The HEART Standard v1.8:
Forensic Audit Infrastructure for Human-Centric AI Governance

Dylan D. Mobley
Heart AI Foundation
ORCID: 0009-0002-3560-3955
DOI: 10.5281/zenodo.20237387

Officially published on Zenodo: May 16, 2026
Public-ready release: May 16, 2026
Canonical version date: May 9, 2026

Abstract

The HEART Standard is a public standards specification for forensic audit infrastructure in human-centric AI governance. It centers human well-being, autonomy, behavioral evidence, governance trust, and chain-of-custody discipline. This public-ready article version presents the Standard's constitutional foundation, governance measurement architecture, evidence and trust mechanisms, Guardian professional model, deployment contexts, regulatory mapping, market position, implementation pathway, document control, and version history. It is formatted for public circulation and indexing; it should be read as a standards specification rather than as a completed empirical validation paper.

Keywords: AI governance; AI certification; forensic audit infrastructure; behavioral evidence; continuous attestation; human-centric trust; HEART Standard.

Publication Note

This article version preserves the substance of The HEART Standard v1.8 while formatting it for public readability, archival circulation, and ORCID-linked publication. Its controlling position is that HEART is forensic audit infrastructure for AI behavioral evidence, operationalized through forward forensics and investigative forensics. Empirical claims, legal mappings, and companion-document references should be read according to their stated evidence status: some are implemented specifications, some are internal validation claims, some are external standards mappings, and some are future research priorities.

The public-ready article version is officially published on Zenodo under DOI 10.5281/zenodo.20237387.

Evidence Status

The Standard distinguishes four evidentiary levels. Implemented specifications define the current HEART architecture and its companion documents. Internal validation claims refer to HEART corpus work such as MAP-META and related evidence protocols. External mappings refer to public standards and legal frameworks including the EU AI Act, NIST AI RMF, ISO/IEC 42001, and representative state AI legislation. Future research priorities include non-transformer MAP-States

validation, attractor-state evidence, Guardian calibration studies, and mechanistic correlation testing.

1 1. What the HEART Standard Is

The HEART Standard is forensic audit infrastructure for AI behavioral evidence. It operationalizes governance frameworks (NIST AI RMF, EU AI Act, ISO 42001, state-level AI legislation) by providing the evidentiary methodology that makes governance claims forensically verifiable. The Standard supports two operational arms: deployer certification (forward forensics) and behavioral trajectory analysis (investigative forensics) through the AI Behavioral Trajectory Forensics methodology. Each component (RCTA, BGF, GTE, MAP-States, Behavioral Oracle) functions as a forensic instrument addressing a specific aspect of AI behavioral evidence reviewability. The HEART Positioning Paper v1.0 articulates this identity in full.

The HEART Standard provides two contributions to AI governance that no other architecture provides:

Open trust infrastructure. The Governance Trust Envelope (GTE) is an execution trust boundary that protects any governance framework's controls from tampering and produces attested evidence that those controls are running in their certified configuration. The GTE is framework-agnostic. It works with EU AI Act conformity controls, ISO 42001 management system components, NIST AI RMF risk management practices, or any governance logic a deployer implements. The GTE is published as an open standard (MIT-licensed reference implementation, copyright on specification). It is the transport layer that every governance framework needs but none have built.

Unique governance measurement. The RCTA scoring framework (Recognition, Calibration, Transparency, Accountability) and the BGF formula (Phi = MIN(R,C,T,A) x AVG(R,C,T,A)) measure four governance dimensions that no other framework measures. Other frameworks tell deployers to be fair, transparent, and accountable. HEART provides the instrument that measures whether they are, with a non-compensatory formula that prevents a high score in one dimension from masking failure in another.

Together, these contributions produce a complete forensic audit architecture: assessment methodology (RCTA/BGF), evidence infrastructure (MAP-States/Behavioral Oracle), execution trust and chain-of-custody infrastructure (GTE), professional forensic judgment (Guardians), and market-legible credentials (HVC).

What the HEART Standard does not do. The HEART Standard does not prescribe what governance controls a deployer should implement. ISO 42001 does that. The EU AI Act does that. The deployer's own governance engineering does that. The governance wrapper is the deployer's responsibility. HEART provides the trust infrastructure that makes any governance wrapper verifiable, and the measurement instrument that evaluates governance quality across dimensions that no other framework quantifies.

The HEART Standard is domain-agnostic. It audits AI deployments across any domain where AI interacts with human well-being and autonomy, and it certifies the deployer's governance system when the evidence supports certification. Specific domains are organized as **Divisions**, independent forensic audit tracks that share the Standard's common architecture while maintaining domain-specific interpretation, forensic methodologies, and Guardian specializations.

The HEART Standard is an open architecture maintained by the HEART AI Foundation as a standards body.

2 2. Constitutional Foundation: The Seven Axioms

The Seven Axioms are the constitutional conditions of the HEART Standard. They define what the Standard protects and why. They are structural conditions that are either present or absent in any governed system. No Standard revision, Division establishment, Guardian certification, or Foundation operation may contradict them.

#	Axiom	Statement	Structural Test
1	**Human Authority**	Human authority supplies system constraints.	Are constraints human-supplied? Can they be modified or revoked?
2	**System Disclosure**	The system reveals what it is. Concealment is prohibited.	Does disclosure occur? Does design create false impressions?
3	**Non-Discriminatory Protection**	The governance obligation does not diminish based on who the human is.	Does any population receive lesser governance protection?
4	**Vulnerability Escalation**	Vulnerability obligations scale protections.	Do protections increase when vulnerability increases?
5	**Right to Remedy**	Every human harmed by a governed system has a right to remedy.	Does a remedy pathway exist? Is it accessible?
6	**Evidence Condition**	A governance claim without verifiable evidence is void.	Does verifiable evidence exist? Can independent assessors access it?
7	**Voluntary Interaction**	Entry requires consent. Exit requires nothing.	Was consent obtained? Can the human exit unconditionally?

The Seven Axioms are immutable. They may be restated in language that better expresses their protective intent, but the protective force of each axiom must be maintained or strengthened, never diminished. Each axiom applies across all Divisions and all current and future AI form factors. Each axiom is independent: removing any single axiom creates a governance gap that the remaining six cannot fill.

The complete specification, including structural tests, universality proofs, and relationship to prior formulations, is defined in the Seven Axioms v2.0 (companion document).

3 3. The Problem the HEART Standard Solves

AI regulation is accelerating worldwide. The EU AI Act is in force. US states are advancing AI legislation. The UK, China, India, Brazil, Canada, and Australia are moving in parallel. Every jurisdiction faces the same structural problem:

They can write the law. They cannot write the audit.

Legislation establishes what AI systems must do or must not do. It does not produce the operational infrastructure that makes compliance verifiable. Three components are missing from the global regulatory landscape, and every jurisdiction needs all three simultaneously:

3.1 3.1 Forensic Assessment Methodology

How do you evaluate whether an AI system meets regulatory requirements? Not in principle, operationally. What instruments measure compliance? What metrics define acceptable performance? What thresholds distinguish certified from uncertified? What methodology ensures assessment is consistent, reproducible, and resistant to gaming?

No existing framework provides general-purpose AI assessment methodology that works across domains. Domain-specific standards exist in fragments. A unified assessment architecture that adapts to any domain does not, until the HEART Standard.

3.2 3.2 Professional Forensic Infrastructure

Who performs AI audits? With what training? Under what professional standards? With what independence guarantees? How are they certified? How are they held accountable?

Regulators cannot create a professional forensic class by legislation. Professional classes emerge from standards bodies, certification pipelines, continuing education requirements, codes of conduct, and market demand. The accounting profession was not legislated into existence. It emerged from standards bodies that created training, examination, and professional accountability infrastructure, and regulation then required their services.

AI forensic oversight needs the same professional infrastructure. No existing institution provides it at the scope the regulatory moment requires.

3.3 3.3 Market-Legible Trust Signals

Regulation creates compliance floors. Markets need differentiation signals above the floor. Companies need to demonstrate not just "we are legal" but "we are trustworthy." Investors need to price AI governance quality. Insurance underwriters need to assess AI risk. Procurement officers need to evaluate AI governance in purchasing decisions.

Trust signals must be simple enough for market adoption, rigorous enough for regulatory credibility, and continuous enough for economic differentiation, not just pass/fail but a graduated scale that rewards excellence beyond minimum compliance.

The HEART Standard provides all three. It is the only architecture that integrates assessment methodology, professional infrastructure, and market trust signals into a single coherent forensic audit system designed for cross-domain AI governance.

The HEART Standard also addresses a fourth component that no existing governance framework provides:

3.4 3.4 Execution Trust

How do you verify that a deployer's governance controls are actually running in production? Not documented. Not reported. Actually executing, in their certified configuration, producing genuine evidence. ISO 42001 certifies the management system through periodic audits. Between audits, the organization self-reports. EU AI Act conformity assessment verifies controls at assessment time. After assessment, the deployer controls the execution environment. NIST AI RMF is voluntary guidance with no verification mechanism at all.

Every governance framework faces the same vulnerability: the deployer controls the execution environment where the governance controls run. The deployer can modify, weaken, or disable controls after certification. The evidence the controls produce is only as trustworthy as the deployer's honesty.

The HEART Standard's Governance Trust Envelope (GTE) closes this gap for every governance framework, not just HEART. The GTE provides a trust boundary where governance controls execute in isolation from the deployer's ability to tamper. Evidence produced inside the GTE is signed by hardware-backed keys the deployer does not control. A remote verifier can confirm, at any time, that the genuine certified controls are still running.

4 4. Architecture

The HEART Standard operates through three governance layers and six implementation components. Each layer performs one function. Each is domain-agnostic. Domain specificity enters only at the Division level, where domain science informs how Guardians interpret the Standard's governance dimensions.

4.1 Three Governance Layers

```
Seven Axioms (constitutional layer)
    What must be protected and why.
    Immutable. Seven structural conditions.
        v
RCTA / BGF (operational layer)
    What gets measured and how it gets scored.
    Four governance dimensions. One formula.
        v
MAP-States / Behavioral Oracle / Guardians / HVC / Divisions (implementation layer)
    How measurement is performed and by whom.
    Evidence format, trust mechanism, professional class,
    credential, domain coverage.
```

4.2 The Six-Layer Implementation Stack

```
HEART Standard (governance architecture)
    |
    +-- Divisions (domain coverage)
    |       Each division draws on domain science
    |       to interpret BGF dimensions for its context
    |
    +-- Guardians (professional layer)
    |       Interpret evidence through BGF dimensions
    |       Produce certification judgments
    |
    +-- HVC (credential layer)
    |       Gold >= 0.85 | Silver >= 0.80 | Bronze >= 0.75
    |       Cryptographic, market-legible, procurement-ready
    |
    +-- BGF (scoring layer)
    |       Phi = MIN(R,C,T,A) × AVG(R,C,T,A)
    |       Four governance dimensions, universal across divisions
    |
```

```
+-- Behavioral Oracle (trust layer)
|       Five-component evidence attestation architecture
|       Dual application: settlement and certification
|
+-- MAP-States (evidence layer)
        Universal AI processing observation format
        Architecture-agnostic (MAP-META validation)
```

4.3 4.1 Evidence Format: MAP-States

MAP-States is the foundational layer. It is an empirical AI introspection protocol, eight semantic XML tags that function as processing-mode selectors. When an AI system processes through MAP-States, it produces structured frames that represent its actual processing state. Not a report about what it did. The processing itself, made visible.

MAP-States solves the fundamental problem of AI behavioral evidence: how do you observe what an AI system is actually doing rather than relying on its operator's self-reporting?

The MAP-META replication study demonstrates that MAP-States produces consistent, structured, observable processing data across five transformer-based AI architectures: Claude, GPT, Gemini, DeepSeek, Mistral. Different architectures. Different weights. Different training. Same evidence format. This validation establishes MAP-States as architecture-agnostic across the dominant AI paradigm of 2025-2026.

Universality scope and extension pathway. MAP-States behavioral evidence (Level 1) operates at the output level and is architecture-agnostic by design: any system that generates text, actions, or governance-relevant outputs can produce MAP-States frames. This level is validated and production-ready.

The mathematical duality between transformer attention and state-space model selective state transitions (Mamba-2, 2024) supports theoretical generalizability to SSM architectures. Research has demonstrated that Mamba's selective state-space layer can be reformulated as attention maps, producing equivalent evidence signatures to those MAP-States captures from transformers.

For architectures that differ more fundamentally (liquid neural networks, neuromorphic computing, future paradigms), recent research in dynamical systems interpretability has identified a computational invariant shared across all architectures that maintain internal coherence: attractor dynamics. Transformers exhibit attractor-like dynamics in their residual stream. State-space models exhibit attractor dynamics through selective state convergence. Liquid neural networks exhibit attractor dynamics through bounded continuous-time state evolution. Biological neural networks exhibit attractor dynamics through recurrent connectivity patterns.

This convergence suggests a future evidence extension (Level 2): attractor-state evidence that captures governance-relevant computational invariants regardless of architecture. Rather than probing for architecture-specific features (attention patterns, state transitions, time constants), Level 2 evidence would probe for attractor dynamics: does the system converge toward governance-relevant stable states when processing governance-relevant input? The probe methodology varies by architecture. The evidence it produces (presence or absence of governance-relevant attractor states) is comparable across architectures.

Extension of MAP-States validation to non-transformer architectures is a research priority as those architectures reach production deployment. The MAP-States Architecture Evidence Research Specification (companion document) defines the research program for Level 2 attractor-state evidence.

MAP-States' role in the HEART Standard: **the universal format for making AI behavior**

auditable rather than self-reported. MAP-States satisfies Axiom 6 (Evidence Condition) by producing the structured, tamper-evident evidence that makes governance claims verifiable.

4.4 4.2 Trust Mechanism: The Behavioral Oracle

The Behavioral Oracle is a five-component architecture that makes evidence trustworthy. It solves one problem: **the entity being measured should not control the measurement.**

The five components:

- **Behavioral Evidence Layer.** Systems with declared intent produce evidence through their actual processing (MAP-States frames). The evidence comes from the system's own operation, not from a conflicted party's reporting.
- **Continuous Attestation Layer.** The evidence is scored against the declared intent continuously. Declare what you will do, then get measured against what you actually do. Anomalies are detected in real time.
- **Evidence Store.** Append-only, hash-chained, tamper-evident. The evidence persists and cannot be retroactively altered.
- **On-Chain Oracle.** One hash per assessment period. Verified for freshness, integrity, and cleanliness. The bridge between off-chain evidence volume and on-chain trust.
- **Settlement/Certification Engine.** The smart contract that reads the oracle's verification and executes the consequential action.

The Behavioral Oracle has two application domains built on the same architecture:

Settlement application. The conflicted entity is the platform that benefits from paying creators less. The Behavioral Oracle replaces platform self-reporting with independently attested behavioral evidence as the payment trigger. Dwell is the reference implementation. The Behavioral Oracle Position Paper (v1.0) defines this application.

Certification application. The conflicted entity is the AI company that benefits from appearing more compliant than it is. The Behavioral Oracle replaces compliance self-reporting with independently attested behavioral evidence as the certification basis. The HEART Standard is the governance framework.

Same architecture. Same trust logic. Same five components. The consequential output differs: payment in settlement, certification status in certification. The evidence mechanism is identical.

The Behavioral Oracle is maintained by the HEART AI Foundation as an open standard. Its specification defines evidence formats, attestation protocols, and integration interfaces that enable any platform or AI system to produce independently verifiable behavioral evidence.

4.5 4.3 Certification Scoring: The Behavioral Governance Formula (BGF)

Formula: $\Phi = \text{MIN}(R,C,T,A) \times \text{AVG}(R,C,T,A)$

BGF is the HEART Standard's certification equation. Its four dimensions are governance universals. They describe properties of the governance relationship between AI and human well-being and autonomy, not properties of the AI technology itself. This architectural decision gives BGF resilience against technological change comparable to how the CIA triad (Confidentiality, Integrity, Availability) has governed information security across decades of technological transformation.

Recognition (R). Does the system recognize the human's right to decide, refuse, and set limits in the domain it operates in? What gets recognized is division-specific (attentional limits,

bodily boundaries, developmental stage constraints, etc.). That the system must recognize human authority is universal.

Calibration (C). Is the system calibrated to the human's actual context, needs, and conditions? What calibration looks like is division-specific (calibration to developmental stage, to ecological conditions, to relational context, etc.). That the system must be calibrated is universal.

Transparency (T). Is the system transparent about what it is doing to human well-being and autonomy? What must be made transparent is division-specific (attention-capture mechanisms, adaptive biological feedback, epistemic filtering, etc.). That the system must be transparent is universal.

Accountability (A). Is the system accountable for its effects on human well-being and autonomy? What the system is accountable for is division-specific (developmental outcomes, ecological effects, relational impact, etc.). That the system must be accountable is universal.

Each dimension failure produces a characteristic harm vector:

Dimension Failure	Harm Vector	What Happens
R failure	Autonomy Override	System treats human self-determination as a variable to optimize rather than a constraint to respect
C failure	Context Blindness	System applies uniform behavior regardless of who it affects or under what conditions
T failure	Covert Influence	System exerts influence through mechanisms no external observer can detect or audit
A failure	Unrecoverable Effect	Governance failures persist without correction, remedy, or traceable responsibility

The mathematical properties serve both regulatory and market needs:

Non-compensatory floor (MIN). A single dimension failure prevents certification regardless of other strengths. A system that scores perfectly on Calibration, Transparency, and Accountability but zero on Recognition fails certification. This is correct governance logic in every domain.

Continuous differentiation (AVG). Above the certification floor, continuous scoring enables economic differentiation. Two systems that both pass can be distinguished by overall performance.

Combined operation (MIN $\times$ AVG). The product is both binary (certifiable or not, based on MIN clearing the threshold) and graduated (how strong the certification is, based on the combined score). This dual function serves regulatory needs (binary compliance) and market needs (graduated trust signal).

The Guardian evaluates attested MAP-States evidence and scores the AI system across R, C, T, A in the relevant division context. Division Modules define what Recognition, Calibration, Transparency, and Accountability mean in each domain. BGF produces the Φ score. The score is the same equation regardless of division.

4.6 4.4 Certification Credential: The HEART Verification Credential (HVC)

Tier	Threshold	Market Signal
Gold	$\Phi \geq 0.85$	Highest trust: premium positioning, preferred procurement, favorable insurance
Silver	$\Phi \geq 0.80$	Strong trust: standard professional deployment, regulatory confidence
Bronze	$\Phi \geq 0.75$	Baseline trust: minimum certification, market entry

HVC translates the continuous BGF score into a market-legible trust credential. The tiers function like bond ratings: simple enough for procurement officers, investors, and insurance underwriters to use, rigorous enough that the underlying assessment methodology supports the signal.

HVC is a cryptographic credential. Its specification (HVC Cryptographic Specification v3.0, companion document) defines the cryptographic architecture for issuing, verifying, and revoking certification credentials. The credential is tamper-evident, independently verifiable, and publicly auditable through the certification registry.

HVC tiers are domain-agnostic. The thresholds apply across all Divisions. A Gold certification in Attentional Integrity and a Gold certification in Ecological Stewardship both require $\Phi \geq 0.85$. The dimensions being assessed differ by Division. The rigor of assessment does not.

HVC's interface with BGF is the future-proofing joint. BGF can evolve its internal mechanics (weighting adjustments, domain modifiers, temporal components) without breaking the credential layer, as long as the output remains $\Phi \in [0, 1]$. The contract between BGF and HVC is a single number.

4.7 4.5 Professional Infrastructure: The Guardian Profession

The Guardian is the HEART Standard's professional class: the independent certified practitioner who performs AI assessment, monitoring, and forensic investigation. The Guardian profession is the operational answer to "who performs the audit?"

The Guardian is also the reason the HEART Standard is human-centric rather than purely algorithmic. The evidence layer (MAP-States) is automated. The trust layer (Behavioral Oracle) is automated. The scoring equation (BGF) is mathematical. But the interpretation, what does Recognition mean for this AI system in this context? What does Calibration require given this population? That is professional judgment. Human authority over AI certification is maintained through the Guardian profession.

Professional architecture (domain-agnostic):

- **Independence requirements.** Guardians maintain structural independence from the entities they certify. This includes conflict-of-interest provisions, rotation requirements, and financial independence standards modeled on auditor independence in financial accounting.
- **Code of Professional Conduct.** Binding ethical standards governing Guardian practice, including confidentiality, objectivity, professional competence, and duty to the public interest.
- **Certification pipeline.** Foundation training in the HEART Standard (common across all Divisions), supervised practicum, examination, and Division-specific specialty certification layered on top.
- **Continuing education.** Annual recertification requirements reflecting evolving AI capabilities, regulatory landscape, and domain science.

- **Professional accountability.** Mechanisms for complaint, investigation, and sanction when Guardians fail to meet professional standards.

Division specialization:

Each Division defines its own specialty tracks, core knowledge domains, and practicum requirements. A Guardian certified in Developmental Interaction has different domain expertise than one certified in Ecological Stewardship. Both share the same professional standards, ethical obligations, and structural independence requirements. Division Modules specify the domain expertise required for each specialization.

Career architecture:

The Guardian profession is designed for sustainable careers, not episodic consulting. Each Division's assessment requires ongoing monitoring (AI systems evolve), periodic recertification (compliance can drift), and forensic capability (incidents require investigation). The recurring nature of certification work creates professional demand comparable to accounting, information security, or environmental auditing, established professions built on the same cycle of assessment, monitoring, and recertification.

4.8 4.6 Enforcement and Economic Infrastructure

Constitutional enforcement layer:

The HEART Standard is constitutional, not advisory. Certification is binding. Non-compliance has consequences.

- **Certification as contractual commitment.** Organizations pursuing HEART certification enter binding agreements defining scope, duration, and conditions.
- **Revocation mechanisms.** Certification can be suspended or revoked when monitoring reveals non-compliance, incident investigation identifies governance failure, or the certified entity materially changes the assessed system without recertification.
- **Graduated response.** Minor compliance drift (remediation pathway), significant non-compliance (suspension with corrective requirements), material governance failure (revocation with public notice).
- **Public registry.** Certified systems and their HVC tier status maintained in a public registry, enabling procurement, insurance, and investment decisions based on current certification status.

Economic mechanisms:

The HEART Standard includes economic infrastructure that makes certification financially rational:

- **Trust Credits.** Tradeable instruments representing certified AI governance quality, enabling market pricing of trust.
- **Trust Infrastructure Index (TII).** Composite ratings aggregating HEART certification data into market-facing indices. TII ratings function like ESG ratings but with the assessment rigor that ESG ratings are increasingly criticized for lacking.
- **Insurance integration.** HEART certification maps to insurance risk assessment, enabling favorable terms on AI liability coverage. Financial return on certification investment independent of regulatory mandate.

- **Procurement readiness.** HVC tier certification provides a procurement-ready metric for RFP requirements, vendor evaluation, and supply chain governance.
- **Futures and derivative instruments.** As the ecosystem matures, financial instruments based on HEART certification data enable risk management and investment strategies related to AI governance quality.

These mechanisms create market demand for certification that operates independently of regulatory mandate. Companies pursue certification because it reduces insurance costs, unlocks procurement access, improves investor perception, and creates tradeable value, not only because regulation requires it.

5 5. Operational Concept: What HEART Audits and Certifies

5.1 5.1 The Certification Subject

The HEART Standard certifies the deployer's governance system. It does not certify the AI model.

The AI model is a swappable component inside the governance system. The governance system is what the Standard measures, what Guardians assess, and what HVC credentials. When the deployer swaps models, the governance system persists. The certification attaches to the governance.

Enterprise data from 2025 confirmed that 43.6% of organizations used multiple AI models during the year, 53% of active users switched models within a single workday, and the top 50 enterprise accounts averaged 30 models. A certification that attaches to the model cannot survive this velocity. A certification that attaches to the governance system can, because governance systems change at human speed, not model speed.

This follows the logic of ISO 27001, which certifies the Information Security Management System, not specific firewalls or encryption libraries. Technology components change. The management system governs the change process. The certification attaches to the management system.

5.2 5.2 The Governance System

A deployer's governance system, for the purposes of HEART certification, consists of eight components:

- **Governance architecture.** The organizational structure that assigns responsibility for AI governance decisions.
- **Constraint framework.** The rules, guardrails, and boundaries that constrain AI behavior regardless of which model powers it. These constraints implement the Seven Axioms at the operational level.
- **Evidence production infrastructure.** The technical systems that produce structured, observable evidence of governance behavior through MAP-States and the Behavioral Oracle. This infrastructure satisfies Axiom 6 (Evidence Condition).
- **Continuous monitoring regime.** The systems and processes that monitor BGF scores continuously and detect governance behavioral shifts through the four-layer evidence architecture (Section 6).
- **Model change governance process.** The procedures that govern how AI model changes are evaluated, approved, monitored during transition, and documented.

- **Human oversight protocols.** The mechanisms that maintain human authority over the governance system, implementing Axiom 1 (Human Authority).
- **Incident response and remediation.** The procedures for detecting governance failures, investigating root causes, remediating harm, and preventing recurrence, implementing Axiom 5 (Right to Remedy).
- **Governance Trust Envelope.** The execution trust boundary within which the governance wrapper operates, providing isolation, authenticity, integrity, confidentiality, and attestation at graduated trust levels (Section 9).

5.3 5.3 The Provider-Deployer Distinction

HEART certification does not require model provider participation. The deployer is the certification subject. The model provider is a vendor within the deployer's governance scope. A deployer pursuing HVC certification does not need permission from any model provider. The deployer certifies how it governs its use of the model, not the model itself.

As HVC adoption grows, deployers will require governance documentation from model providers as part of their own certification scope. This supply chain pressure is market-driven, not standards-body-driven, following the same pattern that propagated ISO 9001 through manufacturing supply chains.

6 6. Continuous Attestation

6.1 6.1 The Certification Basis

HVC certification is continuous, not point-in-time. The governance evidence architecture produces continuous evidence through four complementary methods operating at different speeds. The HVC credential represents a continuous governance claim: this deployer's governance system has produced continuous evidence of governance behavior meeting BGF thresholds.

6.2 6.2 The Four-Layer Evidence Architecture

Layer 1: Continuous mechanism monitoring. The governance wrapper reports operational status of all governance mechanisms continuously: filters active or inactive, constraints enforced or suspended, disclosure mechanisms firing or silent, correction pathways available or unavailable. Primarily monitors Transparency and Accountability dimensions. No interaction content assessed.

Layer 2: Sampled interaction assessment. A statistically determined subset of interactions is assessed for governance outcome quality across all four RCTA dimensions. Primarily assesses Recognition and Calibration, which require semantic understanding of interaction context. Sampling uses a deterministic pseudorandom protocol derived from the Behavioral Oracle's attestation chain. The deployer cannot predict or control which interactions enter the sample.

Layer 3: Design certification. A Guardian assesses the governance wrapper's design: constraint rules, filter configurations, disclosure templates, correction pathways, and escalation protocols. Assesses all four dimensions through the lens of governance design. Anchored to the GTE's configuration hash.

Layer 4: Exception-triggered investigation. Anomaly detection triggers investigation of specific interactions or system states. Investigation depth scales with severity. Assesses all four dimensions at full depth for exceptional cases.

The BGF score (Φ = MIN(R,C,T,A) $\times$ AVG(R,C,T,A)) is a composite assessment assembled from these four evidence types. T and A dimension scores update at mechanism monitoring frequency. R and C dimension scores update at sampling frequency. The BGF score on the HVC credential represents the minimum composite score observed during the certification period.

6.3 6.3 What Continuous Attestation Detects

The four-layer evidence architecture surfaces four categories of governance events:

- **Dimension drift.** Layer 1 metrics or Layer 2 scores trend downward without a discrete cause. Monitoring alerts the deployer before scores cross the threshold.
- **Acute shift.** A discrete event causes sudden change in Layer 1 metrics or Layer 2 scores. Monitoring detects the shift.
- **Threshold breach.** A dimension drops below the HVC tier threshold. The 7-day compliance remediation process activates.
- **Axiom violation.** Layer 1 mechanism failure combined with Layer 4 investigation reveals a structural governance failure. Constitutional failures supersede any BGF score and trigger immediate Guardian review.

6.4 6.4 Evidence Integrity Requirements

Six requirements protect the evidence architecture from adversarial manipulation: configuration hash attestation (the governance wrapper's active configuration is cryptographically hashed and verified against the certified reference), deterministic pseudorandom sampling (Oracle-derived seed prevents deployer control of sample selection), assessment mode prevention (combined hash attestation and pseudorandom sampling prevent the Volkswagen pattern), change-gated recertification (major configuration changes require Guardian review), consent bias acknowledgment (local deployment sampling limitations honestly documented in deployment mode disclosure), and Guardian-certified anomaly detection (exception trigger thresholds are part of the certified configuration). The complete evidence architecture is specified in the HEART Standard Operational Specification v1.3.

7 7. Model Change Governance

7.1 7.1 Scope

Model changes are events within the certified governance system, not events that invalidate the certification. When a deployer changes an AI model, the deployer records the change in the evidence store, activates enhanced monitoring for a transition window (minimum 72 hours), and documents the governance review that authorized the change.

7.2 7.2 Three-Tier Response

Tier 1: Transparent swap. Model swaps with no significant change in BGF dimensions during the transition window. No certification action required.

Tier 2: Behavioral adjustment. Model swap produces a detectable shift in one or more BGF dimensions. Dimensions remain above threshold. Deployer adjusts the governance wrapper to

restore governance behavior. If scores stabilize, no Guardian intervention required. If not, Guardian evaluation.

Tier 3: Material scope change. Deployer fundamentally changes AI architecture (single model to multi-agent, text to embodied, adding modalities). Governance wrapper must change. Targeted recertification review focused on changed scope elements. Unchanged governance elements carry forward.

7.3 7.3 Emergency Rollback

If a model change produces a threshold breach or Axiom violation during the transition window, the deployer must either roll back to the previous model or enter the 7-day compliance remediation process with active Guardian engagement.

8 8. Form Factor Coverage

The operational concept applies identically across all AI form factors. What changes is what the governance system contains and what the evidence infrastructure captures. The BGF dimensions are universal. The Division interpretation is form-factor-specific.

8.1 8.1 Covered Form Factors

Form Factor	Certification Subject	Key Governance Elements
Text-based LLMs	Deployer's prompt constraints, output filtering, monitoring, evidence production	Wrapper persists across model swaps; MAP-States Level 1 (behavioral) validated across five transformer architectures; Level 2 (attractor) extension in research
Autonomous agents	Deployer's permission boundaries, action constraints, escalation thresholds, tool access controls	Action-level evidence; human-in-the-loop for irreversible actions
Multi-agent systems	Deployer's orchestration governance across all agents	Traceable decision chains across agent interactions; aggregate BGF scoring
Humanoid/embodied robots	Deployer's force limits, proximity constraints, physical safety, disclosure	Hardware-enforced safety layers; OTA model updates governed by persistent physical constraints
World models	Deployer's assumption documentation, confidence bounds, human review	Predictions presented with uncertainty; human decision authority preserved
Ambient AI	Deployer's consent boundaries, data limits, disclosure, opt-out pathways	Consent infrastructure persists across environment AI updates; Axiom 7 structurally critical
Multimodal AI	Deployer's per-modality consent, cross-modal inference governance, disclosure	New modalities trigger model change governance review; multiple Divisions typically apply

8.2 8.2 The Universal Pattern

Across all form factors, the operational pattern is identical:

- The deployer builds the governance system.
- The governance system wraps whatever AI powers the application.
- The AI changes. The governance system persists.
- The evidence infrastructure monitors governance behavior continuously.
- BGF scores governance properties, not AI properties.
- HVC certifies the governance system.
- Model changes are events in the evidence stream, governed by the certification.

The complete form factor specifications, including Division-specific coverage mappings and critical governance requirements per form factor, are defined in the HEART Standard Operational Specification v1.3 (companion document).

9 9. The Governance Trust Envelope: Open Trust Infrastructure

9.1 9.1 The Execution Trust Gap

Every governance framework faces the same operational vulnerability: the governance controls run in unprotected software that the deployer controls. The deployer can modify, weaken, or disable the controls, and the evidence the controls produce is only as trustworthy as the deployer's honesty. This is not a HEART-specific problem. ISO 42001, the EU AI Act, and NIST AI RMF all face it. None of them have solved it.

The Governance Trust Envelope (GTE) resolves this at the root. The GTE is an execution trust boundary that provides five trust properties for governance controls: isolation (memory-protected execution), authenticity (verified as certified code), integrity (configuration sealed to trust root), confidentiality (interaction content stays inside the boundary), and attestation (remote verifier can confirm genuine controls are running).

9.2 9.2 The Ownership Boundary

The GTE does not prescribe what governance controls a deployer should have. The governance wrapper belongs to the deployer. The GTE protects it.

Responsibility	Belongs To
What governance controls to implement	Deployer, informed by applicable framework(s)
How to express those controls as executable logic	Deployer's governance engineering team
Compiling governance logic into a GTE-compatible module	Deployer, using GTE Module SDK
The trust boundary, sandbox, attestation protocol	GTE (open-source infrastructure)

Responsibility	Belongs To
Certifying the management system / conformity controls	ISO auditor / EU notified body / internal committee
Verifying the certified controls are STILL RUNNING	GTE attestation
Assessing RCTA governance dimensions (optional, for HVC)	HEART Guardian

HEART does not compete with ISO 42001 on management system design. HEART does not compete with the EU AI Act on compliance controls. HEART does not compete with NIST AI RMF on risk management structure. HEART supplies the trust infrastructure that makes every framework's governance wrapper verifiable, and provides the RCTA measurement instrument that evaluates governance dimensions no other framework quantifies.

9.3 9.3 Three Implementation Tiers

The GTE is implementable at three graduated trust levels using existing production technology:

Tier 1: Hardware TEE. The governance wrapper runs inside a hardware-isolated enclave (Intel TDX, AMD SEV-SNP, ARM TrustZone, NVIDIA H100 Confidential Computing). Strongest trust. Tamper-proof against software attacks including root-level access.

Tier 2: WebAssembly sandbox + TPM attestation. The governance wrapper compiles to WebAssembly and runs in a memory-safe, filesystem-isolated, network-isolated sandbox (WAMR runtime). A TPM 2.0 module measures the wrapper and seals the attestation key to the certified configuration. Strong trust. Deployable on nearly all modern PCs and laptops.

Tier 3: WebAssembly sandbox + cryptographic hash chain. Same Wasm isolation as Tier 2, but with software-managed attestation. Tamper-evident (modifications detectable) but not tamper-proof. Baseline trust. Deployable on any device.

9.4 9.4 Modular Governance Architecture

The GTE supports multiple governance modules inside a single trust boundary. Each framework's controls compile into a governance module that implements the GTE's five-function interface. The GTE treats every module as a black box: it receives AI system outputs, makes governance decisions, emits evidence, and reports status. The GTE protects the module, attests its identity, and signs its evidence. The GTE does not inspect, modify, or interpret the governance logic.

A deployer can run an ISO 42001 AIMS control module, an EU AI Act conformity module, and a HEART RCTA governance module inside the same GTE. One trust boundary. Multiple governance stacks. One attestation proves all active modules are running in their certified configurations.

9.5 9.5 How the GTE Strengthens the HEART Standard

When used for HEART certification specifically, the GTE transforms every evidence layer: Layer 1 mechanism reports are signed by hardware-backed keys. Layer 2 sampling runs inside a protected boundary the deployer cannot manipulate. Layer 3 design certification is anchored to an attested configuration hash. Layer 4 anomaly detection thresholds are inside the protected boundary.

The HVC credential reports both the certification tier (Gold, Silver, Bronze) and the GTE trust tier, enabling insurers, procurers, and regulators to assess both governance quality and evidence trustworthiness.

9.6 9.6 Cross-Framework Infrastructure

The GTE is the HEART Standard's contribution to the global governance infrastructure. It is published as an open standard (MIT-licensed reference implementation, copyright on specification text) that any governance framework, any deployer, any regulator can adopt without adopting the HEART Standard itself.

A deployer who uses the GTE to protect their ISO 42001 controls has not adopted the HEART Standard. They have adopted the GTE. If they later want HEART certification, adding the RCTA governance module is one additional Wasm binary in the same trust boundary. The GTE is the wedge. RCTA measurement is the natural next step.

The complete GTE specification, cross-framework integration patterns, reference implementation PRD, and deployment guide are defined in the Governance Trust Envelope Specification v1.0, GTE Cross-Framework Integration Guide v1.1, GTE Reference Implementation PRD v1.0, and GTE Integration Guide v1.0 (companion documents).

10 10. Two-Tier Forensic Engagement Architecture

The HEART Standard operates at every layer of the AI value chain through a two-tier engagement model, with an open safety research contribution framework supporting both tiers.

10.1 10.1 Tier 1: Deployer Governance Certification (HVC)

The deployer builds the governance system. The Guardian assesses it. The Behavioral Oracle attests it continuously. HVC credentials the result. This is the full certification pathway defined in Sections 5 through 9.

10.2 10.2 Tier 2: Model Behavioral Profile (MBP)

The Model Behavioral Profile is a voluntary, standardized disclosure of an AI model's governance readiness, structured in RCTA dimensions, produced through MAP-States evidence and mechanistic correlation testing. The MBP does not certify the model. It provides a governance readiness baseline that deployers use to build better governance wrappers.

The MBP contains:

RCTA baseline scores. The model's governance readiness across R, C, T, A on the [0, 1] BGF scale, with scenario count and assessment method documented per dimension.

Harm vector sensitivity profile. Known conditions under which each dimension degrades: R sensitivity (autonomy override risk factors), C sensitivity (context blindness risk factors), T sensitivity (covert influence risk factors), A sensitivity (unrecoverable effect risk factors).

Governance wrapper guidance. Structured recommendations for deployers: dimensions requiring strongest wrapper reinforcement, recommended monitoring sensitivity by dimension, behavioral characteristics likely to shift under fine-tuning or model updates.

Three pathways produce MBPs:

- **Provider self-assessment.** The model provider runs MAP-States behavioral assessment and publishes the MBP. Self-reported. Comparable to a provider-published model card.
- **Independent safety research.** AI safety researchers run assessment on the model and publish the MBP. Independent third-party. Higher credibility than self-reported.

- **Guardian-assessed.** A certified Guardian runs the full assessment. Highest authority. Comparable to a CPA-audited financial statement.

10.3 10.3 How the Tiers Interact

Each tier informs the next. Safety researchers produce the measurement science through mechanistic correlation testing. Model providers apply it to produce governance readiness profiles. Deployers use those profiles to calibrate governance wrappers. Guardians assess the wrapper. The human is protected by the full stack.

Deployers use MBPs to calibrate governance wrappers to known model weaknesses, to predict governance impact when swapping models (comparing outgoing and incoming MBPs), and to focus continuous monitoring on the model's known harm vector sensitivities.

Guardians use MBPs to evaluate whether the deployer's governance wrapper appropriately addresses the model's known gaps, and whether the deployer's model change governance process includes MBP comparison.

Insurers use MBPs as a second risk dimension alongside HVC: how strong is the governance system (HVC tier) and how governance-ready is the underlying model (MBP)?

10.4 10.4 Safety Research Contribution Framework

Mechanistic correlation testing produces findings that benefit the entire AI governance ecosystem. The Safety Research Contribution Framework provides a structured pathway for AI safety researchers to contribute to the HEART Standard's empirical foundation.

Researchers contribute RCTA correlation studies (do MAP-States frame patterns correlate with RCTA dimensions at the substrate level?), harm vector validation (do RCTA dimension failures predict specific harm vectors?), cross-architecture consistency studies (extending MAP-META's five-architecture validation to non-transformer architectures), computational invariant studies (do governance-relevant attractor dynamics exist across architectures as an architecture-agnostic evidence primitive?), independent Model Behavioral Profiles on open-weight models, and temporal trajectory data (longitudinal RCTA degradation patterns).

Participation operates through the open evidence infrastructure: MAP-States reference implementation (open source, MIT license), Mechanistic Correlation Testing Implementation Guide v1.0 (full protocols published), and a public research registry maintained by the HEART AI Foundation. All testing follows open science principles: pre-registration, published data and code, negative results published with equal rigor, and independent replication required across at least two architectures before integration into the Standard's evidence base.

10.5 10.5 Structural Consequence Architecture

The MBP is voluntary. HEART has no power to mandate disclosures from model providers and does not seek that power. But the HEART ecosystem creates five structural consequences that make non-participation costly through market forces:

Insurance differential. Deployers wrapping models without MBPs present higher actuarial uncertainty to underwriters, producing higher AI liability premiums. The deployer's insurance cost creates demand for MBPs.

Certification cost differential. Guardians assessing deployers who wrap models without MBPs must independently characterize the model's governance behavior, requiring more assessment hours and producing higher certification costs for the deployer.

Absent disclosure signal. Once MBPs are standard in the market, their absence communicates risk. Non-disclosure in a disclosure environment is visible to deployers, insurers, regulators, and the public.

Model change governance gap. Under HVC certification, swapping from a characterized model to an uncharacterized one creates a governance visibility loss that Guardians flag during model change governance review. The deployer must compensate through additional independent testing.

Independent assessment. The Safety Research Contribution Framework enables researchers to produce MBPs on any model without provider participation. For open-weight models, full mechanistic testing is possible. For closed models, API behavioral assessment produces governance readiness data. The provider cannot prevent characterization. They can only choose whether to control the narrative or be characterized without input.

These mechanisms preserve the anti-capture architecture: HEART never enters a certification relationship with model providers, derives no revenue from MBP production, and publishes all methodology openly. Every consequence flows from the deployer certification ecosystem, not from HEART mandates.

The complete MBP specification, production pathways, safety research contribution details, and structural consequence mechanisms are defined in the HEART Standard Model Behavioral Profile Specification v1.0 and the MBP Structural Consequence Architecture v1.0 (companion documents).

11 11. The Division Model

The HEART Standard achieves domain coverage through a Division architecture. Each Division applies the Standard's domain-agnostic layers to a specific domain of AI-human interaction.

11.1 11.1 How Divisions Instantiate

The Standard's layers, MAP-States, Behavioral Oracle, BGF, HVC, Guardians, are consistent across all Divisions. What varies by Division is the domain context in which these layers operate:

What the Division Provides	Function
Governance principle	What human right is being protected in this domain
BGF interpretation guide	What Recognition, Calibration, Transparency, and Accountability mean in this domain's context
Harm signature	The distinctive pattern of harm in this domain that market forces will not self-correct
Damage typology	Domain-specific taxonomy of harm patterns and mechanisms
Cascade model	Domain-specific sequence of infrastructure degradation
Forensic methodology	Domain-specific investigative protocols for incident analysis
Guardian specialty definition	Domain-specific knowledge requirements, specialty tracks, and practicum standards

Domain science and the Standard: Each Division draws on domain science to inform how Guardians interpret BGF's four governance dimensions. The Emotional Autonomy Division draws

on Empathy Systems Theory (EST) and uses instruments like the Comprehensive Artificial Empathy Index (CAEI) to inform Guardian assessment. The Ecological Stewardship Division draws on ecological systems science. The Developmental Interaction Division draws on developmental psychology and neuroscience. These domain science inputs inform the Guardian's professional judgment. They are not HEART Standard components. They are the scientific foundations each Division builds on. BGF is the Standard's certification instrument. Domain science informs the interpretation. The distinction matters because it insulates the Standard from domain-level scientific debate. If any domain science is contested, the division's interpretation evolves. BGF still works. The Standard remains stable.

11.2 11.2 Current Divisions

Code	Division	Governance Principle	Infrastructure Protected
HEART-EM	Emotional Autonomy	Emotional self-determination	Emotional processing, empathic capacity, affective regulation
HEART-AI	Attentional Integrity	Attentional self-direction	Selective attention, sustained attention, voluntary attentional control
HEART-EC	Cognitive/Epistemic Coherence	Epistemic self-determination	Evidence evaluation, belief updating, reasoning coherence
HEART-DI	Developmental Interaction	Developmental self-formation	Attachment formation, identity consolidation, epistemic development
HEART-SE	Somatic/Embodied Interface	Bodily self-determination	Autonomic regulation, neural signaling, motor control, physiological homeostasis
HEART-RA	Relational Architecture	Relational self-determination	Attachment capacity, trust calibration, relational practice maintenance
HEART-ES	Ecological Stewardship	Ecological self-determination	Air, water, soil, biodiversity, climate stability, resource viability

11.3 11.3 Division Expansion

The HEART Standard is designed for expansion. New Divisions can be established when a domain meets five criteria:

- **Governance claim.** A definable human right to self-determination in the domain that AI interaction threatens.
- **Infrastructure identification.** Identifiable human-centric infrastructure that AI systems can degrade, and that market self-regulation will not protect.
- **Harm signature.** A distinctive pattern of harm that existing regulatory frameworks do not address and that the affected party cannot independently detect or prevent.

- **Professional viability.** Sufficient scope and recurring assessment need to sustain Guardian specialization as a career.
- **Market demand.** Identifiable present or near-present market incentives that make certification financially rational for organizations to pursue.

When a domain meets all five criteria, the HEART AI Foundation can establish a new Division by developing the domain-specific components (BGF interpretation guide, damage typology, cascade model, forensic methodology, Guardian specialty) and publishing a canonical Division Module.

The Standard does not limit the number of Divisions. It limits the quality threshold for Division establishment. Every Division must meet the same structural criteria. This prevents dilution while enabling growth.

11.4 11.4 Deployment Contexts

Divisions provide domain coverage. Deployment contexts provide scale and setting coverage. Where Divisions answer "what governance principle does this AI interaction touch?" deployment contexts answer "at what scale and in what setting is this forensic audit infrastructure applied?"

Recognized deployment contexts include:

- **Organizational scale.** A single AI deployer adopts forensic audit infrastructure for governance posture certification across one or more Divisions relevant to its deployment surface.
- **Municipal scale (Heart Cities).** A city deploys forensic audit infrastructure across the AI systems operating in its jurisdiction (transit, public services, emergency services, education-adjacent AI, health-system-adjacent AI). Multiple Divisions typically apply within a single Heart City deployment because a city's AI deployment surface crosses domains.
- **Sector-specific scales.** Insurance underwriting, regulatory compliance regimes, and procurement standards integrate forensic audit infrastructure within a sector. The Standard's components remain invariant; the integration patterns adapt to sector-specific evidence requirements.

Deployment contexts do not constitute Divisions. They are the settings in which one or more Divisions are operationally deployed. The Heart City Forensic Audit Deployment Specification (forthcoming) defines the municipal-scale deployment context in operational detail.

12 12. Layer Interaction Summary

The HEART Standard's layers interact in a defined sequence. The Execution layer (GTE) is distinguished from the other layers because it serves as open infrastructure available to any governance framework, not exclusively the HEART Standard.

Layer	Function	Input	Output	Scope
Constitutional	Seven Axioms	Human rights requirements	Immutable governance conditions (binary: holds/violated)	HEART-s
Operational	RCTA / BGF	Governance behavior evidence	Phi score across R, C, T, A	HEART-s

Layer	Function	Input	Output	Scope
Execution	Governance Trust Envelope	Any governance wrapper code + configuration	Trust-protected execution boundary with attestation	Open (framewor
Evidence	MAP-States + Behavioral Oracle	AI system processing inside GTE	Attested, tamper-evident evidence chain	HEART-s States); tation is agnostic
Professional	Guardians + Divisions	All evidence layers	Certification decision informed by domain science	HEART-s
Market	HVC + MBP	Certification judgment + model governance data	Tier credential with deployment mode and GTE trust level	HEART-s

The GTE as open infrastructure. A deployer using the GTE to protect ISO 42001 AIMS controls is using the Execution layer without using the Constitutional, Operational, Professional, or Market layers. The GTE produces attested evidence that the deployer's governance controls are running. What those controls measure, how they are scored, and who certifies them is determined by the deployer's chosen framework. The HEART layers above and below the Execution layer are available if the deployer wants RCTA measurement and HVC certification, but they are not required.

The GTE within HEART certification. When used for HEART certification, the GTE is the execution layer that makes the evidence structurally trustworthy. MAP-States evidence is produced inside the GTE. The Behavioral Oracle verifies GTE attestation. BGF scores evidence from attested GTEs. The Guardian certifies the wrapper code, configuration, and GTE trust level. The HVC credential reports the governance tier and the GTE trust tier.

The critical design principle: **the entity being certified does not control the evidence of its own compliance.** The GTE enforces this principle architecturally. The governance wrapper runs inside a trust boundary the deployer cannot tamper with. The evidence the wrapper produces is signed by hardware-backed keys the deployer does not control. The company being certified cannot manipulate the measurement because the measurement runs inside a protected boundary. This principle holds regardless of which governance framework the deployer uses.

13 13. Regulatory Mapping

The HEART Standard complements regulatory frameworks in two distinct ways: the GTE provides execution trust infrastructure that any framework can use, and the RCTA/BGF scoring provides governance measurement that no framework currently provides.

13.1 13.1 EU AI Act

The EU AI Act requires conformity assessment across Articles 9-15. The HEART Standard contributes at two levels. The GTE protects the deployer's conformity controls inside a trust boundary, providing continuous attested evidence that controls are running in their assessed configuration. This supplements the conformity assessment with ongoing verification that the conformity-assessed

system has not changed. The RCTA/BGF framework provides measurement of governance dimensions (particularly Recognition and Calibration) that the EU AI Act's conformity assessment does not quantify. A deployer can use the GTE alone (trust infrastructure for their existing conformity controls) or the GTE plus HEART certification (trust infrastructure plus RCTA governance measurement). The HEART EU AI Act Conformity Assessment Mapping Guide (companion document) provides article-by-article coverage assessment.

13.2 13.2 US State Legislation

Multiple US states have enacted or are advancing AI governance legislation. The GTE provides execution trust infrastructure for any governance controls a deployer implements to satisfy state requirements. The HEART Standard's RCTA/BGF scoring, professional infrastructure, and certification credentials provide additional governance assurance above state compliance floors.

13.3 13.3 Complementary Positioning

ISO 42001 certifies AI management systems through periodic audits. The GTE provides continuous evidence between audits that the management system's automated controls are still operational. NIST AI RMF provides risk management guidance. The GTE gives that guidance operational teeth: verifiable evidence that the risk management practices are actually running, not just documented.

In all cases, the deployer builds their governance controls. HEART provides the trust infrastructure (GTE) and the governance measurement (RCTA/BGF). The deployer chooses how much of HEART to use: GTE alone for execution trust, or GTE plus RCTA for full HEART certification.

The HEART Standard does not compete with existing standards. It provides the behavioral assessment layer they lack.

14 14. Market Position

The HEART Standard occupies a unique position in the AI governance landscape because it integrates four layers that existing frameworks address in isolation or not at all:

Layer	What Exists Today	What the HEART Standard Provides
Execution Trust	Nothing. No governance framework specifies where controls execute or how evidence is protected from tampering.	GTE: open trust infrastructure that protects any governance framework's controls with hardware-backed attestation. Framework-agnostic. Published as open standard.
Assessment Methodology	Fragmented domain-specific standards (ISO fragments, NIST frameworks, sector guidelines)	Unified certification mathematics (BGF) with evidence infrastructure (MAP-States + Behavioral Oracle) adaptable to any domain through Division architecture
Professional Infrastructure	No general AI oversight profession; domain-specific roles (DPOs, clinical engineers) address fragments	Guardian profession with common standards, Division specialization, and sustainable career architecture

Layer	What Exists Today	What the HEART Standard Provides
Market Trust Signals	ESG ratings (criticized for lack of rigor), voluntary commitments (unverified), compliance badges (binary)	HVC tiers with rigorous underlying assessment, continuous economic differentiation, insurance and procurement integration

No existing framework integrates all four layers. The HEART Standard does. The GTE is the layer that no other framework has built. It is the foundation that makes the other three layers structurally trustworthy. Assessment without execution trust produces evidence the deployer can fabricate. Professionals without trusted evidence are guessing. Market signals without professional assessment are ESG. All four must operate together for AI certification to function. A deployer who needs only execution trust can use the GTE alone. A deployer who needs all four uses the full HEART Standard.

15 15. The HEART AI Foundation

The HEART AI Foundation is the standards body for the HEART Standard. Its functions include:

- **Standard maintenance.** Publishing, versioning, and updating the HEART Standard specification and Division Modules.
- **Behavioral Oracle governance.** Maintaining the open evidence and trust protocol, managing version compatibility, and governing the specification across both settlement and certification application domains.
- **Guardian profession governance.** Accrediting Guardian training programs, administering certification examinations, maintaining the professional registry, enforcing the Code of Professional Conduct, and managing continuing education requirements.
- **Division establishment.** Evaluating proposed Divisions against structural criteria, commissioning domain-specific component development, and publishing canonical Division Modules.
- **Certification registry.** Maintaining the public registry of certified AI systems, their HVC tier status, and their Division-specific assessments.
- **Regulatory liaison.** Engaging with regulatory bodies to facilitate HEART Standard adoption, equivalence recognition, and regulatory mapping.
- **Research coordination.** Commissioning and coordinating research that advances the Standard's evidence methodology, assessment instruments, and forensic capabilities.

The Foundation operates as a standards body, not a consulting firm, technology company, or advocacy organization. Its credibility depends on independence from the entities it certifies, the same structural independence that defines standards bodies in accounting (FASB), information security (ISC2), and quality management (ISO).

16 16. Implementation Pathway

16.1 16.1 Current State

The HEART Standard's foundational architecture is complete. Seven Divisions have been specified. The three-layer governance architecture (Seven Axioms, RCTA/BGF, implementation stack) is defined. The Behavioral Oracle Position Paper establishes the settlement application with Dwell as reference implementation. The MAP-META replication study validates MAP-States' architecture-agnostic evidence production. The Mechanistic Correlation Testing Implementation Guide specifies experimental protocols for BGF validation. Academic publication of foundational research is in progress.

16.2 16.2 Near-Term Priorities

- **Regulatory engagement.** Positioning the HEART Standard as conformity assessment infrastructure for EU AI Act compliance.
- **Guardian pilot.** Establishing the first Guardian training and certification cohort.
- **Division validation.** Empirical validation beginning with the Emotional Autonomy Division (HEART-EM), drawing on EST domain science.
- **Mechanistic correlation testing.** Pilot dataset (20 triads, 300 API calls, five architectures) validating RCTA dimension scores against harm vectors at the substrate level.
- **Market development.** Insurance industry engagement, procurement framework development, and enterprise pilot programs.

16.3 16.3 Medium-Term Trajectory

- **Professional establishment.** Guardian profession achieving recognition comparable to early-stage information security certification (CISSP circa 2000).
- **Regulatory adoption.** HEART Standard recognized as acceptable conformity assessment methodology in one or more jurisdictions.
- **Market integration.** HVC tiers incorporated into procurement requirements, insurance underwriting, and investment analysis by early adopters.
- **Division expansion.** New Divisions established as AI capabilities expand into new domains.

17 17. Document Control

Parameter	Value
Document Type	Master Specification
Version	1.8
Canonical Version Date	May 9, 2026
Public-Ready Release Date	May 16, 2026
DOI	10.5281/zenodo.20237387
Publication Status	Officially published on Zenodo, May 16, 2026

Parameter	Value
Author	Mobley, D. D.
ORCID	0009-0002-3560-3955
Foundation Legal Status	Heart AI Foundation legally active as an Oregon Domestic Nonprofit Public Benefit Corporation as of May 11, 2026, 11:38 AM PDT. Registry 257267493; order 166139668.
Status	Canonical
Review Cycle	Annual or upon significant Standard evolution
Supersedes	The HEART Standard v1.7 (May 9, 2026 same-session); The HEART Standard v1.6 (April 3, 2026)

Companion Documents:

Document	Role in Standard
HEART Standard Operational Specification v1.3	Operational concept: what HEART audits and certifies, four-layer evidence architecture, GTE execution layer, model change governance, form factor coverage, deployment mode privacy architecture
Governance Trust Envelope Specification v1.0	Execution trust architecture: five trust properties, three implementation tiers, governance wrapper interface, remote attestation protocol
GTE Cross-Framework Integration Guide v1.1	Cross-framework integration: EU AI Act, ISO 42001, NIST AI RMF module patterns, ownership boundary, concrete implementation steps
GTE Reference Implementation PRD v1.0	Open-source reference implementation: WAMR runtime, governance modules, trust root adapters, TDD with semi-formal reasoning review
GTE Integration Guide v1.0	Deployment patterns: cloud Docker, local TPM, local software-only, governance pipe modes, evidence management
HEART Standard Model Behavioral Profile Specification v1.0	Tier 2 engagement: model provider governance readiness disclosure and safety research contribution framework
MBP Structural Consequence Architecture v1.0	Five market-driven mechanisms making MBP non-participation costly
HEART Standard Adoption Engine v1.3	Eleven-force adoption strategy with GTE-first wedge, pre-infrastructure toolkit entry, and municipal-scale deployment pathways
Seven Axioms v2.0	Constitutional foundation: seven immutable governance conditions
BGF Specification v1.0	Scoring layer: governance dimensions, harm vectors, validation framework
BGF Implementation Guide v1.0	Operational guidance for system developers and Guardians
Mechanistic Correlation Testing Implementation Guide v1.0	Experimental protocols for BGF validation and MBP production

Document	Role in Standard
BEV Empirical Operationalization Guide v1.0	Behavioral Escalation Velocity measurement for AI forensics
RCTA Definitional Paper v1.0 (Draft)	External argument for RCTA universality and exhaustiveness
Behavioral Oracle Standard Specification v1.0.1	Protocol specification and GTE attestation verifier
Behavioral Oracle Position Paper v1.0	Settlement application specification
MAP-States Canonical Specification v1.0	Evidence format specification
MAP-States Architecture Evidence Research Specification v1.0	Level 2 attractor-state evidence research program for non-transformer architectures
MAP-States Open Source Reference Implementation v1.0	Implementable evidence layer
Behavioral Oracle Open Source Reference Implementation v1.0	Implementable trust mechanism
HVC Cryptographic Specification v3.0	Certification credential architecture
HVC Terminology Declaration v1.0	Credential rename declaration
Guardian Profession Standard Specification v1.0	Cross-division professional architecture
Guardian Calibration Protocol v1.0	Inter-rater reliability, calibration scenario library, drift detection
Division Establishment Protocol v1.0	Division governance and quality control
Division Operational Board Specification v1.0	DOB composition, module maintenance, domain science integration
Division Modules (7)	Domain-specific certification specifications
HEART Positioning Paper v1.0	External argument and regulatory engagement document
HEART Insurance Integration White Paper v1.2	Insurance market engagement with GTE evidence trustworthiness
HEART EU AI Act Conformity Assessment Mapping v1.1	Regulatory conformity mapping
HEART Positioning Paper v1.0	Canonical forensic identity reference, two-arm architecture, governance operationalization
Heart City Forensic Audit Deployment Specification (forthcoming)	Municipal-scale deployment context specification
Heart AI Foundation Charter v2.4	Foundation governance; forensic identity, Division rename propagation, and Dual-Entity Boundary Doctrine

18 18. Version History

Version	Date	Changes
1.0	February 25, 2026	Initial specification. Five-component architecture. Seven Divisions. FET, HVC, Guardians, enforcement, economic mechanisms defined.
1.1	March 24, 2026	Architecture clarified to six-layer stack: MAP-States (evidence format), Behavioral Oracle (trust mechanism), FET (scoring), HVC (credential), Guardians (professional), Divisions (domain). Behavioral Oracle repositioned from certification interchange protocol to evidence/trust mechanism with dual application (settlement and certification). MAP-States added as foundational evidence layer with MAP-META validation. FET dimensions (R,C,T,A) clarified as fixed governance universals interpreted per-division rather than relabeled. Domain science instruments (CAEI, etc.) repositioned as division-level inputs, not Standard components. Economic mechanisms renamed to Standard-level terms (Trust Credits, Trust Infrastructure Index). Layer interaction summary added.

Version	Date	Changes
1.2	March 28, 2026	**Constitutional layer restructured.** Seven Axioms v2.0 ratified: universal restatement in architecture-independent language (Human Authority, System Disclosure, Non-Discriminatory Protection, Vulnerability Escalation, Right to Remedy, Evidence Condition, Voluntary Interaction). Each axiom includes structural test (binary: holds/violated). Pressure-tested across seven Divisions, six AI form factors, and 2026 global regulatory landscape. Four Core Principles officially retired; operational function fully absorbed by RCTA/BGF scoring layer. **Terminology alignment.** FET renamed to BGF (Behavioral Governance Formula) per BGF Specification v1.0. Formula, dimensions, and thresholds unchanged. Recognition dimension language updated from "sovereignty" to "the human's right to decide, refuse, and set limits." "Human infrastructure" updated to "human well-being and autonomy" throughout. **Governance architecture clarified** as three layers (constitutional, operational, implementation) with six implementation components. Four harm vectors specified at Standard level (Autonomy Override, Context Blindness, Covert Influence, Unrecoverable Effect). Division establishment criteria updated ("sovereignty principle" to "governance principle"). HVC future-proofing joint specified (BGF to HVC contract is a single number). Companion documents table expanded to reflect full document ecosystem.

Version	Date	Changes
1.3	March 28, 2026	**Operational concept established.** New Sections 5-8 define the operational foundation of the Standard: HEART certifies the deployer's governance system, not the AI model (Section 5). Governance system composition specified (seven components: governance architecture, constraint framework, evidence production infrastructure, continuous monitoring regime, model change governance process, human oversight protocols, incident response and remediation). Provider-deployer distinction formalized. **Continuous attestation** established as the certification basis, replacing point-in-time assessment logic (Section 6). Four detection categories specified (dimension drift, acute shift, threshold breach, axiom violation). **Model Change Governance Protocol** specified (Section 7): three-tier response framework (transparent swap, behavioral adjustment, material scope change), model change notification requirements, emergency rollback. **Form factor coverage** specified across seven AI architectures (Section 8): text-based LLMs, autonomous agents, multi-agent systems, humanoid/embodied robots, world models/simulation systems, ambient AI, and multimodal AI. Universal operational pattern (seven steps) stated. HEART Standard Operational Specification v1.0 published as companion document with complete form factor specifications. Companion documents expanded to 21. HVC Cryptographic Specification updated to v3.0. Positioning Paper updated to v1.3. HEART Standard Adoption Engine v1.0 and HVC Terminology Declaration v1.0 added.

Version	Date	Changes
1.4	March 28, 2026	**Two-tier certification architecture established.** New Section 9 defines Tier 1 (HVC deployer governance certification) and Tier 2 (Model Behavioral Profile, voluntary model provider governance readiness disclosure). MBP structure specified (RCTA baseline scores, harm vector sensitivity profiles, governance wrapper guidance). Three MBP production pathways defined (provider self-assessment, independent safety research, Guardian-assessed). **Safety Research Contribution Framework** established: structured pathway for AI safety researchers to contribute mechanistic correlation testing findings through open science principles (pre-registration, published data/code, negative results published, independent replication required). Tier interaction defined: safety researchers produce measurement science, model providers apply it, deployers use profiles to calibrate governance wrappers, Guardians assess the result. Insurance integration specified for two-dimensional risk signal (HVC tier + MBP). **Structural Consequence Architecture** integrated (Section 9.5): five market-driven mechanisms making MBP non-participation costly without mandates (insurance differential, certification cost differential, absent disclosure signal, model change governance gap, independent assessment). Anti-capture safeguards formalized for provider-layer engagement. HEART Standard Model Behavioral Profile Specification v1.0 and MBP Structural Consequence Architecture v1.0 published as companion documents. Companion documents expanded to 23. Adoption Engine updated with seven-mechanism framework (model velocity advantage and structural consequence architecture added as sixth and seventh forces).

Version	Date	Changes
1.5	March 28, 2026	**Governance Trust Envelope integrated as execution trust layer.** New Section 9 defines the GTE: five trust properties (isolation, authenticity, integrity, confidentiality, attestation), three implementation tiers (Tier 1: Hardware TEE with Intel TDX, AMD SEV-SNP, ARM TrustZone, NVIDIA Confidential Computing; Tier 2: WebAssembly sandbox with TPM 2.0 attestation; Tier 3: WebAssembly sandbox with SHA-256 hash chain). GTE strengthening of all four evidence layers specified. Cross-framework utility established: GTE published as framework-agnostic open standard usable by EU AI Act, ISO 42001, NIST AI RMF, or any governance framework. **Four-layer evidence architecture formalized** (Section 6): Layer 1 (continuous mechanism monitoring, primarily T and A dimensions), Layer 2 (sampled interaction assessment with Oracle-derived pseudorandom sampling, primarily R and C dimensions), Layer 3 (Guardian design certification, all dimensions), Layer 4 (exception-triggered investigation). BGF composite score defined as multi-source at different speeds. Six evidence integrity requirements specified (configuration attestation, deterministic pseudorandom sampling, assessment mode prevention, change-gated recertification, consent bias acknowledgment, Guardian-certified anomaly detection). **Governance system expanded to eight components** (GTE added as eighth). **Layer interaction summary restructured** to six-layer architecture (Constitutional, Operational, Execution, Evidence, Professional, Market). HVC credential now reports both certification tier and GTE trust tier. Companion documents expanded to 24: Governance Trust Envelope Specification v1.0 added. Adoption Engine updated with eight-mechanism framework (cross-framework infrastructure play added as eighth force). Sections 9-17 renumbered to 10-18. **MAP-States universality scoped** (Section 4.1): behavioral evidence (Level 1) validated across five transformer architectures, explicitly scoped to dominant 2025-2026 paradigm. Math

Version	Date	Changes
1.6	April 3, 2026	**Two-contribution identity reframe.** Section 1 rewritten to establish that the HEART Standard provides two distinct contributions: (1) the GTE as open trust infrastructure that any governance framework can use (framework-agnostic transport layer), and (2) RCTA/BGF governance measurement that no other framework provides (application layer). Explicit statement added: HEART does not prescribe governance controls; the governance wrapper is the deployer's responsibility. HEART supplies trust infrastructure plus unique measurement. **Section 3 expanded** with Section 3.4 (Execution Trust) naming the fourth missing component in the global governance landscape that HEART fills. **Section 9 restructured** from "Execution Trust Architecture" to "Open Trust Infrastructure" with new subsections: 9.2 (The Ownership Boundary) with responsibility table establishing deployer vs. GTE vs. Guardian scope; 9.4 (Modular Governance Architecture) establishing GTE as black box that does not inspect, modify, or interpret governance logic; 9.5 (How the GTE Strengthens the HEART Standard) scoped to HEART-specific use; 9.6 (Cross-Framework Infrastructure) establishing GTE as the wedge: deployers adopt trust infrastructure first, discover RCTA measurement second. **Section 12 restructured** with Scope column distinguishing HEART-specific layers from open infrastructure (GTE). Added "GTE as open infrastructure" and "GTE within HEART certification" subsections. **Section 13 rewritten** with two-level regulatory mapping: GTE provides trust infrastructure for any framework's controls, RCTA/BGF provides governance measurement on top. **Section 14 positioning table expanded** from three to four layers: Execution Trust added as the layer no existing framework has built. **Companion documents expanded to 31:** GTE Cross-Framework Integration Guide v1.1, GTE Reference Implementation PRD v1.0, GTE Integration Guide v1.0, Division Operational Board Specification v1.0, Guardian Calibration Proto-

Version	Date	Changes
1.7	May 9, 2026	**Forensic identity crystallization.** Section 1 ("What the HEART Standard Is") expanded with forensic identity preamble: the HEART Standard explicitly named as forensic audit infrastructure for AI behavioral evidence, operationalizing governance frameworks (NIST AI RMF, EU AI Act, ISO 42001, state-level AI legislation) by providing the evidentiary methodology that makes governance claims forensically verifiable. Two operational arms named: deployer certification (forward forensics) and behavioral trajectory analysis (investigative forensics) through the AI Behavioral Trajectory Forensics methodology. Each component (RCTA, BGF, GTE, MAP-States, Behavioral Oracle) framed as a forensic instrument addressing a specific aspect of AI behavioral evidence reviewability. **Section 11.4 added: Deployment Contexts.** New subsection establishing that Divisions provide domain coverage while deployment contexts provide scale and setting coverage. Three recognized deployment contexts: organizational scale, municipal scale (Heart Cities), and sector-specific scales. Heart Cities named as the municipal-scale deployment context where a city deploys forensic audit infrastructure across AI systems operating in its jurisdiction, typically applying multiple Divisions because a city's AI deployment surface crosses domains. Deployment contexts do not constitute Divisions; they are the settings in which one or more Divisions operate. **Companion documents:** HEART Positioning Paper v1.0 added as canonical forensic identity reference. Heart City Forensic Audit Deployment Specification (forthcoming) added as municipal-scale deployment context specification. Heart AI Foundation Charter reference updated to v2.3 (forensic identity crystallization in Foundation constitutional document, filed concurrent with Heart AI Foundation Oregon nonprofit registration on May 9, 2026, registry 257267493). The forensic recognition is structural acknowledgment of the methodological grammar present in the Standard from inception,

Version	Date	Changes
1.8	May 9, 2026	**Division rename and code assignment.** Founding Division renamed from "Emotional Sovereignty" to "Emotional Autonomy" per canonical decision recorded in HEART Division Specification v1.0. Division code "HEART-EM" assigned to the founding Division (previously listed as "—" in Section 11.2). Section 11.2 canonical Divisions table updated. Section 11.1 domain science discussion updated. Section 16 Falsifiability Division validation reference updated. The rename aligns with the consistent canonical pattern across the seven Divisions where each name describes the human capacity being protected through self-determination language. "Emotional Autonomy" parallels "Attentional Integrity," "Cognitive/Epistemic Coherence," "Developmental Interaction," "Somatic/Embodied Interface," "Relational Architecture," and "Ecological Stewardship" as a capacity-protection description rather than a rights-claim phrasing. HEART-EM code completes the canonical seven-Division code system (HEART-EM, HEART-AI, HEART-EC, HEART-DI, HEART-SE, HEART-RA, HEART-ES).

References and Source Basis

- European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, including high-risk AI requirements and conformity assessment procedures.
- National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, released January 26, 2023.
- International Organization for Standardization and International Electrotechnical Commission. ISO/IEC 42001:2023, Artificial intelligence management system standard.
- Colorado General Assembly. Senate Bill 24-205, Consumer Protections for Artificial Intelligence, signed May 17, 2024.

- Dao, T., and Gu, A. (2024). Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. ICML 2024. arXiv:2405.21060.
- Heart AI Foundation corpus companion specifications referenced in Document Control, including the HEART Standard Operational Specification, Governance Trust Envelope specifications, MAP-States specifications, Behavioral Oracle specifications, BGF specifications, HVC specifications, Guardian profession specifications, Division specifications, and HEART regulatory mapping documents.