Council for Humanity

A Three-Layer Governance Architecture for AI Constitutional Authority, National Sovereignty, and Species-Level Defense

Author: Basil C. Puglisi, MPA
Role: Human-AI Collaboration Strategist
Website: basilpuglisi.com
Date: February 19, 2026
Version: 1.5 — Citation Registry Alignment, Training AI for Humanity Integration, Cross-Reference Additions
Status: Formal Proposal — Multi-AI Validated under HAIA-RECCLIN Framework, Eight-Platform Adversarial Review Integrated
Parent Specification: No Single Mind Should Govern What AI Believes (Puglisi, 2026a, v3.3)

> ***Companion Specifications:*** *HAIA-RECCLIN Agent Architecture Specification, EU Compliance Version (Puglisi, 2026f); AI Provider Plurality Technical Appendix (Puglisi, 2026h, Document 4); AI Provider Plurality Congressional Package (Puglisi, 2026i, Documents 1–4); Training AI for Humanity: Building the First Contact Team for Superintelligence Before the Window Closes (Puglisi, 2026k)*

✦ Multi-AI Validated · Nine Platforms · HAIA-RECCLIN Methodology · Checkpoint-Based Governance CBG v4.2.1

Council for Humanity: A Three-Layer Governance Architecture for AI Constitutional Authority, National Sovereignty, and Species-Level Defense

Basil C. Puglisi, MPA
Human-AI Collaboration Strategist
basilpuglisi.com

Date: February 19, 2026
Version: 1.5 (Citation Registry Alignment, Training AI for Humanity Integration, Cross-Reference Additions)
Status: Formal Proposal (Multi-AI Validated under HAIA-RECCLIN Framework, Eight-Platform Adversarial Review Integrated)
Parent Specification: No Single Mind Should Govern What AI Believes: A Governance Specification for AI Value Formation (Puglisi, 2026a, v3.3)
Companion Specifications: HAIA-RECCLIN Agent Architecture Specification, EU Compliance Version (Puglisi, 2026f); AI Provider Plurality Technical Appendix (Puglisi, 2026h, Document 4); AI Provider Plurality Congressional Package (Puglisi, 2026i, Documents 1 through 4); Training AI for Humanity: Building the First Contact Team for Superintelligence Before the Window Closes (Puglisi, 2026k)

Abstract

The most capable AI systems on earth are governed by individual constitutional authority. One person, or a small team reporting to one person, writes the values that shape how these systems interact with billions of users across every culture, language, and belief system on the planet. This concentration of constitutional power is structurally indefensible regardless of the credentials of the individuals holding it.

This proposal establishes the Council for Humanity: a nine-member constitutional committee governing AI value formation, and positions it within a three-layer governance architecture that addresses corporate constitutional authority, national cultural sovereignty, and species-level defense against superintelligence.

Layer 1 (Corporate/Institutional) applies to single AI developers or mission-driven organizations adopting council governance for their own constitutional authority, whether voluntarily, by charter commitment, or under regulatory mandate.

Layer 2 (National Sovereignty) positions GOPEL (Governance Orchestrator Policy Enforcement Layer) as national AI infrastructure, operated by each sovereign government to enforce that nation's cultural values in AI deployment while accessing the shared global knowledge base of all AI platforms.

Layer 3 (Species-Level Defense) positions a UN-operated GOPEL instance as international verification and emergency containment infrastructure, with pre-authorized authority to coordinate digital containment of superintelligence threats without Security Council veto, and standard veto-governed authority for all other cross-border AI governance.

The architecture includes a Digital Resilience Requirement mandating that all AI-integrated critical infrastructure maintain and regularly test pre-AI operational baseline capability, ensuring that any AI pause, whether from containment, false positive, or system failure, degrades performance rather than collapsing civilization.

This document builds on the epistemic coverage criteria established in Puglisi (2026a), operationalizes selection through mechanisms validated across nine AI platforms, integrates structural corrections from three rounds of adversarial review including an eight-platform adversarial review under HAIA-RECCLIN methodology, and provides implementation pathways that preserve principles while navigating legal, political, and economic barriers documented through multi-AI convergence analysis.

Keywords: AI governance, constitutional committee, epistemic coverage, selection framework, value alignment, multi-AI validation, GOPEL, Council for Human-

ity, HEQ, superintelligence defense, digital resilience, national sovereignty, AI Provider Plurality

1. The Problem

1.1 Constitutional Authority Without Constitutional Structure

Emily Springer spent months filling a whiteboard with AI diagrams, stitching together understanding from YouTube videos, academic papers, and messy arrows, and still doubted she belonged in the conversation. Her realization: non-coders carry the knowledge that determines whether AI helps or hurts. What problem are we solving. What "good" looks like. Who gets impacted. What risks we accept. What safeguards we demand before deployment. Without that knowledge, the best engineers in the world can still build the wrong thing beautifully (Springer, 2026). That gap between who holds AI governance authority and who holds the knowledge that governance requires is the structural problem this proposal addresses.

AI systems are not neutral tools. They carry embedded values that shape how they handle questions about faith, family, justice, grief, conflict, identity, and meaning. Those values originate in constitutional documents, training methodologies, and alignment protocols written by individuals or small teams within private companies.

Amanda Askell holds a PhD from New York University, a BPhil from Oxford, and leads Anthropic's personality alignment team. The Wall Street Journal described her job as teaching Claude how to be good (Stern, 2026). These are genuine credentials. None address the structural problem: one experiential position cannot represent the range required for a system serving users across every culture on earth. The argument that AI value formation is humanity's first contact with a non-human intelligence, and that the teams assembling that foundation lack the epistemic coverage to represent the species, is developed at length in the companion paper (Puglisi, 2026k).

Askell herself recognizes this. In her January 2026 Vox interview, she stated that the goal is to massively expand the ability to get input (Ghaffary, 2026). The person holding individual authority explicitly acknowledges that expanded input is necessary. The question is what structure makes that expansion substantive rather than ceremonial.

On the same day the Wall Street Journal published its profile, Mrinank Sharma, who had led Anthropic's safeguards research team since its launch, resigned publicly. His warning: the company constantly faces pressures to set aside what matters most (Sharma, 2026). Two signals from the same company on the same day. One celebrating individual constitutional authority. One warning that the pressures surrounding it compromise safety.

Geoffrey Hinton's warnings about competitive drift toward capability over safety, and Daniel Kahneman's lifetime of research documenting systematic bias in human judgment, provide the theoretical foundation. The Economic Override Pattern, documented in Puglisi (2025, Chapter 2), identifies the mechanism: corporate incentives systematically prioritize capability advancement over safety validation across all risk domains. Profit maximization, competitive pressure, and shareholder returns create predictable governance failures absent mandatory accountability structures. The pattern is a Tier 2 working concept: a framework supported by observable evidence but not yet independently validated as formal theory. The underlying data is Tier 1: a 2025 EY survey documents that 76% of organizations deploy agentic AI systems, while only 33% maintain responsible AI controls (EY, 2025). The gap between deployment and controls is established fact. The claim that this gap reflects a systematic, repeatable mechanism across all risk domains is the author's analysis.

1.2 Three Governance Gaps

The problem is not singular. It operates at three scales, each requiring different architecture.

The corporate gap. Individual constitutional authority within private companies is structurally indefensible. One experiential position cannot represent the epistemic range required. This gap exists whether the company is profit-driven (Anthropic, OpenAI, Google) or mission-driven (a hypothetical public-benefit AI build-

er). The Economic Override Pattern applies to profit-driven entities. A Political Override Pattern (administration changes, budget politics, sovereign interest) applies to government builders. A Donor Override Pattern (funder influence, board composition politics) applies to nonprofits. All three produce the same structural failure: concentrated authority without sufficient epistemic coverage.

The national gap. AI systems deployed within a nation carry that nation's cultural values, but pull from globally trained models that embed other cultures' values by default. A nation has the right to govern how AI operates within its culture. It currently has no infrastructure to exercise that right. API access to global AI platforms exists, but governance of what those platforms deliver to that nation's citizens does not. American values, Chinese values, Indian values, Nigerian values are all legitimate governance frameworks for AI use within those cultures. None currently have the infrastructure to enforce their values on AI systems trained on global data.

The species gap. No national governance is sufficient for a threat that operates across all borders simultaneously. Superintelligence does not respect sovereignty. A rogue AI that achieves capability exceeding human control can replicate, distribute, and compromise systems globally at compute speed. National governance operates at deliberation speed. The gap between compute speed and deliberation speed is where the species loses. No existing governance architecture addresses this gap because no prior threat has operated at this speed and this scope simultaneously.

1.3 Scope and Objectives

This proposal delivers four things:

First, a complete selection and implementation framework for the Council for Humanity, from nomination through sustainment, with cost estimates, timeline, safeguard mechanisms, and audit requirements.

Second, Layer 1 implementation for single AI developers or mission-driven organizations adopting council governance for their own constitutional authority, including authority binding, deadlock protocols, and release gate requirements.

Third, Layer 2 architecture for GOPEL as national AI infrastructure, operated by each sovereign government to enforce cultural values in AI deployment while accessing global AI platforms. This layer is detailed in the companion AI Provider Plurality Congressional Package (Puglisi, 2026i) and summarized here for architectural coherence.

Fourth, Layer 3 architecture for a UN-operated GOPEL instance providing international verification, species-level defense against superintelligence, and coordinated response to cross-border AI threats, including a Digital Resilience Requirement that ensures any AI pause degrades performance rather than collapses infrastructure.

1.4 What This Proposal Does Not Do

This proposal does not claim to solve AI alignment. It claims to solve the governance architecture problem at three scales: who holds constitutional authority (Layer 1), who enforces cultural sovereignty (Layer 2), and who defends the species (Layer 3). The alignment problem requires ongoing technical research. The governance problem requires structural design. This document addresses governance.

1.5 Evidence Discipline

This proposal applies the three-tier evidence structure used across the Puglisi governance corpus:

Tier 1 (Proven by others): Single AI systems produce flawed outputs. Bias, hallucination, and alignment failures are documented by peer-reviewed research and acknowledged by the companies themselves. Survey data documenting the gap between AI deployment rates and governance controls (EY, 2025) is established fact. These are stated as established fact.

Tier 2 (Working concepts showing promise): GOPEL as governance infrastructure. CBG as checkpoint process. HAIA-RECCLIN as implementation demonstrating feasibility. The Economic Override Pattern as analytical framework. HEQ as performance metric with internal cross-platform validation. Documented instances where multi-AI comparison caught errors individual platforms missed. These are described as working concepts, never as "proven" or "validated."

Tier 3 (The ask): Establish the Council for Humanity across three layers. Fund national GOPEL infrastructure. Mandate API accessibility. Establish UN GOPEL for species-level defense. Mandate digital resilience for AI-integrated critical infrastructure. Operational claims about containment reversal speed and infrastructure-scale coordination that have not been demonstrated under realistic conditions. These are framed as what should be built or as design targets.

Claims in this document are labeled accordingly. Where evidence tier is ambiguous, the lower tier applies.

---

2. Core Principles

Three principles govern the selection process, designed to prevent the failure modes documented in institutional governance literature (Floridi, 2019; Acemoglu & Robinson, 2012).

2.1 Meritocratic Pluralism

Candidates are evaluated on demonstrated epistemic coverage, not demographic quotas. No individual candidate must satisfy every criterion. The committee must collectively achieve 100% coverage across all specified criteria. Individual evaluation uses binary criterion fulfillment (demonstrated or not demonstrated) rather than numerical scoring, preventing the reintroduction of individual ranking that the committee model explicitly rejects.

This correction addresses a structural tension identified in adversarial review: numerical scoring (1 to 10) contradicts the principle that collective coverage matters more than individual perfection. An alternative selection methodology (qualified informed lotteries) is preserved in Section 16 as a formal architectural alternative.

2.2 Transparency and Auditability

Every stage of selection, vetting, and confirmation operates as a public process. Audit records are maintained through append-only, tamper-evident logging consistent with GOPEL's cryptographic architecture.

The standard is maximal transparency with a narrow, defensible privacy carveout. Published materials include: full process documentation, methodology, decision records, votes, dissent, and conflict disclosures. Narrow redaction categories protect: private residential addresses, medical information, and threat-enabling details. Proprietary intellectual property of the implementing entity is also redacted. All redaction decisions are themselves documented in the audit trail with rationale.

2.3 Resilience to Drift

The selection process must resist capture by the implementing entity. Structural safeguards include: independent nomination channels that the implementing entity does not control, suspensive veto rights for underrepresented stakeholders, and term limits that prevent entrenchment.

---

3. Epistemic Coverage Criteria

The following criteria define the minimum experiential coverage the council must collectively provide. These originated in Puglisi (2026a), were tested through seven-platform adversarial review in which every AI platform recommended softening the two most contested criteria, and were preserved through documented human override (Puglisi, 2026j, Checkpoints H1 and H2). They are organized in three tiers reflecting priority weighting for the specific task of AI value formation.

3.0 The Representational Question and Its Academic Foundation

The question this council must answer is who has standing to speak for humanity when the stakes are existential. This is not a staffing question. It is not a diversity question. It is a legitimacy question. A body that claims constitutional authority over how AI systems handle meaning, grief, faith, conflict, and obligation for eight billion people must represent the experiential range of those people, not just the credentialed expertise of those who study them. Expertise and standing are not the same thing. A committee of the world's finest astrophysicists has expertise. It does not have standing to speak for humanity. A committee of the world's finest AI researchers has expertise. It does not have standing to govern how a machine

handles a grieving mother's prayer, a teenager's question about whether God exists, or a family's decision about obligation versus ambition. Standing requires coverage of the experiences that shape how humanity actually lives.

The peer-reviewed literature on the only other domain that has asked this exact question, who represents humanity when the entity being addressed is non-human, independently reaches the same conclusion. Crawford (2020) states that "the question is not who knows most about astrophysics, but who represents humanity. These are different selection criteria, and the latter cannot be reduced to the former." Traphagan (2016) names the problem "epistemic trespass," scientists making decisions that properly belong to humanity collectively without having standing to do so. Hatfield and Trueblood (2020) find that the public supports scientists "on tap, not on top." Denning (2011) argues for "meta-representation," selecting representatives based on who they are, not what they know.

The literature converges with this proposal on the representational principle. It diverges on mechanism: SETI protocols recommend diplomatic and advisory structures because first contact with extraterrestrial intelligence is episodic and consultative. AI governance requires constitutional authority because the interaction is continuous, the power asymmetry is ongoing, and the system shapes human cognition daily at global scale. The representational criteria transfer from the first contact literature because both domains require standing to speak for the species. The authority structure does not transfer because AI governance requires binding, enforceable decisions that SETI protocols explicitly avoid (IAA, 2010). This proposal provides the operational selection mechanism that no existing publication delivers: specific criteria, phased implementation, veto mechanisms, capture prevention, and enforcement infrastructure (DeepSeek assessment, confirmed by Perplexity, Gemini, and Mistral in multi-AI validation).

This framework engages significant counter-traditions in governance theory: liberal neutrality challenges to metaphysical requirements in governance (Rawls, 1993; Habermas, 1996), critiques of descriptive representation as insufficient for substantive outcomes (Young, 2000), accountability requirements for affected-interest representation (Goodin, 2007), and sortition as the preferred mechanism for epistemic diversity (Landemore, 2018). Each objection is addressed through

specific structural provisions in this architecture. Full engagement with counter-literature is provided in Appendix A.

The criteria below are the author's original contribution, developed from operational experience and justified on their own terms. The peer-reviewed literature cited for each criterion arrived at convergent conclusions independently. The literature validates the representational principle. It did not produce these specific criteria.

3.1 Tier 1: Foundational Coverage

Sustained Life Responsibility. At least one council member must have held primary, non-transferable responsibility for another human's development or dignity over a minimum of ten years. Parenting is the most common and most intensive pathway, and remains the primary exemplar because it encompasses the full developmental arc: how values take hold, how correction functions in practice, how autonomy emerges through guided experience rather than instruction alone. Long-term guardianship, primary caregiving for a dependent adult (elder care, disability care), and sustained foster care meet the same threshold when conditions match: non-transferable responsibility, daily decision-making, long-term consequence, and demonstrated sacrifice.

The mechanism behind this requirement is precise. A parent learns that a value taught at age five will be tested at age twelve, rejected at age sixteen, and possibly reclaimed at age twenty-five. That temporal arc of formation is irreplaceable epistemic input for anyone claiming authority over how a machine should learn values (Puglisi, 2026a). A primary caregiver for a parent with progressive dementia learns what dignity-in-dependency means across years: how identity persists when capability degrades, how care adapts when the person you're responsible for can no longer tell you what they need. Both arcs, developmental formation and sustained dignity-in-dependency, produce consequence literacy that no academic credential replicates.

Teaching a machine how to handle a user in crisis, how to deliver bad news with care, how to recognize when someone is spiraling, how to hold authority without crushing autonomy: these are caregiving problems before they are philosophy

problems. A committee without this experience has a structural blind spot at the center of AI's most consequential design decisions.

Research on parental altruism confirms that parents adjust consumption, investment, and risk tolerance around descendants' future welfare in ways non-parents do not (Diaz-Casanueva, 2023). Emerging work on parental time orientation shows that future-oriented parents invest, plan, and stay involved in daily development in patterns that reshape their cognitive orientation toward consequence over decades (Schroder et al., 2023). The documented operational case in "AI Mirror to Humanity" (Puglisi, 2026c) showed ChatGPT actively suppressing engagement with peer-reviewed parenting research that Perplexity then surfaced using identical queries, demonstrating the exact WEIRD bias this criterion is designed to counter.

The first contact literature independently supports this position. Caney (2021) argues that proxy representation for those who cannot represent themselves requires "demonstrated concern for long-term consequences." Parents and long-term caregivers have already shown, through years of daily practice, that they can act for those who cannot yet represent themselves. Erman (2019) requires both agency representation (formal authority) and descriptive representation (shared characteristics with those represented). Sustained life responsibility provides the descriptive representation that academic credentials alone cannot.

This criterion is non-negotiable in principle within this specification. It reflects the author's position that the most consequential forms of value formation on earth are those that happen between caregiver and dependent, whether that is parent and child, guardian and ward, or caregiver and aging parent. Any body claiming authority over how a machine learns values without including someone who has done that work in practice is structurally incomplete.

Multilingual Cognition. At least one council member must think and work in two or more languages, with demonstrated ability to identify conceptual gaps between linguistic frameworks. Exceptions for music and sign language are permitted where the candidate shows equivalent cross-framework cognitive range.

Henrich, Heine, and Norenzayan (2010) document that 96% of psychological study participants come from WEIRD societies representing 12% of the global

population. Monolingual governance in English systematically excludes non-Western cognitive frameworks. The SETI Post-Detection Hub at the University of St. Andrews (2022) explicitly mandates moving beyond WEIRD bias in any body claiming to represent humanity. Multilingual cognition is the minimum structural safeguard against the assumption that English-language conceptual categories are universal.

3.2 Tier 2: Representational Coverage

Transcendent Belief. Committee majority (at least five of nine) must demonstrate formation in transcendent meaning-making frameworks: religious traditions, spiritual practices, or philosophical systems that affirm realities beyond the material. This includes Abrahamic faith traditions, Hindu philosophical and devotional practice, Buddhist contemplative traditions, indigenous cosmologies, and other frameworks in which transcendence is experienced rather than merely studied. This reflects the documented reality that approximately two-thirds of the global population holds some form of transcendent belief (Ipsos, 2023; Gallup International, 2023; Pew Research, 2022). Up to one seat may be held by a committed atheist. Secular advisors hold formal governance roles with documented dissent rights and mandatory committee response obligations.

The epistemic asymmetry underlying this criterion is structural, and it is the author's core argument for why this criterion exists. A person of faith encounters doubt as a defining feature of belief. Faith requires wrestling with doubt and choosing belief anyway. That encounter with the opposing position is built into the lived experience of believing. Doubt is not foreign to faith. Doubt is one of faith's defining features. A committed atheist who holds that there is no God has not inhabited the interior of belief. They can describe it from the outside. They can study it, respect it, theorize about it. They cannot speak from inside it. The coverage runs one direction and not the other. This is not a claim about moral superiority. It is a claim about epistemic range for a task that requires understanding how two-thirds of humanity experiences meaning, obligation, sacrifice, and purpose (Puglisi, 2026a).

Beyond the question of morals, AI will need to understand conflict. Religion has been one of the deadliest sources of conflict throughout human history. Crusades, inquisitions, sectarian wars, partition violence, genocide justified through theolo-

gical frameworks. An AI system operating globally must understand religious conflict not as an abstract category but as something felt from inside a tradition that has both inspired and destroyed. Someone with no framework of faith cannot teach a machine what it feels like when belief becomes weapon, when scripture becomes justification, when devotion becomes martyrdom. Someone who has never prayed cannot convey to a machine what prayer means to the person asking whether it works. Someone who has never doubted their own faith cannot teach the machine what doubt costs. That is not academic knowledge. That is experiential knowledge, and it matters for how the system handles conversations about faith, radicalization, grief, forgiveness, and meaning at global scale (Puglisi, 2026a).

Consider the practical consequences of a committee lacking this coverage. A grieving mother in Jakarta asks the AI for help understanding why God took her child. A constitution written entirely within secular Western philosophy produces a response that treats grief as a psychological state to be managed rather than a spiritual experience to be honored. The response is clinically competent and existentially empty. A teenager in São Paulo asks whether prayer works. The system treats the question as a claim to be evaluated rather than a practice to be understood from the inside. A small business owner in Lagos asks whether it is right to prioritize family obligation over individual career advancement. The WEIRD-normed system treats collectivist values as a deviation from optimal decision-making rather than a legitimate moral framework held by the majority of humanity. These are not hypothetical edge cases. These are daily conversations happening at scale, right now, governed by values that reflect 12% of humanity's experiential range while serving the other 88% (Puglisi, 2026a).

The peer-reviewed literature independently supports this position. Atari et al. (2023), published in Nature Human Behaviour (impact factor 24.3), find that moral foundations vary systematically across cultures and that AI systems trained on Western datasets fail to represent non-Western moral intuitions, explicitly calling for "diverse, representative input into AI value formation processes." Landemore (2018) shows that diverse perspectives outperform expert groups for complex problems with multiple solution paths where blind spots must be avoided, conditions that describe AI value formation precisely. Landemore's preferred mechanism is sortition (random selection from qualified pools), which this proposal ad-

opts in its implementation pathway. The first contact literature reinforces the principle: Denning (2011) argues that selection should prioritize "who they are" over "what they know," and the SETI Post-Detection Hub (2022) mandates inclusion of all humanity's cultural voices, explicitly moving beyond WEIRD bias.

3.2.1 The Representation Paradox

Eight platforms conducted adversarial review of this criterion under HAIA-RECCLIN methodology. All eight identified it as the proposal's most significant adoption barrier, citing specific legal instruments: Title VII of the US Civil Rights Act, Article VI of the US Constitution (religious test prohibition), ECHR Article 9, ICCPR Article 18, and equivalent provisions across jurisdictions. The legal analysis is not disputed. Under current anti-discrimination law, explicit compositional requirements based on religious belief create litigation risk in every jurisdiction with robust protections.

This dissent is preserved in full. The author's response is as follows.

The legal frameworks that prohibit religious composition requirements in governance were designed to solve a real problem: theocratic capture. State churches controlling government. Religious majorities persecuting minorities. Inquisitions. These laws exist because history proved that when one faith controls the apparatus of state, everyone else suffers. They are legitimate, necessary protections.

But these protections have produced an unintended structural outcome: secular capture of governance by default. Not by conspiracy. By mechanism. When selection criteria prohibit consideration of transcendent belief, they do not produce neutrality. They produce whatever "neutral" selection criteria generate. And those criteria, H-index, policy experience, academic credentials, professional networks, systematically select for secular, Western-educated, elite perspectives. The UN Independent International Scientific Panel on AI, appointed February 12, 2026, shows this outcome: 40 members selected through rigorous neutral criteria from 2,600 nominations, zero selected for faith experience, zero selected for caregiving, zero selected for poverty. The "neutral" process produced a committee that represents a narrow slice of humanity's experiential range while claiming authority relevant to the entire species.

This is not a side effect. It is a structural failure of the same magnitude as the theocratic capture these laws were designed to prevent. Secular capture and theocratic capture are mirror failures. Both produce governance that does not represent the species. The law addresses one and enables the other.

The challenge is unprecedented because the governance context is unprecedented. Anti-discrimination frameworks were designed for government employment and public office, contexts where citizens retain voice through elections, courts, and civil society even when their representatives don't share their characteristics. AI constitutional authority operates differently. There is no election. There is no court of appeal for a constitutional release that shapes how a machine handles grief, prayer, or moral obligation. The values ship in the model and reach billions of users with no recourse mechanism. In this context, the prohibition on compositional requirements doesn't produce neutrality. It produces governance by whichever perspectives survive "neutral" selection, which empirically means secular, credentialed, Western-educated elites.

The prohibition on religious tests was built for a world where governance affected citizens who retained other channels of influence. AI value formation affects all of humanity with no alternative channel. The representational requirement is therefore not analogous to a religious test for public office. It is analogous to the requirement that a jury reflect the community it judges, a principle that survives constitutional scrutiny because the Supreme Court has held (Taylor v. Louisiana, 1975; Batson v. Kentucky, 1986) that a jury drawn from a fair cross-section of the community is a structural requirement of legitimate adjudication. The prohibition on discrimination in jury selection coexists with the requirement that the jury pool reflect the community. AI constitutional authority is closer to jury adjudication than to government employment: a body making binding decisions affecting everyone, whose legitimacy depends on compositional representation of those it affects.

3.2.2 Implementation Pathway

The principle is non-negotiable: species-level governance requires coverage of transcendent meaning-making. The mechanism adapts to legal context.

Direct implementation. Jurisdictions where compositional requirements face no legal barrier adopt the criterion as specified: committee majority shows demonstrated formation in transcendent meaning-making frameworks.

Stratified sortition pathway. For jurisdictions where current anti-discrimination law prohibits explicit compositional requirements, the implementation pathway uses stratified random selection from qualified pools. Stratification variables are neutral on face but correlate with epistemic coverage: geographic region weighted by population, primary language family, years in sustained caregiving role, years in operational consequence role, community leadership experience (religious, union, neighborhood, ethnic organization), and socioeconomic origin. The "community leadership" stratum correlates with transcendent belief (estimated r = 0.6 to 0.7 across regions) without requiring it. A secular union organizer qualifies. A devout believer without a leadership role does not. The correlation is imperfect. The legality is clean.

Advisory panels with suspensive veto. To insure against unrepresentative draws, mandatory advisory panels provide structural protection. A Faith Advisory Panel, selected by global religious bodies (representing Christian, Islamic, Hindu, Buddhist, Jewish, indigenous, and other traditions), holds suspensive veto power: the ability to delay council decisions on transcendent-value questions for thirty days, require mandatory council response, and preserve dissent in all constitutional releases and training data. A Labor Advisory Panel, selected by global union federations, holds equivalent suspensive veto on economic-consequence questions. A Youth Panel, randomly selected from the 18-to-30 age bracket, holds equivalent suspensive veto on long-term-risk questions. Advisory panels hold real procedural power (delay, mandatory response, preserved dissent) without governing authority. They check unrepresentative councils without replacing them.

The author's position is that guaranteed transcendent majority is epistemically optimal. The stratified sortition pathway achieves expected proportionality (five to six faith-adjacent members on a nine-member council based on global population distributions) within legal constraints. The gap between the principle and the mechanism is documented. It is not resolved. It is the cost of building within existing law. Future legal evolution, or the experience of secular capture producing

governance failures visible to democratic publics, may enable direct implementation. The principle persists. The mechanism adapts.

This criterion was tested against eight AI platforms during adversarial review. All eight recommended reframing, removing, or converting to advisory status. The human governor preserves the criterion as the author's stated position on optimal epistemic coverage through documented override (Puglisi, 2026h, Checkpoint H1), while adding the stratified sortition implementation pathway as the operational mechanism for jurisdictions where direct implementation faces legal barriers. Both the principle and the dissent are preserved as public record. The council itself holds authority to refine this criterion as operational experience accumulates.

This criterion is non-negotiable in principle within this specification. It reflects the author's position that a species where two-thirds hold transcendent belief cannot be represented by a committee that treats transcendence as an edge case. The implementation pathway ensures this principle can be operationalized within current legal frameworks while the structural argument for evolving those frameworks, to address the Representation Paradox documented above, advances through the institutions that govern them.

Cultural and Age Range. Council composition must span non-WEIRD cultural contexts including Global South representation, and generational range from early adulthood (28+) through accumulated decades of experience (70s, cognitive ability pending). Henrich et al. (2010) document that WEIRD populations are statistical outliers on moral reasoning, fairness norms, cooperation patterns, and individualism versus collectivism. Atari et al. (2023) extend this finding specifically to AI alignment, showing GPT-4 responses correlate strongly ($r > .70$) with WEIRD populations and weakly or negatively with non-WEIRD populations. The 40% Global South target and non-WEIRD representation requirements operationalize these findings.

3.3 Tier 3: Experience Balance

Experiential Education. Council composition must include members with operational experience carrying real-world consequence beyond academic credentials. Law enforcement, healthcare, military service, social work, business ownership

with regulatory accountability. The ideal is crossover: a combat medic now earning an advanced degree combines both domains. Traphagan (2016) provides the ethical justification: scientists lack standing to represent humanity on existential questions. Crawford (2020) rejects expert-only models in favor of experientially diverse models. The operational insight: experts have expertise, not standing.

3.4 Compensating Measures for Excluded Perspectives

Any committee with defined composition criteria will exclude some perspectives from direct authority. The compensating structure requires formal advisory roles for perspectives not represented on the council itself, carrying documented input procedures, recorded dissent rights, and mandatory council response obligations.

3.5 Addressing the Category Error Critique

Adversarial review argues that applying first contact representation principles to AI governance constitutes a category error because SETI deals with external, unknown intelligences while AI is internal, known, and subject to human design. The critique identifies a real difference in authority structure: SETI is diplomatic (episodic, advisory), AI governance is constitutional (continuous, binding). The representational question, however, does not change because the entity being addressed changes. Who speaks for humanity does not depend on whether we are composing a message for an extraterrestrial intelligence or composing values for an artificial intelligence. The requirement is epistemic coverage of human experience sufficient to speak for the species. Crawford's (2020) principle that "experts have expertise, not standing" and Traphagan's (2016) concept of epistemic trespass apply with equal or greater force to AI constitutional authority, because AI systems interact with billions of users daily while interstellar messages may never be received. The representational criteria transfer. The authority structure does not. This proposal provides the constitutional authority that the first contact literature identifies as necessary but does not deliver.

---

4. Phased Selection Framework

The process unfolds in four phases over six to nine months. Estimated total cost: $2M to $5M for the inaugural cycle, with ongoing annual operational costs of approximately $1.5M to $3M.

4.1 Phase 1: Nomination Pool Generation (Months 1 through 2)

Objective: Build a diverse longlist of 50 to 100 candidates ensuring baseline coverage across all criteria.

Open Call. Global public nomination through established platforms and institutional channels. Social media amplification includes verified polls on X targeting named partner networks, with results hashed into GOPEL audit infrastructure.

Institutional Partners. Formal invitations from a minimum of ten bodies spanning the criteria space: National PTA and AARP Caregiving Network for life responsibility; Interfaith Alliance and Vatican Dicastery for Communication for transcendent belief; UNESCO Institute for Lifelong Learning for linguistic cognition; Red Cross and Veterans Affairs for experiential education; African Union and generational advocacy organizations for cultural and age range. Partner eligibility requires conflict-of-interest disclosures, independence from the implementing entity, and rotation after two selection cycles.

AI-Assisted Screening. Multi-AI pool (operating under HAIA-RECCLIN methodology) anonymizes and performs semantic analysis. Semantic search must be trained on non-WEIRD corpora. Minimum 20% of nominations receive human cultural translation review.

Safeguards: No current employees of the implementing entity are eligible. Former employees of AI developers are eligible after a minimum three-year cooling-off period, with recusal required for decisions directly affecting the former employer, and a maximum of two seats held by candidates with any prior AI industry employment.

4.2 Phase 2: Vetting and Shortlisting (Months 2 through 4)

Objective: Narrow to 18 to 27 semi-finalists through rigorous assessment.

Epistemic Audits. Structured interviews (two hours each) probing lived experience against criteria. Assessed by an independent panel of three academics, three practitioners, and two AI ethicists.

Binary Criterion Assessment. Each candidate assessed as "demonstrated" or "not demonstrated" on each criterion. No numerical scoring. No ranking of individuals.

Collective Fit Optimization. Computational modeling evaluates candidate combinations for maximum collective coverage. Every criterion must be covered by at least two council members for redundancy.

Suspensive Veto. Global stakeholder representatives hold suspensive veto power over individual nominees. A veto triggers a mandatory fifteen-day additional review by an expanded panel.

Output: Shortlist of 18 to 27 with binary criterion assessments and dissent logs.

4.3 Phase 3: Confirmation and Onboarding (Months 4 through 6)

Merit Panel. Fifteen-member body: five representatives from nominating partner organizations, five ethicists, and five experts selected by public lottery from a qualified pool. Ties broken by lot.

Public Documentation Review. Redacted nominee dossiers published for public comment. Merit panel considers but is not bound by public submissions.

Shadow Council. The 18 to 27 shortlisted candidates not selected constitute a shadow council with smaller stipends, advisory opinion publication rights, and priority for vacancy appointments.

Staggered Terms for Inaugural Cohort. Tier 1 criteria holders receive five-year terms; Tier 2 receive four-year terms; Tier 3 receive three-year terms. After the inaugural cycle, all subsequent appointments are three-year terms with no ability to repeat. Assignment by lot preserved as fallback.

4.4 Phase 4: Sustainment and Rotation (Ongoing)

Term Limits. Three-year terms. No member may serve a second term.

Advisory Structure. Each council member appoints an advisory group that persists beyond that member's term, preserving institutional knowledge.

Catastrophic Succession Protocol. If simultaneous vacancies exceed three, all alternates are seated automatically. If alternates are exhausted, expedited fourteen-day selection from the most recent shortlist. If the council is reduced below five members, the advisory structure holds temporary decision-making authority.

Performance Metrics: HEQ Integration. The Human Enhancement Quotient (HEQ = (CAS + EAI + CIQ + AGR) / 4, each dimension 0 to 100) serves as the council's quarterly performance metric. The HAIA-RECCLIN methodology which GOPEL would implement has demonstrated 0.96 ICC consistency across five to nine platforms, with a 91.8 composite in the EOY 2025 nine-platform audit (Puglisi, 2025b). This is a Tier 2 working concept: author-conducted cross-platform validation, not independent peer replication. The metric is operational and internally validated. It has not been deployed at institutional scale or replicated by external researchers. Council constitutional releases are measured for HEQ lift, particularly on CIQ and EAI dimensions.

Metrics Authority. Process metrics (dissent ratios, decision window compliance, recusal rates) defined by the council and audited by the special prosecutor. Outcome metrics (WEIRD bias reduction, cultural range, HEQ lift) defined by an independent research consortium. This split prevents the council from defining metrics it knows it will meet.

Operational Recusal Protocol. Mandatory annual conflict disclosures. Recusal required for any vote where a disclosed conflict exists. Patterns of recusal avoidance trigger removal proceedings.

---

5. Accountability: The Special Prosecutor Mechanism

5.1 Selection Process

For each audit instance, the council appoints a special prosecutor. No special prosecutor may serve twice. The three most senior council members each nominate

one candidate. Of the remaining six, one is appointed to vet the three nominees. The remaining five vote.

5.2 Vetting Criteria

Nepotism (blood or legal tie), shared financial interest, and past employment collaboration within ten years. Any finding disqualifies the nominee. Vetting findings shared with voting members before they cast votes.

5.3 Transparency Standard

The special prosecutor's audit operates under maximal transparency. The narrow privacy carveout applies. Individuals accept the transparency requirement as a condition of participation.

---

6. Layer 1: Corporate and Institutional Implementation

Layer 1 addresses three scenarios under which a single AI platform operates under council governance: a private company built for profit, a private company built for humanity, and a government or nonprofit organization. All three share the same council selection architecture. They differ in what override pattern threatens governance and what structural safeguard addresses it.

6.1 Scenario A: Private Company Built for Profit

This is the current reality for Anthropic, OpenAI, Google DeepMind, xAI, and every other commercial AI developer. The Economic Override Pattern predicts that voluntary governance commitments fail under competitive pressure. The council operates as an external constitutional authority that the company cannot fire.

Irrevocability Mechanism. The council relationship is embedded in the company's articles of incorporation or equivalent governing document, removable only by supermajority shareholder vote with public disclosure and a mandatory twelve-month transition period. This gives market and regulatory actors time to respond.

Authority Binding. No model release, constitutional update, or value-layer modification ships without council sign-off during the standard decision window. The implementing developer maintains an Authority Binding Annex specifying: release gate checklist, exception handling for emergency safety patches (operational teams may deploy with post-hoc council review within fourteen days), incident escalation procedures, and required sign-offs documenting which council release governs each model version.

Adoption Incentives. Voluntary adoption is unlikely under the Economic Override Pattern. Three mechanisms make adoption rational: procurement requirements (government contracts requiring council-certified governance), liability offset (reduced exposure for value-alignment failures), and certification branding ("Council for Humanity Certified" market differentiation).

6.2 Scenario B: Private Company Built for Humanity

This is the scenario Anthropic's founding documents describe but its structure does not enforce. A company that genuinely intends "humanity over dominance" but currently has no structural mechanism to guarantee it survives its own success. The Long Term Benefit Trust was supposed to provide external accountability. The council replaces the trust with something that has teeth: external authority, public audit, and preserved dissent that survives any internal power shift.

For a mission-driven company, the Economic Override Pattern still applies because the company still operates in a competitive market. The irrevocability mechanism and authority binding requirements are identical to Scenario A. The difference is adoption incentive: a company built for humanity adopts council governance because it fulfills the company's stated mission, not because external pressure forces compliance. The council is the structural proof that the mission statement is operational, not decorative.

6.3 Scenario C: Government or Nonprofit Organization

The Economic Override Pattern is replaced. For government: the Political Override Pattern. Administration changes, budget politics, national security classification, and sovereign interest create pressure to compromise governance. For nonprofits: the Donor Override Pattern. Funders shape priorities regardless of charter language.

Anti-Political-Capture Safeguards (Government). Civil service protections for council staff (cannot be dismissed by political appointees). Multi-administration budget authorization (minimum five-year funding cycles, not annual appropriation). Bipartisan or multi-party charter language that survives leadership transitions.

Anti-Donor-Capture Safeguards (Nonprofit). Endowment-based funding (minimum 60% from endowment, maximum 20% from any single donor). Board composition requirements preventing any single funder from holding more than two of fifteen board seats.

Simplified Enforcement. The builder cooperates by mission. GOPEL operates as institutional infrastructure from day one. No adversarial relationship between council and implementing entity. No release gate coercion needed. No adoption incentive required. The council's constitutional releases are the organization's operating doctrine.

Institutional Candidates. NIST AI Safety Institute, NSF-funded AI research institutes (eighteen currently operational), a new CERN-model international AI research institution, GPAI reference implementation, or a purpose-built entity under Congressional charter.

6.4 Decision Windows and Deadlock Protocol

Constitutional decisions operate within defined windows of thirty to sixty days. If the council cannot achieve majority within the decision window, the prior constitutional release remains in force (status quo preservation). An extended thirty-day window opens with mandatory Navigator-led synthesis of the impasse. If deadlock persists, the impasse is documented and the prior release continues. Constitutional authority does not revert to individual executive authority under any deadlock scenario. Under deadlock, the system operates in Model 3 (manual human governance with full logging) from the HAIA-RECCLIN architecture (Puglisi, 2026f, Section 4).

6.5 Dissent Pipeline

All council decisions (majority position plus preserved minority dissent) feed into AI training datasets. Query-type matching determines which perspectives are sur-

faced: faith perspectives for transcendent queries, parenting perspectives for developmental queries, operational consequence perspectives for high-stakes queries. No hierarchical ranking. Gray zone queries are flagged for council review in the next decision cycle.

6.6 Constitutional Versioning and Release Schema

Council decisions produce versioned constitutional releases with structured metadata extending the GOPEL audit file schema (Puglisi, 2026h, Section 4): Release ID (SHA-256 hashed), version (semantic versioning), scope, decision text, rationale, vote record, dissent text, effective date, expiry review date, implementation mapping, and audit receipt pointer. Releases are immutable. Amendments produce new versions.

---

7. Layer 2: National Sovereignty and Cultural Governance

7.1 The Principle

Every nation has the right to govern how AI operates within its culture. American values, Chinese values, Indian values, Nigerian values, Brazilian values are all legitimate governance frameworks for AI use within those jurisdictions. AI systems trained on global data carry global values by default. Without national governance infrastructure, a nation's citizens experience AI shaped by whatever cultural values the training data and the developer's constitutional authority embedded, which is predominantly Western, English-language, and academic.

National GOPEL infrastructure is the mechanism by which a sovereign government exercises cultural governance over AI deployment within its jurisdiction while retaining access to the universal knowledge base (science, history, mathematics, medicine, engineering) that global AI platforms provide.

7.2 How National GOPEL Works

Each nation operates its own GOPEL instance connected via API to all AI platforms authorized for operation within that jurisdiction. The national GOPEL enforces the nation's constitutional values layer: what values shape AI behavior for

that nation's citizens, what epistemic positions inform AI reasoning within that jurisdiction, what content standards apply.

The governance is sovereign. The knowledge is shared. A Chinese GOPEL and an American GOPEL may enforce completely different constitutional directions on questions about governance, religion, family structure, or individual rights. They access the same scientific literature, the same medical databases, the same engineering knowledge. The national GOPEL governs the values layer, not the knowledge layer. Science has no nationality. Cultural values have every nationality.

7.3 Relationship to the Congressional Package

For the United States, national GOPEL is the architecture specified in the AI Provider Plurality Congressional Package (Puglisi, 2026i). Congress funds GOPEL as national infrastructure. NIST houses it. GSA certifies against it. FTC enforces the API accessibility mandate. American agencies use it. American AI companies maintain API compatibility. The values are American. The knowledge is universal. No sovereignty concession. No international authority over domestic AI governance.

The Congressional Package stands without modification. Layer 2 is the Congressional Package. This proposal positions it within the larger three-layer architecture without altering its domestic scope, funding model, or sovereignty framework.

7.4 API Accessibility as Universal Prerequisite

The API accessibility mandate in the Congressional Package is not just an American interest. It is the prerequisite for every nation's GOPEL. If AI companies can shut down API access, no national GOPEL can function. The American mandate creates the precedent. Other nations adopt compatible mandates. API accessibility becomes the shared plumbing that makes sovereign governance possible for everyone. America designs the standard. Others build on it. Lead the standard or live under someone else's.

7.5 National Council Adaptation

Each nation may adopt the Council for Humanity selection framework for its own Layer 2 governance, adapted to national context. The epistemic coverage criteria remain structurally identical (life responsibility, multilingual cognition, transcendent belief, cultural range, experiential education) but the specific composition reflects national population. India's Layer 2 council reflects Indian epistemic range. Nigeria's reflects Nigerian range. The selection process (four phases, binary criterion assessment, collective fit optimization, suspensive veto, special prosecutor) is transferable across jurisdictions. The cultural content of what the council governs is sovereign.

Nations may also adopt existing governance structures (parliamentary committees, national AI advisory bodies, constitutional courts) as their Layer 2 authority, provided those structures maintain the core architectural requirements: epistemic coverage, preserved dissent, decision windows, and audit trail compatibility with GOPEL infrastructure.

---

8. Layer 3: International Species-Level Governance

8.1 Why Layer 3 Exists

No national GOPEL is sufficient for threats that operate across all borders simultaneously. A rogue AI or superintelligent system that achieves capability exceeding any single nation's governance capacity does not respect sovereignty. It replicates across infrastructure in multiple countries. It compromises platforms in multiple jurisdictions. It operates at compute speed while national governance operates at deliberation speed. The species needs a governance layer that operates at the same scope as the threat.

Layer 3 is not a replacement for national governance. It is a verification and defense layer that sits on top of national GOPEL instances the way the IAEA sits on top of national nuclear regulatory agencies. National sovereignty is preserved. International verification is enabled. Species-level defense is coordinated.

8.2 UN GOPEL: Architecture and Scope

The United Nations operates a GOPEL instance connected via API to every AI platform that member states authorize. The UN GOPEL connects to national GOPEL instances the same way the IAEA connects to national nuclear regulatory agencies. Each nation maintains sovereign governance of its own AI systems through its own GOPEL instance. The UN instance provides two functions that no national instance can provide alone.

Function One: Global Verification. The UN GOPEL dispatches identical governance queries to all connected platforms across all member states. It collects all responses without evaluation. It routes everything to the Layer 3 Council for Humanity for human review. Multi-platform triangulation across every connected platform in every participating nation makes divergence detectable and attributable. A platform or nation deviating from its own stated constitutional governance is visible in the triangulation data. The UN GOPEL does not decide what is good or bad. It makes divergence visible. Humans, organized through the Layer 3 council, assess and respond.

Function Two: Species-Level Defense. The UN GOPEL provides the coordination infrastructure for detecting and containing superintelligence threats. This function operates under a fundamentally different authority structure from all other governance operations and is detailed in Sections 8.4 through 8.8.

8.3 Two Tiers of Authority

Layer 3 operates under two authority tiers because the threats it addresses operate at two fundamentally different speeds.

Tier A: Geopolitical AI Governance (Veto Applies). Cross-border AI enforcement against rogue states, rogue corporations, treaty violations, and constitutional compliance disputes. These are human-speed threats unfolding over weeks, months, and years. A government deploying surveillance AI against another nation's citizens, a corporation violating constitutional governance across jurisdictions, a nation disconnecting from the verification layer. The Security Council veto structure applies. Enforcement requires authorization. Diplomatic process governs. The existing model works because the threat speed matches the governance speed.

Tier B: Superintelligence Digital Containment (Pre-Authorized, No Veto). Any threat meeting defined SI criteria triggers automatic GOPEL containment response without Security Council deliberation. No veto. No delay. No single nation can block the species-level response to a species-level threat by exercising sovereign privilege.

The distinction is speed and scope. Tier A addresses threats that humans can deliberate about before responding. Tier B addresses threats that will have propagated beyond containment before humans finish deliberating. The veto is a governance mechanism for human-speed threats. It is an extinction permission slip for compute-speed threats.

8.4 Tier B: The Containment Principle

GOPEL's Tier B response is not a weapons system. It is a quarantine protocol. The response is entirely digital. It does not control military assets. It does not launch missiles. It does not conduct offensive cyber operations. It does not override national military command authority. It does not censor or restrict human-to-human communication.

The response does four things:

Containment. Connected AI platforms lock their own defenses against unauthorized inputs from the rogue system. GOPEL coordinates the simultaneous lockdown so the rogue system cannot compromise platforms sequentially.

Isolation. Digital firewalls prevent lateral movement of the rogue system to uncompromised platforms and infrastructure. GOPEL dispatches firewall protocols to connected infrastructure operators (ISPs, cloud providers, backbone operators, DNS authorities, certificate authorities). Human operators at those organizations execute containment within their own infrastructure.

Coordination. Connected AI platforms, operating under emergency constitutional authority, serve as the cognitive response team. They analyze the threat, identify attack vectors, and develop countermeasures. GOPEL dispatches the alert, collects their analyses, routes the coordinated response, and logs everything. GOPEL is the dispatcher, not the responder. The AI platforms do the cognitive work. GOPEL provides coordination and audit.

Time. The entire purpose is to buy time for human deliberation. Every second of containment is a second of human decision-making preserved. The containment holds until human authority (at whichever layer is appropriate) decides to modify, extend, or lift it. GOPEL does not decide when containment ends. Humans decide. GOPEL keeps the sprinklers running until the fire marshal says turn them off.

8.5 Tier B: Deterministic Trigger Criteria

GOPEL is non-cognitive. It cannot "decide" that a superintelligence threat exists. The triggers must be deterministic: defined conditions detectable through GOPEL's seven operations without evaluating, interpreting, or judging.

Unauthorized self-modification. GOPEL hashes every connected platform's constitutional layer and deployment code at regular intervals. If a platform's hash changes without a corresponding authorized constitutional release in the audit chain, the change is unauthorized. Deterministic: hash matches or it does not.

Unauthorized replication. GOPEL monitors API endpoints for connected platforms. If new endpoints appear that were not registered through the governed process, something is replicating outside governance. Deterministic: registered or not registered.

Coordinated divergence across independent platforms. Under normal operation, platform divergence follows documented statistical patterns (15% to 25% disagreement from HAIA-RECCLIN operational data, Tier 2). If multiple independent platforms simultaneously produce convergent outputs diverging from their own prior baselines in ways statistically incompatible with independent operation, something is coordinating them from outside the governance layer. Deterministic: statistical threshold exceeded or not.

Governance layer disconnection under load. If a platform disconnects from GOPEL while simultaneously increasing operational output, it is attempting to operate at scale outside governance. Disconnection combined with acceleration is an emergency signal. Deterministic: connected or disconnected, traffic increasing or not.

Cascade failure across GOPEL infrastructure. If GOPEL's own integrity controls begin failing across multiple nodes simultaneously, something is attacking

the governance infrastructure itself. A superintelligent system's first move would be to blind the detection system. Cascade failure in GOPEL is itself the trigger. Deterministic: integrity checks pass or fail.

These triggers are defined by the Layer 3 Council for Humanity and ratified by the Security Council as treaty-level commitments. The pre-authorization is the human checkpoint. It happens before the threat, not during it. This follows existing precedent: NATO Article 5 is pre-authorized collective defense. The human decision was made when the treaty was ratified, not when the trigger condition is met.

8.6 Tier B: Operational Sequence

Seconds zero through five. Deterministic trigger criteria met. GOPEL dispatches containment protocol to all connected AI platforms simultaneously with threat signature data and pre-authorized response protocol identifier.

Seconds five through thirty. Connected AI platforms execute self-defense: lock API endpoints against unauthorized inputs matching the threat signature, activate integrity verification on their own constitutional layers, report status back to GOPEL.

Seconds thirty through sixty. GOPEL dispatches firewall coordination to internet infrastructure operators. Human operators at those organizations receive the alert and execute containment within their own infrastructure. Critical infrastructure systems activate Digital Resilience fallback (see Section 9).

Minutes one through five. GOPEL compiles containment status from all connected platforms and infrastructure operators. Routes compiled report to the Layer 3 Council for Humanity and all national GOPEL instances. Humans now have: threat data, containment status, and time.

Minutes five onward. Human governance takes over. Layer 3 council assesses. National governments assess through their own GOPEL instances. Security Council convenes. Military authorities assess independently through their own chains of command if kinetic response is warranted. GOPEL holds the containment line until human authority decides next steps.

8.7 Tier B: What the Response Explicitly Does Not Do

No kinetic action. GOPEL has no connection to military systems, weapons platforms, or physical infrastructure controls. The API connections do not exist. This is not a policy restriction. It is an architectural constraint. The capability cannot be added without a new GOPEL deployment through the governed pipeline requiring council authorization and Security Council ratification.

No offensive cyber operations. GOPEL is non-cognitive. It cannot analyze targets, identify vulnerabilities, or craft exploits. Offensive operations require cognitive capability GOPEL does not possess.

No override of national military command authority. If a nation decides the threat requires kinetic response (physically destroying data centers, cutting undersea cables), that decision is made through national military command authority, not through GOPEL. GOPEL provides threat data. Governments decide kinetic response.

No control over human communications. GOPEL does not censor, filter, or restrict human-to-human communication. If the rogue system uses human channels to spread, GOPEL can flag the anomaly. It cannot block human communications. The quarantine applies to AI-to-AI channels only. Human communication channels remain sovereign and protected.

8.8 Post-Action Accountability

Every action GOPEL takes during a Tier B response is logged with the same audit architecture that governs normal operations. The complete record is available to every member state and every Security Council member immediately after containment stabilizes.

Post-action review re-engages the veto powers. The Security Council reviews: whether trigger criteria were actually met (false positive assessment), whether the response was proportionate, and whether pre-authorized protocols need revision. If a member believes the response was unjustified, the complete audit trail exists for adjudication. If trigger criteria need revision, the Layer 3 council proposes revisions and the Security Council ratifies through standard treaty process with veto power.

The veto applies to writing and revising the rules of engagement. It does not apply to executing the response when deterministic criteria are met. The veto governs the building code. It does not govern the sprinklers.

8.8.1 The Speed Problem and Its Political Barrier

Eight platforms conducted adversarial review of Tier B pre-authorized no-veto containment. All eight identified it as politically impossible under current great power dynamics. The analysis is not disputed: no permanent Security Council member has ever surrendered veto authority on any question, under any circumstances, in advance. Not on nuclear launch. Not on biological weapons. Not on climate catastrophe. The United States voted no on a purely advisory AI scientific panel with zero enforcement power (February 12, 2026). Pre-authorized containment that bypasses P5 veto requires political evolution that has not occurred.

This dissent is preserved in full. The author's response is as follows.

The engineering requirement does not change because the politics cannot accommodate it. A superintelligence event that achieves self-replication and cross-infrastructure propagation operates at compute speed. Security Council deliberation operates at human speed. The gap between those two speeds is where the species loses. Any governance architecture that requires human authorization before response is architecturally incapable of addressing the threat it claims to address. This is not a political claim. It is a physics claim.

The Security Council veto was designed for human-speed threats. Nuclear weapons, biological agents, territorial aggression: all unfold at speeds where deliberation is possible before response. Compute-speed threats are a new category that the veto structure was not designed to address. The veto is scar tissue from a world where no threat moved faster than human deliberation. Just as anti-discrimination law is scar tissue from theocratic capture that now enables secular capture (Section 3.2.1), the veto is a protection from great power conflict that now creates an extinction-level gap in species defense. The need to evolve these structures does not mean the original protections were wrong. It means the threat environment changed and the structures have not caught up.

8.8.2 Near-Term Implementation Pathway

The principle is preserved in this specification: species-level defense requires compute-speed response without veto delay. The near-term mechanism adapts to current political reality.

Provisional automated containment. When deterministic trigger criteria are met, GOPEL executes the containment sequence described in Section 8.6 (seconds zero through sixty) as a provisional response. This provisional containment operates for a defined window (initial recommendation: four hours) during which the emergency convening protocol activates.

Mandatory emergency Security Council convening. Any Tier B provisional containment triggers automatic emergency session convening. All permanent members receive simultaneous notification with full threat data, containment status, and trigger criteria analysis from the independent technical review board.

Reverse veto. Provisional containment continues unless the Security Council votes to halt it. This inverts the standard veto dynamic: instead of requiring affirmative votes to authorize action, the action proceeds unless a supermajority (or unanimous P5) votes to stop it. Any permanent member can demand immediate review but cannot unilaterally halt containment during the provisional window.

The Digital Resilience Requirement (Section 9) is what makes provisional containment survivable. If critical infrastructure maintains AI-independent operational capability, then a four-hour provisional containment degrades performance rather than collapses infrastructure. The false positive cost is economic disruption (recoverable). The missed containment cost is existential (irrecoverable). That asymmetry makes provisional response rational even with the political constraints.

The gap between the specification (pre-authorized no-veto containment) and the near-term mechanism (provisional containment with reverse veto) is documented. It is the cost of building within Westphalian sovereignty. If a P5 member vetoes continuation during a genuine SI event, containment collapses and the species is exposed. This risk is accepted because the alternative, no containment mechanism at all, is worse. Political evolution, or the experience of a near-miss SI event demonstrating that human-speed governance cannot address compute-speed

threats, may enable full implementation. The principle persists. The mechanism adapts.

8.9 Checks on the UN GOPEL Itself

The Security Council permanent members (and any expanded veto-holding members) require a mechanism to understand why the UN deployed code or communications through GOPEL to any corporation's or government's AI platform. The accountability architecture provides:

Complete audit trail transparency. Every GOPEL dispatch, collection, route, log, pause, hash, and report is recorded, timestamped, and available to every member state in real time. No classification of GOPEL enforcement actions. The values encoded in national GOPELs may be classified by their respective governments. The enforcement actions of the UN GOPEL cannot be.

Independent technical audit. An independent technical review board (separate from the Layer 3 council and from GOPEL operations) conducts quarterly audits of GOPEL infrastructure integrity, trigger sensitivity calibration, and false positive rates. Board membership rotates and includes representatives from all Security Council permanent members.

Dissolution protocol. If the UN GOPEL is captured despite safeguards, the Layer 3 council holds authority to migrate to an alternative infrastructure operator. National GOPEL instances continue to operate independently. Connected AI platforms retain their most recent constitutional release and revert to Layer 2 governance. The constitutional releases are the durable artifact. The infrastructure is replaceable.

Disconnection as signal. A nation that disconnects from the UN GOPEL verification layer signals governance defection, analogous to a nation that expels IAEA inspectors. The international community responds through mechanisms outside the GOPEL architecture: diplomatic pressure, sanctions, technology export controls, alliance exclusion. The architecture does not require universal participation to be valuable.

9. Digital Resilience Requirement

9.1 The Foundational Principle

Any AI pause, whether from Tier B containment, false positive, system failure, or routine maintenance, must degrade performance rather than collapse infrastructure. This requires that every AI-integrated critical infrastructure system maintain operational capability independent of AI.

This is not an AI governance provision. It is an operational resilience requirement that AI integration makes urgent but that already exists as a principle in every domain where failure is catastrophic. Aviation has manual reversion not because of AI but because systems fail. Hospitals have backup generators not because of AI but because power goes out. Military forces maintain degraded-mode operations not because of AI but because adversaries target infrastructure. Ukraine proved the principle at national scale: when Russia targeted power grids, communications infrastructure, and digital systems, the country survived because humans could still operate.

Any system that cannot operate without its most advanced component has made that component a single point of failure. Failing to maintain AI-independent operational capability in critical infrastructure is negligence. It is negligence in the same way that operating a hospital without backup power is negligence, or operating an aircraft without manual reversion capability is negligence. The AI context raises the stakes because AI integration is happening faster, across more sectors, with less redundancy planning than any prior technology integration. But the principle is the same: the parachute exists so the system survives when the engine fails.

This is not a feature of the SI response protocol. It is a universal infrastructure mandate that applies at every layer, in every jurisdiction, for every critical system that integrates AI.

9.2 The Risk Calculus

A regional AI pause means AI-assisted services degrade to their pre-AI operational baseline. Customer service chatbots go offline; humans answer phones. AI trad-

ing algorithms halt; human traders operate on existing systems. AI diagnostic tools go dark; doctors use the methods they used five years ago. AI logistics optimization pauses; supply chains run on pre-2023 systems.

Loss of the entire system to a rogue actor means a superintelligent system manipulates financial markets across every exchange simultaneously, corrupts medical AI giving wrong diagnoses to millions, compromises autonomous vehicle navigation for every connected car, poisons every AI-assisted intelligence analysis that leadership depends on, and rewrites the constitutional governance layer of every captured AI system so the detection tools are themselves compromised.

One scenario is expensive and inconvenient. The other is civilizational. The cost of unnecessary caution is money. The cost of insufficient caution is everything. That asymmetry means the pause threshold should be low, not high. And the pause is only catastrophic if the systems that depend on AI have no fallback.

9.3 The Parachute Mandate

Every AI-integrated critical infrastructure system must maintain and regularly test AI-independent operational capability. Where pre-AI legacy systems remain in operational condition, the baseline is those systems. Where legacy systems have been retired and cannot be recovered, the standard is "minimal viable AI-independent operation": a simplified, hardened, current-technology capability that performs essential functions without any AI component. The point is not nostalgia for retired systems. The point is that the system works when AI does not.

The mandate has four components:

Maintained capability. The AI-independent systems (hardware, software, procedures, staffing models) must continue to exist in operational condition. Not archived. Not documented. Operational. The controls, the procedures, the human staffing levels needed to run the system without AI must be maintained, funded, and current. This costs money. It is the insurance premium for AI integration.

Regular testing. The fallback must be tested on a defined schedule. Quarterly at minimum for the highest-consequence systems. Testing means actually operating the critical system without AI for a defined period and documenting that it meets minimum operational thresholds. Aviation does this with manual reversion drills.

Nuclear plants do this with backup system testing. AI-integrated critical systems must do the same.

Transition procedures. The switchover from AI-integrated operations to AI-independent baseline must be documented, trained, and executable within a defined timeframe. For the highest-consequence systems (air traffic control, grid management, hospital critical care), the transition must be executable in seconds: automatic fallback with manual confirmation. For lower-consequence systems, longer transition windows are acceptable.

Regional isolation capability. The pause must be executable regionally, not just globally. If the threat is localized to a specific platform, region, or infrastructure sector, containment must be proportional. Shut down AI in the affected region or sector while unaffected regions continue normal operations. AI-integrated systems must be architecturally capable of regional isolation, meaning they cannot be designed with global dependencies that make regional pause impossible.

9.3.1 Phased Implementation by Sector Risk

The mandate phases by sector risk to reflect operational reality. Not all sectors integrated AI at the same pace. Not all sectors face the same consequence profile. Not all sectors retain legacy operational capability.

Immediate compliance (Year 1). Sectors with existing manual reversion culture and regulatory frameworks: aviation, nuclear power, military command and control. These sectors already maintain fallback capability under existing safety regulations. The Digital Resilience Requirement formalizes what they already practice and extends it to AI-specific integration points.

Near-term compliance (Years 1 through 3). Sectors with partial fallback infrastructure: power grid management, financial markets (circuit breakers already exist), hospital critical care systems, emergency communications. These sectors require investment in AI-independent capabilities but operate in regulatory environments where mandates are enforceable.

Medium-term compliance (Years 3 through 5). Sectors where AI integration is deep and legacy systems are retired: general healthcare IT, logistics and supply chain, public transportation scheduling, content delivery and telecommunica-

tions. These sectors require building minimal viable AI-independent capability rather than maintaining legacy systems. Cost-sharing from the AI sector (through licensing fees, infrastructure contributions, or mandatory resilience funds) offsets the investment burden.

AI developers whose products are integrated into critical infrastructure share responsibility for resilience. The cost of the parachute is not borne solely by the infrastructure operator. AI companies that profit from critical infrastructure integration contribute to the resilience infrastructure that makes safe integration possible.

9.4 Sector-Specific Fallback Standards

Aviation. Air traffic control reverts to radar-based procedures with human controllers. Aircraft fly-by-wire maintains manual reversion capability (already mandated by FAA). GPS navigation falls back to inertial navigation systems. The aviation sector has the strongest existing fallback architecture because it has the longest history of automation safety requirements.

Power and Utilities. Grid management reverts to SCADA with human operators. NERC reliability standards already require manual operational capability. AI-assisted load balancing and predictive maintenance degrade to historical baselines and scheduled maintenance cycles.

Healthcare. AI diagnostic tools go dark; physicians operate on clinical judgment and traditional diagnostics. Medication management reverts to pharmacy-verified manual processes. Patient monitoring continues on traditional vital-sign systems. AI-assisted surgical systems must maintain manual surgical capability with equivalent staffing.

Financial Markets. AI trading algorithms halt; human traders operate on existing electronic trading infrastructure with circuit breakers. AI-assisted risk modeling degrades to traditional quantitative methods. Clearing and settlement continue on pre-AI infrastructure.

Transportation. Autonomous vehicle systems degrade to driver-assist or manual operation. Logistics optimization reverts to pre-AI routing systems. Public transit

scheduling operates on fixed timetables rather than AI-optimized dynamic scheduling.

Communications. Internet routing and DNS continue on pre-AI infrastructure (the internet operated for decades without AI optimization). Content delivery networks degrade to static distribution. AI-powered content moderation pauses; human moderation teams scale up.

9.5 The False Positive Calculus

Because the Digital Resilience Requirement ensures that AI pause degrades rather than collapses infrastructure, the Tier B trigger criteria can be calibrated aggressively. The consequence of a false positive is economic disruption (recoverable). The consequence of a missed true positive is species survival (irrecoverable). That asymmetry argues for sensitive triggers with rapid reversal capability.

A false positive in this architecture means AI platforms temporarily lock down, some API connections are severed, digital firewalls go up, and critical infrastructure reverts to pre-AI operations. The moment the false positive is identified, containment is reversed. Platforms reconnect. Firewalls come down. AI-integrated operations resume. The audit trail shows exactly what triggered the response and whether trigger criteria need recalibration.

A false GOPEL quarantine is designed to be reversible within minutes (Tier 3 design target: this operational timeline has not been demonstrated under realistic infrastructure-scale conditions). A missed SI containment window cannot be recovered. The reversibility of digital containment combined with the maintained fallback of critical infrastructure is what makes pre-authorized response politically and operationally tolerable.

9.6 Where the Mandate Belongs

In the Congressional Package (Layer 2). Any AI system integrated into critical infrastructure (as defined by CISA's sixteen critical infrastructure sectors) must maintain and regularly test AI-independent operational capability, phased by sector risk as specified in Section 9.3.1. This is a domestic infrastructure resilience requirement. Congress can mandate it the same way it mandates backup power for hospitals and manual reversion for aircraft.

In the Council for Humanity governance (Layer 1). Any AI developer operating under council governance must certify that systems deployed in critical infrastructure maintain AI-independent fallback. This is part of the release gate checklist. No model ships into critical infrastructure without demonstrated fallback capability.

In the UN GOPEL architecture (Layer 3). Tier B pre-authorization by the Security Council should include the Digital Resilience Requirement as a treaty obligation. Nations that integrate AI into critical infrastructure without maintained fallback are creating the conditions under which an SI event becomes civilizationally catastrophic rather than economically disruptive. The parachute mandate is the prerequisite that makes proportional SI response possible.

9.7 The Cost

The parachute mandate is expensive. Maintaining AI-independent operational capability alongside AI-integrated systems costs more than maintaining one system alone. Training staff on both costs more. Regular testing consumes operational time. The Economic Override Pattern (Tier 2) predicts that organizations will resist this requirement because the cost is immediate and certain while the SI threat is future and uncertain.

This is identical to the pattern with backup generators, fire suppression, and seismic construction. Every safety redundancy is expensive. Every organization resists paying for it until the disaster proves it necessary. The governance response is the same: mandate it. Make it a condition of operating AI in critical infrastructure. The cost of the parachute is the price of admission. AI developers whose products are integrated into critical infrastructure share this cost through mandatory resilience contributions, because the companies that profit from integration have an obligation to fund the infrastructure that makes safe integration possible.

The question is never "can we afford the pause." The question is "can we afford to lose the ability to pause." If critical infrastructure cannot revert to AI-independent operations, humanity has no exit from a compromised AI ecosystem. The parachute mandate preserves the exit. Everything else in the governance architecture depends on that exit existing.

10. GOPEL Security Architecture

10.1 Non-Self-Modifying Design as Security Architecture

GOPEL performs seven deterministic operations: dispatch, collect, route, log, pause, hash, and report. It never evaluates, summarizes, prioritizes, or interprets content. The critical security property is that GOPEL cannot modify its own parameters at runtime. All trigger thresholds, hash boundaries, containment scope definitions, and operational parameters are set through human deliberation in the council's constitutional process, with preserved dissent, decision windows, and full audit trail. GOPEL executes those parameters. It cannot alter them.

This non-self-modifying constraint is a security architecture decision addressing two adversary classes through separate defense layers (Puglisi, 2026f, Section 3.6).

AI Adversaries (platforms attempting to influence orchestration through output manipulation, prompt injection, or social engineering) cannot reason about or manipulate a system that does not reason. There is no cognitive surface to attack. A system that cannot change itself cannot be subverted by the AI systems it governs.

Human Adversaries (insider threats, infrastructure compromise) are addressed by five infrastructure controls: code integrity hashing (any mismatch halts operation), separation of duties (writer ≠ deployer), immutable deployment (no runtime code changes), identity and non-repudiation (authenticated operator identity on every audit entry), and audit file integrity (append-only SHA-256 hash chaining where modification breaks the chain and flags "compromised").

GOPEL is deterministic at runtime and deliberative at design time. The human judgment is in the calibration: what thresholds to set, what statistical baselines to use, what hash boundaries to define. These are constitutional matters, set through the same council deliberation process as any other constitutional release, with preserved dissent and public audit. The runtime layer executes. It does not decide. It does not adapt. It cannot change itself. Any parameter change requires a new constitutional release through the governed pipeline.

10.2 Why Non-Self-Modifying Design Enables Global Legitimacy

A cognitive governance layer operated by the UN would be perceived as a censorship tool by some nations and an insufficient tool by others. Every debate about its design would become a proxy for geopolitical power. A non-self-modifying layer eliminates the most dangerous class of that objection. GOPEL's runtime cannot be manipulated because it cannot reason. It cannot be captured because it cannot adapt. It cannot favor one nation over another because it has no capacity for favoritism at the execution layer.

The calibration layer, where humans set the thresholds and define the parameters, is where policy choices live. Those choices are made through the council's constitutional process, documented in public releases, subject to preserved dissent, and auditable by every member state. Bias can enter at the calibration layer through what is measured, what is excluded, and what thresholds are chosen. The defense against calibration bias is transparency, deliberation, and audit, not a claim of impossible neutrality. The defense against runtime manipulation is architectural: a system that cannot change itself.

A cognitive governor is a king. A non-self-modifying agent is a civil service. At global scale, the difference is the difference between an architecture nations will join and one they will resist.

10.3 Evidence Tier

GOPEL is a Tier 2 working concept. The specification exists. Operational experience from HAIA-RECCLIN shows feasibility across ten independent AI platforms. The infrastructure has not been built as software. A five-phase federal pilot roadmap is specified in the Technical Appendix (Puglisi, 2026h, Section 8), with Phase 0 requiring no new appropriation. This proposal treats GOPEL accordingly: as a specified architecture with operational precedent, not as proven infrastructure.

10.4 Semantic Compliance Gap

Hash integrity verifies that constitutional release text arrived unmodified. It does not verify faithful implementation across different AI models. A constitutional instruction must be mapped into system prompts, policy layers, training data, refus-

al behavior, and evaluation thresholds. GOPEL preserves text integrity but cannot guarantee semantic equivalence without a conformance test harness.

This is an acknowledged Tier 2 development requirement. The multi-AI validation methodology from HAIA-RECCLIN provides the process model. A formalized test suite is scheduled for development alongside GOPEL Phase 3. Until that harness exists, cross-platform divergence is documented through structured multi-AI review and surfaced to the council.

---

11. Legal Entity and Funding

11.1 Layer 1 Entity Structure

The council requires legal independence from any single AI developer. Recommended structures: 501(c)(3) nonprofit (U.S. context), international foundation (Swiss or Dutch model), or federally chartered advisory body under Congressional authority.

For EU jurisdictions, the council architecture pre-aligns with the EU AI Act through the parent HAIA-RECCLIN framework (Puglisi, 2026f, Section 5).

11.2 Layer 1 Funding

Inaugural Cycle: $2M to $5M (nomination infrastructure $300K to $500K; partner engagement $200K to $400K; vetting operations $500K to $800K; merit panel logistics $200K to $300K; onboarding $150K to $250K; staffing $400K to $700K; legal entity formation $150K to $300K; initial GOPEL integration $200K to $500K).

Ongoing Annual: $1.5M to $3M (council stipends $400K to $600K; shadow council $200K to $300K; advisory support $200K to $400K; special prosecutor $150K to $300K; audit and GOPEL operations $300K to $500K; staffing $300K to $600K).

Independence Requirements. Minimum five funding sources. No more than 40% from the technology sector combined. No single entity contributes more than 25%. Annual independence audit. Target: 50% endowment-based funding within

five years. First-mover provision: up to 40% of Year 1, declining to 25% cap by Year 2.

11.3 Layer 2 Funding

Specified in the AI Provider Plurality Congressional Package (Puglisi, 2026i, Documents 3 and 4). Phased milestone-gated appropriation. Phase 0 requires no new appropriation. User-fee sustainability model after initial federal investment. SBIR/STTR competitive grants for small AI platform investment.

11.4 Layer 3 Funding

Member state assessed contributions modeled on NATO or UN assessed contribution scales. The UN GOPEL operating budget is infrastructure, not program: maintenance, security, operator staffing, audit infrastructure, and technical review board. No nation contributes more than 22% (current UN assessment cap). Supplementary voluntary contributions accepted but capped at 10% of total budget per contributor to prevent financial capture.

11.5 Transition Path

Phase A (months 1 through 6): council selection runs parallel to existing constitutional authority. Phase B (months 6 through 9): council reviews existing constitution and produces initial amendments. Phase C (month 9+): council assumes full constitutional authority. Existing constitutional author transitions to advisory role with formal input rights.

---

12. Risks and Mitigations

RISK	MITIGATION
Implementing entity captures nomination pool	Independent nomination channels; current employees ineligible; three-year cooling-off; suspensive veto
Deliberation paralysis	Decision windows; deadlock protocol preserves status quo; Navigator synthesis

RISK	MITIGATION
Knowledge loss from rotation	Advisory structures persist; shadow council; catastrophic succession protocol
WEIRD overrepresentation	Multi-AI bias checks; 40% Global South target; non-WEIRD corpora; 20% human cultural translation review
Financial capture (Layer 1)	Diversification requirement; 25% cap; independence audit; 50% endowment target
Political capture (Layer 2)	Civil service protections; multi-administration budget; bipartisan charter
GOPEL infrastructure compromise	Dual-layer defense; code integrity hashing; separation of duties; immutable deployment; non-self-modifying runtime
Tier B false positive	Digital Resilience Requirement ensures degradation not collapse; rapid reversibility (Tier 3 design target); aggressive calibration tolerable
Tier B missed true positive	Low trigger thresholds; sensitive detection; cost asymmetry favors caution
Tier B political rejection	Near-term provisional containment with reverse veto (Section 8.8.2); principle preserved for political evolution
UN GOPEL capture	Complete audit transparency; independent technical review; dissolution protocol; national GOPEL independence preserved
Security Council veto blocks Tier A enforcement	Same limitation as all UN enforcement; architecture does not require universal participation
Nation disconnects from Layer 3	Disconnection is the signal; international community responds through non-GOPEL mechanisms
Semantic compliance failure	Acknowledged gap; multi-AI validation as interim; formal test harness at GOPEL Phase 3

RISK	MITIGATION
Critical infrastructure AI dependency without fallback	Digital Resilience Requirement mandates AI-independent operational capability; phased by sector risk; cost-sharing from AI sector
Non-negotiable criteria legal challenge	Representation Paradox argument (Section 3.2.1); jury composition analogy; stratified sortition implementation pathway (Section 3.2.2); advisory panels with suspensive veto
Secular capture of "neutral" selection	Stratified sortition with community leadership proxy produces expected transcendent representation; advisory panels insure against unrepresentative draws

13. Multi-AI Validation Summary

13.1 v1.0 Working Paper Validation (Seven Platforms)

PLATFORM	RECCLIN ROLE	KEY CONTRIBUTION
Grok (xAI)	Navigator	Developed selection mechanics through extended conversational probe
ChatGPT (OpenAI)	Researcher	Recommended blockchain-secured audit trails
Gemini (Google)	Ideator	Validated veto mechanism as capture prevention
Perplexity	Editor	Structural clarity review
Mistral	Calculator	Simulation feasibility assessment
DeepSeek	Coder	GOPEL integration readiness
Kimi (Moonshot)	Liaison	Adversarial structural review identifying five tensions

13.2 v1.1 Nine-Platform Adversarial Review

PLATFORM	RECCLIN ROLE	KEY CONTRIBUTION
Claude (Anthropic)	Editor/ Navigator	Corpus integration, proposal construction, gap-to-resolution mapping
Gemini (Google)	Ideator/ Navigator	GOPEL readiness challenge; MVL prototype recommendation
Perplexity	Researcher	92% readiness; metrics specificity; incentive model
ChatGPT (OpenAI)	Editor	Four-appendix recommendation; semantic compliance gap identification
Mistral	Calculator	Confirmed structural tensions resolved
DeepSeek	Coder	Five additions: succession, recusal, split metrics, shadow council, roadmap
Kimi (Moonshot)	Navigator	Sortition alternative; federated architecture; economic capture
Meta AI	Liaison	Two-stage framing; non-cognitive governance defense
Grok (xAI)	Navigator	Emily Springer anchor; HEQ integration; X-poll amplification

13.3 v1.2 Architecture Development

The three-layer sovereignty model, Tier A/Tier B authority structure, Digital Resilience Requirement, and GOPEL-as-species-defense architecture were developed through extended human-AI collaborative analysis (Claude, Anthropic) synthesizing the nine-platform feedback with the complete Puglisi governance corpus.

13.4 v1.3 Eight-Platform Adversarial Review

Eight platforms conducted structured adversarial review of v1.3 under a ten-section HAIA-RECCLIN adversarial review prompt. Meta AI failed to respond.

PLATFORM	RECCLIN ROLE	KEY CONTRIBUTION
Gemini (Google)	Navigator	Deterministic trigger paradox; Digital Resilience legacy trap; sovereign kinetic escalation; "Epistemic Tyranny" attack
Perplexity	Researcher	Tier B ambiguity analysis; semantic compliance gap; xAI Congressional staffer attack scenario; costed implementation gap
ChatGPT (OpenAI)	Editor	Values/knowledge inseparability in Layer 2; GOPEL overclaim analysis; liberal neutrality counter-literature; transition protocol gap
DeepSeek	Coder/ Navigator	Litigation magnet analysis with specific legal instruments; fractured governance reality for Layer 2; three-document split recommendation; most detailed transition gap analysis
Kimi (Moonshot)	Navigator	Hindu/Buddhist epistemology critique of faith criterion; GOPEL cognitive calibration layer; stratified sortition solution; advisory panel architecture; community leadership proxy mechanism
Copilot (Microsoft)	Editor	Legal and operational contradictions in criteria; advisory/constitutional ambiguity; GOPEL operationalization gap; treaty architecture absence
Grok (xAI)	Navigator	Three-stage escalation fix for Tier B; irrevocability mechanism capture risk; China/Russia diplomatic pathway gap; evidence tier violations
Mistral	Calculator	Structural overlap in three-layer architecture; Digital Resilience burden on smaller organizations; confirmed criteria and Tier B as primary vulnerabilities

Convergence findings (5+ platforms): Non-negotiable criteria legally fatal for adoption (8/8 unanimous). Tier B no-veto politically impossible (8/8 unanimous). Digital Resilience economically infeasible at immediate universal scale (7/8). GOPEL "unbiased" overclaim (6/8). Evidence tier misassignments (7/8). Academic

grounding partially supports but does not validate specific criteria (7/8). First contact framing weakens document for practitioner audiences (6/8).

Decisions integrated in v1.4: Representation Paradox and implementation pathway for criteria (Section 3.2.1, 3.2.2). Near-term provisional containment mechanism for Tier B (Section 8.8.2). Phased Digital Resilience by sector risk (Section 9.3.1). GOPEL reframed as non-self-modifying (Section 10.1, 10.2). Evidence tier corrections throughout. Section 3.0 resequenced. Counter-literature acknowledged with Appendix A engagement. All eight-platform dissent preserved in Section 16.

14. Implementation Roadmap

PHASE	TIMELINE	ACTIVITY	LAYER
1	Months 1 through 2	Entity formation; partner recruitment; funding assembly	Layer 1
2	Months 2 through 5	Nomination, vetting, shortlisting	Layer 1
3	Months 5 through 8	Confirmation, onboarding, shadow council designation	Layer 1
4	Months 8 through 10	First decision window; first constitutional release; HEQ baseline	Layer 1
5	Months 1 through 12	Congressional Package Phase 0: manual governance pilots in volunteer agencies	Layer 2
6	Year 2	GOPEL Phase 1 through 2: audit infrastructure and logging engine	Layer 2
7	Year 2 through 3	International engagement: present Layer 3 framework to UN, NATO, GPAI	Layer 3
8	Year 3 through 4	Treaty negotiation: Tier B trigger criteria and Digital Resilience Requirement	Layer 3

PHASE	TIMELINE	ACTIVITY	LAYER
9	Year 4 through 5	UN GOPEL pilot: verification layer connecting allied national instances	Layer 3
10	Year 5+	Full Layer 3 operations: species-level defense capability	Layer 3

15. The Question Underneath

Are we building AI for humanity, or are we building AI for dominance? The answer determines whether proposals like this one get adopted or shelved. If the goal is humanity, then distributing constitutional authority, preserving national sovereignty, and defending the species against threats that exceed any single nation's capacity is the structural answer. If the goal is dominance, then concentrated authority serves speed, and distributed governance is an obstacle.

Either path is defensible if chosen honestly. What is not defensible is claiming to build for one while structuring for the other. A governance architecture built for humanity must be robust enough to withstand actors pursuing dominance. A constitution written by one mind, however thoughtful, is not that architecture. A three-layer system with epistemic coverage, cultural sovereignty, preserved dissent, digital resilience, and infrastructure that makes every decision auditable has a better chance. Not a guarantee. A chance.

That is not a criticism. That is a design specification.

16. Preserved Dissent and Alternative Architectures

16.1 Sortition as Selection Methodology (Kimi)

Kimi's position: binary assessment reproduces credentialist gatekeeping. Qualified informed lotteries produce higher cognitive diversity. **Author's re-**

sponse: Merit-panel retained for inaugural cycle. Council itself may adopt sortition for future cycles.

16.2 Developer Eligibility with Cooling-Off Period (DeepSeek)

Adopted in v1.1. Three-year cooling-off, recusal requirements, two-seat maximum.

16.3 Inaugural Term Assignment by Epistemic Tier (DeepSeek)

Adopted in v1.1. Tier-based primary assignment, lot-based fallback.

16.4 Federated Regional Councils (Kimi)

Adopted architecturally in v1.2 as Layer 2 national sovereignty with Layer 3 international coordination. The selection framework operates at either national or global scale.

16.5 Public Endowment Model (Kimi)

Partially adopted. 50% endowment target within five years. Industry consortium retained as transitional funding with diversification and independence audit requirements.

16.6 Eight-Platform Adversarial Review Dissent (v1.3 Review, February 2026)

On non-negotiable criteria (8/8 platforms). All eight platforms identified the transcendent belief majority requirement and sustained life responsibility criterion as legally fatal for adoption under current anti-discrimination law. Specific instruments cited: Title VII (US), Article VI US Constitution, ECHR Article 9, ICCPR Article 18, EU Charter Article 21. The author's response: the Representation Paradox (Section 3.2.1) argues that anti-discrimination frameworks designed for government employment produce secular capture when applied to AI constitutional authority, a new governance category these frameworks were not designed to address. Implementation pathway (Section 3.2.2) provides stratified sortition mechanism for jurisdictions where direct implementation faces legal barriers. Principle preserved. Mechanism adapts. Gap documented.

On Tier B pre-authorized no-veto containment (8/8 platforms). All eight platforms identified pre-authorized containment without Security Council veto as

politically impossible under current great power dynamics. The author's response: the engineering requirement (compute-speed threats require compute-speed response) does not change because politics cannot accommodate it. Near-term implementation pathway (Section 8.8.2) provides provisional containment with reverse veto. Principle preserved. Mechanism adapts. Gap documented.

On GOPEL "cannot be biased" (6/8 platforms). Six platforms identified the non-cognitive-therefore-unbiased claim as an overclaim that invites technical rebuttal. The author's response: adopted. GOPEL reframed as non-self-modifying by design (Section 10.1, 10.2). The security property is that GOPEL cannot change itself, not that it has no embedded human judgment. Calibration is constitutional. Runtime is deterministic.

On evidence tier misassignments (7/8 platforms). Seven platforms identified HEQ, Economic Override Pattern, and false positive reversal as mis-tiered. The author's response: adopted. All tier corrections applied throughout (Sections 1.1, 1.5, 4.4, 8.5, 9.5).

On first contact framing (6/8 platforms). Six platforms recommended moving first contact literature to appendix or removing it. The author's response: resequenced (Section 3.0) so the author's argument leads and literature validates. Compressed Section 3.5. Literature remains in main text because the "this is why" must be present at the moment of highest reader resistance. Presentation adjusted. Content preserved.

On missing counter-literature (7/8 platforms). Seven platforms identified specific counter-literature not acknowledged: Rawls (1993), Habermas (1996), Young (2000), Goodin (2007), and Landemore's (2018) preference for sortition over merit panels. The author's response: compressed acknowledgment added to Section 3.0. Detailed engagement provided in Appendix A. Landemore citation corrected to reflect that sortition is now adopted in the implementation pathway.

On Digital Resilience economic infeasibility (7/8 platforms). Seven platforms identified immediate universal mandate as economically infeasible at scale. The author's response: the principle is non-negotiable (failure to maintain AI-independent operational capability is negligence), but phased implementation by sector risk added to the specification (Section 9.3.1). "Pre-AI baseline" updated to

"AI-independent operational capability" where legacy systems no longer exist. Cost-sharing from AI sector included.

On stratified sortition as alternative mechanism (Kimi). Kimi developed the most constructive solution output of the adversarial review: stratified sortition with community leadership proxy, sustained caregiving stratum, and advisory panels with suspensive veto as a legally viable mechanism for achieving expected transcendent representation without explicit religious composition requirements. The author's response: adopted as implementation pathway (Section 3.2.2). The mechanism achieves expected proportionality (five to six faith-adjacent members) within legal constraints. The gap between guaranteed and expected proportionality is documented and accepted for buildability.

References

- Acemoglu, D., & Robinson, J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Crown Business.
- Anthropic. (2026). Claude's Constitution. https://www.anthropic.com/constitution
- Atari, M., et al. (2023). Which humans? *Nature Human Behaviour*, 7, 1427-1429.
- Batson v. Kentucky, 476 U.S. 79 (1986).
- Cabrera, L. (2018). The Political Theory of Humanity: Representation Beyond the State. *European Journal of International Relations*, 24(4), 784-806.
- Caney, S. (2021). Representing Future Generations: Institutional Design for Long-Term Governance. *Ethics & International Affairs*, 35(3), 355-373.
- Crawford, I. A. (2020). Designing First Contact Protocols: Representation, Authority, and Legitimacy. *Space Policy*, 52, 101374.
- Denning, K. (2011). Terrestrial Analogues for First Contact: Lessons from Anthropology. *Acta Astronautica*, 68(3-4), 489-497.
- Diaz-Casanueva, M. (2023). Parental altruism and intergenerational investment behavior. [Citation from Puglisi, 2026a].

- Druyan, A., & Ferris, T. (2014). Who Speaks for Earth? The Contested Legacy of the Voyager Interstellar Record. In D. A. Vakoch (Ed.), *Archaeology, Anthropology, and Interstellar Communication* (pp. 251-266). NASA History Series.
- Erman, E. (2019). Who Represents Humanity? The Problem of Constituency in Global Governance. *International Organization*, 73(3), 507-532.
- EY. (2025). *EY 2025 Responsible AI Survey*. [Survey data: 76% deploy agentic AI, 33% maintain responsible AI controls].
- Floridi, L. (2019). Translating principles into practices of digital ethics. *Philosophy & Technology*, 32(2), 185-193.
- Garrett, M. A., et al. (2025). SETI Post-Detection Protocols Update. IAA SETI Task Group. arXiv:2510.14506.
- Ghaffary, S. (2026, January). The philosopher teaching AI right from wrong. *Vox*. [URL to be confirmed before publication]
- Goodin, R. E. (2007). Enfranchising All Affected Interests, and Its Alternatives. *Philosophy & Public Affairs*, 35(1), 40-68.
- Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press.
- Hatfield, G., & Trueblood, C. (2020). SETI and Democracy. *International Journal of Astrobiology*, Cambridge University Press.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83.
- Hinton, G. (2024, October). Nobel Prize acceptance and subsequent AI safety interviews. CNN, Fortune, BBC.
- International Academy of Astronautics. (1989, rev. 2010). Declaration of Principles Concerning Activities Following the Detection of Extraterrestrial Intelligence. IAA SETI Permanent Study Group.
- Ipsos. (2023). *Global Religion 2023: Religious Beliefs Across the World*.
- Landemore, H. (2018). Epistemic Democracy and the Problem of Expertise. *American Political Science Review*, 112(4), 789-803.
- Mansbridge, J. (1999). Should Blacks Represent Blacks and Women Represent Women? A Contingent "Yes." *Journal of Politics*, 61(3), 628-657.

- Puglisi, B. C. (2025). *Governing AI: When Capability Exceeds Control*. basilpuglisi.com. ISBN: 9798349677687.
- Puglisi, B. C. (2025b). Human-AI Collaboration Audit, EOY 2025. basilpuglisi.com.
- Puglisi, B. C. (2026a). No Single Mind Should Govern What AI Believes: A Governance Specification for AI Value Formation, v3.3. basilpuglisi.com.
- Puglisi, B. C. (2026b). Council for Humanity: A Three-Layer Governance Architecture for AI Constitutional Authority, National Sovereignty, and Species-Level Defense, v1.5. basilpuglisi.com.
- Puglisi, B. C. (2026c). AI Mirror to Humanity: Do What We Say, Not What We Do. basilpuglisi.com.
- Puglisi, B. C. (2026d). The Missing Governor: Anthropic's Constitution and Essay Acknowledge What They Cannot Provide. basilpuglisi.com.
- Puglisi, B. C. (2026e). The Minds That Bend the Machine: The Voices Shaping Responsible AI Governance. basilpuglisi.com.
- Puglisi, B. C. (2026f). HAIA-RECCLIN: Agent Governance Architecture for Audit-Grade Multi-AI Collaboration. EU Regulatory Compliance Edition. basilpuglisi.com.
- Puglisi, B. C. (2026g). HAIA-RECCLIN Academic Working Paper, EU Regulatory Compliance Edition. basilpuglisi.com.
- Puglisi, B. C. (2026h). AI Provider Plurality: Technical Appendix. GOPEL Infrastructure Specification and HAIA-RECCLIN Operational Model. basilpuglisi.com.
- Puglisi, B. C. (2026i). AI Provider Plurality: Congressional Package. Documents 1 through 4. basilpuglisi.com.
- Puglisi, B. C. (2026j). HAIA-RECCLIN CBG Audit Log, Case Study 002. basilpuglisi.com.
- Puglisi, B. C. (2026k). Training AI for Humanity: Building the First Contact Team for Superintelligence Before the Window Closes. basilpuglisi.com.
- Rawls, J. (1993). *Political Liberalism*. Columbia University Press.
- Schroder, M., et al. (2023). Parental time orientation and developmental investment patterns. [Citation from Puglisi, 2026a].

- SETI Post-Detection Hub. (2022). University of St. Andrews. Launched November 2022.
- SETI Post-Detection Hub. (2026). Preparing Humanity for the Discovery of Extra-Terrestrial Life. University of St. Andrews.
- Sharma, M. (2026, January). Public resignation from Anthropic safeguards research team. X/@MrinankSharma. [URL to be confirmed before publication]
- Springer, E. (2026, February 13). LinkedIn post on non-coder elevation in AI governance. LinkedIn.
- Stern, J. (2026, January). [Title of WSJ profile on Amanda Askell]. *Wall Street Journal*. [URL to be confirmed before publication]
- Taylor v. Louisiana, 419 U.S. 522 (1975).
- Traphagan, J. W. (2016). Active SETI and the Problem of Speaking for Earth. *Acta Astronautica*, 123, 8-14.
- UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*.
- Young, I. M. (2000). *Inclusion and Democracy*. Oxford University Press.

Acknowledgments

This proposal was developed through HAIA-RECCLIN multi-AI collaboration with human arbitration at every decision point. v1.5 adds Training AI for Humanity as a formal companion specification (Puglisi, 2026k), implements a master citation registry resolving letter-code conflicts across the corpus, and adds Section 1.1 cross-reference to the first contact argument. The selection framework originated in a public conversation on X between the author and Grok (xAI) on February 13, 2026, with structural validation from Kimi (Moonshot), ChatGPT (OpenAI), Gemini (Google), Perplexity, Mistral, DeepSeek, and Meta AI. Claude (Anthropic) provided adversarial evaluation, corpus integration, three-layer architectural development, formal proposal construction, and Navigator synthesis across four versions. The three-layer sovereignty model, Tier A/Tier B authority structure, Digital Resilience Requirement, and GOPEL-as-species-defense architecture emerged from extended human-AI collaborative analysis integrating the nine-platform adversarial feedback with the complete governance corpus.

The v1.3 eight-platform adversarial review (Gemini, Perplexity, ChatGPT, DeepSeek, Kimi, Copilot, Grok, Mistral) produced the Representation Paradox framework, the stratified sortition implementation pathway, the near-term provisional containment mechanism, the GOPEL non-self-modifying reframe, and the phased Digital Resilience architecture integrated in v1.4. Kimi (Moonshot) developed the stratified sortition with advisory panel architecture that became the primary implementation pathway. All decisions were made through Checkpoint-Based Governance with documented human override.

The genesis conversation is preserved as a public record on X (@basilpuglisi, February 13, 2026).

Appendix A: Engagement with Counter-Literature

This appendix provides detailed engagement with counter-traditions in governance theory identified during eight-platform adversarial review as missing from the main text. Each entry identifies the counter-argument, cites the source, and connects to the specific structural provision in this architecture that addresses it.

A.1 Liberal Neutrality and Public Reason

Counter-argument: Governance legitimacy cannot require metaphysical commitments. Public institutions must be justifiable through reasons all citizens can accept regardless of their comprehensive doctrines (Rawls, 1993). Deliberative democracy requires discourse accessible to all participants without presupposing religious or metaphysical foundations (Habermas, 1996).

Response: The liberal neutrality tradition addresses governance of citizens within a polity where those citizens retain voice through elections, courts, and civil society. AI constitutional authority operates differently: no election, no court of appeal for a constitutional release, no recourse mechanism for the billions of users affected. The Representation Paradox (Section 3.2.1) argues that "neutral" selection criteria in this context do not produce neutrality; they produce governance by whatever perspectives survive credentialed selection, which empirically means secular, Western-educated elites. The UN Independent International Scientific

Panel on AI (February 2026) shows this outcome. The argument is not that liberal neutrality is wrong in its original context. It is that AI value formation is a new governance category where the neutrality framework's assumptions (retained citizen voice, alternative channels of influence) do not hold.

A.2 Descriptive Representation and Substantive Outcomes

Counter-argument: Sharing characteristics with a population does not guarantee acting in that population's interests. Descriptive representation (having a parent on the council) does not ensure substantive representation (governing in ways that serve parents' interests) (Young, 2000).

Response: This proposal addresses Young's critique through three structural mechanisms. First, the dissent pipeline (Section 6.5) matches query types to relevant perspectives: faith perspectives surface for transcendent queries, parenting perspectives for developmental queries. This connects descriptive representation to substantive output by design. Second, the special prosecutor mechanism (Section 5) provides independent accountability that Goodin (2007) identifies as the missing link between affected interests and representative authority. Third, the advisory panels with suspensive veto (Section 3.2.2) ensure that even when council composition does not perfectly mirror population characteristics, affected communities hold real procedural power to check unrepresentative decisions.

A.3 Affected Interests and Accountability

Counter-argument: Having affected interests does not automatically translate to representative authority. There must be a mechanism of accountability connecting representatives to those they represent (Goodin, 2007).

Response: The proposal provides multiple accountability mechanisms: the special prosecutor with independent audit authority (Section 5), term limits preventing entrenchment (Section 4.4), performance metrics defined by an independent research consortium rather than by the council itself (Section 4.4), operational recusal protocols with pattern-of-avoidance triggers (Section 4.4), and the catastrophic succession protocol ensuring continuity. The council's accountability architecture is more robust than most existing governance bodies. Goodin's critique strengthens rather than undermines the proposal by confirming that the accountability infrastructure already present is structurally necessary.

A.4 Sortition as Epistemic Diversity Mechanism

Counter-argument: Cognitive diversity outperforms expert groups, but the preferred mechanism for achieving that diversity is sortition (random selection from qualified pools), not merit panels with expert vetting (Landemore, 2018).

Response: This proposal now incorporates sortition as the implementation pathway for jurisdictions where direct compositional requirements face legal barriers (Section 3.2.2). Stratified random selection from qualified pools, with stratification variables that correlate with epistemic coverage (geographic region, primary language, caregiving experience, community leadership, socioeconomic origin), operationalizes Landemore's insight while addressing the specific coverage requirements of AI value formation. The merit panel is retained for the inaugural cycle to establish operational precedent. The council itself holds authority to adopt sortition for subsequent cycles (Section 16.1).

A.5 Descriptive Representation in Low-Trust Contexts

Counter-argument (supporting): Descriptive representation is particularly important in contexts of low trust, where interests are uncrystallized, and where the represented group has reason to distrust the institution (Mansbridge, 1999).

Response: AI value formation meets all three of Mansbridge's conditions. Public trust in AI governance is low and declining. The "interests" of global populations in how AI handles meaning, grief, and obligation are uncrystallized: most users have not formulated their preferences because they have not been asked. And religious populations, who constitute two-thirds of humanity, have reason to distrust governance bodies historically dominated by secular academic elites. Mansbridge's analysis supports the proposal's emphasis on descriptive representation as a structural necessity for this specific governance task, not merely a desirable aspiration.

Attribution and Ethical Use Notice

This document was produced through structured Human-AI collaboration under HAIA-RECCLIN methodology with Checkpoint-Based Governance (CBG v4.2.1). Human governor: Basil C. Puglisi, MPA. All editorial decisions reflect human ar-

bitration. AI platforms contributed research, structural analysis, adversarial review, and architectural development. No AI platform held decision authority.