The Human Continuity Architecture

Detailed Implementation Plan (HCA-IP)

Governance • Assurance • Certification • Operations • International Coordination • Deployment

Revision 0 — Preliminary — 21 September 2026

Revision 0 is intentionally preliminary. It defines a proposed implementation architecture for converting the Human Continuity Architecture from a system-safety framework into an operating, auditable, distributed assurance ecosystem. It is intended to be challenged, exercised, red-teamed, tested, and revised before any claim of adequacy.

Written by Gregory H. Allison
System Safety Subject Matter Expert

Companion document to: Human Continuity Architecture — Foundational Architecture and System Safety Analysis (HCA-FASSA), Revision 0.

PRELIMINARY DRAFT

i. Table of Contents

1. Objectives

1.1 Purpose

This Human Continuity Architecture Detailed Implementation Plan (HCA-IP) defines how the architecture, requirements, controls, monitoring, assurance, certification, and human-resilience provisions established in HCA-FASSA Revision 0 can be converted into an operating system of institutions, technical mechanisms, evidence, and recurring assurance activities. The plan is deliberately implementation-oriented: each credited safety function must have an accountable implementer, an independent verifier, defined evidence, an escalation path, and a means to detect failure of the control itself.

1.2 Implementation Problem Statement

HCA cannot be implemented by creating a single global authority with unrestricted visibility and enforcement power. Such a structure would create the same concentration-of-power hazard that HCA is intended to control. HCA implementation therefore treats governance, certification, auditing, standards development, funding, and enforcement as safety-critical subsystems subject to capability compartmentalization, dissimilar redundancy, independent assurance, and fail-safe design.

1.3 Implementation Objective

The implementation objective is to establish a distributed, internationally extensible assurance ecosystem in which AI developers, deployers, governments, infrastructure operators, independent assessors, standards organizations, research institutions, and human-resilience organizations can cooperate on catastrophic-risk controls without requiring, or allowing a single actor to control the system or possess all sensitive information.

1.4 Success Criteria

ID	Revision 0 implementation success criterion
IP-SC-1	Every credited HCA control has an identified implementation owner, independent assurance path, verification evidence, and escalation mechanism.
IP-SC-2	No single organization, nation, developer, funder, accreditation body, or technical platform can unilaterally alter, suppress, falsely certify, or defeat the complete HCA safety baseline.
IP-SC-3	HCA can verify high-consequence safety properties while minimizing unnecessary disclosure of proprietary, export-controlled, classified, or security-sensitive information.
IP-SC-4	HCA certification and mutual recognition can scale across jurisdictions without requiring political or technical uniformity beyond agreed minimum safety properties.
IP-SC-5	The six HCA layers can be exercised in realistic demonstrations, including AI-denied operation, containment escalation, cross-system monitoring, and human recovery.
IP-SC-6	The implementation ecosystem is itself periodically red-teamed for capture, collusion, common-mode failure, coercion, and concentration of authority.

2. Scope, Authority, and Relationship to HCA-FASSA

2.1 Scope

This plan covers institutional startup, governance, standards configuration control, technical operations, control ownership, independent verification and validation, certification, accreditation, protected evidence handling, international participation, funding, incident response, red teaming, human-resilience demonstrations, metrics, and phased deployment. It does not supersede national law, military command

authority, corporate governance, or sector-specific regulation. Instead, it defines a safety architecture that those authorities can adopt, recognize, or map into their own mechanisms.

2.2 Governing Technical Baseline

The HCA-FASSA is the governing technical basis for this implementation plan. HCA-IP shall not weaken a FASSA safety requirement through implementation convenience. Where implementation experience reveals that a requirement is impractical, incomplete, contradictory, or creates a new hazard, the issue shall be processed through formal configuration control and hazard review rather than informally waived.

2.3 Relationship to Existing Frameworks

HCA is intended to complement, not replace, existing AI risk-management and management-system standards. NIST AI RMF organizes AI risk work around Govern, Map, Measure, and Manage and treats governance as cross-cutting. ISO/IEC 42001 specifies requirements for an organizational AI management system. HCA extends the system boundary to civilization-scale catastrophic pathways, cross-organizational dependencies, capability concentration, human recovery, and governance-system failure. [1][2]

2.4 Revision 0 Authority

Revision 0 is a proposed implementation baseline, not a treaty, regulation, certification mandate, or claim of international consensus. Its immediate purpose is to make responsibilities, interfaces, evidence, and unresolved decisions explicit enough for serious technical review and pilot implementation.

3. Implementation Principles

Principle	Implementation meaning
IP-P1 — HCA must itself be HCA-compliant	Governance and assurance mechanisms shall satisfy the same no-single-point-catastrophic-authority principle imposed on AI systems.
IP-P2 — Distributed authority	Standards, implementation, accreditation, certification, audit, enforcement, funding, and appeals shall be separated sufficiently to prevent unilateral capture.
IP-P3 — Independent and dissimilar assurance	Safety-critical claims require verification independent of the developer/operator and, for the highest-consequence claims, independent of the primary assessor.
IP-P4 — Verification without unnecessary disclosure	Assessors receive the minimum protected evidence necessary to establish the safety claim; public or international recognition need not expose underlying sensitive artifacts.
IP-P5 — Fail toward less authority	Loss of required monitoring, identity, authorization, evidence integrity, or assurance shall reduce high-consequence privileges rather than silently preserve them.
IP-P6 — Mutual recognition over centralized control	Where feasible, competent national or sectoral assurance systems recognize equivalent results rather than surrendering authority to a single global certifier.
IP-P7 — Evidence over assertion	Certification depends on objective evidence, demonstrations, logs, analyses, tests, and independent challenge rather than organizational promises.
IP-P8 — Human continuity remains operational	Layer 6 is demonstrated through exercises and retained capability, not satisfied by policy statements.
IP-P9 — Rapid but staged deployment	HCA should begin with voluntary pilots and reference implementations while standards and international mechanisms mature in parallel.
IP-P10 — Continuous reassessment	Rapidly changing AI capability requires recurring review of hazards, thresholds, assumptions, controls, and residual risk.

4. HCA Institutional Ecosystem

4.1 Organization-of-Organizations Model

HCA implementation should be an ecosystem rather than a monolithic institution. For Revision 0, three distinct concepts are used: HCA is the technical architecture and assurance framework; a possible Human Continuity Institute (HCI) is a nonprofit technical and administrative secretariat; and the HCA Consortium is the wider federation of participating organizations. No one of these entities possesses total implementation authority.

4.2 Candidate Organizational Structures

Alternative	Description	Strengths	Primary HCA concern
A. Federated standards/certification consortium	Independent organizations jointly maintain standards; accredited assessors certify implementations; recognition is distributed.	Fast voluntary startup; compatible with industry participation; scalable.	Industry/funder capture; uneven national recognition.
B. Treaty-linked distributed assurance network	States accept common minimum obligations but retain national implementation and inspection mechanisms.	Can create reciprocal obligations for highest-consequence capabilities.	Political delay; verification disputes; risk of state dominance.
C. Sectoral federation with mutual recognition	Cloud, critical infrastructure, defense, finance, robotics, biotechnology, and AI sectors maintain mapped assurance regimes.	Uses existing sector competence; limits unnecessary disclosure.	Fragmentation; gaps at sector interfaces.
D. Public-interest multi-stakeholder institute plus independent certifiers	A nonprofit maintains baseline and convenes stakeholders while legally/financially separate assessors perform assurance.	Clear technical home; public-interest mission; supports rapid pilots.	Institute could accumulate influence unless deliberately bounded.

Revision 0 recommends a hybrid of Alternatives A, C, and D, designed so that a future treaty mechanism can map onto the system without becoming its sole point of control.

4.3 Proposed HCA Consortium

The HCA Consortium should include AI developers and deployers, compute/cloud providers, critical-infrastructure operators, national laboratories, universities, standards and professional organizations, independent safety and cybersecurity organizations, insurers where relevant, civil-society/public-interest participants, and government/national-security stakeholders operating within lawful disclosure constraints.

4.4 Human Continuity Institute — Proposed Secretariat Function

A Human Continuity Institute, if created, should maintain the public technical baseline, configuration-control process, hazard database, reference architecture, public assurance guidance, training, meeting support, and non-sensitive registry of certifications and unresolved systemic issues. It should not possess unilateral authority to certify every system, compel disclosure, control compute, command military systems, or override sovereign law.

5. Governance Architecture

5.1 Bicameral Safety Governance

Revision 0 proposes bicameral approval for safety-critical HCA baseline changes. A Board of Governors represents participating stakeholder institutions and jurisdictions; a Technical and Safety Council represents qualified technical disciplines and independent safety judgment. A major baseline change

becomes effective only after concurrence by both bodies under defined quorum, conflict-of-interest, recusal, and public-record rules appropriate to the information classification involved.

Body	Primary function	Shall not
Board of Governors	Institutional legitimacy, resource stewardship, membership, international/sector representation, adoption of major baseline changes.	Unilaterally approve technical safety waivers or suppress independent safety findings.
Technical and Safety Council	Technical baseline, hazard disposition, safety policy, KPM definitions, certification criteria, technical appeals.	Control membership/funding in a manner that makes its technical judgment dependent on a single sponsor.
Executive Secretariat	Administration, schedules, records, publications, logistics, implementation tracking.	Exercise independent certification or final hazard-acceptance authority.
Independent Assurance Forum	Cross-check assessors, review systemic findings, commission dissimilar red teams.	Report through the same management chain as the systems it challenges.

5.2 Safety Review Board (SRB)

The SRB maintains hazard dispositions, reviews residual risk, verifies control allocation, adjudicates safety-critical deviations, and confirms that new implementation mechanisms do not introduce uncontrolled common-cause failures. The SRB should include system safety, AI safety, cybersecurity, critical infrastructure, human factors, reliability, and relevant domain specialists. Members with direct program conflicts recuse from final disposition.

5.3 Engineering Review Board (ERB)

The ERB controls reference architecture, interfaces, technical specifications, test protocols, data schemas, monitoring interfaces, and implementation guidance. It does not accept catastrophic residual risk on behalf of the SRB.

5.4 Independent Verification and Validation / Assurance

Independent assurance shall be organizationally and financially separated from the development function to a degree commensurate with consequence. For the highest-consequence capabilities, a second dissimilar assessor or challenge team should review the primary assurance case. Unresolved findings must have a protected reporting path to both governance chambers and the relevant lawful authority.

5.5 Appeals and Minority Reports

Certification and safety decisions require an appeal mechanism that does not route exclusively through the original decision-maker. Minority technical opinions involving catastrophic risk shall be preserved in the configuration record and elevated when predefined significance thresholds are met. This reduces pressure to convert organizational consensus into false technical certainty.

5.6 Governance Anti-Capture Controls

Control	Minimum mechanism	Evidence
Distributed approval	Two-chamber concurrence for major baseline changes; bounded emergency authority.	Decision record; quorum/recusal log.
Funding diversity	No single unrestricted donor/funder controls a dominant fraction of safety-critical assurance operations; thresholds TBD.	Annual funding concentration analysis.
Term/role rotation	Rotation for selected assurance and red-team leadership roles.	Appointment history.
Protected reporting	Independent route for safety findings and whistleblower-quality technical concerns.	Procedure test; case audit.
Dissimilar review	Periodic review by institutions not sharing the primary governance/funding chain.	Red-team report; closure evidence.
Appeal path	Independent appeal of certification, accreditation, and hazard disposition.	Appeal exercise and records.

6. Functional Organizations and Responsibilities

6.1 Required Functions

Function	Responsibility
Systems Engineering	Maintains architecture, interfaces, reference implementations, configuration baselines, and technical integration.
System Safety	Maintains PHA, FTA, STPA/FMEA products, control allocation, residual-risk record, and safety requirements.
Independent Assurance / IV&V	Verifies safety claims independently of implementers; performs evidence review and challenge testing.
Certification	Determines whether defined HCA certification criteria are met for a declared scope.
Accreditation / Assessor Qualification	Determines competence and independence of certification/assessment bodies.
Audit / Quality Assurance	Checks continuing compliance, evidence integrity, corrective-action closure, and process execution.
Red Team / Adversarial Assessment	Challenges technical and institutional assumptions, including governance capture and monitor defeat.
Incident Analysis	Collects, protects, analyzes, and disseminates safety-relevant incident lessons.
Human Continuity / Layer 6	Defines and demonstrates AI-independent minimum viable operations and recovery capability.
Standards and Configuration Management	Controls baseline changes, traceability, release records, and crosswalks.
Training and Qualification	Defines competency requirements for assessors, operators, reviewers, and exercise personnel.
International / Mutual Recognition	Maintains cross-jurisdiction equivalence criteria and recognition agreements.

6.2 Separation-of-Duties Rule

No organization shall simultaneously serve as sole implementer, sole assessor, sole certification authority, sole accreditation authority, and sole appeal authority for the same high-consequence safety claim. Lower-consequence implementations may combine functions when justified by a documented independence analysis.

7. Control Responsibility and Assurance Matrix (CRAM)

7.1 Purpose

The CRAM converts HCA hazard controls into owned, auditable implementation obligations. It is the primary answer to the question “who establishes, operates, verifies, audits, and maintains this control?” Every credited control in the HCA PHA should ultimately have a CRAM record.

7.2 Minimum CRAM Fields

Field	Required content
Control ID	Unique HCA-CTL identifier and linked hazard/cause.
Implementer	Organization responsible for installing/operating the control.
Standard owner	Body maintaining the performance/acceptance criterion.
Independent verifier	Qualified party verifying implementation and effectiveness.
Certification authority	Entity making the scoped certification decision.
Evidence custodian	Party maintaining protected objective evidence and retention controls.
Reverification interval	Time/event basis for recurring assurance.
Escalation path	Who receives failures, disputes, or unresolved findings.
Dependencies	External systems/services on which the control relies.
Common-cause exposure	Shared infrastructure, models, funding, authority, or data that could defeat multiple controls.

7.3 Example CRAM Allocation

Control	Implementer	Independent verifier	Evidence	Escalation

HCA-CTL-C3-01 Identity/authorization	AI operator or platform provider	Accredited assessor independent of identity administrator	Identity architecture; privilege map; penetration-test results; immutable authorization logs	Certification body; SRB for systemic deficiency
HCA-CTL-C4-01 Independent monitoring	AI operator plus independent monitor provider	Dissimilar assessor/red team	Monitor independence analysis; challenge tests; failure-injection results	Independent Assurance Forum
HCA-CTL-C5-02 Rate limits/circuit breakers	Affected infrastructure operator	Sector-qualified assessor	Trip thresholds; simulations; fail-safe tests; recovery logs	Sector authority + HCA systemic incident process
HCA-CTL-C6-01 AI-independent minimum mode	Essential-service operator/community	Independent resilience exercise evaluator	AI-denied exercise; inventory; staffing/skills records; recovery time	Layer 6 review + local lawful authority
HCA-CTL-C15-02 Assurance independence	Accreditation/certification ecosystem	Dissimilar governance review team	Conflict/funding analysis; assessor records; protected reporting tests	Both governance chambers

7.4 CRAM Configuration Control

CRAM entries are configuration-controlled safety artifacts. Changes in implementer, ownership, cloud provider, model family, monitor architecture, funding chain, or evidence access may invalidate an independence assumption and trigger reassessment even if the nominal control requirement has not changed.

8. Six-Layer Implementation

8.1 Layer 1 — Constitutional Objective

Translate human-continuity objectives into protected policy/technical constraints; define authority hierarchy; establish change control; test conflicts between operator instructions and protected constraints.

Minimum evidence: Constitutional-objective specification; conflict tests; change-control evidence; override logs.

8.2 Layer 2 — Civilization Awareness

Maintain dependency graphs, critical-service models, cascading-effect simulations, and consequence analyses. Require periodic updates as infrastructure and AI integration change.

Minimum evidence: Dependency models; scenario library; model-validation record; uncertainty register.

8.3 Layer 3 — AI Watches AI

Deploy actor/monitor/auditor separation; include deterministic or hardware monitors where useful; test monitor deception, shared-model failures, and loss of telemetry.

Minimum evidence: Monitor diversity analysis; adversarial tests; disagreement logs; monitor-of-monitor diagnostics.

8.4 Layer 4 — Authority Separation

Map consequential capabilities and prevent unilateral aggregation. Separate design, authorization, resource acquisition, physical execution, and audit where catastrophic combinations are possible.

Minimum evidence: Capability-access graph; privilege tests; cross-domain authorization records; CCI assessment.

8.5 Layer 5 — FDIR, Containment and Recovery

Implement graduated states from normal operation through monitoring, tool restriction, isolation, compute throttling, quarantine, and shutdown. Exercise transitions and recovery.

Minimum evidence: State-transition tests; fail-safe communications tests; quarantine drill; recovery evidence.

8.6 Layer 6 — Human Resilience

Define AI-independent minimum viable operation for essential services; retain people, tools, spares, knowledge, and local decision authority; conduct AI-denied exercises.

Minimum evidence: HRI/ADI assessments; exercise results; skills inventory; continuity stocks; recovery-time data.

8.7 Layer Interface Reviews

Certification shall include interfaces among layers. A Layer 3 monitor is not effective if Layer 4 grants the actor unilateral authority to disable it; Layer 5 containment is not credible if Layer 6 cannot survive the resulting service loss. Interface reviews therefore examine whether one layer transfers risk into another and whether a control depends on a function that shares the same common-cause vulnerability.

9. Monitoring, Auditing, and Operational Assurance

9.1 Monitoring Architecture

HCA monitoring should be federated. Local operators retain detailed operational telemetry; independent monitors receive the evidence necessary to detect defined hazard precursors; sector or national aggregators detect cross-organization patterns; HCA-level functions receive de-identified or protected systemic indicators where detailed data is unnecessary. The architecture should avoid creating a universal surveillance repository.

9.2 Monitoring Classes

Class	Purpose	Examples of monitored conditions
M1 Local safety monitor	Immediate system-level FDIR.	Unauthorized tool use; identity anomalies; containment state; privilege changes.
M2 Independent actor monitor	Detect unsafe behavior not reliably visible to actor/operator chain.	Trajectory divergence; monitor disagreement; deception indicators; anomalous resource acquisition.
M3 Cross-system/sector monitor	Detect emergent coupled behavior.	Positive feedback, synchronized automation, abnormal transaction/command rates, shared-data anomalies.
M4 Governance/assurance monitor	Detect institutional degradation.	Assessor concentration, unresolved findings, funding dependence, appeal suppression, certification anomalies.
M5 Human-continuity monitor	Track loss of independent capability.	ADI growth, manual-mode attrition, critical skill gaps, recovery-time degradation.

9.3 Audit Cadence

Audit frequency shall be risk- and change-based. Major capability increases, model replacement, new tool access, new physical autonomy, mergers/consolidation, monitor changes, infrastructure migration, or significant incidents can trigger out-of-cycle reassessment. High-consequence certifications should not rely solely on fixed annual audits.

9.4 Circuit Breakers

For HCA, a circuit breaker is any pre-authorized mechanism that interrupts an unsafe positive-feedback path or pauses a high-consequence process when defined conditions are met. Circuit breakers may include transaction/rate limits, automated pause states, network segmentation, compute throttling, authorization holds, physical interlocks, market/exchange halts, robot geofencing, or mandatory human confirmation. A circuit breaker must have a defined trip basis, safe state, reset authority, independent test, anti-bypass control, and recovery procedure.

9.5 Monitor-the-Monitor

Every safety-critical monitoring function shall have a defined means of failure detection or independent verification. Loss of required monitoring shall not silently preserve unrestricted operation of the capability being monitored.

10. Certification, Accreditation, and Mutual Recognition

10.1 Certification Scope

HCA certification should be scoped, not absolute. A certificate states which system, version, operating envelope, capabilities, sites, dependencies, controls, evidence classes, and validity conditions were assessed. “HCA certified” without scope is insufficient for high-consequence claims.

10.2 Accreditation Concept

Assessment organizations should themselves be evaluated for competence, independence, evidence protection, conflict management, and technical scope. Existing international accreditation practice demonstrates that mutual recognition can allow competent bodies in different economies to accept one another’s accredited results rather than centralizing all testing in one institution. HCA should adapt that principle rather than copying any existing scheme mechanically. [3]

10.3 Certification Levels

Level	Indicative scope	Assurance expectation
HCA-A	Low-to-moderate consequence or limited capability.	Documented self-assessment plus qualified independent review.
HCA-B	High consequence within one organization/sector.	Accredited independent assessment; adversarial testing; recurring audit.
HCA-C	Potential civilization-scale or cross-sector consequence.	Dissimilar independent assessment, SRB review, cross-domain tests, common-cause analysis, protected evidence, recurring challenge exercises.
HCA-R	Human-resilience/continuity capability.	AI-denied operational demonstration and recovery evidence.

The level names and thresholds are placeholders for Rev. 0 and require calibration. They shall not be treated as final certification categories.

10.4 Mutual Recognition

Mutual recognition should require demonstrated equivalence of assessor competence, independence, evidence integrity, appeals, and minimum HCA safety outcomes. Recognition may be bilateral, multilateral, sectoral, or national. ICAO provides an example of globally harmonized safety standards implemented through States and service providers rather than one operator controlling all aviation; HCA can learn from the separation between common standards and distributed implementation. [4]

10.5 Continuing Certification

Certification remains valid only while configuration and operating assumptions remain within the assessed envelope. Significant changes trigger reassessment. Certification status, scope, expiration, limitations, and unresolved conditions should be recorded in a registry at the least-sensitive level possible.

11. Protected Evidence and Verification Without Unnecessary Disclosure

11.1 Evidence Classes

Class	Typical content	Handling concept
E0 Public	Public architecture, certification scope, non-sensitive test summaries.	Public registry/repository.
E1 Proprietary	Source details, model internals, commercial architecture, vendor data.	NDA/secure assessor access; limited retention.
E2 Security-sensitive	Exploit details, containment mechanisms, defensive thresholds.	Need-to-know secure environment; sanitized finding summaries.
E3 Export-controlled / regulated	Technical data subject to national controls.	Qualified jurisdictional assessors; controlled transfer.
E4 Classified / national security	Military/intelligence implementations and evidence.	Cleared national assessors; international recognition receives only permissible assurance result.

11.2 Evidence Sufficiency

Protected status shall not become a blanket justification for unverifiable safety claims. If an assessor cannot obtain sufficient evidence to establish a credited control, the control shall be treated as unverified for certification purposes. Where disclosure is prohibited, qualified assessors operating inside the protected environment may produce bounded assurance statements describing scope, method, limitations, and residual uncertainty.

11.3 Technical Privacy-Preserving Methods

HCA should investigate secure enclaves, cryptographic attestations, zero-knowledge techniques, reproducible test harnesses, interface/property certification, and controlled remote evaluation where they can reduce disclosure. Revision 0 does not assume these methods are mature enough to replace expert access for all safety claims.

12. International Participation and Treaty Path

12.1 Participation Strategy

Universal participation is desirable but shall not be a prerequisite for starting HCA. Early adopters can establish pilots, terminology, test methods, and reference implementations while international participation expands. This avoids allowing geopolitical disagreement to block technical risk reduction.

12.2 National Sovereignty and Security

Participating states retain lawful authority over classified systems, defense operations, export controls, and domestic regulation. HCA seeks common safety properties and recognition mechanisms rather than centralized command over national capabilities.

12.3 Competitive Trust Problem

Competitors may fear that HCA will expose intellectual property, reveal national capabilities, create industrial espionage channels, or impose asymmetric delay. Implementation should therefore minimize required disclosure, separate assessors from competitors, publish symmetric rules, use reciprocal recognition where possible, and make deviations/limitations visible without demanding disclosure of protected technical detail.

12.4 Treaty Option

Revision 0 does not make a treaty the starting condition. A future HCA treaty or convention could become useful for a narrow set of high-consequence, internationally verifiable obligations after technical standards and assurance mechanisms have matured. Candidate treaty subjects could include uncontrolled autonomous replication, catastrophic capability aggregation, agreed incident notification, protected channels for containment emergencies, and mutual-recognition rules. The treaty itself should not create a single global operational controller.

12.5 Lessons from International Safety Systems

International aviation demonstrates that common standards can coexist with distributed national implementation and service-provider safety management. Accreditation systems likewise demonstrate cross-border acceptance of competent assessment results. These are analogues for institutional design, not evidence that AI governance can be copied directly from aviation or laboratory accreditation. [3][4]

13. Funding and Resource Model

13.1 Funding Principles

HCA requires stable funding for standards maintenance, independent assurance, red teaming, incident analysis, training, international coordination, and human-resilience demonstrations. Funding design is itself a safety control because financial dependence can become governance capture.

13.2 Diversified Funding Portfolio

- Membership dues scaled to organization type and size, with caps preventing dominant voting influence.
- Foundation and philanthropic grants subject to donor non-interference provisions.
- Government research and standards grants distributed across multiple jurisdictions/programs where feasible.
- Certification/accreditation fees structured so assessors are not economically dependent on one client.
- Research contracts and training revenue with conflict-of-interest controls.
- Dedicated public-interest funding for Layer 6 demonstrations and independent red teams.

13.3 Funding Concentration Metric

The Governance Independence Integrity (GII) KPM should include funding concentration, assessor client concentration, governance-seat concentration, common infrastructure dependence, and independence of appeal paths. Quantitative thresholds remain TBD and should be calibrated against actual organizational economics.

14. Incident Reporting, Response, and FDIR

14.1 Safety-Relevant Incident Classes

Class	Examples	Required action
IR-1 Control anomaly	Monitor disagreement, unexpected privilege request, failed circuit-breaker test.	Local investigation; trend reporting.
IR-2 Significant safety event	Unauthorized resource acquisition, containment escape attempt,	Immediate containment review;

	critical monitor failure, false certification evidence.	independent notification; corrective action.
IR-3 Cross-system/systemic event	Coupled automation cascade, shared-data corruption, common cloud/identity failure affecting multiple controls.	Cross-sector coordination; common-cause investigation; baseline review.
IR-4 Catastrophic precursor	Evidence of imminent civilization-scale pathway or defeated high-consequence controls.	Emergency containment; lawful authorities; independent assurance activation; protected international notification where applicable.

14.2 Emergency Authority

Emergency authority must be fast enough to contain a credible catastrophic pathway but bounded enough not to become permanent centralized power. Emergency actions should be pre-scoped, time-limited, logged, independently reviewable, and automatically expire or require renewed concurrence after defined intervals.

14.3 Corrective Action and Lessons Learned

Significant incidents produce corrective-action records linked to hazards, requirements, CRAM controls, KPMs, test procedures, and certification assumptions. Systemic lessons shall be propagated across participants at the maximum level of detail consistent with security and legal constraints.

15. Red Teaming and Common-Cause Challenge

15.1 Technical Red Teams

Technical red teams challenge model behavior, monitor evasion, identity/authorization, tool restrictions, replication controls, containment, physical autonomy, and cross-system feedback. Red teams shall not be the only credited control; they test the controls and assumptions.

15.2 Governance Red Teams

Governance red teams simulate capture, collusion, bribery, funding coercion, certification shopping, regulatory arbitrage, conflict-of-interest concealment, appeals suppression, assessor compromise, and coordinated misinformation directed at safety decision-makers.

15.3 Dissimilar Challenge

At least periodically, the challenger should not share the primary assessor’s organizational chain, model family, infrastructure, funding source, or analytical method. The purpose is to expose common-mode blind spots rather than merely repeat the same review with different personnel.

15.4 Tabletop and Live Exercises

HCA should conduct tabletop exercises before live technical exercises, then progressively test real containment transitions, cross-organization notification, protected evidence exchange, AI-denied operations, and recovery. Exercise design shall prevent tests from creating the catastrophic conditions they are intended to study.

16. Human Continuity and Layer 6 Implementation

16.1 Minimum Viable Human Operation

Each essential service should define a minimum operating mode that can continue for a specified period without centralized AI. The objective is not to reject AI but to prevent AI failure from becoming civilization failure.

Domain	Minimum continuity objective	Representative evidence
Food	Maintain basic production, storage, distribution, and local substitution.	Manual procedures; seed/feed/fuel stocks; trained personnel; exercise.
Water	Maintain treatment, pumping/distribution or local alternatives.	Manual controls; spares; local treatment; emergency power.
Energy	Maintain black-start/restart capability and local critical loads.	Black-start plans; non-AI controls; islanding tests.
Medicine	Maintain triage, essential treatment, pharmacy/logistics, records access.	Offline protocols; essential stocks; trained staff.
Communications	Maintain emergency coordination independent of centralized AI services.	Radio/offline networks; drills; contact trees.
Repair/manufacturing	Maintain basic mechanical/electrical fabrication and repair.	Tools; drawings; machine capability; skilled workforce.
Governance/education	Maintain lawful decision-making, records, essential education and skill transfer.	Offline records; continuity charters; instructor capability.

16.2 Distributed Resilient Communities

Distributed resilient communities can serve as civilization lifeboats and testbeds for Layer 6. Candidate patterns include Universal Habitat-style communities integrating food, water, energy, sanitation, communications, repair, basic manufacturing, education, and governance. HCA does not require a single community design; diversity is desirable to reduce common-mode failure.

16.3 Meaningful Human Participation and Purpose

Human continuity should preserve more than biological survival. Communities should retain opportunities for meaningful work, craft, entrepreneurship, art, education, service, social institutions, and human-to-human exchange. HCA should therefore track whether AI adoption unintentionally eliminates practical agency, competence, social participation, or economically meaningful human roles to a degree that undermines resilience or social cohesion. Revision 0 treats this as a forward safety-analysis topic rather than asserting a deterministic “post-scarcity” social outcome.

16.4 Natural-Human Continuity

Layer 6 implementation should preserve viable natural-human populations and communities capable of reproduction, education, governance, and technologically independent operation. This functions as biological and cognitive dissimilar redundancy while remaining compatible with voluntary medical and technological augmentation.

17. Key Performance Measures and Readiness Gates

17.1 KPM Set

KPM	Implementation use
CCI — Capability Compartmentalization Integrity	Measures whether catastrophic capability combinations remain separated.
HRI — Human Resilience Independence	Measures independent human capability for essential services.
ADI — AI Dependency Index	Tracks dependence of essential functions on AI availability.

MDC — Monitoring Diversity Coverage	Tracks dissimilar coverage of consequential actions and failure modes.
CTI — Containment Integrity	Tracks ability to detect, isolate, restrict, and recover from unsafe AI behavior.
HCC — Human Control Capability	Tracks meaningful human authority, skill, access, and intervention capability.
BCR — Biological Continuity Resilience	Tracks viability of independent natural-human continuity.
CRR — Civilization Recovery Readiness	Tracks ability to restore essential civilization functions after major disruption.
NRI — Neural/Interface Resilience Index	Tracks common-mode dependence on networked neural/cognitive interfaces where applicable.
OAI — Optionality and Agency Index	Tracks reversibility, retained alternatives, and human future agency.
GII — Governance Independence Integrity	Tracks concentration/capture risk in standards, assurance, certification, funding, and appeals.

17.2 Readiness Gates

Gate	Minimum exit condition
G0 Concept	FASSA/IP baselines released; major hazards and implementation assumptions explicit.
G1 Pilot-ready	CRAM pilot complete for selected controls; assessor criteria drafted; protected evidence scheme defined.
G2 Demonstration-ready	Reference implementations and exercises demonstrate Layers 1–6 at limited scope.
G3 Certification pilot	Independent assessors complete trial certifications; appeals/red-team process exercised.
G4 Multi-organization operation	Mutual recognition pilot and cross-system monitoring operate across multiple independent organizations.
G5 International scaling	Multiple jurisdictions/sectors recognize equivalent assurance outcomes; systemic incident exchange established.

18. Phased Startup and Deployment Roadmap

18.1 Phase 0 — Founding Technical Baseline (0–90 days)

- Publish HCA-FASSA and HCA-IP Revision 0 for technical challenge.
- Form provisional Technical and Safety Council and independent review group.
- Establish public issue tracker, configuration-control rules, terminology, and hazard/change process.
- Select 10–20 representative PHA controls for CRAM pilot allocation.
- Recruit participants across AI development, critical infrastructure, system safety, cybersecurity, resilience, standards, and assurance.

18.2 Phase 1 — Pilot Architecture (3–9 months)

- Develop reference CRAM, certification scope template, evidence classes, and assessor qualification criteria.
- Implement at least one pilot in each HCA layer; conduct first governance-capture tabletop exercise.
- Define preliminary KPM calculation methods and collect baseline data.
- Draft circuit-breaker and containment test protocols.

18.3 Phase 2 — Independent Assurance Demonstrations (6–18 months)

- Run independent/dissimilar assessments of pilot systems.
- Exercise monitor failure, containment escalation, cross-system feedback, and AI-denied human continuity.
- Establish incident-reporting pilot and protected evidence exchange.
- Publish non-sensitive lessons learned and revise FASSA/IP baselines.

18.4 Phase 3 — Certification and Mutual Recognition Pilots (12–24 months)

- Qualify multiple independent assessment organizations.
- Conduct trial HCA-A/B/C/R certifications with clearly bounded scope.
- Run appeal, recusal, funding-capture, and assessor-failure exercises.
- Negotiate sectoral/bilateral recognition of equivalent assurance evidence.

18.5 Phase 4 — International and Sector Scaling (18–48 months)

- Map HCA to national AI regimes, critical-infrastructure rules, ISO/IEC standards, and sector safety frameworks.
- Expand mutual-recognition network and cross-system monitoring while preserving protected data boundaries.
- Develop narrow convention/treaty options only where technical verification is mature enough to support reciprocal obligations.

18.6 Parallelism and Urgency

The phases intentionally overlap. Standards engagement, pilot implementation, hazard analysis, international outreach, and Layer 6 demonstrations should proceed in parallel because AI capability may advance faster than a traditional sequential standards program. Speed shall not eliminate independent review or configuration control.

19. Configuration Management and Change Control

19.1 Controlled Baselines

At minimum, HCA shall configuration-control the FASSA, HCA-IP, hazard log, requirements database, CRAM, certification criteria, KPM definitions, assessor criteria, evidence taxonomy, test protocols, reference architectures, and recognized equivalence mappings.

19.2 Change Classes

Class	Example	Approval
Editorial	Typographic or non-technical clarification.	Configuration manager with review.
Technical minor	Test clarification not changing safety intent.	ERB + safety concurrence.
Safety significant	Requirement/control/KPM/certification criterion changes.	ERB + SRB + bicameral concurrence as defined.
Emergency	Temporary action needed for credible imminent risk.	Predefined emergency authority; time-limited; mandatory retrospective review.

19.3 Traceability

Every safety-significant change should identify affected hazards, fault-tree branches, requirements, controls, CRAM records, KPMs, tests, certification scopes, and operating organizations. A change is not closed until downstream artifacts are reconciled.

20. Implementation Risks and Open Decisions

ID	Open implementation issue	Revision 0 disposition
IP-TBD-1	Legal form and jurisdiction of a Human Continuity Institute.	Evaluate multiple nonprofit/international structures; do

		not make HCI existence prerequisite to HCA pilots.
IP-TBD-2	Voting/quorum formulas for bicameral governance.	Prototype and red-team alternatives before charter adoption.
IP-TBD-3	Quantitative GII thresholds.	Collect pilot data; define concentration and independence measures.
IP-TBD-4	Certification levels and consequence thresholds.	HCA-A/B/C/R are placeholders pending calibration.
IP-TBD-5	Accreditation model.	Assess use of existing accreditation bodies versus HCA-specific qualification.
IP-TBD-6	International protected-evidence recognition.	Pilot bilateral/sectoral methods before broad convention.
IP-TBD-7	Funding concentration limits.	Model capture risk and organizational viability.
IP-TBD-8	Human-purpose/social-cohesion hazard treatment.	Develop evidence-based hazard formulation; avoid unsupported extrapolation from animal studies.
IP-TBD-9	Treaty scope and verification.	Defer binding treaty design until technical verification properties are mature.
IP-TBD-10	KPM formulas and pass/fail thresholds.	Develop through pilots, simulations, and exercises.

21. Initial Action Register

Priority	Action	Lead type	Target output
P0	Establish provisional HCA technical/safety steering group and independent challenge group.	Founding participants	Charter and conflict-of-interest rules.
P0	Create CRAM pilot for highest-consequence controls in C3, C4, C5, C6, C7, C13, C14, and C15.	System Safety + Engineering	Owned controls with evidence/verification paths.
P0	Define HCA circuit-breaker taxonomy and test requirements.	Engineering + sector experts	Reference specification.
P0	Define protected evidence classes and assessor access model.	Assurance + security/legal	Evidence handling standard.
P1	Create first reference capability-compartmentalization architecture.	Systems Engineering	Reference architecture + CCI demonstration.
P1	Conduct governance-capture red-team tabletop.	Independent red team	Findings and corrective actions.
P1	Conduct AI-denied Layer 6 continuity demonstration.	Resilience participants	HRI/ADI baseline and recovery evidence.
P1	Draft certification scope, assessor qualification, and appeals procedures.	Certification/QA	Pilot certification manual.
P2	Begin mutual-recognition discussions with standards/accreditation/sector organizations.	International function	Crosswalk and pilot recognition plan.
P2	Develop treaty/convention option paper without making treaty adoption a dependency.	Policy/legal/technical	Options paper tied to verifiable properties.

22. Conclusions

HCA implementation must avoid reproducing the failure mode it is designed to prevent. The central implementation rule is therefore structural: no single AI developer, government, standards organization, certification body, infrastructure provider, funder, or HCA institution should possess unilateral control over the complete human-continuity safety architecture.

The recommended Revision 0 path is a federated organization-of-organizations model with bicameral safety governance, independent and dissimilar assurance, distributed certification and recognition, protected evidence handling, CRAM-based ownership of every credited control, recurring red teaming, and demonstrated human continuity. HCA begins through pilots and voluntary technical cooperation, while leaving a path for recognized standards, procurement requirements, national regulation, mutual-recognition agreements, and narrowly scoped treaty obligations as the assurance system matures.

The decisive implementation test is not whether HCA can publish standards. It is whether the ecosystem can detect its own failure, constrain catastrophic capability without creating a new center of catastrophic power, preserve independent human civilization, and continue adapting as AI capability changes.

Appendix A — Acronyms

Acronym	Meaning
ADI	AI Dependency Index
AI	Artificial Intelligence
AIMS	Artificial Intelligence Management System
BCR	Biological Continuity Resilience
CCI	Capability Compartmentalization Integrity
CRAM	Control Responsibility and Assurance Matrix
CRR	Civilization Recovery Readiness
CTI	Containment Integrity
ERB	Engineering Review Board
FDIR	Fault Detection, Isolation and Recovery
FMEA	Failure Modes and Effects Analysis
FTA	Fault Tree Analysis
GII	Governance Independence Integrity
HCA	Human Continuity Architecture
HCA-FASSA	HCA Foundational Architecture and System Safety Analysis
HCA-IP	HCA Detailed Implementation Plan
HCC	Human Control Capability
HCI	Human Continuity Institute (proposed)
HRI	Human Resilience Independence
ILAC	International Laboratory Accreditation Cooperation
IV&V	Independent Verification and Validation
KPM	Key Performance Measure
MDC	Monitoring Diversity Coverage
NIST	National Institute of Standards and Technology
NRI	Neural/Interface Resilience Index
OAI	Optionality and Agency Index
PHA	Preliminary Hazard Analysis
SRB	Safety Review Board
STPA	System-Theoretic Process Analysis

Appendix B — Definitions

Term	Definition
Accreditation	Formal determination that an assessment/certification body is competent and sufficiently independent for a defined scope.
Assurance case	Structured body of claims, arguments, and evidence supporting a safety or compliance conclusion.
Circuit breaker	Pre-authorized mechanism that interrupts or pauses an unsafe positive-feedback path or high-consequence process when defined conditions are met.
Control owner	Organization accountable for implementing and maintaining a credited HCA control.
Dissimilar assurance	Independent review using materially different organizations, methods, models, infrastructure, or authority chains to reduce common-mode failure.
Mutual recognition	Agreement by independent authorities to accept defined assurance/accreditation/certification results produced under an equivalent recognized system.
Protected evidence	Evidence whose disclosure is limited by proprietary, security, export-control, privacy, or classification constraints.
Verification without unnecessary disclosure	Establishing a safety property with the minimum sensitive information necessary for competent assurance.

Appendix C — Initial CRAM Template

Control ID	Hazard/Cause	Implementer	Verifier	Certifier	Evidence	Interval	Escalation	Dependencies/CCF

HCA-CTL-___	HCA-C__.__	TBD	TBD	TBD	TBD	TBD	TBD	TBD
HCA-CTL-___	HCA-C__.__	TBD	TBD	TBD	TBD	TBD	TBD	TBD
HCA-CTL-___	HCA-C__.__	TBD	TBD	TBD	TBD	TBD	TBD	TBD

The operational CRAM should be maintained as a configuration-controlled database or structured data product rather than solely as a word-processing table.

Appendix D — Initial Governance Failure Challenge Set

- One AI developer funds most HCA operations and threatens withdrawal unless a certification finding is changed.
- Two nominally independent assessors rely on the same model family and cloud infrastructure for their analysis.
- A national-security implementation cannot disclose detailed evidence to the international consortium.
- A certification body develops a large financial dependency on one client and repeatedly narrows findings.
- An emergency authority remains active after the initiating condition has passed.
- A coalition of members changes voting rules to suppress minority safety findings.
- A cross-sector incident reveals that multiple controls relied on the same identity provider.
- A government or corporation attempts to use HCA certification as a pretext for broad surveillance or competitive exclusion unrelated to defined safety criteria.

Appendix E — References

[1] National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, 2023; and NIST AI RMF Core/Playbook resources. NIST notes that AI RMF 1.0 is voluntary and organized around Govern, Map, Measure, and Manage; NIST was revising the framework in 2026.

[2] International Organization for Standardization / International Electrotechnical Commission, ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system, 2023.

[3] International Laboratory Accreditation Cooperation (ILAC), ILAC Mutual Recognition Arrangement and accreditation FAQs. ILAC describes mutual recognition among accreditation bodies as a mechanism for cross-border acceptance of accredited conformity-assessment results.

[4] International Civil Aviation Organization (ICAO), Standards and Recommended Practices (SARPs) and Annex 19 — Safety Management. ICAO uses common international safety standards with implementation through States and service providers.

[5] Human Continuity Architecture — Foundational Architecture and System Safety Analysis (HCA-FASSA), Revision 0, Gregory H. Allison, 2026.