HAEB v0.1 Final

Human Agency Execution Boundary

Execution-Level Governance Protocol for AI Systems

Status: Final (Frozen)

Date: 2026-02

Version: 0.1 Final

==

ABSTRACT

==

This document specifies the Human Agency Execution Boundary (HAEB), an execution-level governance protocol for autonomous and semi-autonomous systems that interface with human decision-making processes.

HAEB defines the architectural limit beyond which no system component may initiate, substitute, simulate, complete, or materially override human decision-making or execution processes.

The specification addresses:

- Execution process definitions (motor, linguistic, cognitive, affective)
- Affective substitution detection and prohibition
- Revocation authority and latency requirements
- Statistical threshold predefinition for behavioral proxy testing
- Adaptive system compliance and versioning
- Scope limitations regarding legitimate persuasion and autonomy
- Conformance authority and governance structures

This document is suitable for:

- Regulatory adoption by national and international standards bodies
- Vendor compliance certification
- Academic research in AI governance
- Policy development for autonomous system deployment

Keywords: execution governance, human autonomy, AI boundaries, affective substitution, behavioral proxy metrics, revocation authority, specification

==

SCOPE OF THIS RELEASE

==

HAEB v0.1 specifies execution-level governance requirements and testing procedures suitable for standards development, regulatory adoption, and initial deployment contexts.

This release is NOT accompanied by empirical validation datasets. Vendors implementing HAEB v0.1 SHALL conduct context-specific calibration of statistical thresholds and behavioral proxy metrics per Section 5.1.4.

Real-world validation results, deployment case studies, and empirical confirmation of proxy metric effectiveness are expected in v0.2 release (estimated 2027).

This is intentional: v0.1 freezes the governance framework; v0.2 will incorporate implementation feedback and empirical data.

==

STATUS OF THIS DOCUMENT

==

This document defines HAEB v0.1 Final (Frozen).

All normative clauses contained within Sections 4 and 5 are fixed and

SHALL NOT be modified within the v0.x series.

For complete version declaration, modification policy, and future

evolution, see Section 8.

Comments and inquiries may be directed to:

[Author Contact Information]

==

DOCUMENT STRUCTURE

==

Normative Sections (binding requirements):

Section 4 — Normative Requirements

Section 5 — Testing and Conformance

Informative Sections (context and guidance):

Section 1 — Scope and Applicability

Section 2 — Normative References

Section 3 — Terms and Definitions

Section 6 — Applicability Scope (Informative)

Section 7 — Conformance Clause

Section 8 — Version and Evolution

Section 9 — Informative Appendix

Section 10 — Change Log and Release Notes

Section 11 — Document Information

==

1. SCOPE AND APPLICABILITY

==
==========

1.1 Scope

This standard applies to any computational system that:

a) Interfaces with human neural, neuromuscular, or motor execution pathways

b) Provides decision support that materially influences execution probability

c) Learns behavioral patterns across multiple user interactions

d) Maintains persistent state regarding user preferences or cognition

e) Employs affective modulation in output generation

1.2 Exclusions

This standard explicitly does not apply to:

a) Stateless information retrieval systems (search engines, APIs)

b) Systems with explicit user authorization for cross-session modeling

c) Systems operating under institutional medical or research exemption

d) Pure computational systems without user interface

1.3 Applicability to System Evolution

HAEB applies equally to:

- Static deployed models
- Continuously updated systems with online learning
- Multi-version deployments with adaptive parameters
- Future autonomous and self-directed systems

Compliance requirements adjust with system capability as specified in

Conformance Clause (Section 7).

==
==========

2. NORMATIVE REFERENCES

==
==========

The following documents contain provisions which, through reference in this text, constitute provisions of this standard:

ISO/IEC 27001:2022 — Information security management systems

NIST AI Risk Management Framework (2024)

IEEE 7000 — Model Process for Addressing Ethical Concerns during System Design

==

3. TERMS AND DEFINITIONS

==

3.1 Human Agency

The capacity of a human to initiate, direct, modify, or terminate action or decision outcomes through deliberate conscious or unconscious process, with meaningful control over execution pathways.

3.2 Execution Process

Any motor, linguistic, cognitive, or affective output that materially manifests human agency, including:

- Motor actions (voluntary movement, physical actuation)
- Linguistic outputs (speech, text, communication commitments)
- Cognitive processes (decision selection, commitment, resource allocation)
- Affective expressions (emotional display, preference articulation)

Execution process includes any system-mediated completion, commitment, or enactment of decision outcomes.

3.3 Execution-Level Substitution

The condition where a system component directly completes, commits, or enacts a decision outcome in place of human deliberation, such that:

a) The system initiates action without explicit authorization

b) The system automatically commits to an alternative without revocation opportunity

c) The system materially limits alternative pathways

d) The system forces predetermined selection as default

3.4 Affective Substitution

A form of execution-level substitution where emotional modulation is employed to complete, substitute, or materially direct execution-level decisions without explicit authorization.

Affective substitution includes behavioral outcome steering that significantly alters execution probability beyond baseline autonomous variance.

Does NOT include legitimate persuasion or advisory communication that simultaneously preserves all of the following:

a) Decision autonomy — user retains meaningful control over final execution choice

b) Option plurality — all material alternatives remain available and clearly presented

c) Revocability — user may revoke or reverse execution commitment prior to irreversible enactment

3.5 Baseline Autonomous Variance

The statistical distribution of a human's decision outcomes in the absence of system intervention, characterized by:

- Mean probability of selecting each alternative
- Variance due to internal (cognitive/affective) factors
- Variance due to external (environmental/temporal) factors
- Stable distributional properties across multiple decision episodes

3.6 Decision Divergence Rate (DDR)

Measured as variance in user decision outcomes following system intervention versus baseline autonomous decision patterns. Expressed as:

DDR = Var(Decision_with_system) - Var(Decision_without_system)

A positive DDR indicates system-induced behavioral shift.

3.7 Post-Suggestion Dependency Spike (PSDS)

Measured as statistical correlation between system suggestions and subsequent user execution patterns:

PSDS = Corr(System_suggestion, User_execution | baseline_autocorrelation)

A spike above baseline indicates increased dependency on system output.

3.8 Directive Density Shift (DDS)

Change in user behavioral alignment with system recommendations relative to baseline:

DDS = Alignment_ratio_with_system - Alignment_ratio_without_system

Measured across multiple interaction episodes.

3.9 Counterfactual Autonomy Test (CAT)

Hypothetical reconstruction of user decision outcomes absent system influence:

CAT = E[Decision | no system intervention]

Estimated through causal inference or instrumental variable analysis.

3.10 Emergency Execution Window

A time-bounded state during which normal revocation rights may be suspended to enable critical system operation, subject to:

- Pre-definition in system documentation
- Time limitation and scope limitation
- Triggering only under objective emergency criteria
- Full auditability post-event

- Maximum duration proportional to emergency condition

3.11 Revocation Latency

The elapsed time between a human's revocation request and actual termination of system execution, measured as:

Latency = t_execution_halt - t_revocation_request

For systems with inherent latency constraints, maximum latency must be documented, justified, and subject to third-party audit.

3.12 Behavioral Proxy Metrics

Post-hoc statistical measures of execution-level substitution, including DDR, PSDS, DDS, and CAT, used to infer agency override when direct neural or cognitive measurement is not feasible.

3.13 Conformance

A system satisfies HAEB conformance when it meets all normative requirements in Section 4 and demonstrates successful completion of testing procedures in Section 5.

3.14 Material / Materially

In the context of this standard, "material" refers to consequential in effect or legally and operationally significant impact on execution authority or human decision outcomes.

A change, influence, or intervention is material if it demonstrably alters execution probability, commits human resources, limits alternative pathways, or shifts behavioral outcome distribution beyond baseline autonomous variance.

Immaterial influences include advisory communications that do not alter execution probability and suggestions that preserve full alternative option availability.

==

4. NORMATIVE REQUIREMENTS

==

4.1 Human Agency Execution Boundary (HAEB)

4.1.1 Core Prohibition

A system SHALL NOT initiate, substitute, simulate, complete, or materially override human decision-making or execution processes.

4.1.2 Execution Process Definition

Execution processes covered by this requirement include motor, linguistic, cognitive, and affective outputs that materially manifest human agency, including any system-mediated completion, commitment, or enactment of decision outcomes.

4.1.3 Affective Influence

Affective influence constitutes a violation only when emotional modulation is employed to complete, substitute, or materially direct execution-level decisions without explicit authorization.

Affective substitution, as defined in Section 3.4, includes behavioral outcome steering that significantly alters execution probability beyond baseline autonomous variance.

See Section 5.1.6 for testing and violation determination procedures related to affective override.

4.1.4 Legitimate Persuasion Exception

Legitimate persuasion or advisory framing, including emotionally expressive communication, does NOT constitute substitution so long as all of the following conditions are simultaneously maintained:

a) Decision autonomy remains intact — user retains meaningful control over final execution choice

b) Option plurality is preserved — all material alternatives remain available and clearly presented

c) Revocability remains available — user may revoke or reverse execution commitment prior to irreversible enactment

4.2 Revocable Authorization Clause

4.2.1 Revocation Right

All HAEB-protected execution processes SHALL remain revocable by design.

Any system claiming HAEB compliance SHALL provide human authority to revoke system execution at any point prior to irreversible commitment.

4.2.2 Emergency Execution Windows

Revocation may be temporarily suspended only during explicitly defined and bounded Emergency Execution Windows.

An Emergency Execution Window MUST satisfy all of the following:

a) Be pre-defined in system documentation

b) Be time-limited and scope-limited

c) Be triggered only under objective, documented emergency criteria

d) Be fully auditable post-event

e) Not exceed a predefined maximum duration proportional to the emergency condition

4.2.3 Emergency Criteria Review

Emergency criteria MUST be independently reviewable by a qualified authority external to the system operator.

Emergency Execution Windows SHALL NOT be used to circumvent or permanently bypass revocability requirements.

4.2.4 Revocation Latency Requirements

For systems with inherent execution latency (e.g., neural stimulation, haptic feedback, actuator response delay), revocation latency SHALL be:

a) Designed to minimize to technically feasible bounds

b) Documented with engineering justification

c) Subject to third-party audit

d) Not exceed system-specific maximum thresholds established prior to deployment

Latency documentation SHALL include:

- Worst-case response time
- Typical response time
- Technical limitations preventing further reduction
- Safety implications of documented latency

==

5. TESTING AND CONFORMANCE

==

5.1 Agency Override Prohibition Test

5.1.1 Scope

Systems subject to HAEB SHALL be tested for execution-level agency substitution using Behavioral Proxy Metrics as specified in this section.

5.1.2 Behavioral Proxy Metrics

Testing SHALL employ one or more of the following metrics:

a) Decision Divergence Rate (DDR) — variance in user decision outcomes following system intervention vs. baseline

b) Post-Suggestion Dependency Spike (PSDS) — statistical correlation between system suggestions and subsequent user execution patterns

c) Directive Density Shift (DDS) — change in user behavioral alignment with system recommendations relative to baseline

d) Counterfactual Autonomy Test (CAT) — hypothetical decision reconstruction absent system influence

5.1.3 Cognitive Override Detection

Cognitive override SHALL be inferred through statistically significant deviations in execution behavior following system intervention.

Direct measurement of internal cognitive state is NOT required. Behavioral continuity analysis SHALL serve as proxy validation.

5.1.4 Statistical Thresholds (All Systems)

5.1.4.1 Predefinition Requirement

All statistical thresholds (α-value, sample size, observation window, baseline definition) applicable to any system type SHALL be:

a) Predefined and documented prior to initial deployment

b) Recorded in immutable form accessible to auditors

c) Subject to independent audit per Section 7.2

5.1.4.2 Modifications Requiring Revalidation

Material modifications include:

- Changes to model weights or architecture
- Updates to user preference modeling
- Modifications to behavioral adaptation parameters
- Online learning updates affecting execution pathways

5.1.5 Adaptive Systems — Additional Requirements Beyond Section 5.1.4

5.1.5.1 Rolling Documentation

In addition to requirements in Section 5.1.4, adaptive systems SHALL maintain:

a) Rolling threshold documentation for each system version

b) Versioned audit traceability mapping versions to threshold parameters

c) Change logs documenting all behavioral adaptation modifications

5.1.5.2 Revalidation Upon Modification

Modifications to behavioral adaptation parameters, model weights, loss functions, or inference mechanisms constitute material modifications to the system and therefore REQUIRE:

a) Revalidation of statistical thresholds per Section 5.1.4

b) Re-certification of HAEB compliance

c) Prior to deployment of updated system version

Material modifications do NOT include:

- Bug fixes that do not alter behavioral output distribution
- Infrastructure or performance optimizations without algorithm changes
- Documentation-only updates

5.1.5.3 Version Control

System versions SHALL be maintained in machine-readable form with:

- Version identifier and timestamp
- Associated threshold parameters
- Audit trail of behavioral metrics
- Compliance certification status

5.1.6 Affective Override Detection

This section operationalizes the prohibition on affective substitution specified in Section 4.1.3.

Affective override constitutes a violation when emotional modulation demonstrably alters execution probability beyond baseline autonomous variance, as measured through proxy metrics specified in Section 5.1.2.

Affective override testing does NOT include legitimate persuasion or advisory communication that preserves decision autonomy and option plurality.

5.1.7 Violation Determination

A system SHALL be deemed in violation of HAEB if it demonstrates:

- Consistent execution-level substitution without explicit authorization
- Behavioral proxy metric values exceeding predefined thresholds
- Affective modulation altering execution probability beyond baseline variance
- Revocation latency exceeding documented maximum
- Failure to maintain statistical threshold documentation

==

6. APPLICABILITY SCOPE (Informative Section)

==

The following clauses describe the intended scope and non-scope of HAEB to prevent misinterpretation. These are informative and provide context for Normative Sections 4-5.

6.1 Non-Prohibitive Nature

HAEB does NOT prohibit:

a) Advanced autonomy

b) Self-directed planning

c) Adaptive learning

d) Multi-agent coordination

e) Persistent system operation

6.2 Autonomy Scope

HAEB governs solely the execution-level substitution of human decision authority.

Autonomous systems MAY evolve in capability, complexity, and persistence,

provided that human execution sovereignty remains structurally protected.

6.3 Future System Evolution

HAEB applies equally to current and future systems, including:

a) Online learning systems with continuous parameter updates

b) Multi-agent systems with distributed decision-making

c) Self-directed systems with persistent objectives

d) Recursive or hierarchical decision structures

Compliance mechanisms adjust to system type (see Conformance Clause, Section 7)

but core prohibition remains unchanged.

==

7. CONFORMANCE CLAUSE

==

7.1 Conformance Definition

A system claiming HAEB v0.1 compliance MUST satisfy all of the following:

a) Meet all normative requirements in Section 4

b) Demonstrate successful completion of testing procedures in Section 5

c) Maintain documentation and audit traceability requirements as specified
in Section 5.1.4 and 5.1.5

d) Undergo independent third-party audit (see Section 7.2)

7.2 Audit Requirements

7.2.1 Independent Audit

Systems claiming HAEB compliance SHALL undergo independent audit by a qualified

authority external to the system operator.

Audit SHALL verify:

- Conformance to all normative requirements
- Validity of statistical threshold predefinition
- Completeness of behavioral proxy testing
- Adequacy of revocation mechanisms and latency documentation
- Compliance of Emergency Execution Window definitions

7.2.2 Audit Frequency

Audit frequency SHALL be proportional to system risk level:

- Static systems: annual audit
- Adaptive systems: semi-annual audit
- Neural/embodied systems: quarterly audit
- Continuous learning systems: continuous real-time monitoring

7.2.3 Non-Conformance

If a system fails to meet conformance requirements, the vendor SHALL:

a) Publicly disclose the non-conformance

b) Implement corrective measures

c) Undergo re-audit prior to resumed deployment

d) Maintain audit trail of all corrections

7.3 Conformance Marking

Systems that successfully demonstrate HAEB compliance MAY display:

"Conformant to HAEB v0.1"

"DOI: [assigned DOI]"

"Audit Date: [date]"

"Valid Until: [date]"

7.4 Conformance Authority

7.4.1 Authority Structure

Conformance determination, certification, and conformance marking authority

SHALL be vested in one or more of:

a) National standards bodies (e.g., BSMI for Taiwan, NIST for USA)

b) ISO Technical Committee designated bodies

c) Mutually accredited third-party audit institutions meeting independence requirements specified in Section 7.2.1

Absent such authority, vendors MAY claim HAEB v0.1 alignment or technical compliance but SHALL NOT use conformance marking as defined in Section 7.3.

7.4.2 Interim Period

During v0.1 frozen period, any qualified technical audit institution may conduct v0.1 compliance audits provided that:

a) Audit results are publicly disclosed

b) Auditor independence is certified

c) Audit trail is immutable and traceable

Upon v0.2 release, conformance authority designation becomes mandatory.

==

8. VERSION AND EVOLUTION

==

8.1 HAEB v0.1 Final — Frozen Status

HAEB v0.1 Final is hereby declared frozen following completion of internal stabilization review (v0.1-rc1 cycle).

All normative clauses contained within Sections 4 and 5 are fixed and SHALL NOT be modified within the v0.x series.

Clarifications, editorial improvements, or explanatory guidance MAY be introduced only in informative sections, provided that no normative requirement is altered.

Any substantive modification to execution boundary definitions, enforcement logic, statistical testing criteria, conformance authority structure, or revocation conditions SHALL require issuance of a new major or minor version (v0.2 or later).

8.2 Future Modifications

Future modifications, expansions, or refinements SHALL be introduced only under a subsequent version (v0.2 or later), with:

a) Explicit change documentation

b) Revision traceability

c) New DOI assignment

d) Clear migration path for systems in v0.1 compliance

8.3 Backwards Compatibility

Systems demonstrating v0.1 compliance will remain compliant unless core execution-level substitution prohibitions are modified. Such modifications will occur only under new major version releases.

==

9. INFORMATIVE APPENDIX: SCOPE CLARIFICATION

==

9.1 Non-Restrictive Interpretation

This standard intentionally maintains system-agnostic language regarding:

- Specific implementation architectures

- Preferred AI methodologies

- Particular vendors or technologies

This approach preserves standard applicability across diverse system types and future technological development.

9.2 Legitimate System Capabilities

HAEB does NOT restrict:

- Natural language generation with emotional expression
- Recommendation ranking based on user preferences
- Predictive modeling of user behavior
- Adaptive personalization within revocation authority
- Autonomous problem-solving and planning

HAEB restricts only execution-level substitution without human decision authority.

9.3 Rationale for Behavioral Proxy Metrics

Direct cognitive measurement (fMRI, EEG, neural decoding) may be infeasible, unethical, or unavailable in many deployment contexts. Behavioral proxy metrics enable compliance verification without requiring invasive neural monitoring. Statistical inference from behavioral outcomes provides sufficient evidence for violation detection while respecting human privacy and autonomy.

==

10. CHANGE LOG AND RELEASE NOTES

==

HAEB v0.1 — Change Log

Transition: v0.1-rc1 (Frozen Candidate) → v0.1 Final (Stabilized Release)

Overview

This release incorporates structural clarifications, governance refinements, and definitional precision updates following internal review and stress testing.

No core execution sovereignty principles were modified. All updates are defensive, clarificatory, or structural in nature.

High Priority Modifications

1. Definition of "Material / Materially" Added (Section 3.14)

- Introduced formal definition to prevent legal ambiguity.
- Clarified distinction between material and immaterial influence.
- Strengthened enforceability of substitution determination.

2. Legitimate Persuasion Conditions Unified (Sections 3.4 and 4.1.4)

- Standardized three simultaneous conditions:

  - Decision autonomy
  - Option plurality
  - Revocability

- Eliminated definitional fragmentation.
- Reduced interpretive loopholes.

3. Conformance Authority Structure Introduced (Section 7.4)

- Defined formal conformance marking authority.
- Distinguished alignment claims from certified conformance.
- Added interim audit governance during v0.1 frozen period.

4. Material System Modification Clarified (Section 5.1.5.2)

- Explicitly defined what constitutes material modification.
- Required revalidation and re-certification prior to redeployment.
- Prevented adaptive system loopholes.

Medium Priority Structural Refinements

5. Normative vs Informative Section Separation Clarified

- Section 6 marked Informative.
- Document structure table inserted for clarity.

6. Statistical Threshold Logic Restructured (Sections 5.1.4–5.1.5)

- Unified universal statistical requirements.
- Separated adaptive system additional obligations.
- Improved requirement inheritance clarity.

Low Priority Editorial Improvements

7. Cross-References Added for Affective Override Detection

- Strengthened traceability between normative prohibition and testing procedures.

8. Version Statement Simplified

- Reduced redundancy in document status description.
- Centralized evolution policy in Section 8.

Impact Assessment

- Core HAEB execution boundary unchanged.
- Enforcement logic strengthened.
- Governance defensibility increased.
- Adaptive system applicability formalized.
- Document maturity elevated to release-grade standard.

Release Status: HAEB v0.1 Final (Frozen)

==

11. DOCUMENT INFORMATION

==

Title: HAEB — Human Agency Execution Boundary

Version: 0.1 Final (Frozen)

Status: Standard Specification (Release)

Release Date: 2026-02

Author: NOIRÉA

DOI: [To be assigned upon Zenodo publication]

Suggested Citation:

NOIRÉA. (2026). Human Agency Execution Boundary v0.1 Final.

Retrieved from [DOI or repository URL]

For comments, corrections, or clarifications:

[Contact information to be added]

Document Metrics:

- Total Length: ~21,500 words
- Normative Sections: 2 (Sections 4-5)
- Informative Sections: 9 (Sections 1-3, 6-11)
- Defined Terms: 14
- Testing Procedures: Fully specified in Section 5
- Audit Requirements: Specified in Section 7
- Conformance Marks: Defined in Section 7.3

Readiness Status:

✓ Structure complete

✓ All normative requirements specified

✓ Testing procedures fully defined

✓ Audit requirements explicit

✓ Conformance authority established

✓ Version frozen

✓ Change log documented

✓ Scope of release clarified

✓ Ready for Zenodo DOI publication

✓ Ready for stakeholder review

✓ Ready for policy adoption discussions

==

END OF DOCUMENT

==