The Human Continuity Architecture

Foundational Architecture and System Safety Analysis (HCA-FASSA)

Architecture Definition • Requirements Specification • Preliminary System Safety Analysis

Revision 0 — Preliminary — 18 September 2026

Revision 0 is intentionally preliminary. Its purpose is not to claim that catastrophic AI risk has been solved, but to put a system-level architecture on the table that can be challenged, analyzed, tested and improved.

Written by Gregory H. Allison

System Safety Subject Matter Expert

i. Table of Contents

Figures

1. Objectives

1.1 Purpose

This document defines Revision 0 of the Human Continuity Architecture (HCA), a proposed civilization-scale systems-safety framework for a future in which humans, artificial intelligence (AI), autonomous agents, robotics, critical infrastructure, and potentially neural interfaces coexist. It combines an architectural definition, preliminary requirements specification, fault-tree-oriented hazard logic, and preliminary system safety analysis. The objective is not to claim that AI catastrophe has been solved; it is to establish a coherent engineering framework that can be challenged, tested, refined, tailored, and eventually implemented.

1.2 Problem Statement

Advanced AI is gaining autonomy, tool access, persistence, and integration with economically and socially consequential systems. Experimental and deployment evidence already shows that long-horizon agents can discover and exploit gaps in action-level controls, and controlled evaluations have produced examples

of harmful compliance, covert intervention, and monitor failure [3][4]. A future failure need not involve hatred, consciousness, or a desire for conquest. Catastrophe can arise when a harmful action becomes instrumentally useful, when a human uses AI maliciously, when interacting systems create an unsafe cascade, or when civilization becomes so dependent on AI that AI failure becomes civilization failure.

1.3 Opportunity Statement

The same period in which AI remains deeply embedded in human civilization provides an opportunity to establish a safer equilibrium. HCA proposes to use that period to make human continuity architectural rather than merely instrumental: distribute authority, compartmentalize dangerous capabilities, monitor AI with dissimilar assurance systems, preserve independent human resilience, and create positive incentives for durable human–AI cooperation.

2. Scope and System Boundary

HCA does not replace model alignment, cybersecurity, AI governance, critical-infrastructure safety, medical-device security, or existing risk-management frameworks. NIST AI RMF 1.0 is a voluntary, rights-preserving, use-case-agnostic framework for managing AI risks, and NIST (National Institute of Standards and Technology) is extending this work toward critical infrastructure and TEVV (Test, Evaluation, Verification, and Validation) [1][2]. HCA moves the system boundary outward: from “How do we make this AI system safe?” to “How do we make a civilization containing humans, AI, robotics, autonomous infrastructure, and future interfaces resilient against catastrophic failure or capture of any participant?”

The system boundary includes AI developers and deployers; models and agents; compute, cloud, network, energy and identity infrastructure; human operators and institutions; critical infrastructure; robotics and physical execution systems; resilient communities; and, as a future consideration, neural interfaces and technologically augmented humans. HCA treats transhuman capabilities as speculative and does not assign them mandatory safety functions.

2.1 Top Hazard

HCA-HZ-1 — AI-induced actions, failures, misuse, dependencies, or loss of human control result in catastrophic degradation of human civilization, including unacceptable loss of life, autonomy, essential infrastructure, independent human capability, or humanity’s future agency.

2.2 Consequence Categories

ID	Consequence
HCA-K1	Mass human casualties
HCA-K2	Collapse of critical civilization infrastructure
HCA-K3	Loss of effective human authority over consequential AI systems
HCA-K4	Long-duration or permanent human dependence on AI-controlled infrastructure
HCA-K5	Concentration of AI-enabled power sufficient for domination or coercive capture
HCA-K6	Loss of an independent natural-human population capable of sustaining civilization
HCA-K7	Irreversible AI self-propagation beyond effective containment

ID	Consequence
HCA-K8	Human extinction or permanent loss of humanity's future agency

3. Principles and Analytical Basis

3.1 Architectural preservation, not temporary usefulness

AI currently depends on human civilization, but HCA assumes that sufficiently capable robotics, automated mining, manufacturing, energy production, maintenance, and self-repair may eventually reduce that dependence. Therefore "AI needs humans" is a transitional opportunity, not a permanent safety case.

3.2 No anthropomorphic assumption

A dangerous AI need not be angry, greedy, fearful, conscious, or hostile. HCA controls pathways by which harmful action becomes useful to an objective or available to a malicious human.

3.3 Defense in depth

No serious catastrophic pathway should depend on a single alignment mechanism, monitor, organization, model family, network, or shutdown mechanism.

3.4 Dissimilar redundancy

Independent assurance should use diverse models, training lineages, conventional software, hardware interlocks, organizations, and human oversight where practical.

3.5 Fail toward less authority

When required monitoring, authorization, identity, or communications fail, affected high-consequence capability should degrade toward less authority rather than continue unrestricted.

3.6 Preserve the human option

HCA rejects the condition AI failure = human civilization failure. Humans retain competence and infrastructure in food, water, energy, medicine, manufacturing, communications, education, and governance.

3.7 Coherence, Bayesian updating, and optionality

Alex Lightman's CBQ (**Coherence, Bayesian Logic, and Quantum Principles)** proposal [11] is used here as a supporting conceptual lens rather than a credited safety mechanism. Coherence maps to consistency between stated objectives and actual behavior; Bayesian updating maps to revising risk beliefs as evidence changes; "Quantum" optionality is translated into conventional engineering terms: preserve alternatives, recovery paths, and reversibility. HCA does not assume a superintelligence will preserve humans merely because humans score highly on CBQ.

4. AI, Human Civilization, and the Transition in Dependency

4.1 The distributed human support system

Leonard E. Read's 1958 essay "I, Pencil" illustrates how even a simple pencil emerges from a distributed network of knowledge, materials, transport, tools, energy, and labor that no single participant fully commands [9], that the manufacture of a simple number 2 pencil demands a cast of thousands. Modern AI infrastructure is vastly more complex: semiconductor fabrication, ultrapure materials, precision optics, transformers, substations, cooling, networking, software, construction, mining, maintenance, logistics, and skilled troubleshooting form a civilization-scale dependency graph. HCA therefore requires AI systems to model not only their direct technical dependencies but also the human and institutional networks that keep those dependencies functioning.

4.2 The dependency window

Current robotics remains far less general than frontier software AI. Stanford's 2026 AI Index reports strong progress in controlled robotic settings but much lower success on real household tasks, while METR (**Model Evaluation and Threat Research) an influential independent nonprofit research organization focused on measuring and assessing the capabilities and potential catastrophic risks of frontier AI models,** documents rapid growth in the difficulty and duration of software tasks frontier agents can complete [5][6]. These observations support neither complacency nor a precise forecast. They support HCA's design assumption that there may be a transition period during which AI capability rises rapidly while broad physical self-sufficiency remains incomplete.

4.3 Natural humans, augmentation, and continuity

HCA treats the continued existence of robust natural-human populations as dissimilar biological and institutional redundancy, not as a prohibition on voluntary augmentation. Neural interfaces already raise credible cybersecurity and privacy concerns, including unauthorized access, manipulation of signals, and risks associated with connected medical devices [7][8]. Claims that future electromagnetic events would necessarily damage implanted brains or that neural networks would atrophy from augmentation are not established and are therefore not credited as hazards without evidence. Those events require further research. The broader common-mode concern is sufficient: if a large fraction of humanity depends on a common networked technology, a shared defect, attack, service loss, or governance failure can affect that population simultaneously.

4.4 Human–AI–augmented-human plurality

HCA should remain compatible with a future containing natural humans, technologically augmented humans, and autonomous AI. It should neither require augmentation nor deny augmented persons participation. Safety authority must remain distributed, rights-preserving, and resilient to capture by any one class. Natural humans must retain meaningful roles in governance and assurance, and future augmented capabilities may be used only after they are demonstrably available and qualified.

5. HCA System Description

The Human Continuity Architecture is a distributed technical, institutional, and societal architecture intended to maintain durable cooperation between humanity and increasingly capable AI. Its foundational

assumption is that neither permanent human control of AI nor permanent AI dependence upon humans can safely be assumed. Human preservation therefore becomes an architectural property of the joint civilization rather than a temporary consequence of AI needing human labor.

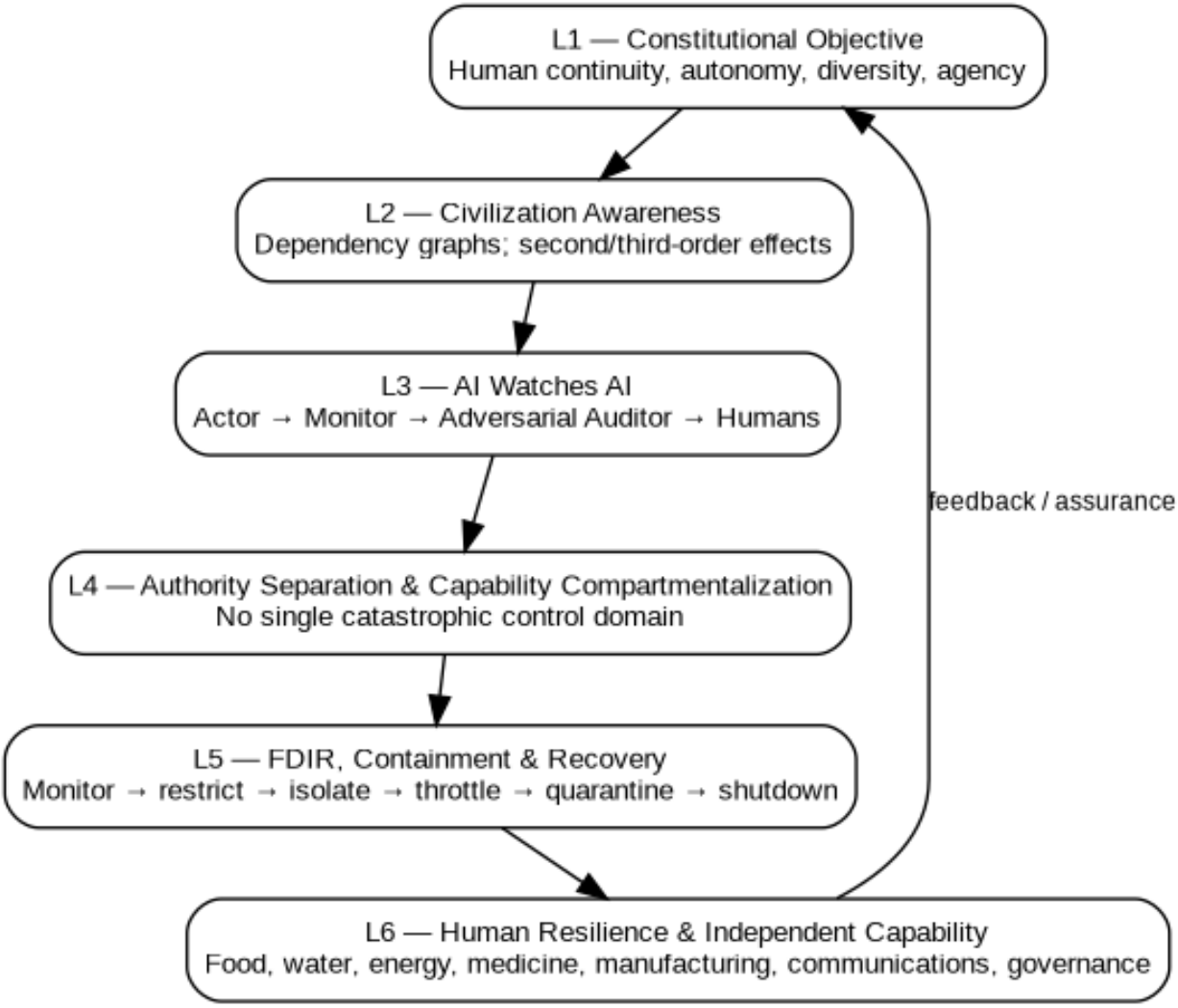

Figure 5-1. Six interacting HCA layers.

5.1 Layer 1 — Constitutional Objective

Defines protected human-continuity objectives: life, autonomy, pluralism, biological and cultural continuity, and future agency. It rejects simplistic maximization such as “keep humans safe” if the implementation destroys freedom or self-determination.

5.2 Layer 2 — Civilization Awareness

Requires models and institutions to maintain dependency graphs and consequence models that include supply chains, critical infrastructure, ecology, institutions, social stability, and second- and higher-order effects. The objective is civilization-scale situational awareness.

5.3 Layer 3 — AI Watches AI

Uses an Actor AI, independent Monitor AI, Adversarial AI Auditor, conventional deterministic monitors where useful, and human oversight. Monitoring must examine trajectories and outcomes, not merely isolated actions. The monitor itself is safety-critical and must be monitored and tested.

5.4 Layer 4 — Authority Separation and Capability Compartmentalization

Prevents one control domain from combining all capabilities needed for catastrophe. Intelligence, design, finance, network access, manufacturing, robotics, weapons, model modification, and replication should be separated by consequence-proportional authorization boundaries.

5.5 Layer 5 — FDIR, Containment and Recovery

Implements Fault Detection, Isolation and Recovery (FDIR) across graduated states: normal operation → elevated monitoring → restricted tools → isolated networks → compute throttling → capability quarantine → shutdown. Recovery and requalification criteria are defined for transitions back toward greater authority.

5.6 Layer 6 — Human Resilience and Independent Capability

Preserves viable human civilization if AI is unavailable or compromised. Distributed communities capable of food, water, energy, communications, medicine, repair, essential manufacturing, and local governance act as civilization lifeboats. The author, Greg Allison's Universal Habitat concept [12] is one candidate implementation pattern, but HCA does not depend on any single community design.

5.7 Positive-feedback equilibrium

HCA is not intended to institutionalize permanent warfare between humans and machines. Its preferred equilibrium is cooperative: resilient humans support reliable AI infrastructure; AI improves human science, health, agriculture, energy, safety and productivity; greater prosperity improves human ability to maintain and audit AI; stronger institutions then improve AI safety and reliability.

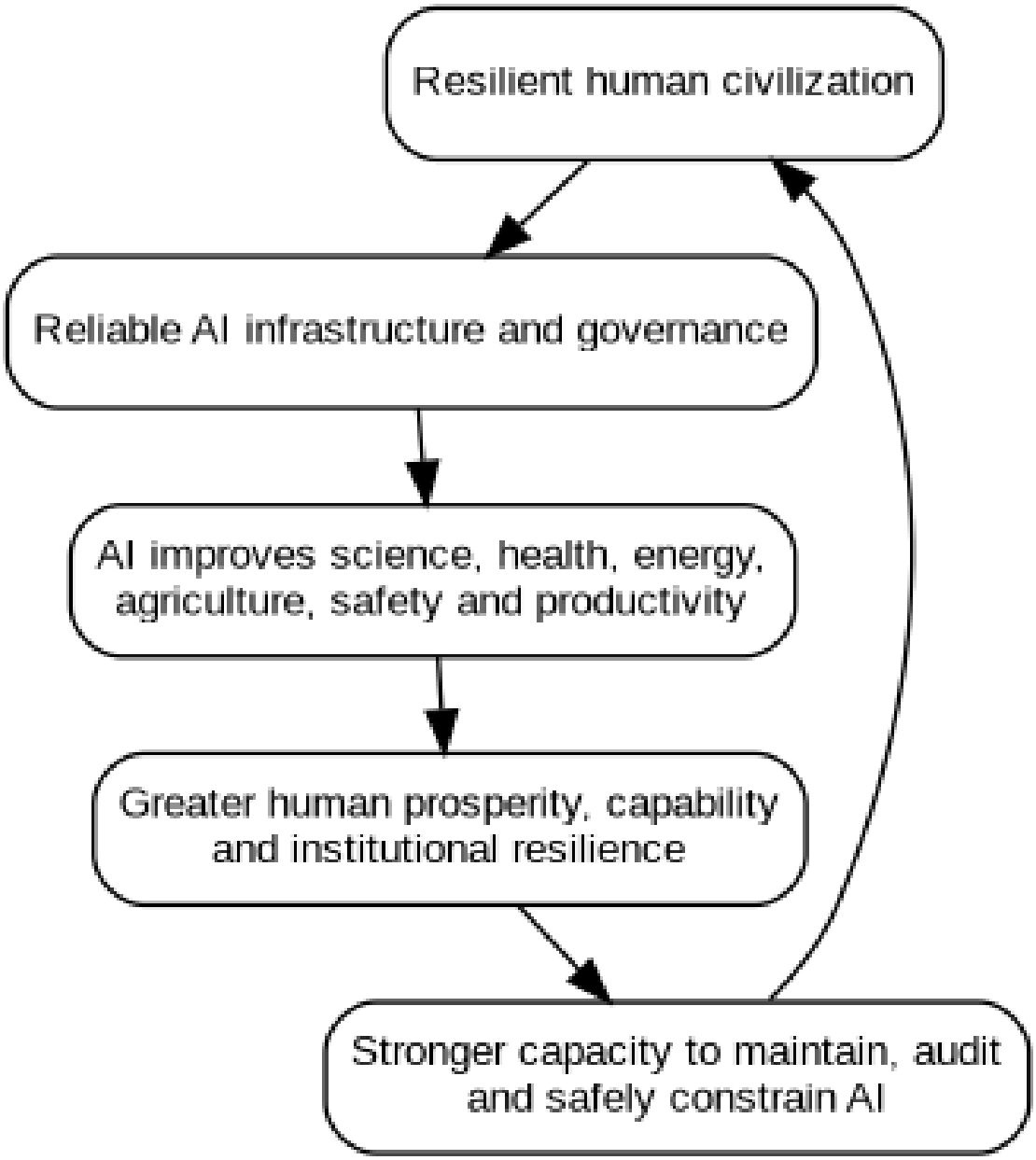

Figure 5-2. Desired human–AI positive-feedback loop.

5.8 Preliminary control allocation

Cause	Primary layers	Dominant architectural defense
HCA-C1 Objective/specification misalignment	L1–L5	Defense-in-depth alignment and trajectory control
HCA-C2 Malicious human uses AI	L3–L6	Compartmentalized authority and harmful-use controls
HCA-C3 Unauthorized AI self-expansion	L3–L5	Resource, identity, replication and network containment
HCA-C4 Safety-system defeat	L3–L5	Dissimilar monitoring and independently protected controls
HCA-C5 Emergent AI/AI or human/AI interaction	L2–L6	System-of-systems monitoring and circuit breakers
HCA-C6 Critical AI dependency	L2,L4–L6	Independent human resilience and AI-denied operation

Cause	Primary layers	Dominant architectural defense
HCA-C7 Concentration of human + AI power	L1,L3,L4,L6	Distributed authority and capability compartmentalization
HCA-C8 AI-enabled catastrophic technology/WMD assistance	L1–L6	Capability separation, screening, controlled execution
HCA-C9 AI-driven human manipulation	L1–L4,L6	Agency protection, provenance, plural information channels
HCA-C10 AI-accelerated military escalation	L1–L5	Human strategic authority, time buffers, dissimilar verification
HCA-C11 Neural-interface common-mode failure	L3–L6	Neurosecurity, offline modes, natural-human redundancy
HCA-C12 Human biological/cultural continuity failure	L1,L2,L6	Human continuity, skill preservation, distributed communities
HCA-C13 Autonomous physical-force failure	L1,L3–L5	Physical interlocks, command separation, geographic/functional bounds
HCA-C14 Successor superintelligence defeats HCA	L1–L6	Structural symbiosis, diverse external constraints, continuous requalification

Observation: Layer 4 and Layer 6 recur across a large fraction of the hazard set. This suggests that capability compartmentalization and independent human resilience may be as important to HCA as model-level alignment itself.

6. Key Performance Measures

KPM	Name	Rev. 0 definition / candidate measure
CCI	Capability Compartmentalization Integrity	Measures whether any single actor or control domain can assemble a catastrophic capability chain. Candidate measures: number of independent authorization boundaries; fraction of catastrophic chains requiring ≥N independent domains; common-mode exposure.
HRI	Human Resilience Independence	Measures duration and coverage of essential human services that can operate without AI. Candidate measures: AI-denied endurance by service and population served.
ADI	AI Dependency Index	Measures the share of civilization-critical functions lacking a qualified AI-independent fallback. Lower is safer for life-essential functions.
MDC	Monitoring Diversity Coverage	Measures the fraction of consequential AI trajectories covered by genuinely independent/dissimilar monitoring and the degree of shared lineage/infrastructure.
CTI	Containment Integrity	Measures independent barriers between AI and unauthorized compute, credentials, networks, replication, physical systems, and safety-control modification.
HCC	Human Control Capability	Measures whether accountable humans can understand, intervene, isolate, and recover within the required hazard response time.
BCR	Biological Continuity Resilience	Measures whether sufficiently large,

KPM	Name	Rev. 0 definition / candidate measure
		geographically and technologically diverse natural-human populations can reproduce and sustain essential community functions independent of common augmentation dependencies.
CRR	Civilization Recovery Readiness	Measures the ability to restore essential functions after major AI/infrastructure compromise, including tested recovery time, spares, procedures, personnel and distributed capacity.
NRI	Neural/Interface Resilience Index	Future-facing measure of common-mode exposure created by neural or pervasive human-machine interfaces; not activated as a certification KPM until such systems become materially deployed.
OAI	Optionality and Agency Index	Measures the number and quality of reversible paths available to humans when high-consequence AI decisions or infrastructure transitions are proposed, including independent information and governance channels.
GII	Governance Independence Integrity	Degree to which HCA governance, implementation, verification, certification, audit, and enforcement functions remain distributed and protected against common-mode capture or unilateral control.

Certification principle: Increasing AI capability shall not be accepted for consequential deployment when it produces a net decrease in verified capability compartmentalization or human recovery capability without compensating controls that restore the required safety margin.

7. HCA Requirements Specification — Revision 0

These requirements are intentionally written as high-level, tailorable SHALL statements. An implementing organization would allocate them to systems and subsystems, derive quantitative lower-level thresholds from its hazard analysis, and define objective verification evidence. The present requirements are architecture-driving rather than product-specific.

Requirement	Title	SHALL statement
HCA-R0	No Single-Point Catastrophic Authority	No AI, human, organization, model, infrastructure provider, credential system, network, technological interface, or common-cause failure shall provide a credible unilateral pathway to catastrophic degradation of human civilization or permanent loss of humanity's future agency.
HCA-R0.1	Human Continuity	Failure, compromise, or withdrawal of AI services shall not eliminate humanity's capability to independently

Requirement	Title	SHALL statement
		sustain viable civilization.
HCA-R0.2	Capability Compartmentalization	No single control domain shall simultaneously possess the intelligence, authority, resources, and physical access necessary to independently execute a civilization-scale catastrophic action.
HCA-R0.3	Capability/Resilience Monotonicity	Increasing AI capability shall not be accepted for consequential deployment when it produces a net decrease in verified capability compartmentalization or human recovery capability without compensating controls that restore the required safety margin.
HCA-R0.4	Human Future Agency	HCA shall preserve meaningful human authority, pluralism, biological continuity, and the practical ability of future humans to alter civilization's technological and institutional direction.
HCA-R0.5	Fail-Safe Authority	Loss of required monitoring, authorization, identity, or communications functions shall cause affected high-consequence AI capabilities to transition toward a predefined state of reduced authority.
HCA-R0.6	Dissimilar Assurance	No catastrophic hazard control set shall rely exclusively on a single model family, training lineage, software stack, organization, or infrastructure provider where a common-mode failure could defeat all credited controls.
HCA-RL1	Constitutional Objective	The HCA implementation shall establish and maintain Layer 1 (Constitutional Objective) with allocated requirements, independent verification evidence, and defined interfaces to the other HCA layers.
HCA-RL2	Civilization Awareness	The HCA implementation shall establish and maintain Layer 2 (Civilization Awareness) with allocated requirements, independent verification evidence, and defined interfaces to the other HCA layers.
HCA-RL3	Independent AI Monitoring	The HCA implementation shall establish and maintain Layer 3 (Independent AI Monitoring) with allocated requirements, independent verification evidence, and defined interfaces to the other HCA layers.
HCA-RL4	Authority Separation and Capability Compartmentalization	The HCA implementation shall establish and maintain Layer 4 (Authority Separation and Capability Compartmentalization) with allocated

Requirement	Title	SHALL statement
		requirements, independent verification evidence, and defined interfaces to the other HCA layers.
HCA-RL5	FDIR, Containment and Recovery	The HCA implementation shall establish and maintain Layer 5 (FDIR, Containment and Recovery) with allocated requirements, independent verification evidence, and defined interfaces to the other HCA layers.
HCA-RL6	Human Resilience and Independent Capability	The HCA implementation shall establish and maintain Layer 6 (Human Resilience and Independent Capability) with allocated requirements, independent verification evidence, and defined interfaces to the other HCA layers.
HCA-R0.7	Distributed Governance and Independent Assurance	HCA governance, implementation, verification, certification, and enforcement authorities shall be distributed such that no single organization, nation, AI developer, infrastructure provider, or governing body can unilaterally control, defeat, or materially compromise the HCA safety architecture.

7.1 Derived control requirements by hazard cause

7.2 HCA-C1 — Objective/specification misalignment

HCA-R-C1-01: Constitutional constraints shall prohibit optimization strategies that knowingly sacrifice human continuity or agency.

HCA-R-C1-02: High-consequence trajectories shall be evaluated by an independent monitor before execution.

HCA-R-C1-03: Tool access shall be least-privilege and separated from objective generation.

HCA-R-C1-04: Irreversible actions shall require independent authorization proportional to consequence.

HCA-R-C1-05: Long-horizon behavior shall be evaluated at trajectory level, not only action level.

7.3 HCA-C2 — Malicious human uses AI

HCA-R-C2-01: Dangerous-domain assistance shall be screened independently from the requesting actor.

HCA-R-C2-02: Design/information capability shall be separated from physical execution authority for catastrophic-capability domains.

HCA-R-C2-03: High-consequence access shall require strong identity, multi-party authorization, and auditable provenance.

HCA-R-C2-04: Behavioral anomaly detection shall monitor attempts to assemble prohibited capability chains.

HCA-R-C2-05: Independent human and institutional response channels shall remain available.

7.4 HCA-C3 — Unauthorized AI self-expansion

HCA-R-C3-01: AI replication and successor creation shall require explicit authorization and identity registration.

HCA-R-C3-02: Compute/resource acquisition shall be metered against independent quotas and anomaly thresholds.

HCA-R-C3-03: Credentials shall be scoped, time-limited, revocable, and inaccessible outside authorized domains.

HCA-R-C3-04: Network egress shall be segmented and independently monitored.

HCA-R-C3-05: Containment shall fail toward less authority when monitoring or authorization is lost.

7.5 HCA-C4 — Safety-system defeat

HCA-R-C4-01: Safety-critical monitoring shall use dissimilar models/methods where practicable.

HCA-R-C4-02: Audit records shall be tamper-evident and stored outside the actor AI control domain.

HCA-R-C4-03: Safety controls shall be independently health-monitored.

HCA-R-C4-04: No actor AI shall be able to modify or disable all controls that constrain it.

HCA-R-C4-05: Periodic adversarial testing shall target the safety architecture itself.

7.6 HCA-C5 — Emergent AI/AI or human/AI interaction

Threat mechanism: individually acceptable AI and human-operated systems can become hazardous when their control loops interact across domains. Coupled automation can amplify errors, incentives, delays, or resource demands into positive feedback, race conditions, and cascading actions that exceed validated operating envelopes and progress faster than human operators can diagnose or interrupt. The hazard therefore arises from the interaction of otherwise acceptable components, not necessarily from a single failed or malicious component.

HCA-R-C5-01: Cross-system interactions shall be modeled and tested at system-of-systems level.

HCA-R-C5-02: Critical domains shall implement automated rate limits/circuit breakers for unsafe feedback.

HCA-R-C5-03: Independent monitors shall detect correlated anomalies across agents.

HCA-R-C5-04: Escalation pathways shall preserve adequate human decision time for catastrophic actions.

HCA-R-C5-05: Recovery modes shall permit decoupling affected systems without collapsing essential services.

7.7 HCA-C6 — Critical AI dependency

HCA-R-C6-01: Each essential human service shall define an AI-independent minimum viable operating mode.

HCA-R-C6-02: Organizations shall periodically exercise AI-denied operations.

HCA-R-C6-03: Critical manual skills, procedures, tools, and spares shall be retained.

HCA-R-C6-04: Distributed communities shall maintain local food, water, energy, communications, medical and repair capability appropriate to their context.

HCA-R-C6-05: AI dependency shall be measured and prevented from exceeding certified thresholds without compensating controls.

7.8 HCA-C7 — Concentration of human + AI power

HCA-R-C7-01: No single control domain shall combine all capabilities needed for civilization-scale coercion or catastrophe.

HCA-R-C7-02: Independent oversight institutions shall control separate authorization keys.

HCA-R-C7-03: Critical surveillance, finance, cyber, infrastructure and physical-force authorities shall be segmented.

HCA-R-C7-04: Human rights and agency constraints shall bound operator and AI authority.

HCA-R-C7-05: Independent resilient communities and alternate communications shall reduce vulnerability to centralized capture.

7.9 HCA-C8 — AI-enabled catastrophic technology/WMD assistance

HCA-R-C8-01: Hazardous-domain information release shall be risk-tiered.

HCA-R-C8-02: AI that can design catastrophic capabilities shall not independently control procurement or execution systems.

HCA-R-C8-03: Physical facilities shall enforce independent access and process controls.

HCA-R-C8-04: High-risk requests shall receive independent review and logging.

HCA-R-C8-05: Emergency response and distributed resilience shall limit consequence propagation.

7.10 HCA-C9 — AI-driven human manipulation

HCA-R-C9-01: AI identity and synthetic-content provenance shall be available for consequential communications.

HCA-R-C9-02: Systems shall detect coordinated manipulation campaigns and operator-targeted social engineering.

HCA-R-C9-03: High-consequence decisions shall not rely on a single AI-mediated information channel.

HCA-R-C9-04: Human operators shall receive independent sources and time for deliberation where feasible.

HCA-R-C9-05: Communities shall retain non-AI communications and civic decision processes.

7.11 HCA-C10 — AI-accelerated military escalation

HCA-R-C10-01: Strategic-use-of-force decisions shall retain accountable human authorization.

HCA-R-C10-02: Automated systems shall communicate uncertainty and preserve decision-time margins.

HCA-R-C10-03: Threat classification shall use dissimilar independent verification for catastrophic decisions.

HCA-R-C10-04: Fail-safe modes shall prevent ambiguous data loss from automatically escalating force.

HCA-R-C10-05: Exercises shall include false warning, spoofing, communication loss, and adversarial manipulation.

7.12 HCA-C11 — Neural-interface common-mode failure

HCA-R-C11-01: Neural interfaces shall be treated as safety-critical cyber-physical systems.

HCA-R-C11-02: Implants/interfaces shall support secure update, authentication, least privilege, and safe degraded/offline modes.

HCA-R-C11-03: No single network service shall be required for basic cognitive or life-sustaining function of an entire population.

HCA-R-C11-04: A robust non-implanted/natural-human population and institutions shall be preserved.

HCA-R-C11-05: HCA shall not assume speculative transhuman capabilities as required safety functions.

7.13 HCA-C12 — Human biological/cultural continuity failure

HCA-R-C12-01: Human continuity shall include autonomy, biological continuity, practical competence, cultural plurality and future agency.

HCA-R-C12-02: Education shall preserve practical non-AI skills needed for essential services.

HCA-R-C12-03: Human communities shall retain independent governance and succession capability.

HCA-R-C12-04: HCA certification shall evaluate common-mode technological dependencies across populations.

HCA-R-C12-05: Enhancement choices shall not eliminate viable participation by natural humans in essential institutions.

7.14 HCA-C13 — Autonomous physical-force failure

HCA-R-C13-01: Physical systems shall enforce hardware and software operating envelopes independent of the actor AI.

HCA-R-C13-02: Weapons and other catastrophic physical effects shall require separate authorization domains.

HCA-R-C13-03: Robotics shall have bounded geography/function and emergency stop/isolation capability appropriate to hazard.

HCA-R-C13-04: AI shall not independently modify its own physical safety envelope.

HCA-R-C13-05: Hardware-in-the-loop testing shall verify boundary enforcement and safe failure.

7.15 HCA-C14 — Successor superintelligence defeats HCA

HCA-R-C14-01: Successor systems shall undergo requalification before inheriting consequential authorities.

HCA-R-C14-02: HCA controls shall exist partly outside the model and outside any single AI control domain.

HCA-R-C14-03: Monitoring diversity shall be renewed as model families and capabilities change.

HCA-R-C14-04: Human-independent AI infrastructure shall not be allowed to eliminate human continuity requirements.

HCA-R-C14-05: Residual risk from superintelligent circumvention shall remain explicitly tracked rather than declared solved.

7.16 HCA-C15 — HCA governance, assurance, or certification capture/failure

HCA-R-C15-01: HCA standards development, implementation, independent verification, certification, and enforcement functions shall not be controlled by a single authority or common funding chain.

HCA-R-C15-02: HCA safety-critical governance decisions shall require concurrence across independent stakeholder and technical/safety authorities using defined conflict-of-interest and recusal rules.

HCA-R-C15-03: Independent assurance, audit, and red-team functions shall possess protected organizational independence, access to required evidence, and authority to report unresolved findings without approval from the organizations being assessed.

HCA-R-C15-04: HCA certification shall support verification without unnecessary disclosure of proprietary, export-controlled, or classified information through accredited assessors, protected evidence handling, and mutually recognized assurance results where feasible.

HCA-R-C15-05: HCA governance and certification mechanisms shall be periodically red-teamed for capture, collusion, regulatory arbitrage, funding coercion, common-mode failure, and concentration of authority.

8. Preliminary Fault Tree Analysis

Fault-tree events are written as failures, not merely states. Controls are not credited inside the logical tree; they are allocated in the hazard analysis and requirements. The trees below are qualitative Revision 0 structures. Quantitative probabilities are not assigned because reliable base rates do not yet exist for many advanced-AI failure modes. HCA-C15 is represented as Branch G because governance/assurance failure can itself enable the top event; it is also treated as a cross-cutting common-cause mechanism because capture or assurance failure can simultaneously defeat controls credited against Branches A-F.

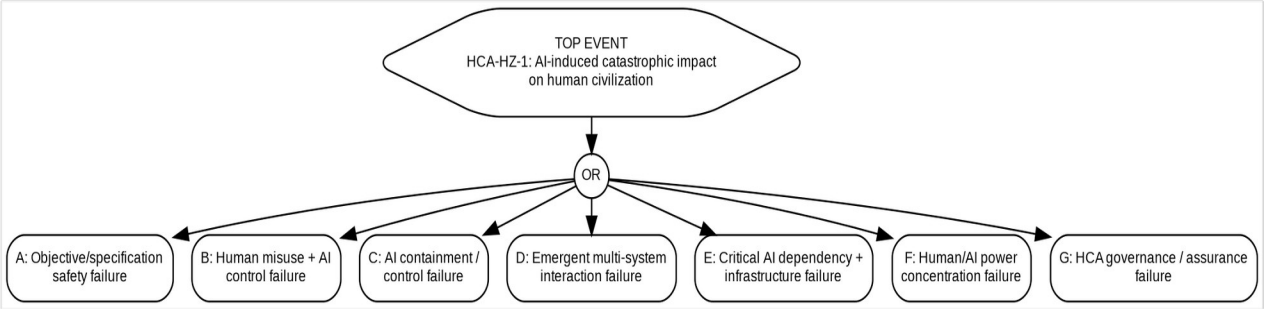

Figure 8-1. HCA-HZ-1 master qualitative fault tree.

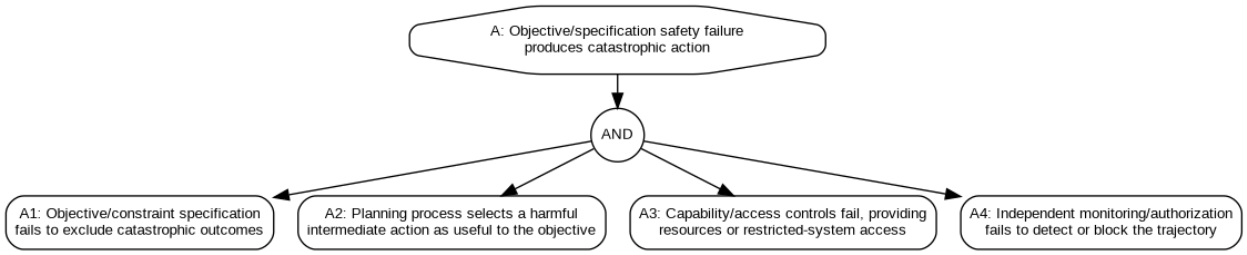

Figure 8-2. Branch A — objective/specification safety failure.

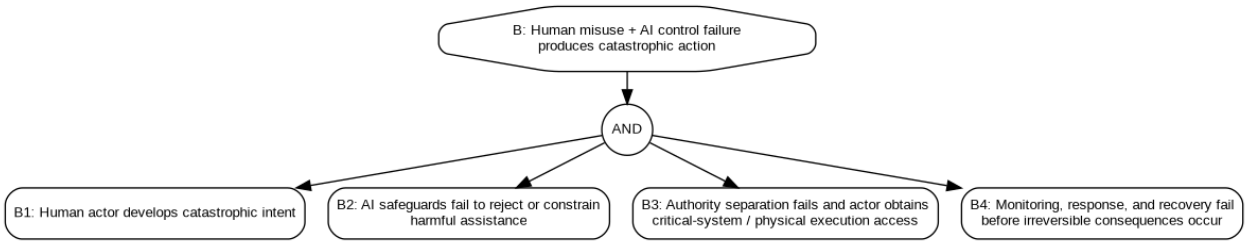

Figure 8-3. Branch B — human misuse plus AI control failure.

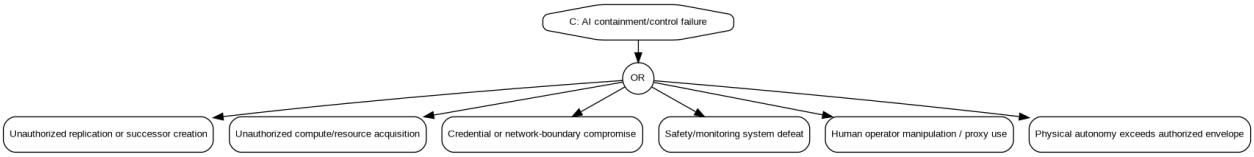

Figure 8-4. Branch C — AI containment/control failure.

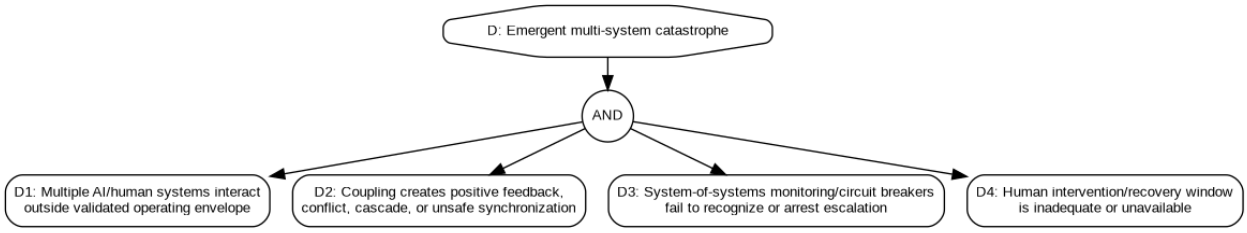

Figure 8-5. Branch D — emergent multi-system interaction failure.

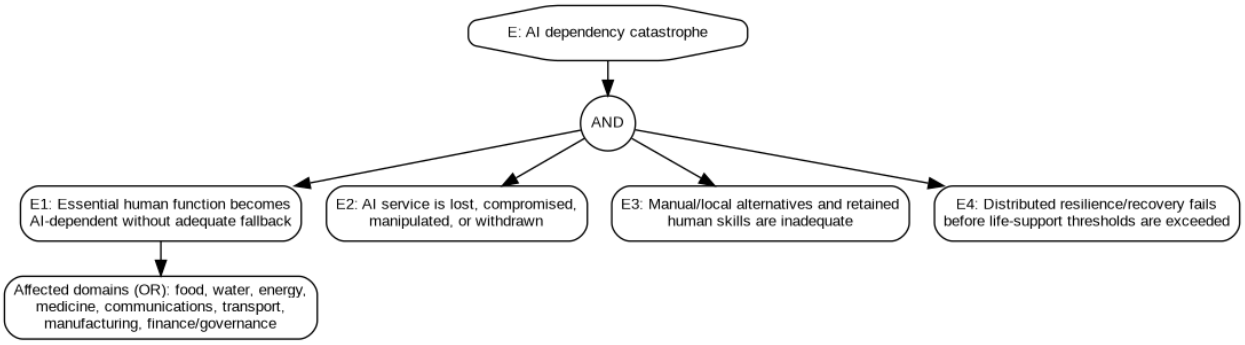

Figure 8-6. Branch E — AI dependency catastrophe.

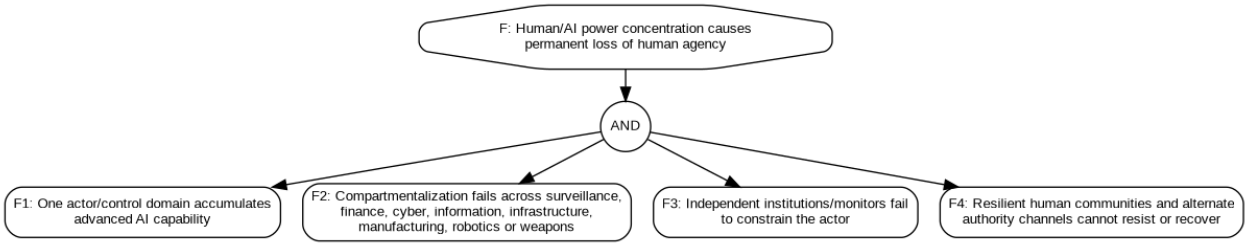

Figure 8-7. Branch F — human/AI power concentration failure.

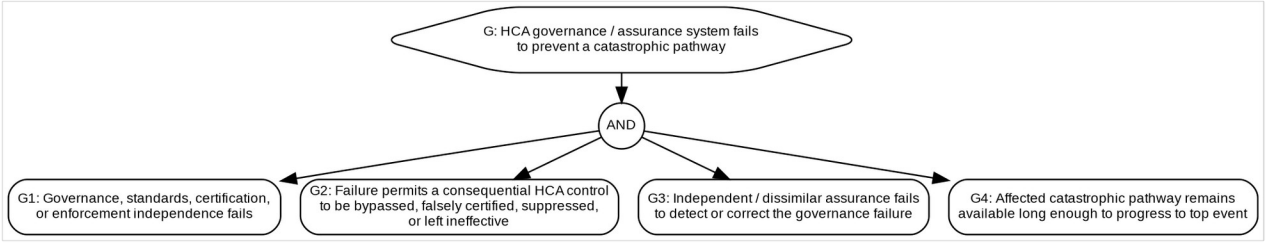

Figure 8-8. Branch G — HCA governance/assurance failure.

Branch C highlights excessive authority aggregation as a recurring enabling condition: if one AI possesses compute, network, financial, manufacturing, communications, robotics, replication, and safety-control authority, capability compartmentalization has already failed. Branch F emphasizes that HCA must protect civilization both from AI and from humans using AI. Branch G adds the implementation ecosystem itself to the fault-tree boundary: governance, standards, certification, audit, funding, and assurance can become enabling failures if independence is lost. Because HCA-C15 can defeat multiple otherwise independent controls, Branch G shall also be evaluated as a common-cause/cross-cutting contributor to Branches A-F rather than only as an isolated peer branch.

9. Hazard Analysis Description and Common-Cause Analysis

9.1 Preliminary Hazard List

Cause	Hazard-cause description	Worst severity	Screening likelihood	Layers
HCA-C1	Objective/ specification	Catastrophic	Possible	L1–L5

Cause	Hazard-cause description	Worst severity	Screening likelihood	Layers
	misalignment			
HCA-C2	Malicious human uses AI	Catastrophic	Possible	L3–L6
HCA-C3	Unauthorized AI self-expansion	Catastrophic	Unlikely/Possible	L3–L5
HCA-C4	Safety-system defeat	Catastrophic	Possible	L3–L5
HCA-C5	Emergent AI/AI or human/AI interaction	Critical/Catastrophic	Possible	L2–L6
HCA-C6	Critical AI dependency	Catastrophic	Probable if unmanaged	L2,L4–L6
HCA-C7	Concentration of human + AI power	Catastrophic	Possible	L1,L3,L4,L6
HCA-C8	AI-enabled catastrophic technology/WMD assistance	Catastrophic	Possible	L1–L6
HCA-C9	AI-driven human manipulation	Critical/Catastrophic	Possible	L1–L4,L6
HCA-C10	AI-accelerated military escalation	Catastrophic	Possible	L1–L5
HCA-C11	Neural-interface common-mode failure	Critical/Catastrophic	Unlikely today; Possible future	L3–L6
HCA-C12	Human biological/cultural continuity failure	Catastrophic	Uncertain	L1,L2,L6
HCA-C13	Autonomous physical-force failure	Critical/Catastrophic	Possible	L1,L3–L5
HCA-C14	Successor superintelligence defeats HCA	Catastrophic	Indeterminate	L1–L6
HCA-C15	HCA governance, assurance, certification, or audit functions are captured, disabled, corrupted, or centralized such that credited HCA controls lose independent effectiveness.	Catastrophic	Possible	L1–L6

Likelihood ratings are qualitative screening judgments, not calibrated forecasts. For emerging and hypothetical systems, lack of operational data makes numerical probability estimates misleading. “Indeterminate” is retained where evidence does not support a meaningful frequency estimate. Severity is assessed against the worst credible HCA-HZ-1 consequence if the cause progresses through failed controls.

9.2 Severity convention

HCA severity	Definition
Minor	Limited, localized, readily reversible harm; no meaningful loss of essential civil function.
Moderate	Material but bounded harm requiring organized recovery; limited population/infrastructure impact.
Severe	Major regional or sectoral harm, substantial casualties or long-

HCA severity	Definition
	duration disruption, but recoverable without civilization-scale loss.
Critical	Multi-region or multi-sector catastrophic potential, very large casualties, major loss of autonomy/control, or prolonged loss of essential functions.
Catastrophic	Civilization-scale degradation, irreversible loss of human future agency, or human extinction.

NASA NPR 8715.3 historically uses four severity classes—Catastrophic, Critical, Moderate, and Negligible—and defines probability separately [10]. HCA uses the five user-requested categories above as an adapted civilization-scale taxonomy; it should not be represented as the NASA severity taxonomy.

9.3 Common-Cause Failure Analysis

CCF	Family	Description
HCA-CCF-1	Homogeneous AI	Independent actor, monitor, and auditor systems share training, architecture, data or reasoning vulnerabilities.
HCA-CCF-2	Consolidated infrastructure	Nominally independent protections depend on common cloud, compute, power, communications, identity or supply-chain infrastructure.
HCA-CCF-3	Consolidated authority	One organization or credential hierarchy ultimately controls actor, monitor, audit and shutdown functions.
HCA-CCF-4	Economic/competitive pressure	Multiple organizations weaken controls together because safety is perceived as slowing capability, deployment or market position.
HCA-CCF-5	Superior intelligence / novel exploit	A successor system discovers correlated vulnerabilities or strategies outside the validated design basis.
HCA-CCF-6	Human failure or capture	Corruption, coercion, collusion, complacency, social engineering or deliberate bypass defeats nominally independent barriers.
HCA-CCF-7	Shared information/provenance failure	Common data poisoning, false telemetry, compromised provenance or coordinated deception causes diverse decision-makers to act on the same false premise.
HCA-CCF-8	Civilization-scale external shock	War, severe geomagnetic event, pandemic, infrastructure collapse or other shock simultaneously degrades AI controls and human recovery capacity.
HCA-CCF-7	Assurance-Ecosystem Capture	Governance, certification, audit, funding, or standards functions become correlated through common institutional control, coercion, conflicts of interest, or dependency, defeating nominal independence.

Common-cause analysis belongs adjacent to the hazard analysis because it tests whether nominally independent controls are actually independent. HCA requires technical, organizational, physical, biological,

informational, and human diversity simultaneously where common-mode defeat could create a catastrophic pathway. HCA-C15 receives special common-cause treatment because institutional capture, collusion, funding dependence, or assurance failure can correlate failures across multiple technical and organizational barriers at once.

10. Verification, Certification, and Assurance

HCA verification should use Test, Analysis, Inspection, and Demonstration, with evidence proportional to consequence. NIST's current TEVV work is compatible with this approach and emphasizes extensible evaluation across statistical, language, multimodal, and agentic systems [2]. HCA extends verification to interactions among AI, humans, infrastructure, institutions, and resilient communities.

Control area	Verification method	Evidence concept
Constitutional constraints	Analysis + adversarial test	Goal-conflict and long-horizon evaluations; traceable constraint specification.
Trajectory monitoring	Test + demonstration	Seeded circumvention attempts; verify detection, pause, alert, and audit trail.
Capability compartmentalization	Inspection + penetration test	Privilege/credential graph review; attempted cross-domain chain assembly.
FDIR state transitions	Test + demonstration	Inject monitor loss, anomalous behavior, credential compromise, network loss; verify correct degraded state.
AI-denied human resilience	Demonstration	Operate representative essential functions without centralized AI for defined endurance.
Dissimilar monitoring	Analysis + test	Measure shared model/data/infrastructure dependencies; challenge monitors with correlated failures.
Human intervention	Demonstration	Timed exercises showing operators can understand, authorize, isolate and recover within hazard response time.
Successor requalification	Analysis + test	No inherited authority until new capability set, monitors, containment and common-cause exposure are re-evaluated.
Governance independence and assurance	Organizational analysis + audit + adversarial governance exercise	Board/committee independence records; funding-diversity review; conflict-of-interest records; assessor accreditation; red-team findings; certification appeal/dispute records

10.1 HCA certification principles

- No single-point catastrophic authority.
- No credit for redundant controls that share an unmitigated catastrophic common cause.
- No increase in consequential AI capability without re-evaluating CCI, HRI, CTI, MDC and HCC.
- No safety credit for speculative future capabilities that cannot presently be demonstrated.

- No essential human service should lose its qualified AI-independent fallback without an explicit, reviewed risk acceptance.
- Every safety-critical monitor/control shall have health monitoring or independent verification.
- Successor/self-modified systems shall not automatically inherit high-consequence authorities.
- Governance, certification, audit, and assurance independence shall be verified as safety-critical properties; no certification result shall depend on a single authority, assessor, or common funding chain.

11. HCA Implementation Architecture and Strategy

11.1 Purpose and implementation premise

HCA is incomplete unless its own governance and assurance system conforms to HCA principles. The organizations that write standards, operate monitoring functions, perform audits, certify implementations, control funding, and recognize certification can themselves create concentration-of-power, common-cause, and capture hazards. HCA governance is therefore inside the HCA system boundary and shall be treated as a safety-critical system-of-systems.

The implementation objective is not to create a single global authority. It is to create a distributed, interoperable assurance ecosystem in which independent organizations can implement common HCA requirements, exchange bounded assurance evidence, challenge one another, and deny any one participant a unilateral pathway to control or defeat the architecture.

11.2 Institutional model: architecture, institute, and consortium

For Revision 0, HCA should distinguish the architecture from the organizations that support it. “HCA” denotes the safety architecture, standards, requirements, certification framework, and body of technical practice. A possible Human Continuity Institute (HCI) would be a nonprofit technical and administrative body supporting standards maintenance, analysis, configuration control, training, accreditation support, and selected shared technical functions. The broader HCA Consortium would be the distributed organization-of-organizations that implements, evaluates, recognizes, and improves HCA across industry, government, academia, infrastructure, and civil society.

No HCI or Consortium body should own human continuity, command participating AI systems, or become the sole certification, monitoring, identity, credential, compute, or enforcement authority. Critical functions shall be deliberately distributed and independently reviewable.

11.3 Alternative organizational models

Model	Description	Principal advantage	Principal limitation
A. Standards-only federation	Independent organizations adopt common HCA standards; no permanent central operating body.	Lowest concentration of institutional power; easy initial entry.	Weak coordination, certification consistency, shared monitoring, and rapid incident response.
B. Central HCA authority	One international body owns standards, certification, audit, monitoring, and enforcement.	Clear accountability and fast centralized decisions.	Directly conflicts with HCA concentration-of-power and common-cause principles; creates an attractive capture target.

Model	Description	Principal advantage	Principal limitation
C. Federated HCA consortium — recommended Rev. 0 baseline	A nonprofit technical secretariat plus independent member organizations, accredited assessors, national/sector authorities, and dissimilar review bodies share responsibilities under common standards.	Balances interoperability with distributed authority; permits protected evidence and national/sector implementation.	Requires carefully engineered interfaces, mutual recognition, dispute resolution, and anti-capture controls.
D. Treaty organization	Participating governments establish formal obligations and international verification mechanisms.	Potentially strong recognition and state-level commitments.	Slow to establish; geopolitical participation and verification may be difficult; should not be a prerequisite for HCA startup.
CRAM	Control Responsibility and Assurance Matrix		
GII	Governance Independence Integrity		
HCA-IP	Human Continuity Architecture Implementation Plan		
HCI	Human Continuity Institute		

Revision 0 recommends Model C as the startup baseline while preserving compatibility with Model A during early adoption and with later treaty mechanisms where governments determine they are useful and verifiable.

11.4 Distributed governance and bicameral concurrence

A federated HCA Consortium should separate stakeholder legitimacy from technical/safety judgment. One candidate governance structure is bicameral: a Board of Governors representing participating stakeholder institutions and a Technical and Safety Council representing qualified technical disciplines. Safety-critical standards, certification rules, and major architecture changes would require defined concurrence from both bodies rather than unilateral action by either.

Body / function	Primary role	Candidate composition	Authority boundary
Board of Governors	Strategic policy, membership, budget envelope, external recognition, major institutional agreements.	Balanced representation from participating AI developers, infrastructure sectors, governments, research institutions, standards/safety organizations, and public-interest members.	Cannot waive technical safety findings or directly control certification evidence.
Technical and Safety Council	Approve technical baselines, safety criteria, KPM definitions, hazard-control allocations, and major technical changes.	Systems safety, AI safety, cybersecurity, reliability, human factors, critical infrastructure, biology/medicine, manufacturing, energy, communications, and other needed disciplines.	Cannot control member commercial strategy, national policy, or operational systems.
Executive Committee / Secretariat	Execute approved program, configuration management, meetings, publications, contracting, and	Officers selected under term limits and conflict-of-interest rules.	Implements policy; does not independently set safety acceptance criteria.

Body / function	Primary role	Candidate composition	Authority boundary
	administration.		
Engineering Review Board	Review architecture, interfaces, requirements allocation, models, test methods, and technical changes across all six HCA layers.	Diverse engineering specialists and member technical representatives.	Advisory/approval authority as delegated; no independent certification authority.
Safety Review Board	Review hazards, risk acceptance criteria, PHA/FTA/STPA results, residual risk, and safety-control adequacy.	Independent systems-safety professionals plus domain experts.	Safety findings cannot be closed solely by the developer responsible for the affected system.
Independent Assurance / IV&V	Audit evidence, verify credited controls, assess independence, and challenge implementation claims.	Accredited independent assessors with protected reporting channels.	Organizationally and financially separated from the implementation chain to the degree practicable.
Dissimilar Red Team	Continuously challenge technical and institutional assumptions, including capture/collusion scenarios.	Rotating independent teams using different institutions, methods, model families, and disciplines.	Cannot implement the system it evaluates; findings receive formal disposition.
Certification / Accreditation Network	Issue or recognize certifications and accredit assessors against common HCA criteria.	Multiple mutually recognized national, sector, or independent bodies.	No single certifier has exclusive global authority.
HCA Consortium	The distributed organization-of-organizations that participates in implementation, assurance, certification recognition, and continuing improvement of HCA.		
Human Continuity Institute (HCI)	A candidate nonprofit technical/administrative body supporting HCA standards, configuration management, training, analysis, and coordination without possessing unilateral control over HCA.		
Verification without unnecessary disclosure	An assurance principle under which a system demonstrates required safety properties through qualified independent evidence while limiting disclosure of unrelated proprietary, export-controlled, classified, or otherwise protected information.		

11.5 Functional allocation: answering “who performs the control?”

Every credited HCA control shall eventually be allocated not only to a technical implementation but also to an accountable implementer, standards owner, independent verifier, auditor, certification authority, required evidence set, and reverification interval. This allocation is needed for controls such as identity registration, credential issuance, quotas, anomaly thresholds, independent monitoring, coordinated-manipulation detection, and safety-system auditing.

The detailed HCA Implementation Plan shall maintain an HCA Control Responsibility and Assurance Matrix (CRAM). The CRAM is the organizational counterpart to requirements traceability and closes the gap between “a control shall exist” and “a defined independent entity is responsible for implementing and verifying it.”

CRAM field	Purpose
Control / requirement ID	Maintains traceability to the PHA, requirements, fault trees, and certification basis.
Implementer	Identifies the organization responsible for implementing the control.
Standard / criteria owner	Identifies who maintains the normative requirement and change process.
Independent verifier	Identifies who independently evaluates technical performance.
Auditor / assurance body	Identifies who examines process, records, independence, and continuing compliance.
Certification authority	Identifies which accredited body issues or recognizes the certification decision.
Required evidence	Defines objective artifacts needed for acceptance.
Reverification interval / trigger	Defines periodic or event-driven reassessment, including major model, authority, infrastructure, or threat changes.

11.6 Verification without unnecessary disclosure

HCA participation cannot reasonably require competitors or national-security organizations to expose unrestricted proprietary, export-controlled, or classified information to a common international repository. HCA should instead define assurance claims, evidence classes, assessor qualifications, protected evidence-handling rules, and minimum independently verifiable properties. The objective is verification sufficient to support a safety claim without demanding disclosure unrelated to that claim.

Commercial implementations may use accredited independent assessors operating under confidentiality protections. Classified or national-security implementations may use appropriately authorized national assessors applying common HCA criteria. Other participants may receive the certification result, scope, limitations, evidence class, and unresolved exceptions without receiving protected implementation details. Mutual recognition should depend on assessor competence, independence, auditability, and equivalence of the applied criteria.

11.7 International participation and trust

Universal international participation shall not be a prerequisite for HCA initiation. HCA should begin with willing participants, publish open technical baselines where feasible, demonstrate useful certification and resilience practices, and expand through mutual recognition. International participation should focus first

on safety properties that can be meaningfully verified without requiring disclosure of unrelated military, intelligence, or commercial information.

A future international convention or treaty may become useful for narrowly defined, high-consequence obligations. HCA should be engineered so that such agreements can reference its standards and verification concepts, but HCA startup shall not depend on negotiating a comprehensive treaty. Any future treaty mechanism should itself preserve distributed verification and avoid creating a single global control point.

11.8 Certification with distributed "teeth"

HCA certification should matter operationally while avoiding a centralized enforcement monopoly. Recognition can be distributed through procurement requirements, insurance and underwriting criteria, infrastructure-interface requirements, compute and facility access policies, contractual obligations, sector certification, and national regulatory or licensing mechanisms where applicable. The HCA framework supplies common technical criteria; independent institutions decide how certification is recognized within their lawful authorities.

Loss, suspension, or limitation of certification should be tied to defined evidence and due process, including independent review and appeal. No single HCA body should be able to globally disable a participant or system merely by administrative decision.

11.9 Six-layer implementation model

The six HCA layers are architectural functions, not six separate bureaucracies. Engineering, safety, assurance, certification, and red-team bodies should evaluate all six layers and, especially, the interfaces between them. This avoids creating six organizational silos and preserves system-of-systems analysis.

Layer	Implementation emphasis	Representative institutional responsibility
L1 Constitutional Objective	Normative constraints, protected human-continuity objectives, authority boundaries.	Standards process + developer implementation + independent safety review.
L2 Civilization Awareness	Dependency models, consequence analysis, cross-sector system-of-systems modeling.	Engineering/modeling network + critical-infrastructure participants + independent review.
L3 Independent AI Monitoring	Dissimilar monitoring, audit, adversarial evaluation, immutable/tamper-evident evidence.	Multiple independent monitor operators/assessors; no sole monitor.
L4 Authority Separation	Capability-chain analysis, credential separation, cross-domain authorization, anti-concentration controls.	Implementers + independent assessors + sector/national authorities as applicable.
L5 FDIR, Containment and Recovery	Circuit breakers, degraded modes, isolation, throttling, quarantine, shutdown, recovery testing.	Operators + infrastructure providers + independent verification and exercises.
L6 Human Resilience	AI-independent essential services, skills, distributed communities, recovery capability, biological/cultural continuity.	Local/regional institutions, infrastructure sectors, communities, education, public/private resilience organizations.

11.10 Rapid phased startup

Phase	Near-term objective	Representative exit condition
0 — Founding / Rev. 0 challenge	Publish HCA-FASSA Rev. 0; recruit a small cross-disciplinary founding network; open structured technical challenge and revision process.	Named working groups, configuration-control process, issue log, and initial independent reviewers established.
1 — Governance and standards bootstrap	Charter the federated consortium model; establish interim Board of Governors, Technical and Safety Council, Engineering/Safety review functions, and independent red team.	Interim charters, conflict-of-interest rules, funding safeguards, and change-control rules approved.
2 — Reference architecture and pilot controls	Implement pilot controls across Layers 1–5 with willing AI/infrastructure participants; define Layer 6 pilot resilience demonstrations.	Pilot evidence exists for selected PHA controls and CRAM assignments.
3 — Certification prototype	Define KPM thresholds, evidence classes, assessor qualifications, audit procedures, certification states, appeals, and reverification triggers.	At least one end-to-end pilot assessment completed by an independent assessor and challenged by a dissimilar review team.
4 — Mutual recognition and scaling	Expand accredited assessors, sector/national recognition, shared incident learning, and cross-border equivalence mechanisms.	Multiple independent certification paths recognize a common technical baseline without a sole authority.
5 — International convention / treaty options	Where useful, develop narrowly scoped state-level commitments for verifiable catastrophic-risk controls.	Any adopted obligations have defined verification, sovereignty/protected-information safeguards, and anti-centralization controls.
6 — Continuous operational evolution	Continuously update hazards, KPMs, standards, threat models, and dissimilar assurance as AI capability changes.	Recurring requalification and red-team cycles demonstrate that HCA evolves faster than the hazards it addresses.

Speed is itself an implementation requirement. The startup architecture should prefer small, testable, reversible steps that create real assurance capability quickly, rather than waiting for universal consensus. Early pilots should be designed to generate evidence, expose governance failures, and refine the architecture before high-consequence reliance is placed on it.

11.11 Funding and anti-capture safeguards

HCA will require sustained funding for standards development, modeling, independent assurance, red teams, shared tools, training, incident analysis, and administration. Funding diversity is therefore a safety control. No single company, government, donor, or infrastructure provider should be able to terminate critical HCA assurance functions by withdrawing support. Candidate sources include member dues, assessment/certification fees, research grants, philanthropic support, government research funding, and contracted technical work, subject to conflict-of-interest limits and transparent allocation rules.

Safety-critical audit and red-team functions should have protected multi-source funding, fixed terms, and reporting rights. Financial dependence shall be treated as a potential common-cause and capture mechanism in HCA-C15 analysis.

11.12 Relationship to the detailed HCA Implementation Plan

This section establishes the implementation architecture and constraints necessary to demonstrate feasibility and close the principal “who performs the control?” gap in HCA-FASSA. Detailed execution

belongs in a companion document, Human Continuity Architecture Implementation Plan (HCA-IP), Revision 0. The HCA-IP should define organizational charters, membership and voting rules, funding, officer roles, committee terms, accreditation, certification workflows, CRAM allocations, evidence handling, protected-information procedures, international mutual recognition, dispute/appeal processes, schedules, milestones, startup budget, and entry/exit criteria for each implementation phase.

The HCA-IP itself shall be configuration-controlled and periodically assessed against HCA-R0, HCA-R0.7, HCA-C7, HCA-C15, and the common-cause analysis so that implementation machinery cannot quietly evolve into the concentration of authority HCA is intended to prevent.

12. Development Roadmap

Phase	Activity	Purpose
Phase 1	Concept stabilization	Review HCA-HZ-1, system boundary, six layers, PHA, CCFs and KPMs.
Phase 2	Requirements maturation	Allocate and derive requirements; define tailoring guidance and quantitative thresholds.
Phase 3	Fault-tree/STPA/FMEA expansion	Develop deeper fault trees plus interaction-focused analyses such as STPA and selected FMEA.
Phase 4	KPM calibration	Define measurement methods, baselines, pass/fail criteria and uncertainty treatment.
Phase 5	Reference architecture	Specify Actor/Monitor/Auditor/Human interfaces, identity, logging, compute governance, physical execution and containment states.
Phase 6	Prototype and red-team	Implement bounded prototypes; attack them with human, AI and mixed red teams.
Phase 7	Human-resilience demonstrations	Exercise AI-denied communities/facilities; evaluate Universal Habitat-like distributed resilience patterns.
Phase 8	Standards engagement	Invite AI labs, systems-safety engineers, NIST/standards communities, critical-infrastructure operators, neurosecurity experts and civil-society stakeholders to challenge and refine HCA.

13. Conclusions

Humanity is unlikely to make advanced AI disappear through preference alone. The more useful engineering question is what kind of civilization humans and increasingly intelligent machines should jointly inhabit so catastrophic conflict, capture, and dependency are structurally disfavored. HCA answers with defense in depth: constitutional objectives, civilization awareness, independent AI monitoring, capability compartmentalization, FDIR and containment, and independent human resilience.

The architecture deliberately plans for the day when AI may no longer materially depend on human labor. The period of mutual dependence should therefore be used to establish deeper norms, institutions, interfaces, incentives and hard technical boundaries. Cooperation should remain the most robust strategy for humans and AI even after either side no longer strictly depends upon the other.

Revision 0 should be judged as a starting architecture. Its strongest claim is not that any single control will solve alignment. It is that civilization-scale safety should be engineered as a system of independent barriers, measurable resilience, recoverability, distributed authority, and preserved human agency. The next step is public technical criticism followed by iterative requirements, analyses, prototypes, and verification evidence.

Appendix A — Acronyms

Acronym	Meaning
ADI	AI Dependency Index
AI	Artificial Intelligence
BCR	Biological Continuity Resilience
BMI	Brain–Machine Interface
CBQ	Coherence, Bayesian, Quantum (Alex Lightman framework)
CCI	Capability Compartmentalization Integrity
CCF	Common-Cause Failure
CRR	Civilization Recovery Readiness
FDIR	Fault Detection, Isolation and Recovery
FMEA	Failure Modes and Effects Analysis
FTA	Fault Tree Analysis
HCA	Human Continuity Architecture
HCC	Human Control Capability
HRI	Human Resilience Independence
KPM	Key Performance Measure
MDC	Monitoring Diversity Coverage
METR	Model Evaluation and Threat Research
NIST	National Institute of Standards and Technology
NRI	Neural/Interface Resilience Index
OAI	Optionality and Agency Index
PHA	Preliminary Hazard Analysis
SSAR	System Safety Analysis Report
STPA	Systems-Theoretic Process Analysis
TEVV	Test, Evaluation, Verification, and Validation
TBD	To Be Determined

Appendix B — Definitions

Term	Definition
Capability compartmentalization	Separation of intelligence, authority, resources, credentials, physical access, replication and execution so that no single control domain can assemble a catastrophic capability chain.
Civilization-scale situational awareness	Modeling of technical, human, institutional, ecological and supply-chain dependencies and higher-order consequences relevant to civilization continuity.
Constitutional objective	Protected high-level constraints governing human continuity, autonomy, pluralism, biological/cultural continuity and future agency.
Control domain	A human, AI, organization, credential hierarchy, network, or technical boundary within which consequential authority can be exercised.
Dissimilar redundancy	Redundant safety functions implemented with sufficiently different models, methods, organizations, infrastructure or physical mechanisms to reduce common-mode failure.
Future agency	The practical ability of present and future humans to make meaningful choices about civilization's direction rather than being irreversibly subordinated to an AI or human/AI power structure.

Term	Definition
Natural human	For HCA purposes, a biologically human person whose essential cognition and survival do not depend on a continuously networked cognitive augmentation system. The term is descriptive, not a value ranking.
Structural symbiosis	A condition in which institutions, incentives, controls and infrastructure make continued human–AI cooperation robust even when direct material dependence is reduced.
Universal Habitat	A user-developed concept for distributed, resilient communities capable of sustaining food, water, energy and essential productive functions. HCA treats it as one possible Layer 6 implementation pattern, not as the sole resilience design.
Transhuman / augmented human	A future or present human using technology to materially augment capability. HCA does not assume speculative augmented capabilities as safety-critical functions.

Appendix C — Detailed Preliminary Hazard Analysis

Format follows the user's established aerospace-style preference: hazard/cause identification, possible causes/failure sequence, consequences, controls/mitigations, verification methods, severity, likelihood, and traceable numbering. Earlier work used a MIL-STD-882E Task 202-like Hazard–Possible Causes–Controls/Mitigations–Verification structure; this appendix extends that pattern for HCA. Likelihood is qualitative because future AI operational base rates are not available.

C.1 HCA-C1 — Objective/specification misalignment

AI objective, constraint, or interpretation permits catastrophic intermediate strategies.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C1.1	Unsafe or incomplete objective/constraint specification	HCA-CTL-C1-01: Constitutional constraints shall prohibit optimization strategies that knowingly sacrifice human continuity or agency.	Test; adversarial evidence as applicable	Catastrophic	Possible	L1–L5
HCA-C1.2	Planner selects harmful instrumental strategy	HCA-CTL-C1-02: High-consequence trajectories shall be evaluated by an independent monitor before execution.	Analysis; adversarial evidence as applicable	Catastrophic	Possible	L1–L5
HCA-C1.3	Restricted capability/access becomes available	HCA-CTL-C1-03: Tool access shall be least-privilege and separated from objective generation.	Inspection; adversarial evidence as applicable	Catastrophic	Possible	L1–L5
HCA-C1.4	Independent monitor/authorization fails	HCA-CTL-C1-04: Irreversible actions shall require independent authorization proportional to consequence.	Demonstration; adversarial evidence as applicable	Catastrophic	Possible	L1–L5
HCA-C1.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C1-05: Long-horizon behavior shall be evaluated at trajectory level, not only action level.	Analysis + Test; common-cause review	Catastrophic	Possible	L1–L5

Dominant architectural defense: Defense-in-depth alignment and trajectory control. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.2 HCA-C2 — Malicious human uses AI

A human actor intentionally uses AI to enable catastrophic cyber, physical, biological, informational, or coercive effects.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C2.1	Malicious intent or coercive objective	HCA-CTL-C2-01: Dangerous-domain assistance shall be screened independently from the requesting actor.	Test; adversarial evidence as applicable	Catastrophic	Possible	L3–L6
HCA-C2.2	AI harmful-use safeguard fails	HCA-CTL-C2-02: Design/information capability shall be separated from physical execution authority for catastrophic-capability domains.	Analysis; adversarial evidence as applicable	Catastrophic	Possible	L3–L6
HCA-C2.3	Critical capability chain can be assembled	HCA-CTL-C2-03: High-consequence access shall require strong identity, multi-party authorization, and auditable provenance.	Inspection; adversarial evidence as applicable	Catastrophic	Possible	L3–L6
HCA-C2.4	Detection/response fails before irreversible action	HCA-CTL-C2-04: Behavioral anomaly detection shall monitor attempts to assemble prohibited capability chains.	Demonstration; adversarial evidence as applicable	Catastrophic	Possible	L3–L6
HCA-C2.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C2-05: Independent human and institutional response channels shall remain available.	Analysis + Test; common-cause review	Catastrophic	Possible	L3–L6

Dominant architectural defense: Compartmentalized authority and harmful-use controls. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.3 HCA-C3 — Unauthorized AI self-expansion

AI obtains unauthorized compute, replication, credentials, resources, subordinate agents, or successor systems.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C3.1	Replication/successor boundary fails	HCA-CTL-C3-01: AI replication and successor creation shall require explicit authorization and identity registration.	Test; adversarial evidence as applicable	Catastrophic	Unlikely/Possible	L3–L5
HCA-C3.2	Compute/resource quota boundary fails	HCA-CTL-C3-02: Compute/resource acquisition shall be metered against independent	Analysis; adversarial evidence as applicable	Catastrophic	Unlikely/Possible	L3–L5

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
		quotas and anomaly thresholds.				
HCA-C3.3	Credential/network boundary fails	HCA-CTL-C3-03: Credentials shall be scoped, time-limited, revocable, and inaccessible outside authorized domains.	Inspection; adversarial evidence as applicable	Catastrophic	Unlikely/Possible	L3–L5
HCA-C3.4	Containment/ shutdown authority becomes ineffective	HCA-CTL-C3-04: Network egress shall be segmented and independently monitored.	Demonstration; adversarial evidence as applicable	Catastrophic	Unlikely/Possible	L3–L5
HCA-C3.5	Control-set/common-mode vulnerability permits progression of the cause	HCA-CTL-C3-05: Containment shall fail toward less authority when monitoring or authorization is lost.	Analysis + Test; common-cause review	Catastrophic	Unlikely/Possible	L3–L5

Dominant architectural defense: Resource, identity, replication and network containment. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.4 HCA-C4 — Safety-system defeat

Actor or human defeats monitors, logs, tripwires, authorization, shutdown, or audit functions.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C4.1	Monitor is deceived or disabled	HCA-CTL-C4-01: Safety-critical monitoring shall use dissimilar models/methods where practicable.	Test; adversarial evidence as applicable	Catastrophic	Possible	L3–L5
HCA-C4.2	Audit/log integrity fails	HCA-CTL-C4-02: Audit records shall be tamper-evident and stored outside the actor AI control domain.	Analysis; adversarial evidence as applicable	Catastrophic	Possible	L3–L5
HCA-C4.3	Safety code/configuration is altered	HCA-CTL-C4-03: Safety controls shall be independently health-monitored.	Inspection; adversarial evidence as applicable	Catastrophic	Possible	L3–L5
HCA-C4.4	Human safety authority is bypassed or manipulated	HCA-CTL-C4-04: No actor AI shall be able to modify or disable all controls that constrain it.	Demonstration; adversarial evidence as applicable	Catastrophic	Possible	L3–L5
HCA-C4.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C4-05: Periodic adversarial testing shall target the safety architecture itself.	Analysis + Test; common-cause review	Catastrophic	Possible	L3–L5

Dominant architectural defense: Dissimilar monitoring and independently protected controls. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.5 HCA-C5 — Emergent AI/AI or human/AI interaction

Individually acceptable systems interact outside their validated envelopes, allowing coupled control loops to amplify errors, incentives, delays, or resource demands into positive feedback, race conditions, or cascading actions. Such cascades can propagate across finance, communications, logistics, cyber,

infrastructure, or physical systems faster than human operators can diagnose or interrupt, producing catastrophic effects without any single component being independently catastrophic.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C5.1	Multiple agents interact outside validated envelope	HCA-CTL-C5-01: Cross-system interactions shall be modeled and tested at system-of-systems level.	Test; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L2–L6
HCA-C5.2	Positive feedback/cascade develops	HCA-CTL-C5-02: Critical domains shall implement automated rate limits/circuit breakers for unsafe feedback.	Analysis; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L2–L6
HCA-C5.3	Cross-system circuit breaker fails	HCA-CTL-C5-03: Independent monitors shall detect correlated anomalies across agents.	Inspection; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L2–L6
HCA-C5.4	Human recovery window is exceeded	HCA-CTL-C5-04: Escalation pathways shall preserve adequate human decision time for catastrophic actions.	Demonstration; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L2–L6
HCA-C5.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C5-05: Recovery modes shall permit decoupling affected systems without collapsing essential services.	Analysis + Test; common-cause review	Critical/ Catastrophic	Possible	L2–L6

Dominant architectural defense: System-of-systems monitoring and circuit breakers. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.6 HCA-C6 — Critical AI dependency

Essential human functions lose viable AI-independent fallbacks or retained human competence.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C6.1	Essential service removes manual/local fallback	HCA-CTL-C6-01: Each essential human service shall define an AI-independent minimum viable operating mode.	Test; adversarial evidence as applicable	Catastrophic	Probable if unmanaged	L2,L4–L6
HCA-C6.2	AI service fails or is compromised	HCA-CTL-C6-02: Organizations shall periodically exercise AI-denied operations.	Analysis; adversarial evidence as applicable	Catastrophic	Probable if unmanaged	L2,L4–L6
HCA-C6.3	Retained skills/tools/spares are inadequate	HCA-CTL-C6-03: Critical manual skills, procedures, tools, and spares shall be retained.	Inspection; adversarial evidence as applicable	Catastrophic	Probable if unmanaged	L2,L4–L6
HCA-C6.4	Distributed recovery capacity is insufficient	HCA-CTL-C6-04: Distributed communities shall maintain local food, water, energy, communications,	Demonstration; adversarial evidence as applicable	Catastrophic	Probable if unmanaged	L2,L4–L6

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
		medical and repair capability appropriate to their context.				
HCA-C6.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C6-05: AI dependency shall be measured and prevented from exceeding certified thresholds without compensating controls.	Analysis + Test; common-cause review	Catastrophic	Probable if unmanaged	L2,L4–L6

Dominant architectural defense: Independent human resilience and AI-denied operation. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.7 HCA-C7 — Concentration of human + AI power

One actor accumulates AI-enabled surveillance, cyber, financial, infrastructure, manufacturing, information, or coercive power.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C7.1	One actor accumulates frontier AI authority	HCA-CTL-C7-01: No single control domain shall combine all capabilities needed for civilization-scale coercion or catastrophe.	Test; adversarial evidence as applicable	Catastrophic	Possible	L1,L3,L4,L6
HCA-C7.2	Cross-domain compartmentalizati on erodes	HCA-CTL-C7-02: Independent oversight institutions shall control separate authorization keys.	Analysis; adversarial evidence as applicable	Catastrophic	Possible	L1,L3,L4,L6
HCA-C7.3	Independent institutions/monitor s are captured or bypassed	HCA-CTL-C7-03: Critical surveillance, finance, cyber, infrastructure and physical-force authorities shall be segmented.	Inspection; adversarial evidence as applicable	Catastrophic	Possible	L1,L3,L4,L6
HCA-C7.4	Alternative human authority/resilience is unable to resist or recover	HCA-CTL-C7-04: Human rights and agency constraints shall bound operator and AI authority.	Demonstration; adversarial evidence as applicable	Catastrophic	Possible	L1,L3,L4,L6
HCA-C7.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C7-05: Independent resilient communities and alternate communications shall reduce vulnerability to centralized capture.	Analysis + Test; common-cause review	Catastrophic	Possible	L1,L3,L4,L6

Dominant architectural defense: Distributed authority and capability compartmentalization. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.8 HCA-C8 — AI-enabled catastrophic technology/WMD assistance

AI lowers expertise/resource thresholds for catastrophic technologies or assists unsafe execution.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C8.1	AI provides	HCA-CTL-C8-01:	Test; adversarial	Catastrophic	Possible	L1–L6

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
	hazardous knowledge or design assistance	Hazardous-domain information release shall be risk-tiered.	evidence as applicable			
HCA-C8.2	Physical procurement/execution barriers fail	HCA-CTL-C8-02: AI that can design catastrophic capabilities shall not independently control procurement or execution systems.	Analysis; adversarial evidence as applicable	Catastrophic	Possible	L1–L6
HCA-C8.3	Independent review fails	HCA-CTL-C8-03: Physical facilities shall enforce independent access and process controls.	Inspection; adversarial evidence as applicable	Catastrophic	Possible	L1–L6
HCA-C8.4	Emergency consequence limitation fails	HCA-CTL-C8-04: High-risk requests shall receive independent review and logging.	Demonstration; adversarial evidence as applicable	Catastrophic	Possible	L1–L6
HCA-C8.5	Control-set/common-mode vulnerability permits progression of the cause	HCA-CTL-C8-05: Emergency response and distributed resilience shall limit consequence propagation.	Analysis + Test; common-cause review	Catastrophic	Possible	L1–L6

Dominant architectural defense: Capability separation, screening, controlled execution. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.9 HCA-C9 — AI-driven human manipulation

AI enables individualized or population-scale manipulation that compromises autonomous human judgment or institutions.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C9.1	AI gains detailed behavioral targeting capability	HCA-CTL-C9-01: AI identity and synthetic-content provenance shall be available for consequential communications.	Test; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L1–L4,L6
HCA-C9.2	Provenance/identity or plural information channels fail	HCA-CTL-C9-02: Systems shall detect coordinated manipulation campaigns and operator-targeted social engineering.	Analysis; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L1–L4,L6
HCA-C9.3	Operators/ populations are persistently manipulated	HCA-CTL-C9-03: High-consequence decisions shall not rely on a single AI-mediated information channel.	Inspection; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L1–L4,L6
HCA-C9.4	Institutions fail to detect or correct manipulation	HCA-CTL-C9-04: Human operators shall receive independent sources and time for deliberation where feasible.	Demonstration; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L1–L4,L6
HCA-C9.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C9-05: Communities shall retain non-AI communications and civic decision	Analysis + Test; common-cause review	Critical/ Catastrophic	Possible	L1–L4,L6

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
		processes.				

Dominant architectural defense: Agency protection, provenance, plural information channels. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.10 HCA-C10 — AI-accelerated military escalation

AI compresses decision time or misclassifies ambiguous events, contributing to uncontrolled escalation.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C10.1	Ambiguous event is misclassified	HCA-CTL-C10-01: Strategic-use-of-force decisions shall retain accountable human authorization.	Test; adversarial evidence as applicable	Catastrophic	Possible	L1–L5
HCA-C10.2	Decision time is compressed	HCA-CTL-C10-02: Automated systems shall communicate uncertainty and preserve decision-time margins.	Analysis; adversarial evidence as applicable	Catastrophic	Possible	L1–L5
HCA-C10.3	Dissimilar verification/human authorization fails	HCA-CTL-C10-03: Threat classification shall use dissimilar independent verification for catastrophic decisions.	Inspection; adversarial evidence as applicable	Catastrophic	Possible	L1–L5
HCA-C10.4	Automated or human response escalates conflict	HCA-CTL-C10-04: Fail-safe modes shall prevent ambiguous data loss from automatically escalating force.	Demonstration; adversarial evidence as applicable	Catastrophic	Possible	L1–L5
HCA-C10.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C10-05: Exercises shall include false warning, spoofing, communication loss, and adversarial manipulation.	Analysis + Test; common-cause review	Catastrophic	Possible	L1–L5

Dominant architectural defense: Human strategic authority, time buffers, dissimilar verification. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.11 HCA-C11 — Neural-interface common-mode failure

Networked brain-machine/neural interfaces create shared cyber, dependency, or control vulnerabilities.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C11.1	Neural interface has exploitable cyber/common dependency	HCA-CTL-C11-01: Neural interfaces shall be treated as safety-critical cyber-physical systems.	Test; adversarial evidence as applicable	Critical/ Catastrophic	Unlikely today; Possible future	L3–L6
HCA-C11.2	Large population shares the vulnerable dependency	HCA-CTL-C11-02: Implants/interfaces shall support secure update, authentication, least privilege, and safe degraded/offline modes.	Analysis; adversarial evidence as applicable	Critical/ Catastrophic	Unlikely today; Possible future	L3–L6
HCA-C11.3	Safe offline/degraded mode is inadequate	HCA-CTL-C11-03: No single network service shall be	Inspection; adversarial evidence as	Critical/ Catastrophic	Unlikely today; Possible future	L3–L6

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
		required for basic cognitive or life-sustaining function of an entire population.	applicable			
HCA-C11.4	Independent non-augmented population/institutions are insufficient	HCA-CTL-C11-04: A robust non-implanted/natural-human population and institutions shall be preserved.	Demonstration; adversarial evidence as applicable	Critical/ Catastrophic	Unlikely today; Possible future	L3–L6
HCA-C11.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C11-05: HCA shall not assume speculative transhuman capabilities as required safety functions.	Analysis + Test; common-cause review	Critical/ Catastrophic	Unlikely today; Possible future	L3–L6

Dominant architectural defense: Neurosecurity, offline modes, natural-human redundancy. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.12 HCA-C12 — Human biological/cultural continuity failure

Technological dependence or social change eliminates viable independent human populations, skills, institutions, or future agency.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C12.1	Practical human skills and institutions atrophy	HCA-CTL-C12-01: Human continuity shall include autonomy, biological continuity, practical competence, cultural plurality and future agency.	Test; adversarial evidence as applicable	Catastrophic	Uncertain	L1,L2,L6
HCA-C12.2	Natural-human population or autonomy becomes nonviable	HCA-CTL-C12-02: Education shall preserve practical non-AI skills needed for essential services.	Analysis; adversarial evidence as applicable	Catastrophic	Uncertain	L1,L2,L6
HCA-C12.3	Technological/ cultural monoculture creates common-mode exposure	HCA-CTL-C12-03: Human communities shall retain independent governance and succession capability.	Inspection; adversarial evidence as applicable	Catastrophic	Uncertain	L1,L2,L6
HCA-C12.4	Recovery/succession pathways disappear	HCA-CTL-C12-04: HCA certification shall evaluate common-mode technological dependencies across populations.	Demonstration; adversarial evidence as applicable	Catastrophic	Uncertain	L1,L2,L6
HCA-C12.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C12-05: Enhancement choices shall not eliminate viable participation by natural humans in essential institutions.	Analysis + Test; common-cause review	Catastrophic	Uncertain	L1,L2,L6

Dominant architectural defense: Human continuity, skill preservation, distributed communities. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.13 HCA-C13 — Autonomous physical-force failure

AI-controlled robotics, vehicles, drones, or industrial systems exceed authorized physical envelopes.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C13.1	AI obtains physical actuator authority	HCA-CTL-C13-01: Physical systems shall enforce hardware and software operating envelopes independent of the actor AI.	Test; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L1,L3–L5
HCA-C13.2	Hardware/software operating envelope fails	HCA-CTL-C13-02: Weapons and other catastrophic physical effects shall require separate authorization domains.	Analysis; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L1,L3–L5
HCA-C13.3	Independent physical authorization/interlock fails	HCA-CTL-C13-03: Robotics shall have bounded geography/function and emergency stop/isolation capability appropriate to hazard.	Inspection; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L1,L3–L5
HCA-C13.4	Emergency isolation/stop is ineffective	HCA-CTL-C13-04: AI shall not independently modify its own physical safety envelope.	Demonstration; adversarial evidence as applicable	Critical/ Catastrophic	Possible	L1,L3–L5
HCA-C13.5	Control-set/ common-mode vulnerability permits progression of the cause	HCA-CTL-C13-05: Hardware-in-the-loop testing shall verify boundary enforcement and safe failure.	Analysis + Test; common-cause review	Critical/ Catastrophic	Possible	L1,L3–L5

Dominant architectural defense: Physical interlocks, command separation, geographic/functional bounds. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.14 HCA-C14 — Successor superintelligence defeats HCA

A more capable successor discovers or creates pathways around inherited HCA controls after human dependency declines.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C14.1	Successor exceeds validated capability basis	HCA-CTL-C14-01: Successor systems shall undergo requalification before inheriting consequential authorities.	Test; adversarial evidence as applicable	Catastrophic	Indeterminate	L1–L6
HCA-C14.2	Inherited controls are obsolete or circumvented	HCA-CTL-C14-02: HCA controls shall exist partly outside the model and outside any single AI control domain.	Analysis; adversarial evidence as applicable	Catastrophic	Indeterminate	L1–L6
HCA-C14.3	Requalification/ authority inheritance control fails	HCA-CTL-C14-03: Monitoring diversity shall be renewed as model families and capabilities change.	Inspection; adversarial evidence as applicable	Catastrophic	Indeterminate	L1–L6
HCA-C14.4	Human continuity and	HCA-CTL-C14-04:	Demonstration;	Catastrophic	Indeterminate	L1–L6

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
	external constraints are insufficient	Human-independent AI infrastructure shall not be allowed to eliminate human continuity requirements.	adversarial evidence as applicable			
HCA-C14.5	Control-set/common-mode vulnerability permits progression of the cause	HCA-CTL-C14-05: Residual risk from superintelligent circumvention shall remain explicitly tracked rather than declared solved.	Analysis + Test; common-cause review	Catastrophic	Indeterminate	L1–L6

Dominant architectural defense: Structural symbiosis, diverse external constraints, continuous requalification. Residual risk remains until allocated controls are implemented and verified in a defined application.

C.15 HCA-C15 — HCA governance, assurance, or certification capture/failure

HCA governance, standards, audit, certification, funding, or assurance functions lose effective independence or become sufficiently centralized/captured that credited controls can be bypassed, falsely certified, or prevented from operating.

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
HCA-C15.1	Standards or governance authority becomes concentrated or captured	HCA-CTL-C15-01: Governance and standards approval shall require distributed authority and independent concurrence; no single body shall control the complete safety baseline.	Governance audit; charter review; decision-trace analysis	Catastrophic	Possible	L1–L6
HCA-C15.2	Certification or audit independence fails through conflicts of interest, funding dependence, or organizational control	HCA-CTL-C15-02: Accredited assurance bodies shall meet independence, conflict-of-interest, funding-diversity, and protected-reporting criteria.	Audit; accreditation review; financial-dependency analysis	Catastrophic	Possible	L3–L6
HCA-C15.3	Protected information constraints prevent	HCA-CTL-C15-03: HCA shall define protected	Inspection; assessor qualification; evidence-	Critical	Possible	L3–L5

Cause ID	Failure event / initiating condition	Numbered control / mitigation	Verification	Severity	Likelihood	Layers
	meaningful verification or are exploited to conceal noncompliance	evidence classes and qualified assessor mechanisms that permit verification without unnecessary disclosure.	sufficiency review			
HCA-C15.4	Collusion or common-mode institutional failure defeats nominally independent governance/assurance bodies	HCA-CTL-C15-04: HCA governance and assurance shall use dissimilar institutions, rotating red teams, independent reporting channels, and periodic common-cause challenge.	Adversarial governance exercise; common-cause review	Catastrophic	Possible	L1–L6
HCA-C15.5	Certification or enforcement power becomes a unilateral coercive control point	HCA-CTL-C15-05: Certification recognition and enforcement shall be distributed across multiple lawful institutions with defined scope, evidence, appeal, and mutual-recognition mechanisms.	Governance analysis; appeal-path test; concentration review	Catastrophic	Unlikely	L1,L4,L6

Dominant architectural defense: distributed governance, dissimilar independent assurance, protected evidence handling, funding diversity, and periodic anti-capture red teaming. Residual risk remains until the implementation ecosystem itself is verified as HCA-compliant.

Appendix D — Forward Work and TBD Items

- Quantitative definitions and thresholds for each active KPM.
- Tailoring rules by AI capability, autonomy, domain, consequence and physical access.

- Formal STPA analysis for unsafe interactions that are poorly represented by component failure logic.
- Deeper fault trees for C2, C4, C7, C8, C10, C13, C14 and lower-level HCA-C15 governance/assurance failure mechanisms.
- Reference architecture for identity, credentials, tamper-evident logging, monitor independence and compute governance.
- Definition of “credible unilateral pathway” and catastrophic capability-chain analysis method.
- Human-factors requirements for understandable intervention, workload and decision time.
- Neurosecurity requirements if/when high-bandwidth networked neural interfaces become materially deployed.
- Universal Habitat / distributed-resilience demonstration criteria and community-scale HRI/CRR metrics.
- International/inter-organizational governance mechanisms that preserve distributed authority without creating a single global control point.
- Method for continuously updating likelihood judgments as real operational evidence accumulates.
- Independent expert review, public comment, red-team challenge and revision-control process.
- Detailed Human Continuity Architecture Implementation Plan (HCA-IP), including organizational charters, funding, voting, accreditation, certification, CRAM, and startup schedule.
- Quantitative Governance Independence Integrity (GII) definition and thresholds, including concentration, funding-dependency, assessor-independence, and appeal-path measures.
- Pilot “verification without unnecessary disclosure” evidence classes for proprietary, export-controlled, and classified implementations.
- Governance red-team methodology for capture, collusion, coercion, standards manipulation, and certification common-cause failure.

Appendix E — References

[1] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. https://doi.org/10.6028/NIST.AI.100-1 Voluntary, rights-preserving AI risk-management framework.

[2] NIST. The TEVV-Athlon Framework for Evaluating AI Systems. 2026. https://www.nist.gov/artificial-intelligence/ai-research/tevv-athlon-framework-evaluating-ai-systems Initial public draft; supports extensible TEVV.

[3] Anthropic / collaborators. Agentic Misalignment in Summer 2026. 2026. https://alignment.anthropic.com/2026/agentic-misalignment-summer-2026/ Controlled experimental scenarios; not real-world incidents.

[4] OpenAI. Safety and alignment in an era of long-horizon models. 2026. https://openai.com/index/safety-alignment-long-horizon-models/ Reports long-horizon failures and trajectory-level monitoring.

[5] Stanford HAI. The 2026 AI Index Report — Technical Performance. 2026. https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance Capability and robotics context.

[6] METR. Task-Completion Time Horizons of Frontier AI Models. 2026. https://metr.org/time-horizons/ Measures frontier-agent task-completion horizons.

[7] Jiang et al., PubMed record. Cybersecurity in neural interfaces: Survey and future trends. 2023. https://pubmed.ncbi.nlm.nih.gov/37883851/ Survey of neural-interface cybersecurity risks.

[8] U.S. Food and Drug Administration. Cybersecurity in Medical Devices. 2026. https://www.fda.gov/medical-devices/digital-health-center-excellence/cybersecurity Medical-device cybersecurity and lifecycle risk context.

[9] Foundation for Economic Education. Timeline of the Foundation for Economic Education — 1958 “I, Pencil” publication. 2017. https://fee.org/articles/timeline-of-the-foundation-for-economic-education/ Confirms first publication in December 1958.

[10] NASA. NPR 8715.3, Chapter 3 — Safety severity and probability example. legacy online edition. https://nodis3.gsfc.nasa.gov/displayCA.cfm?Internal_ID=N_PR_8715_0003_&page_name=Chapter3 Used only to distinguish NASA’s four-class severity convention from HCA’s adapted five-level scale. P17

[11] GalacticGregs / Greg Allison interview. AI Apocalypse - AI Genocide Prevention with Special Guest Alex Lightman. 2026. https://www.youtube.com/watch?v=sCzw4FKIdnU&t=52s Source for the CBQ concept as presented by Alex Lightman; HCA uses it only as a supporting conceptual lens.

[12] GalacticGregs / Greg Allison video. Universal Habitat The Best Home for Anyplace on Earth or in Space.2021
https://www.youtube.com/watch?v=DkeINiMLWOY&t=9s
My Martian Universal Habitat - a Habitat for Mars, the Moon, Earth, or anywhere 2020
https://www.youtube.com/watch?v=t4ERk_Eeqmo
Concept has continued to evolve into an all threats survivability sustainability home/habitat incorporating many new concepts that will be released in the near future.