant	What It Detects	Measurement Basis
AI-8	Entropy Collapse Constraint	Output distribution collapsing to degenerate modes; repetitive or frozen generation	Shannon entropy of output token distribution against certified floor
AI-2	Gradient Starvation Envelope	Training instability where gradient flow to specific layers or attention heads falls below functional thresholds	Layer-wise gradient norm tracking against certified stability envelope
AI-3	Lyapunov Stability Bound	Speculative decoding divergence where draft model predictions become unboundedly unstable	Lyapunov exponent estimation for speculative decoding verification chains
AI-6	Distribution Drift Bound	Model behavior drifting outside its certified operating distribution during deployment	KL divergence or equivalent between current output distribution and certified reference distribution
AI-7	Structural Coherence Requirement	Architectural degradation where internal representations lose structural relationships necessary for correct function	Representation similarity metrics across model layers against certified coherence bounds

In addition to the five mandatory invariants, Layer 2 includes extended specifications that conformant systems **MAY** report for enhanced governance coverage:

- **MoE Routing Attestation.** For Mixture-of-Experts architectures (DeepSeek, Mixtral, Switch Transformer), monitoring of expert selection distributions, load balancing metrics, and routing collapse detection. A system that routes all tokens to a single expert is functionally degraded regardless of output quality metrics.
- **Attention Thermodynamics.** Monitoring of attention entropy across layers and heads, detecting attention collapse (all mass on one token) or attention diffusion (uniform distribution conveying no information). Provides early warning of representational degradation before it manifests in output quality.
- **Inference Latency Bound.** Monitoring of inference latency distributions to detect computational anomalies (e.g., cache poisoning, resource exhaustion, or silent failover to uncertified hardware) that may not be visible in model output but indicate platform compromise.
- **Quantization Integrity.** For quantized deployments (INT8, INT4, mixed precision), monitoring of quantization error bounds to ensure that precision reduction has not exceeded the certified accuracy envelope.

4.3.2 Consumes

Layer 2 consumes Layer 1 evidence (to anchor health measurements in attested hardware) and Layer 3 evidence (to bind measurements to a specific certified model version). Without Layer 1,

health measurements could be fabricated by compromised software. Without Layer 3, health measurements could apply to an uncertified model substituted for the original.

Composition Rule

Layer 2 invariant measurements **SHALL** be cryptographically bound to the Layer 1 TEE attestation report that was active at measurement time. A health measurement not bound to a valid platform attestation **SHALL** be treated as unverified by any Relying Party.

4.3.3 Guarantees

Layer 2 guarantees that the specified health invariants have been measured, that the measurements were performed on attested hardware (via Layer 1 binding), and that the measurement results are within or outside certified bounds as reported. The binary determination—invariant satisfied or violated—is deterministic given the measurement and the certified threshold.

4.3.4 Cannot Guarantee

Honest Framing

Layer 2 cannot guarantee that the set of monitored invariants is complete—that no unmonitored failure mode exists. The five mandatory invariants represent the current state of measurement science for model health. As the field advances, new invariants will be identified and added to the stack. A model that passes all five mandatory invariants may still exhibit failure modes not captured by any current invariant.

Layer 2 also cannot guarantee the correctness of threshold calibration. The certified bounds for each invariant are derived from formal analysis and empirical validation, but they represent conservative estimates. A threshold that is too loose permits degraded models to pass; a threshold that is too tight produces false failures. Threshold calibration is an active area of research within each invariant specification.

The honest framing is this: Layer 2 provides the best currently available set of runtime health indicators, cryptographically anchored in hardware, with formally derived bounds. It is not omniscient. It is better than nothing, and dramatically better than the status quo of no runtime monitoring at all.

4.3.5 Standards Composition

Layer 2 composes with the following:

- **IETF RATS** (RFC 9711): Layer 2 invariant measurements **SHALL** be encodable as EAT claims. Each invariant maps to a custom claim within the EAT profile defined by MAI-1. The AR4SI Trustworthiness Vector's "runtime-opaque" and "sourced-data" dimensions provide the closest existing RATS analog to model health attestation.
- **NIST AI RMF Measure Function**: Layer 2 operationalizes the Measure function defined in AI RMF 1.0 by providing specific, quantitative, continuously monitored metrics with formal thresholds. Where AI RMF says "measure," Layer 2 specifies *what*, *how*, and *against what bound*.
- **EU AI Act Article 15**: Layer 2 invariants provide the technical evidence base for Article 15's accuracy, robustness, and cybersecurity requirements. A system that maintains all five mandatory invariants within certified bounds has quantitative evidence of operational robustness.

- **MLCommons AILuminate**: Layer 2 measurement methods are complementary to AILuminate benchmark evaluation. AILuminate measures model capability at evaluation time; Layer 2 monitors model health at deployment time.

4.3.6 Auburn Documents

Table 6: Layer 2: Model State Invariant Documents

Clause	Title	Status	Role
AI-8	Entropy Collapse Constraint	Published	Mandatory invariant: output distribution entropy floor
AI-2	Gradient Starvation Envelope	Published	Mandatory invariant: gradient flow stability
AI-3	Lyapunov Stability for Speculative Decoding	Published	Mandatory invariant: speculative decoding divergence bound
AI-6	Distribution Drift Bound	Published	Mandatory invariant: deployment-time distribution drift
AI-7	Structural Coherence Requirement	Published	Mandatory invariant: internal representation integrity
—	Attention Thermodynamics	Published	Extended: attention entropy monitoring
—	MoE Routing Attestation	Published	Extended: expert routing health for MoE architectures
—	Inference Latency Bound	In Development	Extended: computational anomaly detection
—	Quantization Integrity	In Development	Extended: precision reduction error bounds

4.4 Layer 3: Provenance Binding (Supply Chain Integrity)

Layer 3 Definition

Layer 3 proves where the model came from and what happened to it. It contains specifications for AI bills of materials, training provenance chains, decision receipts for forensic reconstruction, contamination detection, and model lineage tracking. Layer 3 evidence answers the question: *Can you trace this output back to a certified origin?*

4.4.1 Produces

Layer 3 produces the following evidence categories:

1. **AI Bill of Materials (AI-BOM).** A normative manifest documenting the model's composition: base architecture, training datasets (with provenance hashes), fine-tuning history, alignment procedures, dependency chain (tokenizer, embedding layers, adapter modules), and version identifiers. The AI-BOM **SHALL** follow a structured format extending CycloneDX and SPDX conventions for AI-specific components.

2. **Training Provenance Chain.** A cryptographically linked sequence of records documenting each stage of model development: pre-training data selection, training hyperparameters, checkpoint hashes, fine-tuning datasets, RLHF reward model provenance, and post-training modifications. Each record **SHALL** be signed by the entity that performed the operation and **SHALL** include a timestamp bound to the signing key's validity period.

3. **Decision Receipts.** Structured records enabling forensic reconstruction of individual model outputs. A decision receipt binds a specific input, the model version that processed it, the platform attestation active at processing time, and the health invariant measurements at processing time into a single, cryptographically signed artifact. Decision receipts enable after-the-fact investigation of model behavior without requiring continuous output logging.

4. **Contamination Detection Results.** Evidence from automated scanning for training data contamination—benchmark leakage, personally identifiable information, copyrighted material, or other regulated content that should not appear in training data. Contamination detection results **SHALL** include the scanning methodology, coverage estimate, and a structured report of findings.

5. **Model Lineage Records.** A directed acyclic graph documenting the derivation relationships between model versions: which base model was used, which fine-tuning was applied, which merging operations were performed, and which distillation or pruning steps produced the deployed artifact. Lineage records enable Relying Parties to assess the trustworthiness of derived models based on the trustworthiness of their ancestors.

4.4.2 Consumes

Layer 3 consumes model artifacts (weights, configurations, tokenizers), training infrastructure logs, dataset metadata, and signing key material from model developers and training pipeline operators. In supply chain integrity terms, Layer 3 extends the software bill of materials concept to the AI model lifecycle.

4.4.3 Guarantees

Layer 3 guarantees that a provenance record exists for the attested model, that the provenance chain is cryptographically linked (each record signed and hash-chained to its predecessor),

and that the AI-BOM reflects the declared composition at the time of signing. It guarantees traceability: given a decision receipt, a Relying Party can reconstruct which model version, on which platform, with which health state, produced a specific output.

4.4.4 Cannot Guarantee

Honest Framing

Layer 3 cannot guarantee the completeness or accuracy of provenance records provided by the model developer. If a developer omits a training dataset from the AI-BOM, the AI-BOM is incomplete but may still be cryptographically valid. Provenance attestation proves that *the declared provenance chain is intact*; it does not prove that *the declaration is truthful*.

This is the fundamental limitation of supply chain integrity for AI: unlike software, where a reproducible build can verify that source code matches a binary, model training is stochastic and generally not reproducible. Two training runs with identical data and hyperparameters may produce different weights. Layer 3 therefore attests the *process record* rather than the *computational output*.

Contamination detection is similarly bounded. Current scanning methods detect known patterns (benchmark overlap, PII formats, copyright signatures) but cannot guarantee the absence of all contamination. Layer 3 reports what was scanned and what was found; it does not certify absence.

4.4.5 Standards Composition

- **SCITT** (draft-ietf-scitt-architecture): Layer 3 provenance records **SHOULD** be registrable in SCITT transparency services. The SCITT append-only ledger with Merkle tree inclusion proofs provides a natural infrastructure for AI-BOM and training provenance registration.
- **OpenSSF Model Signing (OMS)**: Layer 3 model artifact signatures **SHOULD** use OMS-compatible signing formats (Sigstore-based, detached signatures). This ensures interoperability with the emerging model signing ecosystem already adopted by NVIDIA NGC.
- **SLSA v1.2**: Layer 3 training provenance chains map to SLSA's Build Track levels. A training pipeline that produces signed provenance attestations at each stage achieves SLSA Build Level 2 or higher.
- **in-toto**: Layer 3 training pipeline attestation **MAY** use in-toto layout/link attestation models for multi-step supply chain verification.
- **C2PA**: For AI-generated content, Layer 3 decision receipts **MAY** be embedded as C2PA Content Credentials, binding model provenance to output provenance.
- **CycloneDX / SPDX**: The AI-BOM specification **SHALL** extend established SBOM formats rather than defining an incompatible alternative.

4.4.6 Auburn Documents

Table 7: Layer 3: Provenance Binding Documents

Designation	Title	Status	Role
—	AI-BOM Specification	In Development	Normative format for AI bills of materials
—	Training Provenance Chain	In Development	Cryptographically linked training history records
—	Decision Receipt Specification	In Development	Forensic reconstruction artifacts for individual outputs
—	Contamination Detection Protocol	In Development	Scanning methodology and structured reporting for training data contamination
—	Model Lineage Specification	In Development	Derivation DAG for model version relationships

4.5 Composition Layer: MAI-1 + AGS-1 (The Narrow Waist)

Composition Layer Definition

The Composition Layer is the narrow waist of the hourglass. MAI-1 defines the canonical interface through which all lower-layer evidence is delivered as a single, verifiable attestation artifact. AGS-1 (this document) defines the architecture within which that interface operates. Together, they are the universal composition point of the Auburn Governance Stack. Every lower-layer document feeds evidence into this waist. Every upper-layer document consumes evidence from it. No document operates independently of the composition waist.

4.5.1 Produces

The Composition Layer produces two categories of output:

1. **MAI-1 Attestation Artifacts.** A single, cryptographically signed token containing all lower-layer evidence in a canonical format. The MAI-1 attestation artifact **SHALL** include:

   - Layer 1 platform attestation summary (TEE report hash, firmware version, thermal status, isolation tier)
   - Layer 2 invariant measurements (all five mandatory invariants, plus any extended invariants reported by the system)
   - Layer 3 provenance binding (AI-BOM hash, training provenance chain head, model version identifier)
   - Freshness evidence (timestamp, nonce, or epoch counter establishing measurement recency)
   - Composition metadata (AGS version, MAI-1 profile version, conformance level claimed)

   The token format **SHALL** be CBOR-encoded, signed using COSE (RFC 9052), and structured as an EAT (RFC 9711) profile.

2. **CRSA-1 Composition Certificates.** For multi-model systems (pipelines, ensembles, agent networks), CRSA-1 (Clause AI-9) defines how individual MAI-1 attestation artifacts

compose into a system-level safety certificate. The composition algebra specifies how invariant measurements propagate across model boundaries under sequential, parallel, and recursive composition patterns.

4.5.2 Consumes

The Composition Layer consumes all evidence from Layers 1, 2, and 3. It also consumes the Cryptographic Binding Specification (which defines how inter-layer evidence is hash-chained into a single unforgeable artifact) and the Versioning Policy (which defines how attestation token formats evolve without breaking backward compatibility).

Composition Rule

The MAI-1 composition waist enforces a critical architectural property: every document in Layers 1–3 **SHALL** produce evidence in a format that MAI-1 can consume. Every document in the Enforcement and Application layers **SHALL** consume evidence exclusively through the MAI-1 interface. No upper-layer document **SHALL** bypass the composition waist to access lower-layer evidence directly.
This rule is the architectural decision that transforms a document series into an infrastructure specification. Violation of this rule—a sector profile that directly reads hardware attestation reports without going through MAI-1—breaks the composition property and is non-conformant with AGS-1.

4.5.3 Guarantees

The Composition Layer guarantees that all lower-layer evidence is bound into a single artifact with a single cryptographic verification path. A Relying Party that verifies the MAI-1 attestation token has verified the platform, health, and provenance evidence simultaneously. There is no need to verify each layer independently—the composition waist handles integration.

For multi-model systems, CRSA-1 guarantees that the composition certificate reflects the safety state of the composed system, not merely the weakest individual component. The composition algebra defines precisely how invariant bounds propagate under composition.

4.5.4 Cannot Guarantee

Honest Framing

The Composition Layer cannot guarantee that the evidence it binds is correct—only that it is internally consistent and cryptographically bound. If Layer 1 produces a fraudulent TEE attestation (because the TEE itself is compromised), the MAI-1 token will faithfully include that fraudulent evidence. The composition waist is a *binding layer*, not a *validation layer*. Validation is the Enforcement Layer's responsibility.
For multi-model systems, CRSA-1 composition is conservative: composed guarantees are weaker than individual guarantees. A pipeline of two models, each individually healthy, may exhibit emergent behavior that neither model's invariants capture. CRSA-1 defines the composition algebra but cannot eliminate compositional risk. Section 10 addresses this limitation in detail.

4.5.5 Standards Composition

- **IETF RATS** (RFC 9334, RFC 9711): MAI-1 is designed as a RATS profile. The MAI-1 attestation artifact is an EAT token with AI-governance-specific claims. The MAI-1

endpoint implements the RATS Attester role. Verifiers evaluate MAI-1 tokens using the RATS appraisal model.

- **COSE** (RFC 9052): All MAI-1 token signing uses COSE Sign1 or COSE Sign envelopes.
- **CBOR** (RFC 8949) / **CDDL** (RFC 8610): Token encoding and schema validation.
- **AIGA** (draft-aylward-aiga): MAI-1 is a strict superset of the AIGA attestation profile. Systems conformant to MAI-1 are conformant to AIGA; the reverse is not necessarily true.

4.5.6 Auburn Documents

Table 8: Composition Layer Documents

Clause	Title	Status	Role
AI-5	MAI-1: Model Attestation Interface	Published	Canonical attestation interface; the narrow waist
—	AGS-1 (this document)	This Document	Architectural specification; names the stack
AI-9	CRSA-1: Compositional Runtime Safety Attestation	Published	Multi-model composition algebra
—	Cryptographic Binding Specification	In Development	Inter-layer hash-chain and Merkle binding mechanisms

4.6 Enforcement Layer: Conformance and Testing

Enforcement Layer Definition

The Enforcement Layer contains the documents that make compliance checkable by strangers. It defines binary pass/fail rules, reference test vectors, known bad states, conformance level profiles, freshness requirements, adversarial testing specifications, and non-conformance consequences. These documents transform the attestation architecture into an enforceable standard. Once published, no organization can quietly diverge.

4.6.1 Produces

The Enforcement Layer produces conformance verdicts: determinations of whether a system's MAI-1 attestation artifact satisfies the requirements for a specific conformance level. MAI-1 defines three conformance levels:

Table 9: MAI-1 Conformance Levels

Level	Name	Requirements
MAI-C0	Structural Conformance	Valid token format, correct CDDL schema, required fields present, valid COSE signature. No evaluation of invariant *values*—only structural correctness.
MAI-C1	Measurement Conformance	MAI-C0 plus all five mandatory invariant measurements present, within certified bounds, bound to valid Layer 1 attestation, and meeting freshness requirements.
MAI-C2	Full Conformance	MAI-C1 plus Layer 3 provenance binding present and verifiable, extended invariants (if claimed) within bounds, adversarial test suite passed, and known bad state checks cleared.

4.6.2 Consumes

The Enforcement Layer consumes MAI-1 attestation artifacts exclusively through the composition waist. It **MUST NOT** access lower-layer evidence directly. This constraint ensures that the Enforcement Layer tests the *integrated attestation*, not individual components in isolation.

4.6.3 Guarantees

The Enforcement Layer guarantees that conformance determination is deterministic, reproducible, and binary. Two independent evaluators applying CTS-1 test assertions to the same MAI-1 attestation artifact **SHALL** reach the same verdict. There is no evaluator discretion in the pass/fail determination.

4.6.4 Cannot Guarantee

Honest Framing

The Enforcement Layer cannot guarantee that conformance implies safety, correctness, or fitness for purpose. A system that achieves MAI-C2 Full Conformance has demonstrated structural correctness, measurement compliance, provenance binding, adversarial resilience, and clean known-bad-state checks. It has not demonstrated that its outputs are accurate, fair, unbiased, or appropriate for any specific use case.

Conformance testing is necessary infrastructure for accountability. It is not sufficient for trust. The distinction matters: an organization that achieves MAI-C2 can demonstrate *due diligence*—that it deployed and maintained a certified governance architecture. It cannot demonstrate *infallibility*.

4.6.5 Standards Composition

- **Common Criteria** (ISO/IEC 15408): The CTS-1 conformance methodology draws on Common Criteria evaluation assurance levels for structured testing rigor, adapted to the AI attestation domain.
- **EU AI Act Article 43**: CTS-1 conformance levels are designed to produce evidence usable in Article 43 conformity assessment. A system at MAI-C2 provides technical documentation evidence for Articles 9, 12, 15, and 17 simultaneously.
- **DORA Article 24**: CTS-1 adversarial testing requirements align with DORA's threat-led penetration testing provisions for financial sector AI systems.

4.6.6 Auburn Documents

Table 10: Enforcement Layer Documents

Designation	Title	Status	Role
CTS-1	MAI-1 Conformance Test Suite	Published	Binary pass/fail test assertions for all three conformance levels
—	Reference Test Vectors	In Development	Known-good and known-bad attestation artifacts for verifier validation
—	Known Bad States Registry	In Development	Catalog of attestation artifacts representing specific failure modes
—	Conformance Level Profiles	In Development	Detailed requirements per level with rationale
—	Freshness Rules Specification	In Development	Maximum attestation age by context and risk tier
—	Adversarial Testing Requirements	In Development	Mandatory adversarial evaluation for MAI-C2
—	Non-Conformance Consequences	In Development	Graduated response framework for conformance failures

4.7 Application Layer: Sector-Specific Compliance Profiles

Application Layer Definition

The Application Layer provides sector-specific compliance profiles that map MAI-1 attestation artifacts and CTS-1 conformance results to the requirements of specific regulatory regimes, insurance underwriting criteria, and procurement eligibility standards. Each profile is designed to be dropped directly into RFPs, conformity assessment reports, or underwriting questionnaires without modification.

4.7.1 Produces

Application Layer documents produce compliance mappings: structured tables showing how specific MAI-1 attestation fields, Layer 2 invariant measurements, and CTS-1 conformance levels satisfy specific articles, clauses, or requirements in a target regulatory framework. Each profile includes:

- A regulatory requirement inventory for the target framework.
- An article-by-article (or clause-by-clause) mapping from regulatory requirement to AGS evidence.
- Gap analysis identifying regulatory requirements that AGS evidence partially satisfies or does not address.
- Sample procurement language, conformity assessment report excerpts, or underwriting questionnaire responses.
- A risk tier classification specifying which MAI-1 conformance level (C0, C1, C2) is appropriate for which risk categories within the target framework.

4.7.2 Consumes

Application Layer documents consume Enforcement Layer outputs exclusively. A sector profile **MUST NOT** reference individual invariant measurements or platform attestation details directly; it **SHALL** reference conformance levels and attestation artifact fields as defined by MAI-1 and verified by CTS-1.

Composition Rule

Application Layer profiles **SHALL** be updated independently of the lower layers. When a new regulation is enacted or an existing regulation is amended, a new or revised Application Layer profile is created. No changes to Layers 0–3, the Composition Layer, or the Enforcement Layer are required. This independence is the architectural property that enables the stack to serve multiple regulatory jurisdictions simultaneously without cross-contamination.

4.7.3 Guarantees

Application Layer profiles guarantee a structured, auditable mapping between AGS conformance evidence and specific regulatory requirements. They guarantee that the mapping is complete (every applicable regulatory requirement is addressed) and honest (requirements that AGS cannot satisfy are explicitly identified in the gap analysis).

4.7.4 Cannot Guarantee

Honest Framing

Application Layer profiles cannot guarantee regulatory acceptance. A compliance profile that maps MAI-1 attestation to EU AI Act Article 15 is an argument that the attestation evidence satisfies the regulatory requirement. It is not a legal determination. Regulatory acceptance depends on the decisions of notified bodies, market surveillance authorities, and courts. The profiles are designed to make the compliance argument as strong, structured, and auditable as possible, but the final determination is not within the stack's control.

4.7.5 Standards Composition

Application Layer profiles compose directly with their target regulatory frameworks:

- **EU AI Act**: Profile maps attestation evidence to Articles 9, 10, 12, 14, 15, 17, and Annex IV technical documentation requirements, plus Article 43 conformity assessment evidence.
- **DORA** (EU 2022/2554): Profile maps to ICT risk management (Chapter II), incident reporting (Chapter III), testing (Chapter IV), and third-party oversight (Chapter V) for AI-as-ICT systems.
- **FDA SaMD**: Profile maps to PCCP requirements, TPLC management, and QMSR alignment for AI-enabled medical device governance.
- **Federal AI / OMB**: Profile maps to M-25-21 risk management requirements, NIST AI RMF alignment, and FedRAMP-adjacent attestation for federal AI procurement.
- **Defense / CMMC**: Profile maps to CMMC Level 2/3 requirements and FY2026 NDAA Section 1513 AI security framework provisions for defense contractor AI systems.

- **Insurance Underwriting**: Profile provides evidence artifacts for AI risk assessment in underwriting questionnaires, following Armilla AI and Munich Re aiSure evaluation patterns.
- **Enterprise VRM**: Profile provides vendor risk management artifacts for enterprise procurement of AI systems.

4.7.6 Auburn Documents

Table 11: Application Layer Documents

Designation	Title	Status	Role
CRSA-1 EU Edition	EU AI Act Compliance Profile	Published	EU AI Act article-by-article mapping
—	Autonomous AI Agents in Regulated Financial Services	Published	DORA + AI Act dual-compliance for agentic systems
—	Federal AI Compliance Profile	In Development	OMB/NIST/FedRAMP mapping
—	FDA SaMD Compliance Profile	In Development	PCCP/TPLC/QMSR mapping
—	Defense / CMMC Profile	In Development	CMMC + NDAA AI security mapping
—	Insurance Underwriting Package	In Development	Risk assessment evidence for AI underwriting
—	Financial Services Profile	In Development	SR 11-7 / MiFID II / DORA mapping
—	Enterprise VRM Profile	In Development	Vendor risk management artifacts

4.8 Bridge: Cross-Cutting Documents

Bridge Category Definition

Bridge documents connect the Auburn Governance Stack to the broader ecosystem. They do not belong to a single layer; they span multiple layers or address concerns that cut across the architecture. Bridge documents handle interoperability with external standards, open-source verification infrastructure, multi-model composition challenges, open-weight model attestation limitations, and regulatory deadline mapping.

Bridge documents include:

Table 12: Bridge / Cross-Cutting Documents

Title	Status	Role
IETF RATS Interoperability Profile (AIGA Alignment)	In Development	Formal mapping between AGS attestation artifacts and IETF RATS architecture; demonstrates that MAI-1 is a valid RATS profile and a strict superset of AIGA
Open-Source Verification Infrastructure	In Development	Specifications enabling open-source implementation of MAI-1 verifiers, leveraging Veraison and related projects
Open-Weight Model Attestation Limitations	In Development	Honest analysis of what attestation can and cannot guarantee for open-weight models where the deployer controls the full stack
Compound Threat Matrix	In Development	Cross-layer threat analysis identifying attack vectors that span multiple layers and require coordinated defenses
Regulatory Deadline Mapping	In Development	Continuously updated mapping of regulatory enforcement deadlines to AGS layer readiness, creating institutional urgency

Bridge documents are intentionally not assigned to a single layer. Their value lies precisely in crossing layer boundaries—identifying interoperability requirements, cross-cutting threats, and ecosystem integration points that layer-specific documents cannot address independently.

5 The Composition Principle

The architectural property that distinguishes the Auburn Governance Stack from a collection of independent specifications is the composition principle: every document in the stack either feeds evidence into the MAI-1 composition waist or consumes evidence from it. No document operates independently of this waist. This section formalizes that principle and its consequences.

5.1 The Formal Rule

Composition Rule

Future clauses in the Auburn Patent Family (AI-6 and beyond) **SHALL** assume MAI-1 conformant attestation as a prerequisite for their governance guarantees. Systems that do not expose an MAI-1 compliant endpoint are outside the scope of all downstream Auburn governance clauses.

This rule has three implications:

1. **Lower-layer documents produce for MAI-1.** Every Layer 1, 2, and 3 specification defines its evidence in terms that MAI-1 can consume. An invariant specification that produces measurements in a format incompatible with the MAI-1 token structure is non-conformant with the architecture, regardless of the scientific quality of the invariant itself.

2. **Upper-layer documents consume from MAI-1.** Every Enforcement and Application Layer document accesses evidence exclusively through the MAI-1 interface. A sector profile that directly reads Layer 1 hardware attestation reports—bypassing the composition waist—violates the architecture. This constraint is not bureaucratic; it ensures that upper-layer documents can operate without knowledge of lower-layer implementation details, enabling independent evolution.

3. **Systems without MAI-1 are out of scope.** The Auburn Governance Stack does not make governance claims about systems that lack an MAI-1 endpoint. This is not a limitation—it is a design boundary. The stack governs systems that participate in the attestation architecture. Systems that do not participate are governed by whatever other mechanisms their operators choose to deploy.

5.2 Composition for Single-Model Systems

For a single model deployed on a single platform, the composition principle is straightforward. The platform produces Layer 1 evidence. The model runtime produces Layer 2 measurements anchored in Layer 1. The model developer provides Layer 3 provenance. MAI-1 binds all three into a single attestation token. CTS-1 evaluates the token. A sector profile maps the result to regulatory requirements.

The composition is *vertical*: evidence flows upward through the layers, converging at the waist. No horizontal composition is required.

5.3 Composition for Multi-Model Systems

Multi-model systems—pipelines, ensembles, agent networks, recursive architectures—require horizontal composition in addition to vertical composition. CRSA-1 (Clause AI-9) defines the composition algebra for this case.

The fundamental challenge is that system-level properties do not follow trivially from component-level properties. Two models that individually maintain entropy above the certified

floor may, when composed in a pipeline, produce outputs with entropy below the floor (e.g., if the second model's conditioning on the first model's output restricts its output distribution). CRSA-1 addresses this through three mechanisms:

1. **Composition Operators.** CRSA-1 defines formal operators for sequential composition (pipeline), parallel composition (ensemble), conditional composition (routing/gating), and recursive composition (self-invocation). Each operator specifies how individual MAI-1 attestation artifacts combine into a system-level composition certificate.

2. **Conservative Propagation.** Composed guarantees are always weaker than or equal to individual guarantees. If Model A has entropy floor H_A and Model B has entropy floor H_B, the composed system's certified entropy floor under sequential composition is $\min(H_A, H_B)$ minus a composition penalty term derived from the interaction analysis. This conservatism ensures that composition certificates never overstate system-level guarantees.

3. **Composition Certificates.** The output of CRSA-1 composition is a composition certificate that references the individual MAI-1 attestation artifacts of each component, declares the composition topology, and states the composed invariant bounds. A Relying Party that verifies a CRSA-1 composition certificate has verified the system-level governance state.

Composition Rule

Multi-model systems that do not expose a CRSA-1 compliant composition certificate are outside the scope of system-level Auburn governance guarantees. Individual model attestation via MAI-1 remains valid for each component; system-level attestation requires CRSA-1.

5.4 Why Composition Matters

The composition principle is the design decision that transforms a paper series into an infrastructure specification. Without it, the Auburn documents are independent analyses—each valuable, but collectively no more than the sum of their parts. With it, the documents form a *system*: a coherent architecture where each component has a defined role, defined interfaces, and defined relationships to every other component.

The TCP/IP analogy is precise. IP (Internet Protocol) became the universal composition point of the internet not because it was the best protocol at any individual layer, but because it was the protocol that *everyone agreed to compose through*. Better physical layer technologies (ATM, Token Ring) failed to displace Ethernet not because they were inferior, but because they could not displace IP as the composition waist. MAI-1 serves the same architectural function: it is the interface that all lower-layer evidence composes through and all upper-layer applications consume from.

The consequence is that the value of the Auburn Governance Stack is superlinear in the number of components. Each new invariant specification added to Layer 2 increases the value of every Application Layer profile, because every profile automatically gains access to the new invariant's evidence through the MAI-1 interface. Each new sector profile added to the Application Layer increases the value of every lower-layer specification, because each specification now serves an additional regulatory audience. This network effect—mediated by the composition waist—is the architectural property that creates infrastructure value.

6 Standards Composition Map

The Auburn Governance Stack does not compete with existing standards. It composes with them. Each external standard addresses a specific layer or function; the AGS provides the architectural binding that connects them into an end-to-end verification system. This section maps the composition relationships.

6.1 Principle: Compose, Do Not Reinvent

The AGS follows the PCI-DSS adoption model rather than the ISO standardization model. PCI-DSS did not invent new cryptographic primitives; it specified how existing primitives (TLS, AES, RSA) **MUST** be configured and deployed for payment card security. Similarly, the AGS does not invent new attestation protocols; it specifies how existing protocols (RATS, EAT, COSE, SCITT) **MUST** be configured and deployed for AI governance.

This design choice has three benefits:

1. **Implementation leverage.** Organizations that have already deployed RATS-compatible attestation infrastructure can adopt the AGS without replacing their existing stack. They need only configure their attestation to produce MAI-1 conformant tokens.
2. **Standards body alignment.** The AGS can be presented to IETF RATS, CEN/CENELEC JTC 21, and ISO SC 42 as a *profile* of their work rather than a competitor. This is the difference between "we built something that replaces your standard" and "we built something that uses your standard in a specific, high-value domain."
3. **Reduced attack surface.** By using battle-tested cryptographic primitives (COSE, CBOR, Merkle trees) rather than inventing new ones, the AGS inherits the security analysis of those primitives. Novel cryptography is a risk; composition of proven cryptography is engineering.

6.2 Composition Map by External Standard

6.2.1 IETF RATS (RFC 9334, RFC 9711, CoRIM, AR4SI)

RATS provides the attestation substrate of the Auburn Governance Stack. The composition is deep and structural:

Table 13: IETF RATS Composition Map

RATS Component	AGS Usage	Relationship
RFC 9334 Architecture	AGS attestation topology	MAI-1 implements the Attester role; CTS-1 verifiers implement the Verifier role; sector profiles implement the Relying Party role
RFC 9711 EAT	MAI-1 token format	MAI-1 attestation artifacts are EAT tokens with AI-governance-specific claims defined as an EAT profile
CoRIM	Layer 1 reference values	Hardware vendor reference values for TEE attestation **SHOULD** be distributed as CoRIM manifests
AR4SI Trustworthiness Vector	Layer 2 health mapping	The eight AR4SI appraisal dimensions map to AGS layer evidence; "runtime-opaque" and "sourced-data" dimensions are closest to model health attestation
EAR	Attestation result format	CTS-1 conformance results **MAY** be encoded as EAT Attestation Results (EAR) tokens

The AGS is designed such that MAI-1 *is* a RATS profile. This is not a loose analogy; it is a normative design constraint. A MAI-1 attestation artifact **SHALL** be parseable by any RATS-aware verifier as a valid EAT token. The AI-governance-specific claims are custom claims within the EAT extensibility mechanism, not violations of the EAT structure.

This positioning matters for standards adoption. When the IETF RATS working group evaluates AI attestation proposals—and the appearance of draft-messous-eat-ai-00 and draft-huang-rats-agentic-eat-cap-attest-00 confirms this is already happening—MAI-1 can be presented as a mature, published, comprehensive RATS profile rather than a competing architecture.

6.2.2 SCITT and Transparency Infrastructure

SCITT (Supply Chain Integrity, Transparency, and Trust) provides the transparency log infrastructure for Layer 3 provenance:

- **AI-BOM Registration.** AI bills of materials **SHOULD** be registered as SCITT signed statements, providing append-only, cryptographically verifiable records of model composition declarations.
- **Training Provenance.** Training provenance chain records **SHOULD** be registered as sequential SCITT entries, with Merkle tree inclusion proofs enabling efficient verification of chain integrity.
- **Attestation History.** MAI-1 attestation artifacts themselves **MAY** be registered in SCITT transparency logs, providing a tamper-evident history of a system's attestation state over time.

The AGS does not require SCITT. Layer 3 provenance can be implemented with any transparency log that provides append-only semantics and cryptographic inclusion proofs. SCITT is the recommended infrastructure because it is the IETF's designated architecture for this function, ensuring maximum interoperability.

6.2.3 C2PA and Content Provenance

C2PA v2.3 provides content provenance—cryptographic binding between a piece of content and its creation history. The AGS and C2PA are complementary, not overlapping:

- **C2PA scope:** Proves that a specific output (image, text, audio) was generated by a specific tool, with a specific editing history, at a specific time. Answers: *Where did this content come from?*
- **AGS scope:** Proves that the AI system that generated the output was operating in a certified state—attested hardware, healthy model, intact provenance—at the time of generation. Answers: *Was the system that produced this content trustworthy?*

The composition point is the Decision Receipt (Layer 3). A decision receipt that is embedded as a C2PA Content Credential binds *system-level governance state* to *output-level provenance*. This creates an end-to-end chain: from hardware trust root, through model health, through provenance, through attestation, to the individual output—all cryptographically linked.

6.2.4 OpenSSF Model Signing, SLSA, and in-toto

The software supply chain integrity ecosystem provides building blocks for Layer 3:

- **OpenSSF Model Signing (OMS):** Model artifact signatures produced by OMS **SHALL** be accepted as valid Layer 3 model identity evidence. The AGS does not define a competing model signing format; it consumes OMS signatures as input to the AI-BOM.
- **SLSA v1.2:** Training pipeline provenance attestations that achieve SLSA Build Level 2 or higher satisfy Layer 3 training provenance requirements for the corresponding pipeline stages.
- **in-toto:** Training pipeline attestation **MAY** use in-toto layout/link models. The AGS Layer 3 training provenance chain is semantically compatible with in-toto's multi-step supply chain verification approach.

6.2.5 NIST AI RMF

The relationship between the AGS and NIST AI RMF is complementary rather than compositional at the protocol level:

- **Govern function:** AGS-1 (this document) and the Application Layer profiles provide the architectural framework and regulatory mappings that operationalize AI RMF's Govern function for organizations deploying the stack.
- **Map function:** The Layer 2 invariant specifications provide the specific, measurable risk indicators that operationalize AI RMF's Map function. Where AI RMF says "identify risks," Layer 2 says "monitor these five invariants against these certified bounds."
- **Measure function:** Layer 2 invariants, CTS-1 conformance testing, and the Enforcement Layer operationalize AI RMF's Measure function with quantitative, continuously monitored, binary-verdict metrics. This is the most direct composition point: the AGS provides the *technical verification architecture* that AI RMF's Measure function describes but does not specify.
- **Manage function:** The Enforcement Layer's non-conformance consequences and the Application Layer's sector-specific response protocols operationalize AI RMF's Manage function.

An organization that deploys the Auburn Governance Stack can demonstrate AI RMF alignment with specific, technical, verifiable evidence rather than process documentation alone.

6.2.6 ISO/IEC 42001

ISO 42001's management system requirements are satisfied by AGS deployment in the following areas:

- **Clause 6 (Planning):** AGS Layer 2 invariant specifications provide the risk treatment measures with quantitative thresholds that Clause 6 requires but does not specify.
- **Clause 8 (Operation):** MAI-1 attestation endpoints and CTS-1 conformance testing provide the operational controls that Clause 8 requires but does not prescribe technically.
- **Clause 9 (Performance Evaluation):** Layer 2 continuous monitoring provides the measurement and evaluation evidence that Clause 9 requires. Unlike process-level monitoring, AGS monitoring is continuous, quantitative, and cryptographically bound to hardware.
- **Clause 10 (Improvement):** The Enforcement Layer's non-conformance consequences and versioning policy provide the corrective action and continual improvement mechanisms that Clause 10 requires.

The AGS does not replace ISO 42001; it provides the technical substrate that makes 42001 compliance *verifiable* rather than merely *attestable by process documentation*.

6.2.7 EU AI Act

The EU AI Act composition is the most commercially significant mapping in the stack. Table 14 provides the article-level mapping.

Table 14: EU AI Act Article-to-AGS Evidence Mapping

AI Act Article	Requirement	AGS Evidence
Article 9	Risk management system	Layer 2 invariant monitoring provides continuous, quantitative risk measurement; Application Layer EU profile maps invariants to risk categories
Article 10	Data governance	Layer 3 AI-BOM and training provenance chain document data lineage, quality, and contamination scanning
Article 12	Record-keeping	Layer 3 decision receipts and SCITT transparency logs provide tamper-evident, cryptographically verifiable event records
Article 13	Transparency	MAI-1 attestation artifacts are machine-readable, structured, and designed for automated consumption by Relying Parties
Article 14	Human oversight	Application Layer profile specifies how attestation evidence integrates with human oversight mechanisms; CRSA-1 Agent Oversight Architecture for autonomous systems
Article 15	Accuracy, robustness, cybersecurity	Layer 2 invariants (entropy, gradient, drift, coherence) provide robustness evidence; Layer 1 TEE attestation provides cybersecurity evidence; CTS-1 adversarial testing provides resilience evidence
Article 17	Quality management	CTS-1 conformance levels provide auditable quality management evidence; prEN 18286 alignment through Application Layer profile
Article 43	Conformity assessment	CTS-1 conformance results at MAI-C1 or MAI-C2 provide technical documentation for notified body evaluation
Article 50	AI-generated content marking	Layer 3 decision receipts combined with C2PA Content Credentials provide machine-readable provenance for AI outputs

The EU AI Act mapping is detailed further in the CRSA-1 EU Edition and the forthcoming dedicated EU AI Act Compliance Profile. Table 14 provides the architectural-level mapping; the sector profiles provide implementation-level detail.

7 Dependency Structure

The Auburn Governance Stack is not a flat collection of documents. It is a directed graph with explicit dependency relationships. This section defines those relationships, identifies the critical path, and provides the full dependency graph.

7.1 Dependency Rules

Three rules govern dependencies within the stack:

1. **Downward dependencies only.** Documents **MAY** depend on documents in the same layer or in lower layers. Documents **MUST NOT** depend on documents in higher layers. This ensures that the evidence production chain (bottom-up) is never circular. A Layer 2 invariant specification may depend on a Layer 1 platform specification (to define hardware anchoring) but **MUST NOT** depend on an Application Layer profile (which consumes its output).

2. **All upper-layer documents depend on the composition waist.** Every Enforcement Layer and Application Layer document **SHALL** declare a dependency on MAI-1. This dependency is the structural enforcement of the composition principle: upper-layer documents access lower-layer evidence exclusively through the MAI-1 interface.

3. **Cross-layer dependencies are mediated by Bridge documents.** When a document in one layer requires awareness of a document in a non-adjacent layer (e.g., an Application Layer profile that needs to reference specific Layer 1 TEE capabilities for a defense procurement context), the dependency **SHOULD** be mediated through a Bridge document or through the composition waist rather than creating a direct cross-layer reference.

7.2 The Critical Path

The critical path through the dependency graph determines the minimum set of documents required for the stack to function as an end-to-end governance system:

$$\textbf{MSAF} \longrightarrow \textbf{MAI-1} \longrightarrow \textbf{CTS-1} \longrightarrow \textbf{Sector Profiles}$$

1. **MSAF** (Layer 0) provides the theoretical foundation: why attestation is necessary, the three-tier architecture, the impossibility bounds. Without MSAF, the stack lacks intellectual authority.

2. **MAI-1** (Composition Layer) defines the interface: the canonical endpoint, the token format, the mandatory invariants, the encoding rules. Without MAI-1, there is no composition waist and the stack is a disconnected collection.

3. **CTS-1** (Enforcement Layer) defines the tests: binary pass/fail assertions, conformance levels, structural and measurement validation. Without CTS-1, conformance is not checkable by strangers.

4. **Sector Profiles** (Application Layer) deliver the value: regulatory mappings, procurement language, conformity assessment evidence. Without sector profiles, the stack is technically complete but commercially inert.

Every other document in the stack either strengthens this critical path (Layer 1–3 documents that feed richer evidence into MAI-1) or extends it (additional sector profiles, enforcement refinements, bridge interoperability). The critical path documents are all published.

7.3 Layer-by-Layer Dependency Summary

Table 15 summarizes the primary dependency relationships by layer.

Table 15: Dependency Structure by Layer

Layer	Depends On	Feeds Into
Layer 0 (Foundation)	External literature and standards	All layers (intellectual authority)
Layer 1 (Platform)	Layer 0 (theoretical basis); hardware vendor specifications	Layer 2 (hardware anchoring for invariant measurements); Composition Layer (platform evidence in MAI-1 token)
Layer 2 (Invariants)	Layer 0 (theoretical basis); Layer 1 (hardware anchoring); Layer 3 (model version binding)	Composition Layer (health evidence in MAI-1 token)
Layer 3 (Provenance)	Layer 0 (theoretical basis); supply chain standards (SCITT, SLSA, OMS)	Layer 2 (model version binding); Composition Layer (provenance evidence in MAI-1 token)
Composition (MAI-1 + AGS-1)	Layers 1, 2, 3 (all evidence); IETF RATS (token format)	Enforcement Layer; Application Layer
Enforcement (CTS-1)	Composition Layer (MAI-1 artifacts); Common Criteria methodology	Application Layer (conformance verdicts)
Application (Profiles)	Enforcement Layer (conformance results); target regulatory frameworks	End users (compliance evidence)
Bridge	Multiple layers as applicable	Multiple layers; external ecosystem

7.4 Dependency Graph

Figure 2 presents the full dependency graph. The critical path is highlighted in red. Standard dependencies are shown in blue. Bridge connections are shown in dashed gray.

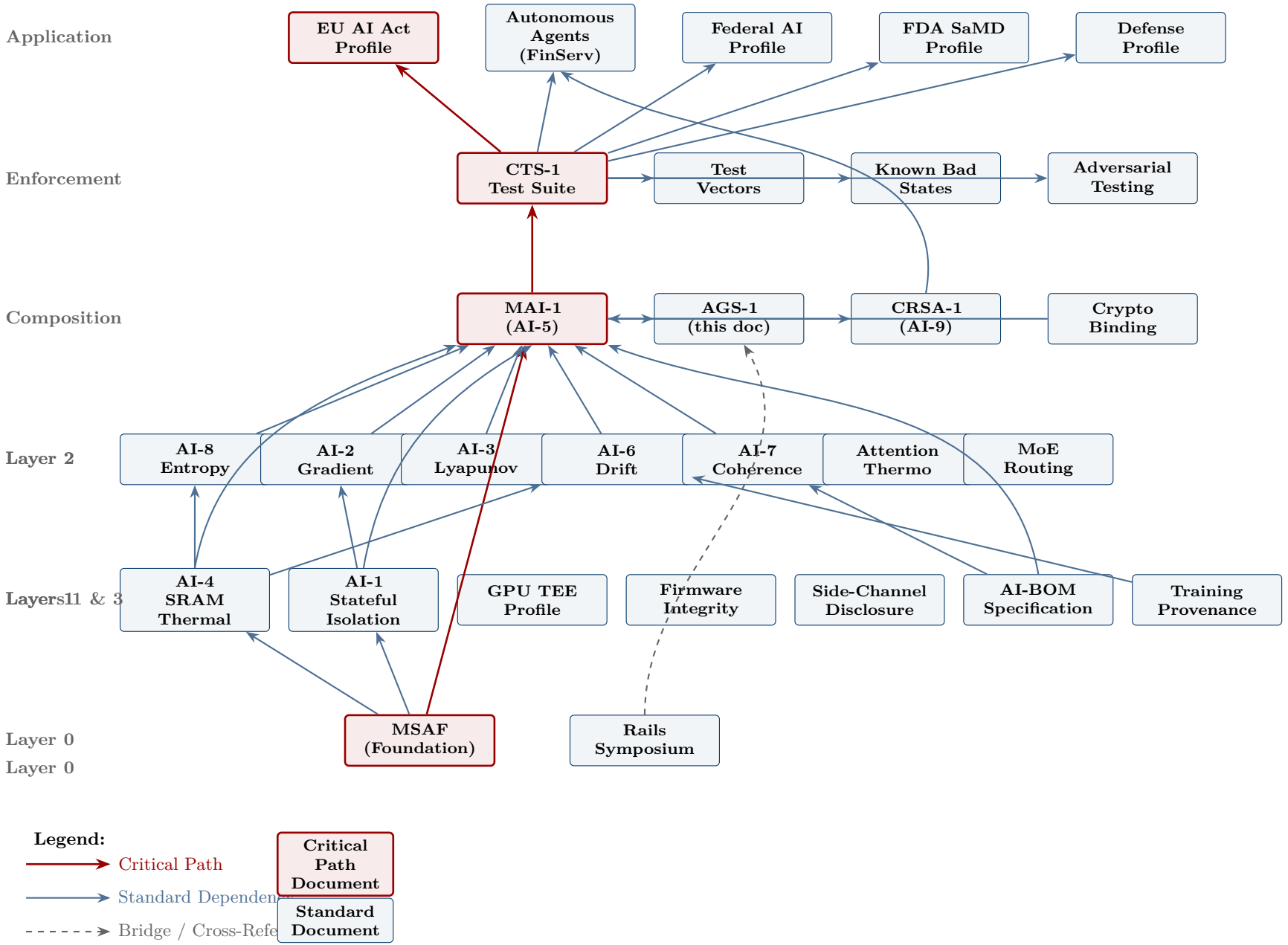

Figure 2: Auburn Governance Stack: Full Dependency Graph. Red arrows and nodes indicate the critical path (MSAF → MAI-1 → CTS-1 → Sector Profiles). Blue arrows indicate standard dependencies. Dashed gray arrows indicate bridge and cross-reference relationships. Layers 1 and 3 are shown at the same vertical level to reflect their parallel evidence production.

7.5 Modularity and Extension

The dependency structure is designed for modular extension. Adding a new document to the stack requires only:

1. **Identify the layer.** Determine which layer the new document belongs to based on what it produces and what it consumes.
2. **Declare dependencies.** Identify which existing documents the new document depends on (downward or same-layer only).
3. **Declare interface conformance.** If the new document produces evidence consumed by MAI-1 (Layers 1–3), it **MUST** produce that evidence in a format compatible with the MAI-1 token structure. If the new document consumes evidence from MAI-1 (Enforcement or Application layers), it **MUST** consume through the MAI-1 interface.
4. **Register in AGS-1.** The new document is added to the document registry (Section 8) with its layer assignment, clause designation, dependencies, and status.

No existing documents require modification when a new document is added, provided the new document conforms to the dependency rules. This is the modularity property that enables the stack to grow without architectural revision.

The one exception is AGS-1 itself: this architectural specification **SHALL** be updated to reflect new documents in the registry. However, the architecture—the layers, the composition waist, the dependency rules—does not change. Only the registry contents change. This is the distinction between architectural revision (changing the layers or composition rules) and registry update (adding a new document to an existing layer).

8 Document Registry

This section provides the authoritative registry of every document in the Auburn Governance Stack. Each entry records the document's designation, title, Auburn clause number (if applicable), layer assignment, current status, and role within the architecture.

The registry is the canonical reference for the stack's composition. When a new document is added to the Auburn Patent Family, it is registered here with its layer assignment and dependency declarations. The registry is maintained as part of AGS-1 and is updated with each AGS-1 version.

8.1 Registry Format

Each entry follows a standardized format:

- **Designation:** The Auburn clause number (AI-n) or document identifier.
- **Title:** The full document name.
- **Layer:** Position within the hourglass architecture.
- **Status:** *Published* (available on Figshare under CC BY-NC-ND 4.0), *In Development* (specification in progress), or *This Document*.
- **MAI-1 Role:** The function this document serves within the MAI-1 attestation interface—which invariant it defines, which evidence it produces, or which consumption pattern it implements.
- **Figshare DOI:** Digital Object Identifier for published documents, enabling permanent citation.

8.2 Layer 0: Foundation

Table 16: Document Registry — Layer 0: Foundation

Clause	Title	Status	MAI-1 Role
—	The Model State Attestation Framework (MSAF)	Published	Theoretical foundation: three-tier attestation architecture, impossibility bounds, regulatory landscape survey. Establishes why the remaining layers are necessary.
—	Rails Symposium	Published	Accessible capstone tutorial making the architecture legible to policy audiences. Depends on the architecture it describes; designed to be read last but published for independent discoverability.

8.3 Layer 1: Platform Attestation

Table 17: Document Registry — Layer 1: Platform Attestation

Clause	Title	Status	MAI-1 Role
AI-4	SRAM Thermal Integrity Bound	Published	Mandatory attestation: defines the temperature envelope within which silicon computations remain reliable. Produces thermal integrity evidence for the MAI-1 token.
AI-1	Stateful Isolation Law	Published	Mandatory attestation: defines the contamination functional, six isolation clauses, and composability theorems for multi-tenant GPU environments. Produces isolation tier evidence for the MAI-1 token.
—	GPU TEE Profile	In Dev.	Vendor-specific attestation profiles for Intel TDX, ARM CCA, AMD SEV-SNP. Produces TEE attestation report for the MAI-1 token.
—	Firmware Integrity Specification	In Dev.	Measured boot chain requirements following the DICE model. Produces firmware measurement chain for the MAI-1 token.
—	Side-Channel Disclosure Protocol	In Dev.	Structured honest reporting of known hardware vulnerabilities. Produces disclosure evidence for the MAI-1 token.

8.4 Layer 2: Model State Invariants

Table 18: Document Registry — Layer 2: Model State Invariants

Clause	Title	Status	MAI-1 Role
AI-8	Entropy Collapse Constraint	Published	Mandatory invariant: output distribution entropy floor. Detects degenerate generation modes. Produces entropy measurement for the MAI-1 token.
AI-2	Gradient Starvation Envelope	Published	Mandatory invariant: layer-wise gradient flow stability. Detects training instability and attention head starvation. Produces gradient norm evidence for the MAI-1 token.
AI-3	Lyapunov Stability for Speculative Decoding	Published	Mandatory invariant: speculative decoding divergence bound. Detects unbounded instability in draft-verify chains. Produces Lyapunov exponent evidence for the MAI-1 token.

Clause	Title	Status	MAI-1 Role
AI-6	Distribution Drift Bound	Published	Mandatory invariant: deployment-time distribution drift. Detects model behavior shifting outside certified operating distribution. Produces KL divergence evidence for the MAI-1 token.
AI-7	Structural Coherence Requirement	Published	Mandatory invariant: internal representation integrity. Detects architectural degradation where layer representations lose structural relationships. Produces coherence metric for the MAI-1 token.
—	Attention Thermodynamics	Published	Extended invariant: attention entropy across layers and heads. Detects attention collapse and attention diffusion.
—	MoE Routing Attestation	Published	Extended invariant: expert selection distribution health for Mixture-of-Experts architectures. Detects routing collapse and load imbalance.
—	Inference Latency Bound	In Dev.	Extended invariant: computational anomaly detection via latency distribution monitoring.
—	Quantization Integrity	In Dev.	Extended invariant: precision reduction error bounds for quantized deployments.

8.5 Layer 3: Provenance Binding

Table 19: Document Registry — Layer 3: Provenance Binding

Clause	Title	Status	MAI-1 Role
—	AI-BOM Specification	In Dev.	Normative format for AI bills of materials. Extends CycloneDX/SPDX for AI-specific components. Produces model composition manifest for the MAI-1 token.
—	Training Provenance Chain	In Dev.	Cryptographically linked records of each training stage. Produces provenance chain head hash for the MAI-1 token.
—	Decision Receipt Specification	In Dev.	Forensic reconstruction artifacts binding specific inputs, model versions, platform attestation, and health measurements. Enables post-hoc investigation.
—	Contamination Detection Protocol	In Dev.	Scanning methodology and structured reporting for training data contamination: benchmark leakage, PII, copyrighted material.

Clause	Title	Status	MAI-1 Role
—	Model Lineage Specification	In Dev.	Directed acyclic graph documenting derivation relationships between model versions. Enables ancestor-based trust assessment.

8.6 Composition Layer

Table 20: Document Registry — Composition Layer

Clause	Title	Status	MAI-1 Role
AI-5	MAI-1: Model Attestation Interface	Published	The composition waist. Defines the canonical endpoint, token format, mandatory invariants, encoding rules, RATS profile alignment, and conformance level structure.
—	AGS-1: Auburn Governance Stack Architecture	This Doc.	Architectural meta-document. Names the stack, defines the layers, declares the dependency structure, specifies composition rules.
AI-9	CRSA-1: Compositional Runtime Safety Attestation	Published	Multi-model composition algebra. Defines how individual MAI-1 attestation artifacts compose under sequential, parallel, conditional, and recursive patterns.
—	Cryptographic Binding Specification	In Dev.	Inter-layer hash-chain and Merkle tree binding mechanisms. Specifies how Layer 1, 2, and 3 evidence is bound into a single unforgeable MAI-1 artifact.

8.7 Enforcement Layer

Table 21: Document Registry — Enforcement Layer

Desig.	Title	Status	MAI-1 Role
CTS-1	MAI-1 Conformance Test Suite	Published	Binary pass/fail test assertions for structural (C0), measurement (C1), and full (C2) conformance. The document that makes compliance checkable by strangers.
—	Reference Test Vectors	In Dev.	Known-good and known-bad attestation artifacts for verifier validation and interoperability testing.

Desig.	Title	Status	MAI-1 Role
—	Known Bad States Registry	In Dev.	Catalog of attestation artifact patterns representing specific failure modes, enabling pattern-matching detection.
—	Conformance Level Profiles	In Dev.	Detailed requirements, rationale, and implementation guidance for each MAI-1 conformance level.
—	Freshness Rules Specification	In Dev.	Maximum attestation age requirements by deployment context and risk tier. Prevents stale attestation reuse.
—	Adversarial Testing Requirements	In Dev.	Mandatory adversarial evaluation methodology for MAI-C2 Full Conformance. Aligned with DORA threat-led penetration testing.
—	Non-Conformance Consequences	In Dev.	Graduated response framework: notification, remediation period, conformance suspension, and public disclosure for persistent non-conformance.

8.8 Application Layer

Table 22: Document Registry — Application Layer

Desig.	Title	Status	MAI-1 Role
—	CRSA-1 EU Edition: EU AI Act Compliance Profile	Published	Article-by-article mapping of AGS evidence to EU AI Act requirements. Conformity assessment evidence for notified bodies.
—	Autonomous AI Agents in Regulated Financial Services	Published	DORA + AI Act dual-compliance governance framework for agentic systems. Three autonomy tiers, six agentic risk categories, four architecture-specific profiles.
—	Federal AI Compliance Profile	In Dev.	OMB M-25-21/M-25-22 mapping, NIST AI RMF alignment evidence, FedRAMP-adjacent attestation for federal AI procurement.
—	FDA SaMD Compliance Profile	In Dev.	PCCP, TPLC, QMSR mapping for AI-enabled medical device governance. Continuous monitoring evidence for post-market surveillance.
—	Defense / CMMC Profile	In Dev.	CMMC Level 2/3 requirements and FY2026 NDAA Section 1513 AI security framework mapping for defense contractor AI systems.

Desig.	Title	Status	MAI-1 Role
—	Insurance Underwriting Package	In Dev.	Risk assessment evidence artifacts for AI underwriting questionnaires. Aligned with Armilla AI and Munich Re aiSure evaluation patterns.
—	Financial Services Profile	In Dev.	SR 11-7 model risk management, MiFID II algorithmic trading, DORA ICT risk management mapping for financial institution AI systems.
—	Enterprise VRM Profile	In Dev.	Vendor risk management artifacts for enterprise procurement of AI systems. Designed for integration into third-party risk assessment workflows.

8.9 Bridge / Cross-Cutting

Table 23: Document Registry — Bridge / Cross-Cutting

Desig.	Title	Status	Role
—	IETF RATS Interoperability Profile	In Dev.	Formal demonstration that MAI-1 is a valid RATS profile and a strict superset of AIGA. Enables standards body engagement.
—	Open-Source Verification Infrastructure	In Dev.	Specifications for open-source MAI-1 verifier implementation. Leverages Veraison and related attestation verification projects.
—	Open-Weight Model Attestation Limitations	In Dev.	Honest analysis of attestation boundaries for open-weight models where the deployer controls the full stack.
—	Compound Threat Matrix	In Dev.	Cross-layer threat analysis identifying attack vectors spanning multiple layers requiring coordinated defense.
—	Regulatory Deadline Mapping	In Dev.	Continuously updated mapping of enforcement deadlines to AGS layer readiness. Creates institutional urgency for adoption.

8.10 Registry Summary

Table 24: Auburn Governance Stack: Registry Summary by Layer

Layer	Total Documents	Published	In Development
Layer 0: Foundation	2	2	0
Layer 1: Platform Attestation	5	2	3
Layer 2: Model State Invariants	9	7	2
Layer 3: Provenance Binding	5	0	5
Composition Layer	4	3	1
Enforcement Layer	7	1	6
Application Layer	8	2	6
Bridge / Cross-Cutting	5	0	5
Total	**45**	**17**	**28**

Of the 17 published documents, all are available on Figshare under CC BY-NC-ND 4.0 with Auburn Patent Family IP declarations. The published documents include the complete critical path (MSAF, MAI-1, CTS-1, and two Application Layer profiles), all five mandatory Layer 2 invariants, both Layer 1 platform specifications, both extended Layer 2 specifications (Attention Thermodynamics and MoE Routing), and the multi-model composition protocol (CRSA-1). The remaining 28 documents extend, operationalize, and apply the architecture defined by the published core.

9 Regulatory Timeline Mapping

The Auburn Governance Stack is not an academic exercise. It is infrastructure designed to be operational before the regulatory enforcement windows that create demand for it. This section maps the major enforcement deadlines to AGS layer readiness, establishing the temporal argument for adoption.

9.1 The Enforcement Landscape

As of March 2026, the following regulatory enforcement deadlines are active or imminent:

Table 25: Regulatory Enforcement Timeline (March 2026)

Regulation	Enforcement Date	Status	Scope
EU AI Act — Prohibited practices	February 2, 2025	**Active**	Unacceptable-risk AI systems banned
DORA (EU 2022/2554)	January 17, 2025	**Active**	ICT risk management for financial entities, including AI-as-ICT
PCI-DSS v4.0.1 future-dated requirements	March 31, 2025	**Active**	All 51 future-dated requirements now mandatory
FDA QMSR alignment	February 2, 2026	**Active**	21 CFR Part 820 aligned with ISO 13485 for medical devices including AI/ML SaMD
EU AI Act — High-risk (Annex III)	August 2, 2026[a]	**Imminent**	High-risk AI systems in Annex III categories must comply with Articles 8–15
DoD NDAA §1513 AI framework	June 16, 2026	**Imminent**	Status update to Congress on AI/ML cybersecurity and physical security framework
CMMC Phase 2	Mid-2026	**Imminent**	CMMC Level 2 certification required in select DoD solicitations
CEN/CENELEC harmonized standards target	Q4 2026	Projected	Full suite of EU AI Act harmonized standards
EU AI Act — High-risk (Annex I)	August 2, 2027[a]	Approaching	High-risk AI embedded in regulated products
CMMC Phase 4 (full)	November 2028	Future	Full mandatory CMMC implementation across DoD contracts

[a] The Digital Omnibus proposal (COM(2025) 836, November 2025) proposes extending Annex III to December 2, 2027 and Annex I to August 2, 2028. The proposal is under legislative negotiation and not yet adopted.

9.2 AGS Readiness Against the Timeline

The critical question for any organization evaluating the Auburn Governance Stack is whether the architecture is operational in time for the enforcement deadlines that affect them. Table 26 maps AGS layer readiness to the regulatory timeline.

Table 26: AGS Layer Readiness vs. Regulatory Enforcement

Regulatory Deadline	Required AGS Layers	Readiness
DORA (active)	Layers 1–2, Composition, Application (Financial Services)	Core published. Autonomous Agents in Financial Services profile published. Financial Services profile (SR 11-7, MiFID II) in development.
EU AI Act Annex III (Aug/Dec 2026–2027)	Full stack: Layers 1–3, Composition, Enforcement, Application (EU profile)	Critical path published (MSAF, MAI-1, CTS-1, CRSA-1 EU Edition). Provenance layer and enforcement extensions in development.
FDA SaMD / QMSR (active)	Layers 1–2, Composition, Application (FDA profile)	Core published. FDA SaMD profile in development. Layer 2 invariants provide continuous monitoring evidence aligned with PCCP requirements.
CMMC / NDAA (2026–2028)	Layers 1–2, Composition, Application (Defense profile)	Core published. Defense profile in development. Layer 1 TEE attestation directly relevant to CMMC cybersecurity requirements.
PCI-DSS analogy (adoption model)	N/A — analogical	PCI-DSS enforcement model validates the procurement-pressure adoption path. AGS adoption follows the same logic: cloud providers and model marketplaces requiring attestation creates de facto mandate.

9.3 The Standards Gap Window

A critical timing dynamic reinforces the AGS positioning. The CEN/CENELEC harmonized standards for the EU AI Act are behind schedule. The first standard to enter public enquiry (prEN 18286, October 2025) addresses quality management systems. The technically substantive standards—accuracy, robustness, cybersecurity, risk management—remain in drafting. The target of Q4 2026 for the full suite is optimistic given the October 2025 acceleration measures were adopted precisely because progress was too slow.

This creates a window: between now and the availability of finalized harmonized standards, organizations preparing for EU AI Act compliance have no published technical standard to implement against. The Auburn Governance Stack occupies this window. It is not a substitute for harmonized standards (which, once published, will have legal presumption of conformity under Article 40). It is the most comprehensive technical architecture available *now*, designed to be compatible with the harmonized standards when they arrive.

Organizations that adopt the AGS architecture during the standards gap window will be positioned to demonstrate conformity assessment readiness to notified bodies before harmonized standards are formally referenced in the Official Journal. Organizations that wait for harmonized standards will begin their implementation effort after the standards are published, with less time before enforcement deadlines.

Honest Framing

The Auburn Governance Stack is not a harmonized standard under the EU AI Act. It does not carry the legal presumption of conformity that harmonized standards provide under Article 40. Organizations that deploy AGS infrastructure do so as a technical compliance measure, not as a legal safe harbor. The Application Layer EU AI Act profile explicitly maps AGS evidence to AI Act articles but does not claim that AGS conformance constitutes regulatory compliance. That determination rests with notified bodies and market surveillance authorities.

The honest framing is this: AGS provides the strongest available technical evidence base for conformity assessment. Whether that evidence is accepted is a regulatory and legal determination, not a technical one.

10 Honest Framing: What the Auburn Governance Stack Cannot Guarantee

The design principles established in Section 3 require that every document in the Auburn Governance Stack maintain an honest framing of its capabilities and limitations. This section consolidates the limitations that apply to the architecture as a whole.

10.1 The Financial Auditing Analogy

The Auburn Governance Stack is analogous to a financial auditing framework. A financial audit certifies that:

- Financial statements conform to accounting standards (GAAP, IFRS).
- Internal controls are designed and operating effectively.
- The audit was conducted in accordance with auditing standards (PCAOB, ISA).

A financial audit does **not** certify that:

- The company will remain solvent.
- Management decisions are wise or ethical.
- Fraud will never occur.
- The company's products or services are good.

The Auburn Governance Stack provides the AI equivalent. A system that achieves MAI-C2 Full Conformance has certified that:

- Its hardware platform has been attested through a cryptographic chain rooted in silicon.
- Its model state invariants are within certified bounds at attestation time.
- Its provenance chain is intact and verifiable.
- Its conformance has been verified against binary pass/fail criteria by an independent test suite.

A system that achieves MAI-C2 Full Conformance has **not** certified that:

- Its outputs are accurate, fair, unbiased, or appropriate for any specific use case.
- It will never produce harmful, misleading, or dangerous content.
- Its behavior aligns with any specific ethical framework or value system.
- It is safe to deploy in any specific context without additional domain-specific evaluation.
- Its training data was ethically sourced, representative, or free of all contamination.

This distinction is not a weakness of the architecture. It is a design constraint that ensures intellectual honesty. Governance infrastructure that claims behavioral guarantees it cannot provide is worse than no governance infrastructure at all, because it creates false confidence.

10.2 Specific Architectural Limitations

10.2.1 Attestation Is Point-in-Time

MAI-1 attestation artifacts represent the system state at the moment of attestation. Between attestation events, the system state may change. Freshness rules (defined in the Enforcement Layer) mitigate this by requiring attestation recency appropriate to the deployment context, but they cannot eliminate the gap entirely. A system that was healthy one second ago may not be healthy now.

Continuous monitoring (Layer 2 invariants measured at inference time) reduces this gap to the measurement interval. But even continuous monitoring has a sampling frequency. Events that occur between samples are unobserved. The architecture provides the best available approximation of continuous verification; it does not provide true continuous verification.

10.2.2 Hardware Trust Is Foundational but Not Absolute

Layer 1 hardware attestation rests on trust in the hardware vendor's endorsements, the TEE implementation's correctness, and the physical security of the deployment environment. All three can be compromised:

- Hardware vendors may issue incorrect endorsements (due to bugs, supply chain compromise, or negligence).
- TEE implementations have historically contained vulnerabilities (Spectre, Meltdown, Foreshadow, and their descendants).
- Physical access to hardware enables attacks that remote attestation cannot detect (decapping, voltage glitching, electromagnetic side channels).

The Side-Channel Disclosure Protocol (Layer 1) requires honest reporting of known vulnerabilities. But "known" is the operative word. Unknown vulnerabilities remain, by definition, undisclosed.

10.2.3 Compositional Risk Is Not Eliminated

CRSA-1 defines the composition algebra for multi-model systems, but compositional risk is inherently harder to bound than individual-component risk. Two models that individually satisfy all invariants may, when composed, exhibit emergent behavior that neither model's invariants capture. CRSA-1's conservative propagation (composed bounds are always weaker than individual bounds) is a mitigation, not an elimination.

The fundamental challenge is that the space of possible multi-model interactions is combinatorially large. CRSA-1 addresses the structured composition patterns (sequential, parallel, conditional, recursive) that dominate current deployments. Novel composition patterns not covered by CRSA-1 operators are outside the scope of system-level attestation until the composition algebra is extended.

10.2.4 Provenance Attestation Is Process-Based

Layer 3 provenance attestation certifies that a process record exists and is cryptographically intact. It does not certify that the process record is truthful. A model developer who omits a training dataset from the AI-BOM produces an incomplete but cryptographically valid provenance chain. Layer 3 detects tampering with declared provenance; it does not detect omission of undeclared provenance.

This is a fundamental limitation shared with all supply chain integrity frameworks (SLSA, in-toto, SCITT). Cryptographic integrity proves that what was declared has not been modified. It does not prove that everything relevant was declared.

10.2.5 The Invariant Set Is Incomplete

The five mandatory invariants and four extended invariants in Layer 2 represent the current state of measurement science for model health. They are not exhaustive. Failure modes exist that no current invariant detects. As the field of AI safety measurement advances, new invariants will be identified and added to Layer 2. The architecture is designed for this evolution (Principle 1: standardize the protocol, let the content evolve), but at any given moment, the invariant set is necessarily incomplete.

10.3 The Value Proposition Despite Limitations

The limitations enumerated above are real. They are also shared, in various forms, by every governance and compliance framework in existence. Financial audits do not prevent fraud. Building codes do not prevent structural failure. Medical licensing does not prevent malpractice. Safety inspections do not prevent accidents.

What these frameworks provide—and what the Auburn Governance Stack provides for AI—is *accountability infrastructure*. They create a verifiable record of due diligence. They establish quantitative baselines against which degradation can be detected. They enable third-party verification without requiring trust in the operator's self-assessment. They make non-compliance visible.

The alternative to imperfect governance infrastructure is not perfect governance. It is no governance—the current status quo, where AI systems are deployed without runtime health monitoring, without hardware-anchored attestation, without verifiable provenance, and without binary conformance testing. The Auburn Governance Stack does not solve the AI governance problem. It provides the infrastructure within which the problem can be systematically addressed.

11 Versioning and Evolution

The Auburn Governance Stack is designed to evolve without breaking. This section defines the versioning policy that enables that evolution.

11.1 The Separation Principle

Versioning follows directly from Design Principle 1 (Section 3): standardize the protocol, let the content evolve. The stack contains two categories of specification with different change rates:

1. **Infrastructure specifications** change slowly. These include: the MAI-1 token format (CBOR/COSE/EAT structure), the canonical endpoint specification, the inter-layer binding mechanism (Cryptographic Binding Specification), the conformance level definitions (MAI-C0/C1/C2), the composition algebra operators (CRSA-1), and the architectural layer definitions (AGS-1). Infrastructure specifications **SHALL** maintain backward compatibility across minor versions. Breaking changes require a major version increment and a migration path.

2. **Science specifications** change rapidly. These include: invariant threshold values (entropy floors, gradient stability bounds, drift detection sensitivity), measurement algorithms (specific entropy estimators, divergence metrics, coherence measures), monitoring frequencies, and calibration data. Science specifications **MAY** change between minor versions without backward compatibility requirements, provided the changes are documented and the MAI-1 token format accommodates the new values without structural modification.

11.2 Version Numbering

Each document in the Auburn Governance Stack carries an independent version number following semantic versioning (MAJOR.MINOR.PATCH):

- **MAJOR** increments indicate breaking changes to the document's normative interface. For infrastructure specifications, this means changes that require existing implementations to modify their integration. For science specifications, this means changes to the invariant definition itself (not just threshold recalibration).

- **MINOR** increments indicate additions that do not break existing interfaces. New optional fields in the MAI-1 token, new extended invariants in Layer 2, new sector profiles in the Application Layer.

- **PATCH** increments indicate corrections, clarifications, and editorial improvements with no normative impact.

11.3 AGS-1 Versioning

AGS-1 is versioned independently of all other documents. The architecture—the layer definitions, composition rules, dependency rules, and design principles—changes only at AGS-1 major version increments. The document registry (Section 8) is updated at AGS-1 minor version increments as new documents are added to the stack.

The current document is AGS-1 Version 1.0. Future versions will reflect the addition of new documents, the resolution of in-development specifications, and any architectural refinements identified through implementation experience and community feedback.

11.4 Deprecation Policy

Documents in the Auburn Governance Stack **MAY** be deprecated when they are superseded by improved specifications. Deprecated documents **SHALL** remain in the registry with a "Deprecated" status and a reference to the superseding document. Deprecated documents **SHALL** remain valid for conformance purposes for a minimum of one year after deprecation, ensuring that deployed systems have time to migrate.

No document on the critical path (MSAF, MAI-1, CTS-1) **SHALL** be deprecated without a major version increment of AGS-1 and a minimum two-year migration period.

12 Conclusion

The Auburn Governance Stack is the first published layered architecture specification for verifiable AI compliance. It defines seven architectural layers plus a cross-cutting bridge category, a narrow composition waist (MAI-1 + AGS-1) through which all evidence flows, normative composition rules that bind the layers into a coherent system, and explicit dependency structures that enable modular extension without architectural revision.

The architecture addresses a systematic gap in the global AI governance ecosystem. As documented in Section 2, every major framework, regulation, and standard operates at a single layer—governance process, measurement science, or cryptographic attestation—with no specification connecting them into an end-to-end verification system. The AGS provides that connection by composing with existing standards (IETF RATS, SCITT, C2PA, SLSA, NIST AI RMF, ISO 42001) rather than replacing them, following the PCI-DSS adoption model of specifying how existing primitives must be configured for a high-value domain.

The stack is operational. The critical path documents—MSAF (theoretical foundation), MAI-1 (composition waist), CTS-1 (conformance testing), and two Application Layer profiles (CRSA-1 EU Edition and Autonomous AI Agents in Regulated Financial Services)—are published. All five mandatory Layer 2 invariants are published. Both Layer 1 platform specifications are published. The multi-model composition protocol (CRSA-1) is published. Seventeen of forty-five documents are available on Figshare under CC BY-NC-ND 4.0.

The regulatory enforcement windows documented in Section 9—EU AI Act high-risk obligations, DORA enforcement, FDA PCCP requirements, CMMC phased rollout, and the FY2026 NDAA AI security framework—create demand for the technical verification infrastructure that the AGS provides. The CEN/CENELEC harmonized standards gap creates a window in which the AGS is the most comprehensive technical architecture available for organizations preparing for conformity assessment.

The honest framing applies to this conclusion as it applies to every section of this document: the Auburn Governance Stack provides accountability infrastructure, not behavioral safety guarantees. It makes AI governance verifiable, auditable, and enforceable by third parties. It does not make AI systems safe. The distinction matters, and maintaining it is a structural commitment of the architecture.

Document Status: AGS-1 Version 1.0 — March 2026.

A Glossary

Table 27: Auburn Governance Stack Glossary

Term	Definition
AGS	Auburn Governance Stack. The full layered architecture specification for verifiable AI compliance, comprising documents across seven layers plus a Bridge category. Named and defined in AGS-1 (this document).
AGS-1	The architectural meta-document of the Auburn Governance Stack. Names the stack, defines the layers, declares the dependency structure, and specifies the composition rules. Analogous to Cerf & Kahn (1974) for TCP/IP.
AI-BOM	Artificial Intelligence Bill of Materials. The normative format for documenting foundation model components: training data, architecture, fine-tuning history, alignment record, and dependencies. Layer 3 specification.
Appraisal Policy	The rules a Verifier applies when evaluating attestation Evidence against Reference Values to produce an Attestation Result. In the AGS context, appraisal policies evaluate invariant measurements against certified thresholds.
Attester	RATS architecture role (RFC 9334). The entity that generates attestation Evidence from a Target Environment. In the AGS context, the AI inference server (TEE + model runtime) is the Attester.
Auburn Clause	A numbered specification in the Auburn Patent Family. Clauses AI-1 through AI-9 (and future extensions) define specific governance invariants, security properties, or architectural requirements.
Auburn Patent Family	The intellectual property portfolio containing all methods, logic structures, and certified constant registries associated with the Auburn Governance Stack. Sole property of Ryan Fields.
Certified Constant	A threshold value derived through formal mathematical analysis that defines the boundary between healthy and unhealthy model states. Part of the Auburn Patent Family IP.
Composition Certificate	A CRSA-1 output artifact that binds individual MAI-1 attestation tokens from multiple models into a system-level governance statement, reflecting the composed safety state.
Composition Waist	The narrow point in the hourglass architecture through which all lower-layer evidence flows and from which all upper-layer applications consume. Comprises MAI-1 (interface) and AGS-1 (architecture).
Conformance Level	One of three tiers of AGS compliance: MAI-C0 (structural), MAI-C1 (measurement), MAI-C2 (full). Defined by CTS-1 and evaluated through binary pass/fail test assertions.

Term	Definition
CRSA-1	Compositional Runtime Safety Attestation Protocol (Clause AI-9). Defines the composition algebra for multi-model systems: how individual MAI-1 attestation artifacts compose under sequential, parallel, conditional, and recursive patterns.
CTS-1	MAI-1 Conformance Test Suite. Defines binary pass/fail test assertions for all three conformance levels. The Enforcement Layer document that makes compliance checkable by strangers.
Decision Receipt	A Layer 3 artifact binding a specific input, the model version that processed it, the platform attestation active at processing time, and the health invariant measurements into a single signed record for forensic reconstruction.
EAT	Entity Attestation Token (RFC 9711). The IETF standard token format for attestation claims, using CWT/JWT with COSE/JOSE security envelopes. MAI-1 attestation artifacts are EAT tokens with AI-governance-specific claims.
Evidence	RATS architecture concept (RFC 9334). Cryptographically signed data produced by an Attester describing the state of a Target Environment. In the AGS context, Evidence includes hardware attestation reports, invariant measurements, and provenance records.
Hourglass Architecture	The structural design pattern in which lower layers produce evidence, upper layers consume evidence, and all evidence flows through a narrow composition waist. Used by TCP/IP, PCI-DSS, IEC 62443, and the AGS.
Invariant	A measurable property of a model's internal state that must remain within certified bounds for the model to be considered healthy. The five mandatory invariants are: entropy floor, gradient stability, Lyapunov stability, distribution drift, and structural coherence.
MAI-1	Model Attestation Interface (Clause AI-5). The canonical interface through which all lower-layer evidence is delivered as a single, cryptographically signed attestation artifact. The narrow waist of the hourglass.
MSAF	The Model State Attestation Framework. The Layer 0 foundational document establishing the three-tier attestation architecture, theoretical justification, and impossibility bounds.
Relying Party	RATS architecture role (RFC 9334). The entity that consumes Attestation Results to make authorization or trust decisions. In the AGS context, Relying Parties include regulators, auditors, insurers, and procurement officers.
Verifier	RATS architecture role (RFC 9334). The entity that evaluates Evidence against appraisal policies to produce Attestation Results. In the AGS context, CTS-1 verifiers evaluate MAI-1 tokens against conformance level requirements.