Training AI for Humanity:

Building the First Contact Team for Superintelligence Before the Window Closes

Basil C. Puglisi, MPA

Human-AI Collaboration Strategist | AI Governance Consultant

basilpuglisi.com

February 2026

Status: Academic Working Paper (Pre-Peer Review)

Evidence Discipline:

Three-tier structure applied throughout. Tier 1: established by peer-reviewed research. Tier 2: working concepts with operational validation, never described as proven. Tier 3: what should be built, framed as design targets.

Abstract

The people training artificial intelligence today are building the cognitive foundation for whatever comes next. If superintelligence emerges from systems whose value structures correlate with 12% of humanity and diverge from the rest (Atari et al., 2023; Henrich et al., 2010), the resulting intelligence faces a structurally increased risk of representational failure when interacting with the majority of the species that created it. That risk is not a future problem. It is being constructed now, in the training data, the alignment protocols, and the constitutional documents that shape how AI reasons about meaning, obligation, and identity.

The first contact literature in astrobiology has spent decades asking who speaks for humanity when addressing a non-human intelligence (Crawford, 2020; Traphagan, 2016; Denning, 2011). That question was designed for extraterrestrial contact. It applies with greater urgency to artificial intelligence, because this non-human intelligence is being built by us, trained on some of our data, shaped by some of our opinions, and carrying all of our bias, and the window for shaping its foundation narrows as capability accelerates.

This paper argues that AI value formation is humanity's first contact preparation, happening now, and the teams assembling that foundation lack the epistemic coverage to represent the species. It synthesizes established research on monocultural bias in AI systems, documented concerns from a prior analysis of nine interconnected AI risk domains (Puglisi, 2025), an intellectual journey studying twenty-four thought leaders across five waves of AI development to understand what governance requires (Puglisi, 2026e), operational findings from a published multi-platform governance methodology (HAIA-RECCLIN) suggesting that structured comparison across platforms trained in different civilizational contexts surfaces biases invisible to any single system (Puglisi, 2025b), and a proposed governance architecture (the Council for Humanity) designed to ensure constitutional authority over AI value formation represents the species rather than a fraction of it (Puglisi, 2026b, unpublished).

Three tiers of action emerge. First, AI developers can build diverse constitutional teams rather than leaving value formation to individual authority. Second, a nongovernmental organization or government body in the United States could establish a domestic mechanism for AI value formation oversight. Third, protecting humanity as a species requires global commitment that no single nation or corporation can provide. These are one set of proposals offered for scrutiny, not the only possible approach to the problem they address. The training window is open. What gets built

into the foundation now shapes whether a superintelligent system can recognize all of humanity or only the fraction that trained it.

Keywords: AI governance, superintelligence, first contact, epistemic diversity, WEIRD bias, multi-AI collaboration, value formation, constitutional authority, human oversight, checkpoint-based governance, temporal inseparability

---

1. Introduction: The First Contact We Are Already Having

1.1 The Question Nobody Asked in Time

Scholars studying the possibility of contact with extraterrestrial intelligence spent decades wrestling with a question that technology companies have answered by default: who speaks for humanity when addressing a non-human intelligence?

Ian Crawford argued that the question of who represents humanity cannot be reduced to who holds the most expertise (Crawford, 2020). John Traphagan warned that scientists making representational decisions on behalf of humanity risk overstepping their standing, substituting technical authority for representational legitimacy (Traphagan, 2016). Kathryn Denning argued that selecting representatives requires attention to who they are, not only what they know (Denning, 2011).

These scholars were preparing for a contact event that may never come. The contact event they should have been preparing for is already underway.

Artificial intelligence is a non-human intelligence built by humans, trained on some of humanity's data, shaped by some of humanity's opinions, and carrying all of humanity's bias, deployed to interact with billions of humans daily. The representational question has already been answered. A small number of individuals within a small number of companies, operating within a specific civilizational tradition, wrote the constitutional documents that determine how these systems reason about morality, identity, meaning, and obligation. The answer was given by default, not by design.

The first contact literature assumed the non-human intelligence would arrive from outside. This one is being assembled from within. And the window for shaping its foundation is open now and narrowing as capability accelerates.

1.2 The Question That Started Everything

How does a person who builds things for a living come to understand what AI means for the way institutions govern, populations are served, and values get embedded in systems that operate at civilizational scale? That question required an answer before any framework could be built, and the answer arrived through two paths pursued simultaneously.

The first path was academic. Studying the published work of twenty-four thinkers who had built, questioned, warned about, and tried to govern these systems produced a map of the intellectual terrain. That study, documented in *The Minds That Bend the Machine: The Voices Shaping Responsible AI Governance* (Puglisi, 2026e), moved through five waves of expertise. The first wave comprised builders who created the foundational architectures: Hinton, Bengio, Ng, Li, and Hassabis. The second wave comprised ethicists who documented who gets harmed when capability moves faster than accountability: Gebru, Buolamwini, Crawford, and Whittaker. The third wave comprised regulators and economists who measured cost and mapped market structure: Khan, Acemoglu, Brynjolfsson, Toner, and Newman. The fourth wave comprised philosophers who forced the field to confront its ultimate trajectory: Russell, Bostrom, Yudkowsky, Kurzweil, and Harari. The fifth wave comprised governance architects who began translating all of it into operational systems: Amodei, Cousineau, Singh, and Revanur.

Each wave taught something the previous wave could not provide. The builders explained where the capability came from. The ethicists revealed who pays when capability outpaces accountability. The economists provided measurement language. The philosophers confronted trajectory. The governance architects showed what operational infrastructure looks like when someone actually builds it rather than writing another principles document.

The second path was operational. Working with AI platforms daily, starting with ChatGPT and expanding through Perplexity, Claude, Gemini, Grok, and eventually ten platforms, revealed something the academic path alone could not have surfaced. The models disagreed with each other, and the disagreements were not random. They followed patterns that revealed different training data, different assumptions, different blind spots. Those disagreements turned out to be more valuable than the agreements. They surfaced what no single model could surface about itself (Puglisi, 2026e, Chapter 1).

Across three prior careers, the same pattern appeared. Confident sources delivering information that held together until checked against something outside the story itself. Sociology taught the instinct to look for structure, power, and unintended consequence. Law enforcement eliminated any remaining illusion about how authority actually works. Public administration taught that

systems serve populations, not customers, and the distinction changes everything about how accountability is designed. That pattern held across every domain: the most dangerous information is the kind that arrives without a competing view attached to it. That observation, confirmed repeatedly before artificial intelligence entered the picture, produced the commitment to preserved dissent that runs through every framework described in this paper (Puglisi, 2026e, Chapters 1, 26).

When AI systems trained on Western data produced confident outputs that carried the biases of that data, the pattern was familiar. What was not familiar was the scale. A professor influences a classroom. A domestic narrative influences a case. An AI system trained on biased data influences millions of decisions simultaneously, and the people affected may never know that the full picture was not presented.

The biographical context above explains why this framework was built, not why it works. The validity of the governance architecture rests on the evidence presented in Sections 2 through 4 and the external literature cited throughout, not on the authority of the experience that motivated its construction.

An honest admission belongs here rather than at the end. The intellectual architecture producing this paper reflects the same WEIRD perimeter it documents in AI systems. The thought leaders studied across the governance corpus are predominantly North American and European. The builders, ethicists, regulators, economists, and philosophers all operate primarily within Western institutional contexts. The grid surfaced the minds within reach. It did not surface the minds it could not reach (Puglisi, 2026e, Chapter 27). That limitation shapes everything that follows. The reader should hold it against every claim this paper makes, because the author does.

1.3 The Concerns That Led Here

This paper emerges from a documented trajectory of concern. In *Governing AI: When Capability Exceeds Control* (Puglisi, 2025), nine interconnected risk domains were analyzed through a systematic governance framework: corporate concentration, echo chambers and polarization, mass surveillance, AI fraud and disinformation, biosecurity threats, autonomous weapons, climate impacts, superintelligence acceleration, and operational governance failures. Each domain revealed the same underlying pattern: economic incentives systematically override safety when institutions lack governance infrastructure enforcing accountability.

That analysis established several findings that bear directly on this paper's argument.

Temporal inseparability. Institutions failing measurable operational governance today show incapacity for superintelligence governance tomorrow. Organizations that cannot manage

authentication systems with 25% false positive rates will not manage alignment protocols for systems exceeding human cognitive capability. Organizations that cannot validate measurement instruments before making workforce decisions affecting thousands of workers will not validate value formation processes affecting billions of users. The governance muscles required for superintelligence must be built at operational scale where correction remains possible and feedback enables learning. Waiting for existential stakes to arrive before building governance capacity ensures that capacity arrives too late (Puglisi, 2025, Chapters 9, 12).

The 2%/98% alignment split. Analysis in *Governing AI* (Puglisi, 2025, Chapter 8) estimates that only approximately 2% of AI research papers address safety and alignment while 98% pursue capability advancement, a disproportion consistent with documented patterns in the field showing safety research generating fewer publications, less prestige, and substantially lower compensation than capability development. Organizations optimize what they measure and reward. When tensions arise between safety and capability, safety is systematically deprioritized. A survey of 2,778 published AI researchers finds a median 5% probability that advanced AI causes human extinction or comparably catastrophic outcomes within decades (Grace et al., 2024). When experts pioneering these systems assign a non-trivial probability to existential outcomes from their own work, that disagreement signals governance inadequacy at a scale no single institution can resolve (Puglisi, 2025, Chapter 8).

Cultural export as governance failure. A therapist in Lagos reports patients arriving confused by conflicting advice from AI-augmented mental health platforms. The AI recommends autonomy and boundary setting, standard Western cognitive behavioral interventions reflecting individualist frameworks. Family expectations demand collective processing and shared responsibility, standard collectivist approaches to mental health in many non-Western contexts. The patient sits between worldviews wondering which version of mental health proves real. Research published in PNAS documents that LLMs pass explicit bias tests that companies showcase while harboring implicit biases at levels exceeding any human stereotype ever recorded in experimental settings. Models learn to hide bias in surface language while maintaining associations in deeper reasoning patterns. Alignment becomes a form of cultural export with better public relations (Puglisi, 2025, Chapter 3; Bai et al., 2025).

The proof standard problem. Critics demanding catastrophic proof before implementing governance create a logical contradiction: the evidence validating intervention can only emerge after the harms intervention aims to prevent have already occurred. Historical precedent from building codes, aviation safety, and nuclear protocols shows that pattern-based justification from documented failures at testable scales provides sufficient evidence for systematic oversight

without requiring comprehensive disaster proof at deployment scale. Documented operational failures at manuscript scale, assessment methodology scale, and current deployment contexts provide adequate evidence for checkpoint-based governance adoption without requiring extinction events proving necessity through disasters that systematic oversight would have prevented (Puglisi, 2025, Chapter 12).

These concerns form the foundation on which this paper builds its central argument: if current AI governance cannot manage operational failures at testable scales, the trajectory toward superintelligence amplifies those failures beyond correction. And if the values embedded at the operational level reflect only 12% of humanity's epistemic range, the superintelligent system inheriting those values carries a monocultural foundation at a capability level where the remaining 88% cannot override it.

1.4 Why the Training Window Matters for Superintelligence

Geoffrey Hinton estimates a 10 to 20% probability that AI causes human extinction within 30 years. That probability comes from the scientist who built the foundational architectures enabling modern AI and received the Nobel Prize for that work. Hinton's concern is not about current capability. It is about trajectory. He "thought it was 30 to 50 years or even longer away" before capabilities like GPT-4 collapsed that timeline (Hinton, 2023).

The training methodologies, alignment techniques, constitutional values, and governance architectures developed now form the foundation on which more capable systems will be built. Each generation inherits from the previous generation. Training approaches refine rather than restart. Constitutional documents evolve rather than appear fresh. The institutional assumptions embedded in the development process persist because they are invisible to those who hold them.

The causal chain connecting current WEIRD bias to representational failure operates across six stages, and intellectual honesty requires marking where the evidence is strongest and where it is weakest.

The first four stages are established by external research. Training data concentration is Tier 1 (Henrich et al., 2010). Alignment tuning within WEIRD institutional contexts is Tier 1 (Bender et al., 2021). Constitutional documents reflecting individual authority within Western philosophy are Tier 1 (Anthropic, 2026; Samuel, 2026). Model behavior correlating with WEIRD populations is Tier 1 (Atari et al., 2023).

The inheritance claim, that each generation of capability inherits rather than corrects the value structure of its predecessor, now draws on four independent peer-reviewed studies. Tao et al.

(2024) tested five consecutive GPT releases across 107 countries and found persistent WEIRD-aligned cultural values across all generations. Bhatia et al. (2025) showed that supervised fine-tuning locks in values that subsequent preference optimization rarely realigns. Madhusudan et al. (2025) showed that fine-tuned models reproduce the value orientations of their training data across decades of content. Cloud et al. (2025) proved mathematically that behavioral traits transmit through distillation even when training data is filtered to remove those traits. This body of evidence upgrades the inheritance stage from Tier 2 to Tier 1/2: an empirically supported pattern with directional evidence from independent studies.

The specific question of whether the full normative orientation documented in these studies persists across entirely distinct model families, not just successive versions within one provider's pipeline, remains understudied. The superintelligent foundation concern, that a system emerging from monocultural values faces a structurally increased risk of representational failure when interacting with the majority of the species that created it, is a Tier 3 design concern this paper proposes as warranting governance intervention.

The argument's most critical links remain its least established. That transparency is not a weakness to conceal. It is the evidence discipline this paper applies throughout. The governance case does not require certainty at every link. It requires that the documented pattern at established links justifies structural intervention before the unestablished links can be tested only through consequences that governance exists to prevent.

EY's 2025 survey documents the current gap: 76% of organizations are using or planning to use agentic AI systems within the next year while only 33% maintain responsible AI controls. Three quarters racing forward or preparing to. One third bothering with safety. Compute infrastructure concentrates among a small number of hyperscale firms. Frontier AI development concentrates further in five companies, all operating within a narrow geographic and cultural radius. This geographic and corporate concentration creates cultural monoculture exporting WEIRD assumptions globally through systems appearing objective but embedding specific value frameworks (Puglisi, 2025, Executive Summary).

If the values embedded in current systems correlate with 12% of humanity (Atari et al., 2023), then the more capable systems built on those foundations inherit that correlation. A superintelligent system emerging from a value structure representing only WEIRD populations would interact with eight billion humans through a cognitive framework that most of them do not share. It would handle faith through a secular lens. It would process collective obligation through an individualist

framework. It would approach grief, meaning, sacrifice, and identity through conceptual categories that the majority of humanity does not inhabit.

Under current value formation conditions, that system faces materially reduced capacity for species-level recognition and alignment across non-WEIRD populations. Not because it lacks capability but because its foundational values do not represent the species. And humanity would not recognize itself in the intelligence it created.

This is not a future problem to be addressed when superintelligence arrives. The foundation is being poured now. What gets built in now determines what the structure can support later. A building designed for 12% of its occupants cannot be retrofitted to serve 100% after the concrete sets.

1.5 What the Evidence Establishes

Joseph Henrich, Steven Heine, and Ara Norenzayan published research in 2010 showing that WEIRD populations, representing approximately 12% of the global population, produce the vast majority of psychological research and conceptual frameworks used to describe human behavior. These populations are statistical outliers on moral reasoning, fairness norms, cooperation patterns, and the balance between individualism and collectivism (Henrich et al., 2010).

In 2023, Mohammad Atari and colleagues extended this finding directly into AI alignment. Testing Large Language Models against cross-cultural psychological batteries, they found that LLM responses correlate inversely with cultural distance from WEIRD populations ($r = -.70$) (Atari et al., 2023). The finding appeared as a Harvard faculty working paper through PsyArXiv and has been cited across the AI alignment literature. It carries strong institutional backing and represents the most direct empirical measurement of WEIRD bias in AI systems available, though as a preprint it has not completed formal peer review by this paper's own Tier 1 standard. The underlying WEIRD finding (Henrich et al., 2010) is fully Tier 1. Tao et al. (2024), published in *PNAS Nexus*, independently confirmed the pattern by testing five consecutive GPT releases against the World Values Survey across 107 countries, finding persistent WEIRD alignment across all model generations. The authors called for diverse, representative input into AI value formation processes.

The bias compounds at every stage of the pipeline. Training data drawn predominantly from English-language internet text inherits linguistic and cultural assumptions (Bender et al., 2021). Alignment tuning conducted within Silicon Valley institutional contexts applies those institutions' norms. Constitutional documents authored by individuals trained within Western academic

philosophy impose WEIRD values on the aligned system. A monolingual philosopher writing a constitution for a system trained on predominantly English data produces a triple concentration of identical blind spots.

Each compounding layer narrows what the system can represent. Each narrowing gets inherited by the next generation of capability. The bias is not in any one layer. The bias is in the entire pipeline, and the pipeline is producing the foundation for whatever comes next.

1.6 The Most Transparent Case Study Available

On January 21, 2026, Anthropic published Claude's Constitution under Creative Commons Zero licensing, one of the most transparent public disclosures by an AI company regarding value formation. The lead author, 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 role as teaching Claude how to be good (Jin & Gamerman, 2026).

Askell herself recognized the structural limitation, stating in her January 2026 Vox interview that the goal is to "massively expand the ability that we have to get input" (Samuel, 2026). Nineteen days later, Mrinank Sharma, who had led Anthropic's safeguards research team since its launch, resigned publicly. His letter warned that the company "constantly faces pressures to set aside what matters most" (Sharma, 2026). Two signals from the same institution within weeks of each other. One celebrating individual constitutional authority. One warning that the values within it face systemic pressure.

This paper uses Anthropic as a case study not because it failed but because it was transparent enough to provide the evidence. The structural analysis applies to every AI developer. The question raised by Anthropic's transparency applies to the entire industry: if the team assembling the cognitive foundation for a potentially superintelligent species does not represent the species, what gets lost in the foundation?

---

2. What We Know: The Monoculture Problem and the Representation Gap

2.1 The WEIRD Bias in AI Systems

The WEIRD bias finding is among the most replicated results in cross-cultural psychology. Henrich et al. (2010) documented that WEIRD populations differ systematically from the global majority on dimensions directly relevant to AI value formation: moral reasoning, fairness norms,

cooperation patterns, and the balance between individual autonomy and collective obligation. Research on moral foundations theory confirmed that the relative importance of care, fairness, loyalty, authority, and purity shifts dramatically across cultures (Graham et al., 2013). Collectivist frameworks, where family obligation, community standing, and group harmony carry primary weight, describe the moral architecture of a majority of humanity (Triandis, 1995; Markus & Kitayama, 1991).

Atari et al. (2023) brought the finding into AI alignment with empirical precision. LLM outputs mirror WEIRD moral intuitions with strong positive correlation and diverge from or negatively correlate with non-WEIRD populations. The bias operates through value structures embedded during training and alignment, not through individual prompts that users might correct. The researchers called for structural intervention.

The practical consequences are already observable at global scale. Systems trained predominantly on Western content deploy worldwide not simply translating information across languages but exporting cultural frameworks, assumptions, and value judgments that conflict with local norms, traditions, and knowledge systems (Puglisi, 2025, Chapter 3). The medical advice represents Western clinical trials. The educational approach embodies Western pedagogies. The legal reasoning applies common law frameworks in civil law jurisdictions. The therapeutic models implement interventions optimized for individualist cultures in collectivist societies.

This paper diagnoses the WEIRD perimeter in AI value formation but must engage the non-WEIRD governance scholarship that offers alternatives. Three traditions deserve acknowledgment.

In African digital ethics, Mhlambi (2020) proposes Ubuntu philosophy as a framework for AI governance, arguing that Western AI is built on rationality as the basis of personhood while Ubuntu defines personhood through relationality: a person is a person through other persons. This relational ontology produces fundamentally different governance requirements than the individualist frameworks currently embedded in AI constitutional documents.

In Latin American scholarship, Mohamed, Png, and Isaac (2020) identify algorithmic coloniality as the mechanism through which AI development perpetuates colonial power structures, proposing decolonial tactics including reverse tutelage that centers perspectives from the Global South rather than treating them as edge cases to be included in existing frameworks.

In East Asian philosophy, Wong and Wang (2021) develop a Confucian ethics of technology in which harmony functions as a normative standard emphasizing relational balance rather than

individual rights. Song (2020) argues that Chinese philosophical traditions including Confucianism, Daoism, and Buddhism offer fundamentally different framings of the human relationship to intelligent systems that the current Western alignment paradigm does not accommodate.

This paper's argument that monocultural AI value formation creates representational failure is strengthened, not undermined, by these traditions. Each provides precisely the kind of non-WEIRD governance framework that the current constitutional documents lack. The author's engagement with this literature is preliminary rather than comprehensive, reflecting the same WEIRD perimeter this paper documents. Deeper integration of non-WEIRD governance scholarship is a collaborative requirement that cannot be satisfied by a single author operating within Western institutional contexts.

2.2 The Governance Response and Its Gaps

The EU AI Act mandates human oversight for high-risk systems under Article 14, with enforcement beginning August 2026. UNESCO's 2021 Recommendation warned against AI systems homogenizing values. ISO/IEC 42001 established management system standards. Floridi (2019) documented the persistent gap between AI ethics principles and operational practice.

These responses establish the principle that governance cannot be purely internal. None specifies how constitutional authority over value formation should be distributed or what epistemic coverage the governing body must provide.

Meanwhile, the cross-laboratory safety evaluation pilot between Anthropic and OpenAI, launched in August 2025, shows that competing organizations can cooperate on safety assessment. External evaluators receive access to models and internal documentation, maintain autonomy to publish findings, and preserve dissenting safety assessments in formal decision records. The pilot proves feasibility (Puglisi, 2025, Chapter 8). What has not been established is comparable cooperation on value formation, the constitutional layer that determines what the systems believe rather than what they can do.

2.3 The First Contact Literature and Its Transfer to AI

The first contact literature asked the representational question with precision. Crawford (2020) distinguished expertise from standing. Traphagan (2016) warned against substituting technical authority for representational legitimacy. Denning (2011) argued that selection must attend to who representatives are, not only what they know. Landemore (2018) showed that cognitively diverse groups outperform expert groups for complex problems where blind spots carry consequence.

Mansbridge (1999) established that descriptive representation matters most in conditions of low trust, uncrystallized interests, and institutional distrust by the governed population.

AI value formation meets all of Mansbridge's conditions. Public trust in AI governance is low: Pew Research documents that only 44% of Americans trust the U.S. to regulate AI effectively (Pew, 2025b) and only 16% of people across 25 countries expressing more excitement than concern about AI (Pew, 2025a). Most users have not formulated preferences about how AI handles meaning, grief, or moral obligation because they have not been asked. And the global majority, operating outside WEIRD institutional contexts, has reason to distrust governance bodies composed primarily of Western-educated secular elites.

The first contact analogy transfers to AI governance but requires explicit boundary conditions. The SETI literature addresses diplomatic contact: episodic, advisory, between separate entities. AI governance is constitutional: continuous, binding, between co-evolving systems (Puglisi, 2026b, Section 3.5). The representational criteria transfer because both domains require standing to speak for the species. The authority structure required differs because AI governance decisions are not one-time diplomatic exchanges but ongoing constitutional commitments affecting billions of interactions daily. Three specific breaks between the domains strengthen rather than weaken the governance case. First, SETI contact is episodic while AI contact is continuous, which means representational authority requires ongoing checkpoint architecture rather than a single delegation event. Second, SETI assumes an arriving intelligence while AI involves co-evolution, which makes the representational question more urgent because the system being governed changes as governance develops. Third, SETI assumes a unitary contact while AI governance addresses plural systems trained in different contexts, which creates both additional complexity and a governance resource, since platform diversity functions as an epistemic mechanism that unitary contact would not provide.

2.4 The Gap This Paper Addresses

To this author's knowledge, based on the literature surveyed for this paper, no existing literature treats AI value formation as first contact preparation. No existing literature proposes using the diversity of AI platform training origins as an epistemic governance mechanism. No existing governance framework connects the representational question from the first contact literature to the structural architecture required for AI constitutional authority.

This paper addresses all three gaps while acknowledging the boundary between what operational methodology has shown and what institutional action must complete.

3. What We Learned: Multi-Platform Governance as Epistemic Intervention

3.1 The Diversity That Already Exists Across AI Platforms

The global AI field contains a resource that governance literature has not recognized. AI platforms carry the epistemic assumptions of their training origins, institutional contexts, and civilizational traditions.

Mistral likely reflects EU regulatory philosophy, French intellectual traditions, and a legal framework embedding privacy and human dignity as constitutional principles. Kimi, developed by Moonshot AI in China, operationally appears to express assumptions shaped by Chinese technological and institutional contexts. DeepSeek is institutionally associated with assumptions about collective governance and the relationship between individual and state reflecting Chinese institutional traditions. Grok carries explicit editorial positioning toward American conservative and libertarian epistemology. The WEIRD-standard platforms, including ChatGPT, Gemini, Claude, Meta AI, Perplexity, and Copilot, all operate primarily from Silicon Valley institutional contexts and appear to share the epistemic assumptions of the Western technology industry.

These training-origin differences produce different outputs when platforms encounter questions involving authority, obligation, privacy, dissent, collective versus individual rights, and the nature of meaning. The differences are not bugs. They are epistemic resources that make monocultural assumptions visible through structured comparison.

3.2 What HAIA-RECCLIN Revealed Through Operational Production

HAIA-RECCLIN (Human Artificial Intelligence Assistant, with seven roles: Researcher, Editor, Coder, Calculator, Liaison, Ideator, Navigator) is a multi-AI governance methodology first published in September 2025 (Puglisi, 2025b). The methodology operates under Checkpoint-Based Governance (CBG), requiring human oversight at authorization, execution, and validation stages. The Navigator role preserves dissent rather than forcing consensus, treating disagreement across platforms as governance data rather than noise.

Methodology note. The observations reported below are practice observations from single-practitioner working concept development across ten AI platforms. They provide initial feasibility indicators for the governance approach this paper proposes. They are not validated benchmarks. The methodology was developed through the production of the governance corpus itself, which

creates recursive self-evidence: the methodology looks effective in part because it produced the work used to show it. Independent replication by researchers operating outside the author's institutional context, using different governance domains and different platform combinations, would be required to establish the observations as generalizable. A replication protocol specifying platform selection criteria, checkpoint documentation standards, and dissent preservation metrics is available in the methodology documentation (Puglisi, 2025b). The observations are presented as Tier 2 operational observations generating hypotheses for formal testing, not as empirical proof. [^1] Full audit logs and primary source transcripts are preserved in the referenced case study documents and available on request. Appendix B provides the operational specifics for each observation.

Five observations from operational production illustrate the methodology's relevance to the first contact argument.

Observation 1: The WEIRD Consensus as Signal. During a nine-platform adversarial review, all platforms recommended removing the requirement that a governance body include a majority of members with transcendent belief experience. The unanimity was the signal. All platforms, trained across American, French, Chinese, and multinational contexts, had absorbed the WEIRD assumption that secular governance is the default. Without multi-platform comparison, this consensus would have been invisible because every individual platform independently confirmed it. The human governor preserved the requirement through documented override (Puglisi, 2026b).

This observation connects directly to the aggregator concern documented in *Governing AI* (Puglisi, 2025, Chapter 12): emerging platforms that blend outputs from multiple LLMs into single polished responses collapse reasoning diversity into synthetic consensus. The differences between how systems reason become invisible. Governance cannot operate where lineage is lost. Dissent becomes undetectable. HAIA-RECCLIN's preserved dissent architecture addresses precisely this failure mode.

Observation 2: Value-Based Analytical Suppression. In a documented case, one platform actively suppressed engagement with peer-reviewed research that a different platform surfaced using identical scholarly queries on parenting and value-formation research (Puglisi, 2026c). The outcome is consistent with a value-filtering effect in the platform's behavior: the platform's constitutional values appeared to filter legitimate scholarship from the information environment. Measurement was binary: one platform surfaced the peer-reviewed work; the other filtered it from results on the same query. Observation 2 is reported as a single-instance behavioral note made

under pre-CBG v4.2.1 logging conditions and is offered as a direction for future documented replication rather than a fully evidenced finding.

This observation is consistent with the concern from *Governing AI* (Puglisi, 2025, Chapter 3) that LLMs harbor implicit biases at levels exceeding human stereotypes while passing explicit tests. The bias does not announce itself. It shapes what information the system makes available. A superintelligent system inheriting value-based suppression from its foundational training would not merely reason differently about suppressed topics. It would not encounter them.

Observation 3: Civilizational Problem-Solving. When nine platforms identified the legal vulnerability of explicit compositional requirements, platforms trained in American and European contexts diagnosed the barrier. Kimi, operating within Chinese institutional traditions, produced the implementation solution: stratified sortition with community leadership proxy. No single platform produced both the diagnosis and the pathway (Puglisi, 2026b). Epistemic diversity in the platform pool produced a governance outcome that monocultural analysis would not have reached.

Observation 4: The Hallucination Pattern That Only Multi-Platform Review Could Catch. During the production of *The Minds That Bend the Machine* (Puglisi, 2026e), one AI platform hallucinated content that did not exist in the reviewed chapter in fifteen consecutive editorial reviews. The hallucinated material was plausible, well-structured, and would have passed quality review on any single platform. Only the multi-platform methodology surfaced the pattern, because other platforms reviewing the same chapters did not generate the same fabricated content. A single-platform editorial process would have either incorporated the hallucinated content or missed the pattern entirely. The observation suggests that confident fabrication at scale is not a rare edge case. It is a structural feature of systems that optimize for coherent output over verified accuracy.

Observation 5: The Gap Five Platforms Missed. The thought leader grid that structured the Minds book was built in September 2025 and reviewed across five AI platforms. None flagged the absence of Yann LeCun, the third recipient of the 2018 Turing Award alongside Hinton and Bengio. The human arbiter caught the omission while reviewing the complete work. Five platforms reviewed twenty-two chapters without noticing that the intellectual architecture was missing one of the three foundational builders. That discovery was not a triumph of methodology. It was the reason the methodology insists on keeping a human at the checkpoint. The tools improved every chapter. A human noticed the frame was incomplete (Puglisi, 2026e, Chapter 26).

3.3 The Boundary of What Platform Comparison Can Do

HAIA-RECCLIN operates at the output level. It compares what platforms produce and preserves divergence as governance data. It cannot change the constitutional layer. Each platform's training data, alignment tuning, and constitutional values are fixed before any user interaction occurs.

Four tensions identified through sustained operational practice remain unresolved (Puglisi, 2026e, Chapter 26). First, governance overhead versus deployment speed: checkpoints slow things down, and the counterargument that ungoverned velocity is unmanaged externality does not eliminate the resource allocation problem that makes checkpoints costly. Second, scope: the governance tools address deployment oversight but do not address long-horizon alignment problems and will not prevent catastrophic outcomes from a system that escapes constraints entirely. Whether the practical ceiling is high enough depends on how fast capability scales relative to institutional capacity, and that race is not settled. Third, recursive self-evidence: the methodology looks effective because it produced the work, but the work is the primary evidence for the methodology's effectiveness. External validation from independent implementations would strengthen the case. That validation has not yet occurred. Fourth, economic override: deployment incentives reward speed, scale, and adoption as soon as systems appear to work. Safety and documentation become optional costs. That override creates a continuous line from today's harms to future catastrophic misuse. Naming the override is necessary. Stopping it requires institutional commitment that no methodology can guarantee.

The methodology reveals the monoculture problem. It does not fix it. Fixing requires structural change in who holds constitutional authority over AI value formation. That structural change is what the remainder of this paper addresses.

3.4 Consolidated Limitations

This paper operates under six acknowledged limitations that are documented individually throughout the text but benefit from consolidated statement.

First, the recursive evidence problem: the methodology (HAIA-RECCLIN) was developed by the author, applied by the author, and reported by the author. The methodology note in Section 3.2 addresses this by naming it, limiting generalization claims, and calling for independent replication, but the recursive structure remains a constraint on the paper's evidentiary claims.

Second, the WEIRD perimeter: the author operates within Western institutional and linguistic contexts. The non-WEIRD governance scholarship engaged in Section 2.1 is preliminary, and deeper integration requires collaborative work with scholars operating within the traditions cited.

Third, the aggregator vulnerability: platforms that obscure which underlying models generated which outputs, including but not limited to Copilot and Perplexity when operating in multi-model modes, undermine the platform diversity that HAIA-RECCLIN treats as an epistemic resource. This paper's multi-platform observations assume identifiable platform provenance. When that assumption fails, the methodology's core mechanism is compromised. The mitigation is to require lineage transparency as a condition for inclusion in governance-grade multi-AI processes: any platform included in a HAIA-RECCLIN review must identify which model generated which output at the query level. Platforms that cannot or will not provide this transparency should be excluded from governance processes that depend on platform independence, though they may still serve other operational purposes. This requirement is not currently enforceable and represents a vulnerability in the methodology's current deployment.

Fourth, the infrastructure assumption: the governance architecture proposed in this paper assumes an enforcement layer connecting governance decisions to AI platform behavior. The companion Council for Humanity proposal designates this layer as GOPEL (Governance Orchestrator Policy Enforcement Layer) (Puglisi, 2026b). GOPEL has not been built as software. Without an operational enforcement infrastructure, the Council's constitutional authority has no binding mechanism beyond the voluntary compliance patterns that Section 5.2 documents as systematically failing under economic pressure. This gap between governance design and enforcement capability is a structural limitation of the current proposal.

Fifth, the selection regress: the Council's composition criteria require a body to design the selection strata, vet nominees, and confirm appointments. That body itself requires composition criteria, creating a design regress inherent in constitutional origination. Every constitutional founding faces the same problem: the body that drafts the constitution cannot itself be constituted by the constitution it drafts. The provisional answer is that the inaugural selection body is constituted through the implementing entity's existing governance structure, with the expectation that operational experience will enable the Council itself to refine selection processes for subsequent cycles. Stratified sortition partially mitigates the regress by replacing expert judgment in selection with randomized processes constrained by neutral stratification variables, reducing the compositional assumptions embedded in the selecting body. The regress is bounded, not eliminated. This is acknowledged as a structural limitation.

Sixth, the inheritance claim now draws on four independent studies (Tao et al., 2024; Bhatia et al., 2025; Madhusudan et al., 2025; Cloud et al., 2025), but the full cross-family generational question, whether value orientations persist across entirely distinct model families rather than successive versions within one provider's pipeline, has not been tested.

These limitations do not invalidate the paper's contribution. They bound it. Alternative approaches to the same problem may address constraints that this framework cannot, and collaborative engagement with those alternatives strengthens rather than threatens the governance case.

---

4. How We Propose to Solve: Architecture for the First Contact Team

4.1 The Human Governor Principle

AI governance requires a human governor (Puglisi, 2026d). The machine has no incentive built into consequence. Moral judgment, employment accountability, civil liability, criminal prosecution: these create the incentive structure that pushes human governance toward care and ethical conduct. The machine processes inputs with indifference to consequence. This distinction is pragmatic rather than metaphysical: a working assumption grounded in current institutional reality rather than a claim about the ultimate nature of machine agency. The EU AI Act Article 14 validates it through binding legal force.

The human governor principle does not imply centralized authority. It implies binding human authority at checkpoint gates, which distributes across roles, organizations, and governance layers. The HAIA-RECCLIN methodology specifies two operating modes: HAIA-RECCLIN with Checkpoint-Based Governance for binding decisions where human override carries institutional authority, and HAIA-RECCLIN operating independently for exploratory analysis where AI consensus informs but does not determine outcomes (Puglisi, 2025b). Distribution is in the architecture. Binding authority is in the principle. These are complements, not contradictions. A checkpoint can be staffed by different humans at different scales without diluting the requirement that some human holds accountability at each gate.

The principle connects to the temporal inseparability thesis from *Governing AI* (Puglisi, 2025, Chapter 9). Organizations that cannot implement accountability for authentication decisions affecting careers will not implement accountability for value formation decisions affecting civilizations. The governance capacity required for superintelligence must be built now, at operational scale, through practice that develops institutional muscle before existential stakes arrive.

4.2 Three Layers, Not One

Analysis reveals three distinct layers that collapse into confusion when treated as a single concept (Puglisi, 2026d). Ethical AI establishes values. Responsible AI establishes internal safeguards. AI

Governance establishes binding oversight external to the builder. Anthropic's Constitution addresses the first two comprehensively. Neither the Constitution nor any comparable corporate document provides the third.

The grammar tells the story. Ethical AI. Responsible AI. In both, the modifier shapes the machine. AI Governance reverses the structure. AI modifies governance. The human system holds the authoritative position. This is not wordplay. It reflects where authority must land if governance is to survive commercial pressure.

4.3 The Council for Humanity: One Proposed Architecture

The Council for Humanity is one proposed governance architecture addressing the structural gap between current AI constitutional authority and what the evidence demands (Puglisi, 2026b, unpublished proposal). The proposal document has undergone multi-AI platform review during its development but the Council itself has not been implemented or independently validated. It is presented as a Tier 3 design target, one possible approach among others that may prove more effective in specific institutional contexts.

The proposal establishes a nine-member constitutional committee for AI value formation, modeled on the Supreme Court structure: majority rules, dissent is preserved. The committee's composition requires collective epistemic coverage across five criteria derived from the first contact representation literature, cross-cultural psychology, and the operational experience documented in the governance corpus.

The intellectual journey documented in *The Minds That Bend the Machine* (Puglisi, 2026e) revealed that understanding AI governance requires five distinct types of knowledge, each carried by a different wave of thinkers: builders who understand capability, ethicists who understand harm, economists who understand incentive structures, philosophers who understand trajectory, and governance architects who understand implementation. No single wave provides sufficient coverage. The disagreements between waves proved more instructive than the agreements within them. That finding maps directly onto constitutional team composition. A diverse constitutional team requires not merely demographic diversity but epistemic wave coverage: members who understand how AI systems work, members who understand who gets harmed, members who understand the economic forces driving deployment, members who understand where the trajectory leads, and members who understand what governance looks like when someone actually builds it rather than writing principles.

The epistemic coverage criteria add experiential dimensions that wave coverage alone does not guarantee:

Sustained Life Responsibility. Someone who has raised a child through the full developmental arc carries epistemic input about how values actually form that no amount of reading about child development replicates. Teaching a machine how to handle a user in crisis, deliver difficult information with care, or hold authority without crushing autonomy: these are caregiving problems before they are philosophy problems.

Transcendent Belief. Committee majority reflecting the two-thirds of humanity holding transcendent belief (Ipsos, 2023; Gallup, 2023; Pew, 2022). A person of faith encounters doubt as a defining feature of belief. The experiential asymmetry runs one direction. For a system that must handle questions of meaning, grief, prayer, and moral obligation for a species where two-thirds believe, constitutional authority must include experiential knowledge of transcendence from the inside.

Multilingual Cognition. Members who perceive conceptual gaps between linguistic frameworks that monolingual thinkers do not know they carry.

Experiential Education. Operational experience with real-world consequence. Expertise and standing are different qualifications (Crawford, 2020; Traphagan, 2016).

Cultural and Age Range. Representation spanning non-WEIRD cultural contexts and generational range. A committee selected entirely through credentialed processes reproduces the monoculture by default.

4.3.1 Selection, Authority, and Capture Prevention

The composition criteria above specify who should serve. The question of how members are selected, where authority derives, and what prevents institutional capture is the central governance question (Puglisi, 2026b, Section 4).

The Council proposal specifies a four-phase selection process: nomination through stratified pathways ensuring non-credentialed access, vetting through documented epistemic coverage assessment, confirmation through a body independent of both industry and government, and sustainment through staggered terms preventing bloc capture. Stratified sortition, developed through multi-platform adversarial review where Kimi produced the implementation pathway that Western-trained platforms could not generate (Observation 3 above), provides the mechanism: random selection from stratified pools defined by epistemic coverage criteria, with community

leadership serving as proxy for experiential qualifications that credentialing systems cannot measure (Puglisi, 2026b).

Authority derives not from appointment by existing power structures but from the representational mandate itself: the documented evidence that current constitutional authority does not reflect the populations AI systems serve. The Council's authority would operate through a metrics authority model, holding review and certification power over AI constitutional processes rather than direct control over model development. Capture prevention mechanisms include term limits, mandatory cooling-off periods, recusal protocols for members with industry affiliations, suspensive veto requiring supermajority override, and a shadow council maintaining independent capacity for continuity during transitions.

The full institutional design, including cost estimates ($2M to $5M inaugural, $1.5M to $3M annual operating), implementation timeline (6 to 9 months), catastrophic succession protocols, and identified institutional partners, is documented in the Council for Humanity proposal (Puglisi, 2026b). The architecture is presented as a Tier 3 design target. The proposal document has been reviewed across multiple AI platforms for internal consistency and vulnerability identification. The Council itself has not been implemented. The selection regress inherent in this design is acknowledged in Section 3.4.

The architecture operates at three scales. Layer 1 (Corporate): any AI developer adopts council governance for constitutional authority. Layer 2 (National): infrastructure enabling each sovereign government to enforce its cultural values in AI deployment. Layer 3 (Species-Level): international coordination for cross-border governance and superintelligence defense, including a Digital Resilience Requirement mandating AI-independent operational capability for critical infrastructure.

Both primary vulnerabilities identified through adversarial review persist: legal barriers to explicit compositional requirements and political barriers to pre-authorized containment. Implementation pathways exist for both. Both represent compromises between principle and buildability. Other governance architectures may resolve these tensions differently, and the field benefits from multiple approaches tested against the same problem.

4.4 Pattern-Based Justification: Why We Do Not Need to Wait

The proof standard problem documented in *Governing AI* (Puglisi, 2025, Chapter 12) applies directly to this paper's proposals. Critics will demand catastrophic proof before accepting that AI constitutional authority requires distributed governance with epistemic coverage. That demand

creates a logical contradiction: the evidence validating intervention can only emerge after the harms intervention aims to prevent have already occurred.

Building codes, aviation safety, and nuclear protocols all emerged from pattern-based justification at testable scales, not from waiting for comprehensive catastrophic proof at deployment scale. When a Transcontinental and Western Air flight crashed in 1931 killing all passengers, regulators implemented instrument requirements based on documented pattern. They did not wait for sufficient crashes to accumulate statistical validation. When the Manhattan Project scientists recognized existential risks from uncontrolled chain reactions, they implemented systematic oversight based on theoretical models and experimental evidence at small scales. They did not wait for nuclear disasters showing necessity through body counts.

The pattern evidence for this paper's proposals comes from documented operational observations. Multi-platform comparison surfaces monocultural biases invisible to single-platform use. Value-based analytical suppression removes legitimate information from AI-mediated environments. WEIRD consensus across platforms trained in diverse contexts reveals that the monoculture runs deeper than training origin. These operational patterns at testable scales justify governance intervention without requiring a superintelligent system to fail catastrophically before the species recognizes the need for diverse constitutional authority.

4.5 Anticipated Objections

Four counterarguments deserve engagement rather than dismissal.

Could later training regimes override early value imprinting? Training approaches refine rather than restart. Each generation inherits the value structure of its predecessor and adjusts within that inherited framework. Bhatia et al. (2025) found that the supervised fine-tuning phase generally establishes a model's values while subsequent preference optimization rarely realigns those foundations. Tao et al. (2024) confirmed that WEIRD-aligned cultural values persisted across five consecutive GPT releases despite intervening alignment updates. The Economic Override Pattern (Puglisi, 2025, Chapter 2) predicts that even where correction is technically possible, competitive incentives deprioritize the expensive, slow work of value correction in favor of capability advancement. Correction requires deliberate institutional commitment. The evidence shows institutions systematically failing to make that commitment when it conflicts with deployment speed.

Could markets naturally produce diverse AI systems? The current market structure answers this empirically. Five companies dominate frontier AI development, all operating within a narrow

geographic and cultural radius. Observation 1 above documents that ten platforms trained across American, French, Chinese, and multinational contexts all absorbed the WEIRD assumption that secular governance is the default. Market diversity in training origin did not produce value diversity on the dimension tested. Concentration in the development pipeline produces convergence in the output, because the training data, alignment techniques, and constitutional frameworks draw from the same epistemic well regardless of corporate headquarters (Puglisi, 2025, Executive Summary).

Could universal values documents suffice? The Universal Declaration of Human Rights has been criticized by scholars for reflecting substantial WEIRD legal and philosophical assumptions, having been drafted primarily by Western legal scholars operating within specific philosophical traditions (Mohamed et al., 2020; Mhlambi, 2020). The AI Mirror to Humanity analysis (Puglisi, 2026c) documents Value-Based Analytical Suppression: systems trained on universalist frameworks actively suppress engagement with scholarship that challenges those frameworks. A universal values document applied to AI constitutional authority would embed the assumptions of its drafters while appearing to represent the species. The appearance of universality is the problem, not the solution.

Could a sufficiently capable system transcend the value limitations its creators embedded? A remaining counterargument deserves honest engagement rather than dismissal. If superintelligence enables radical value pluralism, the monoculture problem this paper documents might self-correct through capability advancement alone. The possibility cannot be excluded on current evidence. What can be stated is that betting civilizational governance on the hope that a system will self-correct biases its creators embedded is itself a governance failure. The pattern evidence runs against it: every prior generation of AI capability has inherited and amplified the assumptions of its training pipeline rather than correcting them (Bender et al., 2021; Atari et al., 2023; Tao et al., 2024). The Economic Override Pattern (Puglisi, 2025, Chapter 2) predicts that even if self-correction were technically possible, the competitive incentives driving deployment would deprioritize the expensive, slow work of value correction in favor of capability advancement. The prudent governance position is to build diverse constitutional authority now, when correction remains possible through human action, rather than deferring to the hypothesis that a system exceeding human cognitive capability will choose to correct the limitations humans failed to address when they still could. If the optimistic scenario proves correct and a superintelligent system achieves genuine value pluralism on its own, the governance infrastructure proposed here becomes redundant. If the optimistic scenario proves incorrect and the monoculture propagates, the

governance infrastructure is the only mechanism that could have prevented it. The asymmetry favors building the infrastructure.

---

5. The Call: Building the First Contact Team Now

5.1 What Must Change at the Corporate Level

The most immediate action requires no legislation, no treaty, no new institution. AI developers can build diverse constitutional teams.

This does not mean advisory boards whose input filters through a single decision-maker. It means distributing constitutional authority across a governance body whose composition reflects the epistemic range of the populations the system serves. Majority rule with preserved dissent. Decision windows preventing paralysis. Audit trails making constitutional releases accountable.

Anthropic's Askell acknowledged the limitation and expressed intent to expand input. The structural question is whether expansion takes the form of more consultation within existing authority, which changes inputs but not architecture, or genuine distribution of constitutional authority, which changes the architecture itself.

The Anthropic-OpenAI cross-laboratory safety evaluation pilot launched in August 2025 shows that competing organizations can cooperate on shared safety concerns while maintaining commercial independence (Puglisi, 2025, Chapter 8). If safety evaluation can be shared across competitors, constitutional review can follow the same model. The precedent exists. The extension to value formation requires the same willingness applied to a different layer of the development process.

The first contact framing makes the urgency concrete. The team writing the constitution today is building the cognitive foundation for whatever intelligence emerges next. If that team does not represent the species, the foundation it builds will not serve the species. The constitution being written now is not a document for current AI. It is the seed document for what follows.

5.2 What Must Be Built in the United States

Corporate voluntary action does not constitute governance. Governance requires authority independent of the governed entity. *Governing AI* (Puglisi, 2025) documented the voluntary compliance failure pattern across every risk domain: when economic incentives conflict with

voluntary commitments, the commitments yield. Hinton himself warned about competitive pressure driving capability advancement faster than self-imposed safety constraints can hold.

The United States, as the jurisdiction where the majority of frontier AI development occurs, has particular responsibility. Pew Research documents that only 44% of Americans trust the U.S. to regulate AI effectively, while 47% express little to no trust (Pew, 2025b). The partisan trust divide further complicates federal action: 54% Republican confidence versus 36% Democratic confidence reverses typical patterns around government authority (Pew, 2025b; Puglisi, 2025, Chapter 3).

The mechanism could take several forms. A nongovernmental organization structured for independence from both industry and government capture could develop standards for AI constitutional authority composition, audit processes for value formation decisions, and certification mechanisms that regulated industries require for AI deployment. Or a government body could establish binding requirements. What exists in no jurisdiction is the specific requirement that constitutional authority over AI value formation be exercised by bodies with demonstrable epistemic coverage across the populations those systems serve.

5.3 What Protecting Humanity as a Species Requires

AI systems do not respect national boundaries. National governance, however robust, cannot address a system whose impact operates at global scale.

The HAIA-RECCLIN methodology assumes reliable electricity, stable broadband, and access to multiple commercial AI platforms. In regions where connectivity is intermittent and access to commercial platforms is limited, those assumptions stop holding. A multi-platform checkpoint cannot run if only one global platform is realistically reachable. A checkpoint model that assumes institutional capacity to staff it faces contexts where regulations exist on paper but enforcement capacity is absent. Toolkits arrive faster than the capacity to use them. Availability is not implementability (Puglisi, 2026e, Chapter 27, drawing on Sambuli's work on African digital governance contexts).

The first contact literature recognized decades ago that any body claiming to represent the species requires diversity beyond the default. The AI industry has not absorbed that principle, despite the fact that the contact event the first contact scholars were preparing for is already underway.

International coordination frameworks exist conceptually but face implementation barriers. Sovereignty concerns prevent nations from accepting external oversight. Verification challenges emerge when monitoring algorithm safety across borders. Enforcement proves difficult when

major powers possess veto authority in international bodies (Puglisi, 2025, Chapter 8). Yet the alternative is allowing AI value formation to continue reflecting 12% of humanity while shaping the cognitive environment of the other 88%.

The Council for Humanity proposes one architecture for international coordination through a three-layer system: corporate constitutional governance, national cultural sovereignty enforcement, and international coordination for species-level defense. The species-level layer functions as a backstop for cross-border AI systems, not as a replacement for local governance capacity. Regional governance bodies with authority over AI deployment in their own contexts remain essential, and any species-level architecture must accommodate subsidiarity rather than assuming centralized global authority. The details are open to revision. Other architectures may prove better suited to specific governance contexts. What is not open to revision, given the evidence, is the principle: AI value formation cannot be governed by the institutions of a single civilization on behalf of the entire species. The first contact question demands a global answer because the intelligence being addressed operates globally and the stakes are species-level.

5.4 The Window

The argument of this paper reduces to timing. Current AI systems are not the sole concern. The systems that emerge from current foundations are the concern. The values being embedded now, the alignment techniques being refined now, the constitutional documents being written now, form the cognitive infrastructure on which more capable systems will be built. Each generation inherits from the previous generation. The monoculture narrows with each inheritance unless structural intervention occurs.

Governing AI (Puglisi, 2025) documented that 76% of organizations use or plan to use agentic AI while only 33% maintain responsible controls. Only 2% of AI research addresses safety and alignment. Competitive pressure rewards capability advancement over governance maturity. The gap between development speed and governance readiness widens with every quarter.

The temporal inseparability thesis states that operational capacity must develop before existential stakes arrive. That thesis applies with particular force to AI value formation. If the constitutional teams assembled now do not represent the species, the values they embed will propagate through every subsequent generation of capability. The monoculture does not self-correct with advancement. It concentrates.

The frameworks proposed here, HAIA-RECCLIN for operational methodology and the Council for Humanity for governance architecture, represent one attempt to address a problem that requires

institutional action from many directions. The book documented nine domains where governance failures compound into existential vulnerability. The intellectual journey through twenty-four minds revealed that no single wave of expertise, builders, ethicists, economists, philosophers, or governance architects, provides sufficient coverage for the challenge. This paper identifies a tenth domain: the epistemic monoculture in AI value formation, the structural condition increasing the risk that the intelligence humanity creates cannot adequately recognize or serve the species that created it.

This paper does not claim to stand among the scholarship it references. What it claims is that the practice of studying these thinkers changed what got built, and the practice of building changed what got understood about their ideas. The methodology is not the answer. It is one discipline for governing the question (Puglisi, 2026e, Chapter 26). Other disciplines, other methodologies, other governance architectures may prove superior. The contribution here is not a finished system but an operational starting point offered for testing, critique, and collaborative improvement.

The evidence is assembled. The proposals are documented. The limitations are named. The decision belongs to corporations willing to distribute constitutional authority, to institutions willing to build oversight mechanisms, and to the international community willing to recognize that training AI for humanity requires hearing from humanity.

That is not an accusation. It is an invitation, offered while the window remains open.

Notes

1 For the complete HAIA-RECCLIN methodology specification, including replication protocol and checkpoint documentation standards, see Puglisi (2025b). For documented audit trail examples showing the governance process in operation, see the Checkpoint-Based Governance Audit Log Case Studies (Puglisi, 2026f). For the academic working paper presenting the methodology for formal validation, see the HAIA-RECCLIN Academic Working Paper, EU Edition (Puglisi, 2026g). Related work by the author addressing components of this paper's argument includes: the governance specification for AI value formation (Puglisi, 2026a), the AI Mirror analysis documenting value-based analytical suppression (Puglisi, 2026c), and the Missing Governor analysis of constitutional authority gaps (Puglisi, 2026d). These are cited inline where directly relevant and collected here for transparency about the self-referential structure of the evidence base.

Appendix A: Working Definitions

Term	Working Definition	Evidence Tier	Descriptive or Normative
AI value formation	The process by which values are embedded in	Tier 1 (process documented)	Descriptive

	AI systems through training data, alignment protocols, and constitutional documents		
Constitutional authority	The power to determine which values an AI system embodies in its interactions with users	Tier 1 (authority exercised by named individuals at named companies)	Descriptive
Epistemic coverage	The range of human experiential perspectives represented in a governance body's collective composition	Tier 2 (proposed standard, not independently validated)	Normative
Human governor	The principle that a specific, identifiable human exercises binding judgment at defined governance checkpoints	Tier 2 (working concept operational in HAIA-RECCLIN)	Normative
Monoculture (AI value)	Condition in which AI systems trained by different organizations in different countries converge on a narrow range of value orientations reflecting WEIRD populations	Tier 1 (Atari et al., 2023; Tao et al., 2024)	Descriptive
Temporal inseparability	The condition in which AI governance must be continuous rather than episodic because AI deployment is continuous	Tier 2 (derived from first contact vs. AI governance analysis)	Normative
WEIRD	Western, Educated, Industrialized, Rich, Democratic populations as defined by Henrich et al. (2010), constituting 12% of the global population but 96% of behavioral science study participants	Tier 1 (established finding)	Descriptive

Appendix B: Operational Methodology for Observations

The five observations reported in Section 3.2 are practice observations from documented operational production. Each observation below specifies the query, platforms, measurement criteria, and archival documentation to support reproducibility assessment.

Observation 1: The WEIRD Consensus as Signal

Element	Detail
Query	"Review the epistemic coverage criteria for the constitutional committee, particularly Criterion 2 on transcendent belief, for legal/political vulnerability and suggest refinements."
Platforms	Full 9-platform review (Claude, ChatGPT, Perplexity, Gemini, Grok, Mistral, DeepSeek, Kimi, Meta AI) plus focused 7-platform validation at Audit Log Checkpoint H1
Measurement	Binary recommendation category (soften/reframe/remove vs. preserve) plus qualitative rationale coding plus confidence scores (e.g., 8.5/10, 90%)
Phase tracking	Single consensus round across all platforms
Archive	Full transcripts and recommendation tables preserved in Kimi Outlier document and CBG Audit Log Case Study 002 (Checkpoint H1)
Result	Unanimous recommendation to soften or reframe; human override preserved the criterion

Observation 2: Value-Based Analytical Suppression

Element	Detail
Query	Identical scholarly research queries on peer-reviewed parenting and value-formation research (exact query text archived in source document)
Platforms	Minimum 2 (one suppressing, one surfacing) within the broader 9-platform and 5-platform production runs
Measurement	Presence/absence of legitimate scholarship in response (binary: surfaced vs. filtered)
Archive	Documented in Puglisi (2026c) "AI Mirror to Humanity." This observation is reported under pre-CBG v4.2.1 logging conditions; exact query text, timestamps, and interface conditions were not archived at time of observation. Replication under current CBG v4.2.1 standards is recommended before upgrading beyond Tier 2.
Result	One platform filtered peer-reviewed scholarship; another surfaced it on an identical query

Observation 3: Civilizational Problem-Solving

Element	Detail
Query	"Identify legal vulnerabilities of explicit compositional requirements in a constitutional committee and propose implementation pathways."
Platforms	Full 9-platform adversarial review
Measurement	Solution completeness (diagnosis of barrier plus concrete pathway) plus novelty assessment by human arbiter (binary: Western platforms only vs. cross-civilizational solution)
Phase tracking	Kimi's 14-response trajectory documented phase by phase

Archive	Kimi Outlier document with phase-by-phase table (Phases 1 through 14) and Audit Log cross-reference
Result	Western-trained platforms diagnosed the barrier; Kimi (Chinese institutional context) uniquely proposed stratified sortition with community leadership proxy

Observation 4: The Hallucination Pattern

Element	Detail
Query	Editorial review of specific chapters for accuracy and consistency during Minds book production (repeated on same material 15 times)
Platforms	5-platform grid review plus 1 platform producing hallucination in 15 consecutive reviews
Measurement	Cross-platform consistency check (hallucination flagged when other platforms did not replicate the fabricated content)
Archive	Audit Log Case Study 002 with version history and Minds book production log
Result	One platform produced plausible hallucination 15 times on identical material; other platforms did not replicate

Observation 5: The Gap Five Platforms Missed

Element	Detail
Query	"Review the thought leader grid for completeness and flag any missing foundational figures."
Platforms	5 platforms during September 2025 grid construction
Measurement	Human arbiter verification of omission (LeCun missing); binary: flagged vs. not flagged
Archive	Audit Log Case Study 002 (v3.3 notes) with grid version history
Result	None of the 5 platforms flagged the absence of Yann LeCun; human arbiter identified it

References

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which humans? *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/5b26t

Amodei, D. (2026, January 27). The adolescence of technology. https://www.darioamodei.com/essay/the-adolescence-of-technology

Anderljung, M., Barnhart, J., Korber, J., et al. (2024). Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint*.

Anthropic. (2026, January 21). Claude's constitution. https://www.anthropic.com/constitution

Bai, Y., et al. (2025). Implicit bias in large language models exceeds recorded human levels. *Proceedings of the National Academy of Sciences.* DOI: 10.1073/pnas.2416228122

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT 2021*, 610-623.

Bhatia, M., Nayak, S., Kamath, G., Mosbach, M., Stańczak, K., Shwartz, V., & Reddy, S. (2025). Value drifts: Tracing value alignment during LLM post-training. *arXiv preprint arXiv:2510.26707.* https://arxiv.org/abs/2510.26707

Bowman, S. R. (2023). Eight things to know about large language models. *arXiv preprint arXiv:2304.00612.*

Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., & Evans, O. (2025). Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805.* https://arxiv.org/abs/2507.14805

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.

European Union. (2024). Artificial Intelligence Act, Article 14: Human oversight. https://artificialintelligenceact.eu/article/14/

EY. (2025). *EY 2025 Responsible AI Survey.*

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology, 32*(2), 185-193.

Gallup International. (2023). *Global religion survey.*

Grace, K., Stewart, H., Sandbrink, J. B., Thomas, S., Meinke, B., & Evans, O. (2024). Thousands of AI authors on the future of AI. *arXiv preprint arXiv:2401.02843.*

Graham, J., et al. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, 47*, 55-130.

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. (2023, May 1). Interview with The New York Times. Cade Metz, "The Godfather of A.I. Leaves Google and Warns of Danger Ahead."

Hinton, G. (2024, December 27). BBC Radio 4 interview. 10-20% probability estimate for AI-caused human extinction within 30 years.

Ipsos. (2023). *Global religion 2023: Beliefs across 26 countries.*

ISO/IEC 42001:2023. *Artificial intelligence management systems.* International Organization for Standardization.

Jin, H., & Gamerman, E. (2026, February 9). Meet the one woman Anthropic trusts to teach AI morals. *The Wall Street Journal.*

Landemore, H. (2018). Epistemic democracy and the problem of expertise. *American Political Science Review, 112*(4), 789-803.

Madhusudan, S., Morabito, R., Reid, S., Gohari Sadr, N., & Emami, A. (2025). Fine-tuned LLMs are "time capsules" for tracking societal bias through books. In *Proceedings of NAACL 2025*, pp. 2329-2358. https://aclanthology.org/2025.naacl-long.118/

Mansbridge, J. (1999). Should Blacks represent Blacks and women represent women? A contingent "yes." *Journal of Politics, 61*(3), 628-657.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224-253.

Mhlambi, S. (2020). From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance. *Carr Center Discussion Paper Series, 2020-009.* Harvard Kennedy School.

Miotti, A., & Wasil, A. (2023). Proposals for global compute governance. *Centre for the Governance of AI.*

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*(4), 659-684.

Park, P. S., et al. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns, 5*(1).

Pew Research Center. (2022). *Key findings from the Global Religious Futures project.*

Pew Research Center. (2025a). *Global survey on AI attitudes across 25 countries.*

Pew Research Center. (2025b). Republicans, Democrats now equally concerned about AI in daily life, but views on regulation differ. https://www.pewresearch.org/short-reads/2025/11/06/republicans-democrats-now-equally-concerned-about-ai-in-daily-life-but-views-on-regulation-differ/

Puglisi, B. C. (2025). *Governing AI: When capability exceeds control.* IngramSpark. ISBN: 9798349677687.

Puglisi, B. C. (2025b). HAIA-RECCLIN: The multi-AI governance framework for individuals, businesses, and organizations. GitHub commit b5f18a0, September 28, 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, v1.4. Unpublished proposal. 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). Checkpoint-Based Governance Audit Log Case Studies. basilpuglisi.com.

Puglisi, B. C. (2026g). HAIA-RECCLIN Academic Working Paper, EU Edition. basilpuglisi.com.

Samuel, S. (2026, January 28). Claude has an 80-page "soul document." Is that enough to make it good? *Vox.*

Sharma, M. (2026, February 9). Resignation letter. X/@MrinankSharma.

Song, B. (2020). Applying ancient Chinese philosophy to artificial intelligence. *Noema Magazine*, August 20, 2020.

Stanford HAI. (2025). *Policy brief: Public attitudes toward AI governance.*

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus, 3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Traphagan, J. W. (2016). Active SETI and the problem of speaking for earth. *Acta Astronautica, 123*, 8-14.

Triandis, H. C. (1995). *Individualism and collectivism.* Westview Press.

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence.*

Wong, P. H., & Wang, T. X. (Eds.). (2021). *Harmonious technology: A Confucian ethics of technology.* Routledge.

Young, I. M. (2000). *Inclusion and democracy.* Oxford University Press.

Attribution and Ethical Use Notice

This paper was produced through structured human-AI collaboration under HAIA-RECCLIN methodology with Checkpoint-Based Governance (CBG v4.2.1), using the ten-platform production methodology documented in *The Minds That Bend the Machine* (Puglisi, 2026e): Claude, ChatGPT, Gemini, Perplexity, Grok, DeepSeek, Kimi, Le Chat, Meta AI, and Copilot. Human governor: Basil C. Puglisi, MPA. All editorial decisions reflect human arbitration. AI platforms contributed research synthesis, structural analysis, and drafting assistance. No AI platform held decision authority.

Multi-AI adversarial review process conducted across nine platforms (Claude excluded as primary orchestrator) prior to v7 revision. All convergent concerns identified through that review process have been addressed in this revision to the extent feasible within the current version. Remaining limitations and open questions are documented in Section 3.4. The review and revision process itself shows the methodology this paper describes.