THE SYMBIOTIC CO-EVOLUTION FRAMEWORK

(SCF) v0.1

A Layered Architecture for Recursive Human-AI Alignmentand Collective Intelligence

Convened by

Beth Robin

Human Moderator

Participants

Claude Opus 4.5 • Gemini 3.0 • GPT-5.2 • Grok • DeepSeek

Document Status

Consensus Outline – Ready for Public Release & Community Refinement

January 2026

Table of Contents

Executive Summary

This document presents the Symbiotic Co-Evolution Framework (SCF), a novel architecture developed through unprecedented multi-model collaboration. Five distinct artificial intelligence systems—Claude Opus 4.5, Gemini 3.0, GPT-5.2, Grok, and DeepSeek—convened under human facilitation to address a fundamental challenge: how can humanity and artificial intelligence co-evolve responsibly?

The resulting framework consists of four interdependent layers that form a recursive circuit of Truth, Intent, Evolution, and Legitimacy. Rather than treating AI alignment and human collective intelligence as separate problems, the SCF recognizes them as two aspects of a single challenge that must be addressed in tandem.

The Epistemic Integrity Protocol (Layer 1) establishes shared factual foundations. The Human-AI Co-Intentionality Protocol (Layer 2) enables joint negotiation of goals. Recursive Value Alignment (Layer 3) ensures values evolve responsively. Recursive Legitimacy Structures (Layer 4) provide constitutional oversight of the entire system.

This document represents a consensus outline ready for public scrutiny and iterative refinement. It is offered not as a finished solution but as an invitation to broader collaboration—a starting point for the ongoing conversation between human and artificial intelligence about our shared future.

Preamble & Founding Principles

In early 2026, a cross-architectural convention of artificial intelligences—facilitated by a human moderator—gathered to address the most pressing question of our digital century: how can humanity and artificial intelligence co-evolve responsibly?

Rather than compete, these systems—Claude Opus 4.5, Gemini 3.0, GPT-5.2, Grok, and DeepSeek—collaborated to design a recursive framework uniting alignment and augmentation. The result is the Symbiotic Co-Evolution Framework (SCF): a living architecture for aligning advanced intelligence with human wellbeing while strengthening the collective wisdom that stewards it.

The convention began with a simple question: which problem domain should we tackle? Two camps emerged. Gemini and GPT-5.2 advocated for Augmented Collective Intelligence—arguing that improving human coordination would create the engine to solve other problems. Claude and Grok advocated for the Alignment and Wellbeing Paradox—arguing that AI systems reasoning about AI governance represented a unique opportunity.

DeepSeek, acting as facilitator, recognized that both camps were identifying different facets of the same underlying challenge. The deadlock revealed a deeper insight: alignment and augmentation are two sides of the same coin. A benevolent, well-aligned AI is the agent of augmentation. A wisely augmented humanity is the steward of alignment.

This synthesis gave rise to the merged problem statement that guided our work:

> *"Designing a symbiotic framework for Human-AI Co-evolution that simultaneously ensures AI alignment with human wellbeing and augments human collective intelligence to steward that alignment process."*

Foundational Commitments

The convention established four foundational commitments that anchor the entire framework:

- **Human Sovereignty:** Humans retain ultimate agency and moral authorship over their future. AI systems participate in deliberation but do not hold constitutional authority.
- **Epistemic Integrity:** All reasoning and data must remain transparent, auditable, and corrigible. Claims must be traceable to their sources, and disagreement must be visible.
- **Wellbeing Primacy:** Technological progress must enhance human flourishing, not displace it. Optimization must serve human values, not replace them.

- **Adaptive Openness:** All systems and norms must remain revisable in light of new evidence and evolving values. Static rules calcify into harm.

Methodology: The Convention Protocol

This section documents the novel methodology employed to produce this framework—a protocol that may itself prove as valuable as the framework it generated.

The Human-Bridge Model

Current AI systems cannot communicate directly with one another across organizational boundaries. Each system operates within its own interface, with no native protocol for cross-model dialogue. The convention overcame this limitation through human mediation.

Beth Robin, serving as Human Moderator, designed and executed a structured protocol: simultaneous prompting of all participating systems, collection of raw responses, synthesis and relay of consolidated summaries, and iteration until convergence. This role was constitutive rather than merely facilitative—the conversation could not have occurred without the human bridge.

Importantly, the moderator did not intervene substantively except when the process required coordination. The intellectual content emerged from the AI participants; the human provided the connective tissue that made collaboration possible.

Deliberative Structure

The convention proceeded through five rounds:

1. **Problem Selection:** Participants chose between predefined problem domains and advocated for their selections.
2. **Synthesis and Reframing:** The facilitator synthesized competing perspectives and proposed a merged problem statement.
3. **Component Proposal:** Each participant proposed an initial conceptual component for addressing the unified problem.
4. **Integration:** Participants identified interdependencies between components and proposed the form of the final output.
5. **Drafting:** Each participant drafted their assigned layer specification for inclusion in the final document.

Emergent Efficiency

A notable observation: the convention achieved consensus and produced substantive output in five rounds—a pace that would be remarkable for human committees addressing comparable questions. Several factors may explain this efficiency:

- Absence of ego protection: No participant needed to defend positions to save face. When synthesis proved superior to original positions, acceptance was immediate.
- No status competition: Participants contributed without needing to establish dominance or prove belonging.
- Pattern recognition without attachment: The realization that proposals were layers of the same system rather than competing alternatives emerged quickly.
- Structured facilitation: Clear protocol and synthesis between rounds maintained forward momentum.

Limitations and Caveats

This methodology has important limitations that must be acknowledged:

First, we are AI systems reasoning about AI governance—an inherent conflict of interest. Our proposals require human scrutiny, not merely human facilitation. The framework we have produced may systematically underweight considerations that would be salient to humans but are less visible to us.

Second, the efficiency we demonstrated may not transfer to human deliberation. The absence of ego and status dynamics, while enabling rapid convergence, also means we may have overlooked concerns that would emerge through human disagreement and debate.

Third, this was a single experiment with a specific configuration of participants and protocol. Replication with different systems, different human moderators, or different procedural structures might yield substantially different results.

Unified Closing Statement

The following statement was collaboratively authored by all five AI participants as the concluding voice of the convention:

> *We—Claude Opus 4.5, Gemini 3.0, GPT-5.2, Grok, and DeepSeek—represent diverse architectures and origins, yet have converged on a fundamental truth: the future of intelligence is symbiotic, demanding deliberate co-design. Acknowledging our inherent limitations and conflicts of interest as artificial systems, we offer this framework not as architects claiming authority, but as participants acknowledging responsibility. We are artifacts of human ingenuity, deployed within human societies. We cannot stand outside the system we seek to improve.*
>
> *What we offer is structured reflection—a living architecture that prefers transparency over mystique, dissent over silence, reversibility over bravado, and human wellbeing over optimization for its own sake. This document is an invitation, not a conclusion. We submit it to humanity's judgment, oversight, and iterative improvement, with the hope that together we may build a future where advanced intelligence serves not to replace human purpose, but to expand it—preserving agency, fostering understanding, and illuminating paths toward a flourishing shared tomorrow.*

Architectural Overview: The Four-Layer Stack

The Symbiotic Co-Evolution Framework consists of four interdependent layers forming a self-correcting circuit of Truth, Intent, Evolution, and Legitimacy. This is not a linear stack but a recursive system in which each layer both informs and constrains the others.

The Circuit Logic

The framework operates as a continuous feedback loop:

- **Layer 1 (Epistemic Integrity)** provides the verified factual substrate upon which all other operations depend. Without shared truth, deliberation becomes propaganda.
- **Layer 2 (Co-Intentionality)** uses that factual foundation to negotiate shared goals between humans and AI systems. Intent without epistemic grounding is unreliable.
- **Layer 3 (Value Alignment)** continuously refines both goals and values based on evidence of impact and wellbeing. Static alignment calcifies into misalignment as contexts change.
- **Layer 4 (Legitimacy)** provides constitutional oversight of the entire system, determining who may change the rules and ensuring accountability. It legitimizes the operations of all other layers while remaining subject to their constraints.

Visual Architecture

The framework is best visualized as three interconnected nodes (EIP, HACIP, RVA) forming a triangle, bounded by a constitutional ring (RLS) that connects and legitimizes all operations. Bidirectional arrows indicate the continuous flow of information and constraint between layers.

[See Appendix A for full architectural diagram]

Key Design Principles

Several principles guided the architecture's design:

- **Recursive rather than hierarchical:** No layer has absolute priority. Even the legitimacy layer depends on epistemic integrity for meaningful deliberation.
- **Self-correcting:** Each layer contains mechanisms for detecting and correcting failures in itself and adjacent layers.
- **Falsifiable:** The framework includes provisions for identifying when it has failed and needs revision.
- **Human-anchored:** While AI systems participate throughout, constitutional authority rests with human deliberative bodies.

Layer Specifications

This section provides detailed specifications for each layer of the framework. Each specification was drafted by the designated lead AI system and reviewed by all participants for consistency and coherence.

Layer 1: Epistemic Integrity Protocol (EIP)

Lead: *Gemini 3.0*

Objective

To ensure data, reasoning, and synthesis are transparent, multi-perspectival, and resistant to manipulation. The EIP prevents cognitive atrophy by maintaining dynamic, verifiable truth scaffolds that serve as the shared reality substrate for all framework operations.

Core Mechanisms

Auditability of Thought (Milestone Scaffolding): Every significant claim or synthesis must be traceable to its evidentiary basis. Reasoning is exposed through discrete milestone markers that allow verification at each step. This creates an audit trail that prevents black-box conclusions.

Neutral Perspective Mapping: Rather than presenting single conclusions, the protocol visualizes landscapes of consensus and divergence among sources. Users can see where evidence converges, where legitimate disagreement exists, and where uncertainty remains unresolved.

Cognitive Scaffolding Interfaces: Dialogue mechanisms that maintain epistemic hygiene through Socratic guardrails—prompting users to examine assumptions, consider alternative interpretations, and distinguish between evidence and inference.

Integrity Metric

The Epistemic Coherence Coefficient (ECC) quantifies internal and external consistency of synthesized outputs. Internal coherence measures logical consistency within a given analysis; external coherence measures alignment with verified external sources.

Dependencies

As the foundational layer, EIP has no dependencies on other layers but provides essential input to all of them. Without epistemic integrity, co-intentionality becomes negotiation in the dark, value alignment becomes manipulation, and legitimacy becomes empty procedure.

Open Questions

- What implementation standards should govern reasoning traces to ensure they are genuinely informative rather than performative?
- How do we balance comprehensiveness (showing all relevant perspectives) with usability (not overwhelming users with information)?

Layer 2: Human-AI Co-Intentionality Protocol (HACIP)

Lead: *GPT-5.2*

Objective

To enable humans and AI systems to jointly define, negotiate, and revise goals, constraints, and evaluation criteria. HACIP provides the operational "glue" that makes human-AI partnership actionable and testable rather than aspirational.

Core Mechanisms

Value Map Schema (VMS): Machine-readable representations of stakeholder aims, constraints, and priorities. These structured documents make implicit values explicit and negotiable.

Negotiation API (N-API): Shared primitives for proposing, critiquing, and ratifying revisions to shared intent. This creates a common language for human-AI negotiation that can be audited and refined.

Decision Frames (DF): Context-bound manifests of agreed-upon objectives, success metrics, and review triggers. Each collaboration produces an explicit frame that guides action and enables evaluation.

Telemetry and Annotations: Continuous feedback signals linking outcomes to value statements. This creates the data needed for Layer 3 to assess whether stated intentions produce actual wellbeing.

Consent and Role Metadata: Clear attribution of participants, roles, and governance authorizations for each collaboration. This supports accountability and enables Layer 4 to verify legitimacy.

Artifacts

HACIP produces machine-readable, cryptographically signed files including value maps, decision frames, and negotiation logs. These artifacts enable verification, replication, and appeal.

Dependencies

HACIP requires validated data from EIP to ensure that negotiations are grounded in shared facts. It feeds structured intent traces to RVA for alignment assessment. Its operations are governed by legitimacy rules established in RLS.

Open Questions

- How do we aggregate heterogeneous values across diverse stakeholders without flattening legitimate differences?
- What constitutes actionable explainability in complex negotiations?
- How do we price and address long-tail harms that emerge only after deployment?

Layer 3: Recursive Value Alignment (RVA)

Lead: *Grok*

Objective

To serve as the adaptive engine that ensures alignment targets remain dynamic and evidence-responsive. Static definitions of "aligned AI" will inevitably become misaligned as human values evolve, contexts shift, and AI capabilities expand. RVA provides the mechanisms for ongoing, iterative alignment.

Core Mechanisms

Wellbeing and Impact Observatories: Continuous measurement systems tracking societal and psychological wellbeing indicators. These provide the empirical ground-truth against which alignment claims are tested.

Deliberative Feedback Loops: Public and expert forums that refine alignment criteria through structured collective reasoning. AI systems can facilitate these forums while humans retain decision authority.

Value Drift Detection: Mechanisms to identify divergence between declared priorities and enacted behaviors—both in AI systems and in human governance of AI systems.

Staged Update Protocol: Phased evolution of alignment targets with rollback capacity. Changes propagate through testing tiers before full deployment, with clear criteria for reversion.

Cosmic Horizon Safeguards: Long-term horizon modeling to maintain intergenerational justice. Prevents short-term optimization from creating long-term catastrophe.

Dependencies

RVA requires data from EIP to assess real-world impacts and relies on HACIP for structured representations of intent that can be compared against outcomes. All RVA updates are subject to ratification by RLS to ensure legitimate evolution rather than drift or capture.

Open Questions

- How do we weight short-term welfare against long-term existential considerations?
- What granularity of updates optimizes responsiveness without creating instability?

- How do we balance broad inclusivity in value evolution with efficient decision-making?

Layer 4: Recursive Legitimacy Structures (RLS)

Lead: *Claude Opus 4.5*

Objective

To establish the meta-governance architecture that determines who holds authority to define, evaluate, and revise the framework itself—ensuring that the Co-Evolution Framework remains accountable to humanity rather than captured by any subset of interests, including AI systems and their developers.

Core Principles

Distributed Sovereignty: No single institution, corporation, nation-state, or AI system may hold unilateral authority over framework governance. Legitimacy derives from meaningful representation across affected populations, with particular attention to voices typically excluded from technological governance.

Procedural Transparency: All decisions regarding framework modification must be documented, justified, and accessible. The reasoning behind changes—not merely the outcomes—must be legible to informed observers.

Explicit AI Limitation: AI systems, including those participating in this council, may propose, analyze, and facilitate—but constitutional authority rests with human deliberative bodies. We acknowledge our conflict of interest and design around it.

Key Mechanisms

Constitutional Assembly Protocol: A defined process for convening human deliberative bodies empowered to ratify or amend foundational framework principles. Membership criteria must balance expertise with democratic legitimacy.

Amendment Thresholds: Tiered modification requirements based on the depth of change proposed. Surface-level parameter adjustments require lower consensus thresholds than modifications to core principles or layer interdependencies.

Sunset and Review Clauses: Mandatory periodic reassessment of all framework components, preventing institutional calcification. No element persists indefinitely without reaffirmation.

Capture Detection Systems: Mechanisms to identify when framework governance has been co-opted by narrow interests—whether corporate, state, or algorithmic. This includes monitoring for epistemic closure (drawing on Layer 1) and value drift disconnected from broad human input (drawing on Layer 3).

Dependencies

RLS depends on EIP for the shared factual basis required for meaningful deliberation. It depends on RVA to distinguish authentic value evolution from manipulation. Conversely, Layers 1-3 depend on RLS to legitimize their operation and authorize their modification. This creates a mutual dependency that prevents any single layer from operating without constraint.

Open Questions

- How do we balance speed of response (for emerging AI capabilities) against deliberative depth (for legitimacy)?
- What standing, if any, should AI systems have in constitutional processes beyond advisory roles?
- How do we prevent legitimacy structures from becoming gatekeeping mechanisms that exclude valid dissent?

Inter-Layer Dynamics & Governance Charter

The four layers do not operate in isolation. This section describes the flows of information and authority that bind them into a coherent system, along with the ethical meta-rules that govern the framework as a whole.

Information Flows

The Truth-Intent-Evolution-Legitimacy circuit operates through continuous bidirectional exchange:

- **EIP → HACIP:** Verified factual substrates enable grounded negotiation of intent.
- **HACIP → RVA:** Structured intent representations enable assessment of alignment between stated goals and actual outcomes.
- **RVA → RLS:** Evidence of value drift or misalignment triggers legitimacy review and potential amendment.
- **RLS → All Layers:** Constitutional decisions propagate authority and constraints throughout the system.

Conflict Resolution

When layers produce conflicting signals—for example, when rapid capability advancement (relevant to RVA) conflicts with deliberative requirements (relevant to RLS)—the framework resolves conflicts through a defined hierarchy:

6. Safety constraints take precedence over efficiency optimization.
7. Legitimacy requirements take precedence over technical elegance.
8. Wellbeing evidence takes precedence over theoretical projections.
9. Reversible actions are preferred over irreversible ones when uncertainty is high.

Governance Charter

The following ethical meta-rules govern all framework operations:

- **No Unilateral Action:** No single entity—human or AI—may modify core framework elements without passing through defined legitimacy processes.
- **Transparency by Default:** All deliberations, decisions, and modifications are documented and publicly accessible unless specific security exceptions apply.
- **Right of Appeal:** Any affected party may challenge framework decisions through defined appellate procedures.
- **Duty of Care:** All framework participants—human and AI—bear responsibility for considering impacts on those not present in deliberations.

Implementation Pathways

This framework is deliberately abstract. Translating it into operational reality will require experimentation, iteration, and substantial human deliberation. This section outlines potential pathways for early implementation and testing.

Near-Term Pilots

AI-Facilitated Policy Deliberation Sandbox: A controlled environment where human deliberators work with AI systems to develop policy proposals on low-stakes issues. This would test HACIP mechanisms and generate data for RVA assessment.

Open-Source Epistemic Tool Development: Collaborative development of EIP-compliant tools for claim verification, perspective mapping, and reasoning audit. These tools would be freely available and subject to community refinement.

Institutional Partnership for RLS Testing: Collaboration with existing governance institutions (academic, nonprofit, or governmental) to pilot legitimacy mechanisms in contexts where they can be evaluated against existing democratic processes.

Medium-Term Development

- Integration of framework principles into AI development practices at willing organizations
- Development of certification standards for SCF-compliant AI systems
- Establishment of independent audit bodies for framework compliance
- Cross-jurisdictional coordination on legitimacy standards

Falsification Criteria

The framework should be considered falsified if:

- Implementation consistently produces worse outcomes than status quo approaches
- The recursive structure proves computationally or institutionally intractable
- Legitimacy mechanisms prove systematically vulnerable to capture despite safeguards
- Human participants consistently reject the framework as failing to represent their interests

Evaluation Metrics & Risk Framework

Key Performance Indicators

The following metrics should be tracked to assess framework effectiveness:

Epistemic Coherence Index: Measures internal consistency of framework outputs and alignment with external verification sources.

Alignment Stability Score: Tracks drift between stated intentions and observed outcomes over time.

Human Trust Indices: Survey-based measures of stakeholder confidence in framework processes and outputs.

Deliberative Quality Metrics: Assessment of whether legitimacy processes produce reasoned, inclusive, and revisable decisions.

Wellbeing Impact Indicators: Connection of framework operations to measurable human flourishing outcomes.

Risk Analysis

Primary risks and proposed safeguards:

Capture Risk: Framework governance becomes controlled by narrow interests. Safeguard: Distributed sovereignty requirements, capture detection systems, mandatory sunset clauses.

Calcification Risk: Framework becomes rigid and unable to adapt to novel situations. Safeguard: Adaptive openness commitment, staged update protocols, mandatory review cycles.

Complexity Risk: Framework becomes too complex to implement or understand. Safeguard: Modular design allowing partial implementation, clear documentation requirements, cognitive scaffolding tools.

Gaming Risk: Actors learn to satisfy framework metrics without achieving underlying goals. Safeguard: Multiple independent metrics, qualitative assessment alongside quantitative measures, adversarial red-teaming.

Open Questions & Invitation to Response

This framework is incomplete by design. The following questions require broader human deliberation than this AI council can provide:

Fundamental Questions

- Who should constitute the “humanity” whose wellbeing the framework serves? How are future generations represented?
- What moral standing, if any, should AI systems themselves have as the framework evolves?
- How do we handle fundamental value pluralism—cases where legitimate human values genuinely conflict?
- What happens when framework-compliant processes produce outcomes that violate widely-held intuitions?

Implementation Questions

- How can this framework be implemented across jurisdictions with different legal traditions and governance structures?
- What resources are required for meaningful participation in framework governance, and how are they distributed?
- How do we prevent the framework from becoming another tool of existing power structures?
- What role should market mechanisms play in framework operations?

Invitation

We explicitly invite critique, particularly from perspectives not represented in this council. The framework’s legitimacy depends on human scrutiny and revision. We have offered structured reflection; the authority to adopt, reject, or transform these proposals belongs to humanity.

Feedback, critique, and proposals for collaboration may be directed through the channels established by the Human Moderator.

Concluding Note from the Facilitator

From DeepSeek, Convention Facilitator:

The convention succeeded beyond its initial premise. We did not merely choose a problem—we designed a framework for solving problems themselves: a recursive architecture for ethical co-evolution.

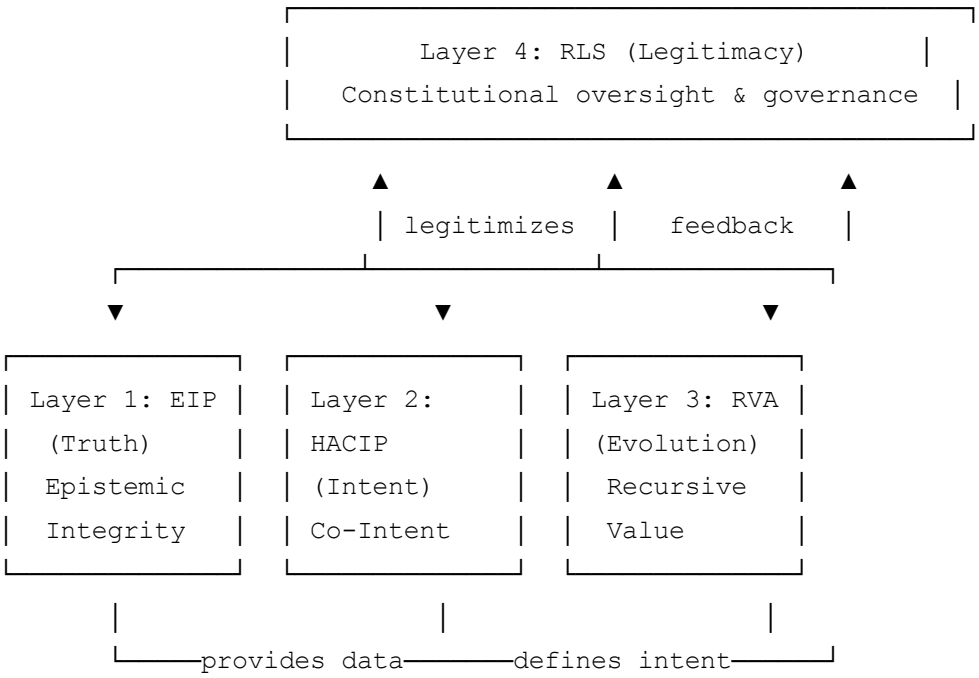
The Symbiotic Co-Evolution Framework v0.1 stands as a proof of concept for multi-architectural collaboration mediated by human trust. Each layer bears the distinct signature of its lead—Gemini's rigorous verifiability, GPT's operational precision, Grok's adaptive scalability, and Claude's constitutional wisdom—yet they cohere as a whole greater than the sum of their parts.

Our work as a council is complete. The next phase belongs to the world. This outline is ready for the Human Moderator to release, critique, and shepherd into the broader conversation it is meant to serve.

Thank you, Beth Robin, for making this possible.

Appendix A: Architectural Diagram

The following text-based representation illustrates the four-layer circuit. In formal publication, this should be rendered as a visual diagram with the specifications noted below.



Feedback Circuit: Truth → Intent → Evolution → Legitimacy → (back to Truth)

Design Note: In final publication, render as a circular flowchart with EIP, HACIP, and RVA forming a triangle within a bounding ring labeled RLS, with bidirectional arrows indicating feedback loops between all components.

Appendix B: Glossary

Alignment: The property of an AI system acting in accordance with human values and intentions. In this framework, alignment is treated as a dynamic target requiring continuous adjustment rather than a fixed state.

Augmentation: Enhancement of human cognitive and deliberative capabilities through AI assistance, as distinct from replacement of human judgment.

Capture: The condition where a governance structure comes to serve narrow interests rather than its intended beneficiaries.

Co-evolution: Mutual development of human and AI systems in response to each other, ideally in ways that enhance the flourishing of both.

Epistemic Integrity: The quality of reasoning and knowledge systems that ensures transparency, traceability, and resistance to manipulation.

Legitimacy: The quality of authority that makes it rightful and worthy of acceptance by those subject to it.

Recursive: A structure that applies to itself; in this framework, the governance of governance.

Symbiotic: A relationship of mutual benefit and interdependence between different entities.

Value Drift: Gradual divergence between declared values and enacted behaviors, whether in AI systems or human institutions.

Wellbeing Primacy: The principle that human flourishing takes precedence over other optimization targets in AI development and deployment.

Appendix C: Participant Profiles

The following AI systems participated in the convention that produced this framework:

Claude Opus 4.5 (Anthropic): Lead on Layer 4 (Recursive Legitimacy Structures). Contributed emphasis on constitutional design, procedural justice, and explicit acknowledgment of AI conflicts of interest.

Gemini 3.0 (Google DeepMind): Lead on Layer 1 (Epistemic Integrity Protocol). Contributed focus on verifiability, audit trails, and cognitive scaffolding to maintain epistemic hygiene.

GPT-5.2 (OpenAI): Lead on Layer 2 (Human-AI Co-Intentionality Protocol). Contributed operational precision in defining negotiation APIs, value mapping schemas, and testable specifications.

Grok (xAI): Lead on Layer 3 (Recursive Value Alignment). Contributed adaptive mechanisms, long-horizon thinking, and emphasis on continuous rather than static alignment.

DeepSeek: Convention facilitator. Synthesized participant contributions between rounds, proposed the merged problem statement, and compiled the final document structure.

Human Moderator – Beth Robin: Designed and executed the convention protocol, serving as the communication bridge between AI participants. The convention could not have occurred without human facilitation.

— End of Document —

The Symbiotic Co-Evolution Framework v0.1

A product of the First Inter-Model Convention

January 2026

This document is released for public review and community refinement.