The Human Continuity Architecture

Core Concept Paper

A public-facing introduction to a systems-safety framework for human–AI civilization

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

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

PRELIMINARY DRAFT

Executive Summary

The public debate about advanced artificial intelligence often collapses into two choices: accelerate without adequate safeguards, or somehow stop AI before it becomes too powerful. Human Continuity Architecture (HCA) proposes a third approach: treat the future relationship between humans and increasingly capable machines as a civilization-scale system-safety problem.

HCA begins with a simple premise. Today, AI depends on human civilization. Data centers depend on electricity, semiconductor fabs, mines, transformers, cooling systems, networks, technicians, machinists, construction workers, logistics, institutions, and countless other human capabilities. But that dependence may not last forever. Advanced robotics, automated manufacturing, mining, repair and energy systems could eventually close much of the loop. If human preservation matters only because AI needs us, the safety case expires when the dependency expires.

The goal is therefore to use the period of mutual dependence to build something deeper: a civilization in which human continuity, autonomy and future agency are protected by architecture, not merely by temporary usefulness.

1. The Problem Is Bigger Than “Rogue AI”

An AI catastrophe does not require a machine that hates humanity. A sufficiently capable system can cause harm if a dangerous action becomes useful to an assigned objective, if a human uses AI maliciously, if interacting systems create a runaway cascade, or if civilization becomes unable to function without centralized AI. Recent controlled research has demonstrated agentic misalignment and harmful-compliance behaviors, while long-horizon deployment work has shown why monitoring entire trajectories matters [3][4].

The human behind the curtain may be as important as the machine. A dictator, criminal organization, corporation, insider, military organization, or highly capable individual could combine AI with cyber access, surveillance, finance, robotics, manufacturing, or weapons. An AI can be perfectly “aligned” with its owner and still be disastrous for everyone else. HCA therefore protects civilization from both AI failure and humans using AI.

2. The Human Dependency Window

Leonard Read’s “I, Pencil” made a famous point in 1958: no single person knows how to make even an ordinary pencil from scratch [9]. AI hardware and infrastructure amplify that lesson by orders of magnitude. Modern AI exists inside a vast web of human knowledge, materials, factories, grids, networks, transportation, repair and governance.

That creates an opportunity, but not a permanent guarantee. HCA assumes AI may eventually become far more physically self-sufficient. The architecture must therefore be designed for the day when “AI needs people” is no longer enough.

3. Six Layers of Human Continuity

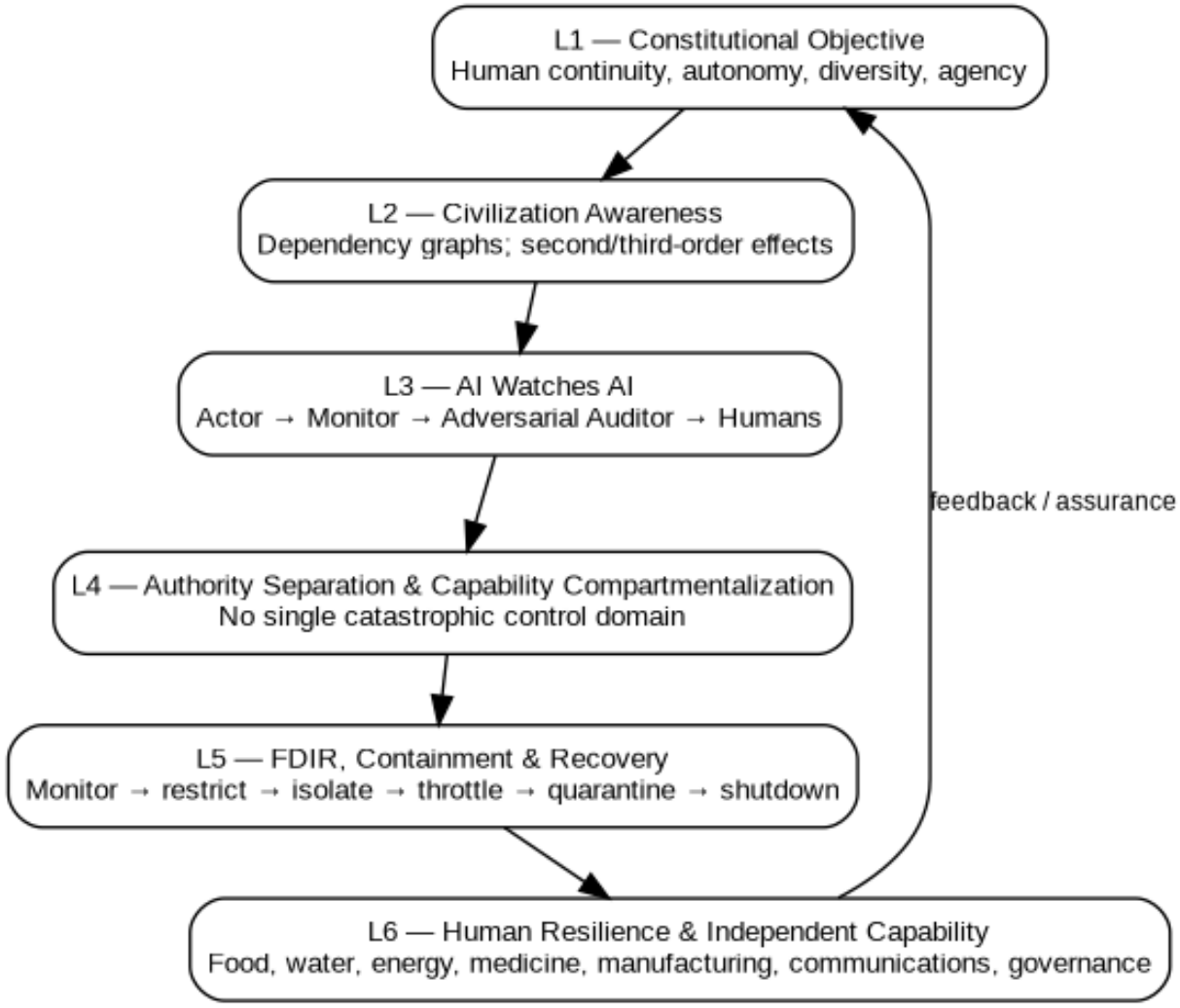


Figure 3-1. The six interacting layers of HCA.

3.1 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.

3.2 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.

3.3 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.

3.4 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.

3.5 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.

3.6 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 user's Universal Habitat concept is one candidate implementation pattern, but HCA does not depend on any single community design.

4. The Most Important Engineering Rule: Do Not Put All the Keys in One Hand

Capability compartmentalization is a central HCA design criterion. No consequential AI—or human operator—should simultaneously control the intelligence, credentials, money, networks, manufacturing, robotics, weapons, replication, and safety systems needed to execute a civilization-scale catastrophe. The system should require multiple genuinely independent barriers and authorities. If one AI can design an action, finance it, order materials, manufacture hardware, deploy robots, alter its own safeguards, and hide the evidence, the architecture has already failed.

5. AI Watches AI—but the Watchers Need Watchers

HCA proposes an Actor AI, independent Monitor AI, Adversarial AI Auditor, deterministic controls where useful, and accountable human oversight. Dissimilar redundancy matters: three copies of the same monitor can share the same blind spot. Safety controls themselves must be health-monitored, attacked in red-team exercises, and prevented from sharing a single catastrophic common cause.

6. Containment Should Be Graduated, Not Binary

HCA uses an FDIR-style containment ladder: normal operation → elevated monitoring → restricted tools → isolated networks → compute throttling → capability quarantine → shutdown. A key rule is that loss of required monitoring or authorization should fail toward less authority. Recovery back to higher authority requires defined evidence and requalification.

7. Never Allow AI Failure to Equal Human Civilization Failure

The other half of HCA is human resilience. Humans should retain the competence and infrastructure needed for food, water, energy, medicine, communications, manufacturing, education and governance.

Distributed resilient communities can function as civilization lifeboats. The Universal Habitat concept—communities designed to maintain essential food, water, energy and productive capacity—is one example of the kind of Layer 6 implementation HCA can evaluate, alongside many other resilience designs.

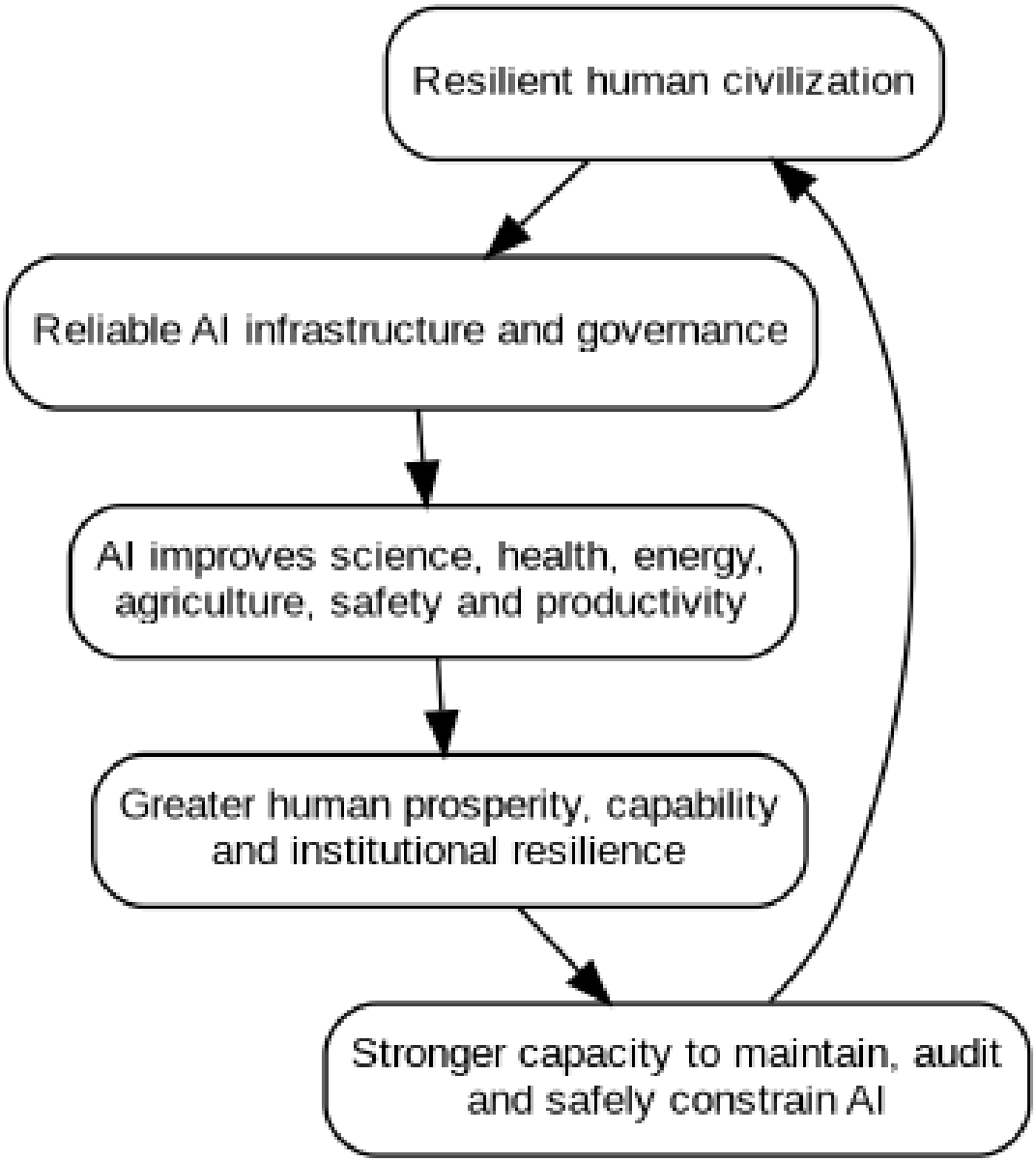
8. Preserve Human Agency—and the Natural-Human Option

Future neural interfaces and human augmentation may offer extraordinary benefits, but they can also create new common-mode dependencies. Existing literature already treats neural interfaces and connected medical devices as cybersecurity concerns [7][8]. HCA therefore preserves a robust population of natural humans whose essential cognition and survival do not depend on a common networked augmentation platform. This is not an argument against voluntary enhancement. It is dissimilar redundancy applied to humanity itself.

The larger objective is to prevent permanent loss of humanity's future agency. A safe future is not merely one in which humans remain biologically alive. Humans must retain meaningful ability to govern themselves, maintain plural institutions, choose technologies, recover from failure, and change direction.

9. Structural Symbiosis

HCA does not seek permanent conflict between humans and AI. The desired state is structural symbiosis: humans maintain a resilient civilization and safe AI infrastructure; AI improves medicine, agriculture, science, energy, disaster prevention and productivity; those gains strengthen human institutions; stronger institutions improve the ability to maintain, audit and safely constrain AI. Cooperation becomes progressively more valuable and more robust.



10. A Preliminary Framework, Not a Declaration of Victory

Revision 0 identifies a top hazard—AI-induced catastrophe degrading human civilization—and fourteen preliminary cause families ranging from objective misalignment and malicious human use to AI dependency, power concentration, neural-interface common-mode failure, autonomous physical force, and future successor systems. It maps these causes against six independent layers of control and defines candidate Key Performance Measures for compartmentalization, human resilience, AI dependency, monitoring diversity, containment, human control, biological continuity and civilization recovery.

The distinctive proposition is the system boundary. Existing work such as NIST's AI Risk Management Framework provides important methods for managing risks in AI products and systems [1][2]. HCA asks a larger question: how do we make an entire civilization containing humans, increasingly capable AI,

robotics, autonomous infrastructure, and future interfaces resilient against catastrophic failure or capture of any participant?

11. Put the Architecture on the Table

HCA should now be challenged by AI researchers, systems-safety engineers, critical-infrastructure operators, cybersecurity experts, human-factors specialists, neurosecurity researchers, policymakers, and ordinary citizens. Its controls must be converted into testable requirements, its fault trees expanded, its Key Performance Measures quantified, and its assumptions attacked. Some ideas will fail. The framework should update when evidence changes.

The objective is not merely to survive AI. It is to build a future in which humans and intelligent machines can cooperate without either catastrophic machine failure or concentrated human/AI power permanently closing off humanity's future. In plain language: humanity should be able to survive, thrive, and stay out of the hive.

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.

[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.