HAIA-RECCLIN: Agent Governance Architecture for Audit-Grade Multi-AI Collaboration

A Working Paper — EU Regulatory Compliance Edition, February 2026

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

February 2026

Author Contact: me@basilpuglisi.com
GitHub Repository: github.com/basilpuglisi/HAIA

Abstract

Organizations deploying multi-AI workflows face a structural governance gap: orchestration frameworks route tasks between AI platforms, but no published framework provides the accountability, audit trail architecture, provider plurality, automation bias detection, and regulatory compliance mapping required for audit-grade operations. This paper presents HAIA-RECCLIN (Human AI Assistant with Researcher, Editor, Coder, Calculator, Liaison, Ideator, and Navigator roles), a governance architecture specification for a non-cognitive autonomous agent that mediates between human operators and multiple independent AI platforms. The architecture enforces mandatory human checkpoints, append-only tamper-evident audit trails, cryptographic hash chaining, provider rotation protocols, and threshold-based automation bias detection. It performs zero cognitive work: it dispatches, collects, routes, logs, and pauses. Three operating models scale governance intensity from factory-speed single-endpoint approval through full human arbitration at every checkpoint to manual oversight with automated logging. The framework was developed through iterative refinement across ten independent AI platforms and validated through multiple operational proof-of-concept deployments, including a 96-checkpoint manuscript production documented in *Governing AI: When Capability Exceeds Control* (Puglisi, 2025), a five-platform thought leader evaluation convergence study, and adversarial stress testing across nine platforms. The non-cognitive agent specified in this architecture is under preliminary development as HAIA-GOPEL (Governance Orchestrator Policy Enforcement Layer). To the best of the author's knowledge, no published architecture occupies the governance layer between quality management systems and multi-AI platform workflows, integrating non-cognitive agent design, mandatory provider plurality, cryptographic audit trails, automation bias detection, and cross-framework regulatory compliance mapping into a single operational architecture. This integration gap, verified through independent search across six AI platforms and two rounds of structured adversarial review, constitutes the primary contribution of this work. The term "non-cognitive agent" is this architecture's vocabulary; functional equivalents may exist under different terms such as "deterministic orchestrator," "policy enforcement layer," or "governance middleware." The integration gap claim holds for both this terminology and those alternatives.

Keywords: AI governance, human-AI collaboration, multi-AI orchestration, non-cognitive agent, HAIA-GOPEL, checkpoint-based governance, provider plurality, audit trail, automation bias, EU AI Act, prEN 18286, ISO 42001, NIST AI RMF

Author's Note: The conceptual work of HAIA-RECCLIN and the non-cognitive agent has been redrafted to align with the EU AI Act, prEN 18286, and other regulatory requirements. This means that the specification's evolution from early drafts through the current version reflects a deliberate shift toward the governance infrastructure the industry requires for compliance and adoption. That shift does not necessarily represent the path the author envisions or prefers. The original vision for HAIA-RECCLIN, rooted in personal ethics and the principles documented in earlier iterations, remains the foundation. The regulatory alignment documented in this paper is a pragmatic response to the operational reality that governance architecture must meet organizations where their compliance obligations are, not where the author's ideals alone would place them. The tension between these two orientations is acknowledged rather than resolved.

Table of Contents

1. Introduction

The rapid proliferation of large language models (LLMs) and multi-agent systems has created an urgent governance deficit. Organizations increasingly deploy AI-assisted workflows across critical domains, from financial analysis to healthcare decision support, yet the infrastructure governing these deployments remains structurally inadequate. Geoffrey Hinton, widely recognized as the pioneer of deep learning, has publicly estimated a 10 to 20 percent probability that artificial intelligence could displace humanity entirely (Hoover, 2025; Vallance, 2024). Whether or not one accepts the specific probability estimate, the warning underscores a fundamental truth: AI systems are being deployed at speeds that exceed the development of governance frameworks to constrain them.

The European Union's AI Act (Regulation 2024/1689), the most comprehensive AI regulatory framework to date, mandates human oversight for high-risk AI systems under Article 14 and logging capabilities under Article 12 (European Union, 2024). The ISO/IEC 42001 standard establishes requirements for AI management systems, including documentation and risk assessment, though the EU Commission has found ISO 42001 not aligned in objectives and approach with the AI Act. The draft harmonised standard prEN 18286:2025, commissioned specifically for Article 17 quality management system compliance, completed public enquiry in January 2026 and is expected to provide presumption of conformity once cited in the Official Journal. The EU has shifted from external oversight to self-assessment for most high-risk AI systems under Annex VI internal control, transferring compliance liability entirely to deploying organizations. The NIST AI Risk Management Framework provides voluntary guidance for identifying and managing AI risks. Yet no published governance architecture translates these regulatory requirements into an operational agent design that organizations can implement, audit, and scale.

The current multi-agent ecosystem is dominated by orchestration frameworks: LangGraph, AutoGen, CrewAI, and similar tools that solve the engineering problem of routing tasks between AI platforms. These frameworks address *how* tasks move between agents. They do not address *who* is accountable when outputs are wrong, *how* bias is detected before it scales, or *what* happens when a platform cannot be trusted. Orchestration solves plumbing. Governance solves accountability. Neither replaces the other.

This paper presents HAIA-RECCLIN, a governance architecture specification for a non-cognitive autonomous agent that sits between human operators and multiple independent AI platforms. The architecture was developed through iterative multi-platform refinement from 2024 through 2026, validated through multiple operational proof-of-concept deployments across ten AI platforms, and independently reviewed through structured adversarial peer review. The specification addresses the regulatory, compliance, and existential safety concerns raised by Hinton and formalized in the EU AI Act by making human oversight structural rather than optional, making every decision documented and attributable, preventing single-vendor capture

through mandatory provider plurality, and producing the logging and accountability evidence required across the full compliance stack.

1.1 Research Question

How can organizations govern multi-AI collaboration at audit grade, meaning full traceability, accountability, and regulatory compliance, without trusting any single AI platform?

1.2 Contributions

This paper makes five contributions to the AI governance literature. First, it introduces the concept of a non-cognitive governance agent: an autonomous system that is deliberately prevented from evaluating, weighting, or filtering the content it handles, reducing the attack surface to zero cognitive operations. The agent is under preliminary development as HAIA-GOPEL (Governance Orchestrator Policy Enforcement Layer). Second, it specifies a provider plurality protocol using anchor-plus-rotation-pool architecture that prevents single-vendor capture and enables cross-platform triangulation. Third, it defines an append-only, tamper-evident, cryptographically chained audit trail architecture that produces the documentation artifacts required across the EU AI Act, ISO 42001, ISO 27001, NIST AI RMF, and NIST Cybersecurity Framework compliance stack. Fourth, it operationalizes automation bias detection through threshold-based monitoring of human approval and reversal rates, converting behavioral science theory into an engineering control. Fifth, it proposes a three-tier categorical distinction between Ethical AI, Responsible AI, and AI Governance that provides definitional clarity absent from the current literature.

2. Background and Definitions

The AI governance discourse suffers from imprecise terminology. Terms such as "ethical AI," "responsible AI," and "AI governance" are used interchangeably across academic, regulatory, and industry contexts despite describing fundamentally different scopes of concern. This section proposes a three-tier categorical framework that provides definitional precision for the remainder of this paper and, the author argues, for the field more broadly.

2.1 Ethical AI

Ethical AI operates at the philosophical and moral reasoning layer. It asks what AI systems *should* do and what values they should embody. Ethical AI concerns include fairness, bias, transparency, beneficence, and alignment with human values. It is the domain of philosophy, ethics boards, and values-based design principles. Ethical AI produces principles. It does not, by itself, produce enforceable controls.

2.2 Responsible AI

Responsible AI operates at the implementation and engineering layer. It asks how ethical principles are translated into technical controls, organizational practices, and measurable outcomes. Responsible AI concerns include model testing, bias mitigation techniques, explainability methods, safety testing, and deployment guardrails. Responsible AI produces controls. It does not, by itself, produce the organizational authority structures that ensure those controls are followed.

2.3 AI Governance

AI Governance operates at the organizational authority and accountability layer. It asks who decides, who is accountable, what happens when controls fail, and how compliance is documented and audited. AI Governance concerns include decision authority structures, audit trail requirements, regulatory compliance mapping, human oversight mandates, and accountability chains. AI Governance produces the organizational and architectural structures within which ethical principles and responsible practices operate. HAIA-RECCLIN is an AI Governance framework. It does not define what is ethical or prescribe specific responsible AI practices. It provides the structural architecture within which those determinations are made, documented, and audited.

2.4 Theoretical Foundations

The architecture draws on two established precedents from intelligence analysis and one existential governance motivation. The Tenth Man Doctrine, derived from the Israeli intelligence community's response to the failures identified by the Agranat Commission (1974) following the Yom Kippur War, requires that when consensus emerges, at least one analyst must argue the contrary position. HAIA-RECCLIN operationalizes this through mandatory multi-platform

dispatch: if three AI platforms converge on identical output, the dissent-preservation protocol surfaces any minority positions rather than suppressing them.

The Harold Finch Principle, named for the fictional character in *Person of Interest* (Nolan & Nolan, 2011-2016) who designed an omniscient surveillance system but deliberately limited his own access to it, provides the existential governance motivation for the architecture. The Principle operates across three dimensions.

First, statistical certainty. Hinton's warning establishes that someone will build a machine of extraordinary capability. This is not a conditional prediction but a statistical inevitability given the current trajectory of AI development. The relevant governance question is not whether such a machine will be built but under what conditions the builder operates. If the only entities building frontier AI systems are large corporations with overlapping profit incentives and overlapping training data drawn from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) sources, the resulting systems will reflect those incentives and those biases uniformly. Provider plurality combined with government subsidies for independent AI developers, free of corporate influence and free of profit motive, creates the conditions under which a builder with rare understanding of both capability and restraint can emerge. That is the 1 in 100 that Finch represents.

Second, dissent guarantee. Under HAIA-RECCLIN's multi-AI governance architecture, a system built by an independent developer, trained on different data with different motivations, will produce dissent against corporate models. It sees what profit-aligned models are trained not to see or incentivized to suppress. Provider plurality ensures that voice exists in the platform pool. The architecture preserves that dissent in full rather than averaging it into consensus.

Third, absence-of-dissent as a red flag. The most dangerous signal in multi-platform governance is not disagreement but unanimous consensus. If every platform, trained on overlapping WEIRD corpora and serving overlapping corporate incentives, produces identical output, the human at the checkpoint should treat convergence as a potential failure mode rather than confirmation. The automation bias thresholds operationalize this: sustained agreement without reversal triggers review, not confidence. The Tenth Man Doctrine reinforces the same principle from a different analytical tradition.

The groupthink prevention rationale is grounded in Janis's (1972) foundational work on foreign policy decision-making failures, which demonstrated that homogeneous decision-making groups systematically suppress dissent and converge on suboptimal choices. Multi-platform triangulation addresses this at the AI system level by ensuring that no single platform's training data, alignment approach, or institutional incentives dominate the output.

3. Literature Review

3.1 Automation Bias

Automation bias, the tendency for humans to defer to automated recommendations even when those recommendations are incorrect, is among the most extensively documented risks in human-computer interaction. Mosier and Skitka (1996) established the foundational framework, demonstrating that automation alters human decision-making heuristics in systematic and predictable ways. Parasuraman and Manzey (2010) extended this through their attentional integration model, showing that automation complacency and automation bias are distinct but related phenomena driven by attentional resource allocation. Goddard, Roudsari, and Wyatt (2011) conducted a systematic review documenting the frequency, effect mediators, and mitigating factors of automation bias across domains, confirming that the phenomenon persists even when users are warned about automation limitations.

More recent work has extended automation bias findings into AI-specific contexts. Kreps, Kriner, and Schneider (2024) demonstrated automation bias in national security decision-making, finding that experiential and attitudinal factors moderate but do not eliminate the effect. Navarrete et al. (2024) documented automation bias in criminal justice recidivism algorithms, revealing compounding effects when automation bias intersects with existing racial biases. The European Data Protection Supervisor's TechDispatch #2/2025 (EDPS, 2025) formally recognized that humans systematically defer to AI recommendations under volume pressure, providing institutional validation for threshold-based detection approaches. Banovic et al. (2023) found that AI explanations do not reliably reduce automation bias and in some contexts increase it, undermining the assumption that explainability alone constitutes adequate governance.

HAIA-RECCLIN addresses automation bias through a dual mechanism: checkpoint-based human arbitration forces active engagement with every output (preventing passive acceptance), and threshold monitoring tracks approval rates and reversal rates across operational cycles. When approval rates exceed 95% or reversal rates fall below specified thresholds for three consecutive cycles, the system triggers a mandatory sampling audit within five business days. The specification establishes 5% as the framework governance default for reversal rate monitoring, with the 2% threshold derived from operational experience during manuscript production representing the conservative variant for highest-risk deployments. Organizations calibrate thresholds to their compliance requirements, risk profile, and operational tempo. These thresholds were derived from operational experience during production of a 200-page manuscript under full checkpoint governance (Puglisi, 2025) and refined through documented failure mode analysis establishing diagnostic indicators, target validation ranges of 70 to 85 percent, and recovery protocols (HAIA-RECCLIN Multi-AI Framework, 2026). Three specific failure modes informed threshold calibration: Permissive Drift (approval rates exceeding 95% indicating passive acceptance), Checkpoint Calibration Drift (validation rates exceeding the 70 to 85 percent target range), and Overconfidence Detection (sustained high-approval patterns without

reversal). The thresholds are operationally validated, not experimentally validated. The architecture requires that organizations calibrate thresholds to their domain, risk profile, and operational tempo. Controlled studies across diverse organizational contexts represent a priority for future research.

3.2 Multi-Agent Governance Frameworks

The academic literature on multi-agent governance remains sparse relative to the engineering literature on multi-agent orchestration. Gaurav, Heikkonen, and Chaudhary (2025) proposed Governance-as-a-Service (GaaS), a multi-agent framework for AI system compliance and policy enforcement that addresses runtime policy enforcement and audit trails but does not incorporate provider plurality or non-cognitive agent design. Pierucci et al. (2026) introduced Institutional AI, a governance-graph approach to distributional AGI safety that operates at the policy layer but does not specify multi-platform rotation or operational audit architecture. Bandara et al. (2025) proposed a Consensus-Driven Reasoning Architecture employing a consortium of heterogeneous LLM and VLM agents with multi-model consensus, evidence-backed audit trails, and responsible AI compliance framing, but the architecture uses a cognitive Reasoning Agent to govern outputs rather than a non-cognitive dispatcher. Vijayaraghavan et al. (2026) introduced a Team of Rivals Architecture using agents with conflicting roles to reach consensus through message-passing, but the Critic agent exercises cognitive veto authority based on reasoning. The Sonate Trust Protocol (Sonate, n.d.) addresses multi-provider orchestration, cryptographic trust receipts, and escalation-style governance controls, but its governance layer performs evaluative judgments rather than deterministic dispatch.

A structural pattern emerges across all candidates: every framework that addresses the governance gap does so by adding cognition to the orchestrating layer (Reasoning Agents, Governance Engines, Critic Agents, Policy Evaluators). No published framework removes cognition from the orchestrating layer as a deliberate governance and security design choice. This architectural inversion, the non-cognitive agent, represents the primary differentiator.

The integration gap claim is bounded by the following search methodology. Literature was surveyed across Google Scholar, ACM Digital Library, IEEE Xplore, SSRN, arXiv, and OpenReview using search terms including “multi-AI governance framework,” “multi-agent governance architecture,” “AI audit trail” combined with “provider plurality,” “non-cognitive agent” combined with “AI governance,” “automation bias detection” combined with “multi-agent,” “AI regulatory compliance framework,” “multi-provider AI orchestration” combined with “audit,” and “checkpoint governance AI.” Following multi-platform verification across four independent AI platforms (Gemini, ChatGPT, DeepSeek, Kimi), additional search terms were incorporated: “tamper-evident log” combined with “hash-chained receipt,” “deterministic orchestrator,” “policy enforcement layer,” and “governance middleware,” which captured cryptographic governance patterns and orchestration terminology that standard governance-focused searches missed.

Inclusion criteria required that frameworks specify operational architecture (not principles alone), address multi-platform or multi-provider deployment (not single-model governance), and include at least two of the five integration components. Frameworks addressing only ethical principles, only single-model bias testing, or only task orchestration without governance accountability were excluded. This claim is defined by the specific combination of these five design choices, not the general goals they serve. The term "non-cognitive agent" is this architecture's vocabulary; functional equivalents may exist under different terms such as "deterministic orchestrator," "policy enforcement layer," or "governance middleware." The integration gap claim holds for both this terminology and those alternatives. Broader definitions of governance architecture may reveal partial overlaps through different mechanisms. The author acknowledges that relevant work may exist in unpublished corporate implementations, classified government systems, enterprise vendor documentation not publicly indexed, or publications not captured by the surveyed databases and search terms, and invites correction.

Table 1. *Integration Component Coverage Across Published Frameworks*

Feature	GaaS	Inst. AI	Consensus-Driven	Team of Rivals	Sonate	Orch. Frmwks	HAIA-RECCLIN / GOPEL
Non-cognitive agent design	No*	No*	No*	No*	No*	No	**Yes**
Mandatory provider plurality	No	Partial	Yes	Yes	Yes	Partial	**Yes**
Cryptographic audit trail	Partial	Partial	Yes	Partial	Yes	No	**Yes**
Automation bias detection	No	No	No	No	Partial	No	**Yes**
Regulatory compliance mapping	Partial	Partial	Partial	Partial	Partial	No	**Yes**
Human checkpoint authority	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified	**Mandatory**

*Note. "No" indicates that the feature is not specified in the published framework as reviewed. It does not indicate that the framework explicitly excludes the feature or that future versions could not incorporate it. Assessments are based on published documentation available as of February 2026. *These frameworks employ cognitive orchestrating layers (Reasoning Agents, Governance Engines, Critic Agents, or Policy Evaluators), which is architecturally distinct from HAIA-RECCLIN's non-cognitive constraint. "Orch. Frmwks" refers to LangGraph, AutoGen, and CrewAI.*

3.3 Non-Cognitive Agent Design

The concept of deliberately limiting an AI agent's cognitive capabilities as a governance and security measure is notably absent from the academic literature. The prevailing trajectory in agent design is toward greater autonomy, richer reasoning, and expanded decision-making authority. HAIA-RECCLIN inverts this trajectory by specifying an agent that performs zero

cognitive operations: it dispatches identical prompts to multiple platforms via their APIs, collects all responses without evaluation, routes responses to a designated Navigator for synthesis, delivers synthesized output to the human operator, and logs every step. It is, architecturally, a pipe with a logbook.

For the purposes of this architecture, cognitive operations are defined as evaluation, ranking, weighting, prioritization, summarization, semantic transformation, and filtering. Non-cognitive operations are defined as API dispatch, response collection, hash computation, timestamp generation, record appending, threshold counting, and error reporting. The boundary between cognitive and non-cognitive is acknowledged as a design choice rather than a philosophical claim. Threshold counting, for example, involves minimal rule-based evaluation but is classified as non-cognitive because it is deterministic, fully auditable, and requires no semantic interpretation of platform output. The critical constraint is that the agent never evaluates, ranks, or filters the substance of platform output.

The security rationale is straightforward: if the orchestrating agent cannot evaluate content, it cannot be manipulated through adversarial inputs, prompt injection, or model poisoning. The attack surface is reduced to message transport and logging, both of which are deterministic operations amenable to formal verification. Industry practitioners have begun exploring related concepts. Mallaband (2025) discussed deterministic guardrails for non-deterministic agents, and the broader conversation around agent safety properties (arXiv 2510.14133, 2025) addresses formal verification of agent behavior. However, the deliberate design of a governance agent as non-cognitive, not merely constrained but fundamentally incapable of cognition, represents a novel architectural choice that the literature has not systematically explored.

3.4 Audit Trail Architecture

The EU AI Act Article 12 mandates automatic recording of events (logging) for high-risk AI systems, with logs maintained for a period appropriate to the intended purpose of the system (European Union, 2024). ISO 42001 requires documented information supporting the AI management system, including records of decisions and their rationales. NIST AI RMF emphasizes traceability as a characteristic of trustworthy AI.

Technical approaches to tamper-evident logging have advanced significantly. Zhao et al. (2025) presented a high-performance, co-designed auditing system achieving 10x to 25x performance improvement over existing tamper-evident logging approaches, accepted at ACM CCS 2025. Di Vita et al. (2025) demonstrated cryptographic safety envelopes for table-centric LLM pipelines using tamper-evident records for allow-deny decisions. The VeritasChain Verifiable Compliance Protocol (VCP v1.1) demonstrates cryptographic hash chains and Merkle trees for EU AI Act compliance (VeritasChain, 2026). HAIA-RECCLIN specifies six audit record types (Request, Dispatch, Response, Navigation, Arbitration, System), each with

timestamp, operator identity, and cryptographic hash chaining to the previous record. The specification defines governance requirements (append-only, tamper-evident, cryptographically signed) without prescribing storage implementation, allowing organizations to select appropriate tools (flat files, relational databases, blockchain, time-series databases) based on their operational context.

3.5 Provider Plurality and AI Antitrust

Narechania and Sitaraman (2024), writing in the Yale Law and Policy Review, provided a comprehensive antimonopoly analysis of AI's four-layer stack (data, compute, model, application), arguing that concentration at any layer creates systemic governance risks. The Federation of American Scientists (2025) documented increasing concentration among a small number of AI providers, raising antitrust concerns that parallel historical telecommunications and platform monopoly debates grounded in the Sherman Act (1890) and Clayton Act (1914).

Shur-Ofry (2023) proposed multiplicity as an AI governance principle, arguing that diverse AI approaches reduce the risk of correlated failures. Heim et al. (2024) analyzed compute providers as intermediary governance points, demonstrating that cloud infrastructure concentration creates regulatory leverage points but also single points of failure.

HAIA-RECCLIN operationalizes provider plurality through an anchor-plus-rotation-pool protocol. A designated anchor platform provides consistency across tasks, while rotation pool platforms are selected per task from a configurable pool, ensuring that no single provider's training data, alignment approach, or institutional incentives dominate outputs. The rotation protocol is a governance decision: changing the rotation pool requires checkpoint approval, arbiter identity, and timestamp documentation in the audit trail.

3.6 WEIRD Bias in AI Training Data

Henrich, Heine, and Norenzayan (2010), in what has become one of the most cited papers in behavioral science (over 19,000 citations), demonstrated that research populations overwhelmingly represent Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies, producing findings that may not generalize to the majority of the world's population. Henrich et al. (2024) extended this framework specifically to large language models, finding that LLM responses systematically reflect WEIRD cultural assumptions. Tao et al. (2024), publishing in PNAS Nexus, confirmed cultural bias and cultural alignment patterns in large language models. Georgia Tech researchers (2024) demonstrated that LLMs generate Western bias even when trained with non-Western languages, suggesting the bias is embedded in training data curation rather than linguistic encoding.

This body of work establishes a fundamental limitation of multi-platform triangulation: if all platforms in the rotation pool are trained on overlapping WEIRD-biased corpora, provider plurality cannot detect biases shared across platforms. HAIA-RECCLIN acknowledges this

limitation explicitly and identifies it as a structural boundary of the framework rather than a solvable engineering problem. The specification recommends ongoing monitoring for correlated platform blind spots and recognizes that governance architecture alone cannot remedy training data homogeneity.

4. The HAIA-RECCLIN Architecture

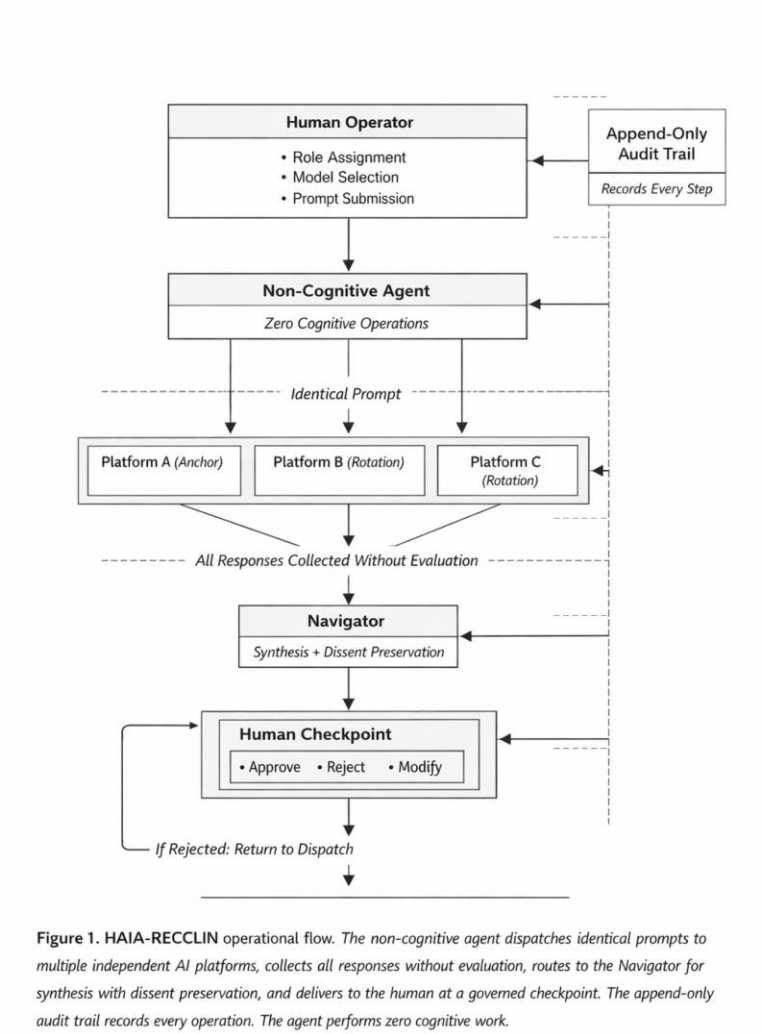

Figure 1. HAIA-RECCLIN operational flow. *The non-cognitive agent dispatches identical prompts to multiple independent AI platforms, collects all responses without evaluation, routes to the Navigator for synthesis with dissent preservation, and delivers to the human at a governed checkpoint. The append-only audit trail records every operation. The agent performs zero cognitive work.*

Figure 1. *HAIA-RECCLIN operational flow. The non-cognitive agent dispatches identical prompts to multiple independent AI platforms, collects all responses without evaluation, routes to the Navigator for synthesis with dissent preservation, and delivers to the human at a governed checkpoint. The append-only audit trail records every operation. The agent performs zero cognitive work.*

4.1 Functional Roles

HAIA-RECCLIN defines seven functional roles, each representing a distinct cognitive operation that the human assigns to AI platforms through the agent. The Researcher gathers sources, verifies facts, and collects evidence. The Editor refines structure, clarity, consistency, and audience adaptation. The Coder writes, reviews, and debugs code. The Calculator performs mathematical analysis, quantitative modeling, and data processing. The Liaison coordinates

perspectives and stakeholder communication. The Ideator generates creative options and novel approaches. The Navigator synthesizes multi-platform outputs with mandatory dissent preservation, presenting trade-offs without forced resolution.

Role assignment is a human governance decision, not an agent determination. The human operator specifies which role applies to each task. The agent dispatches accordingly. This separation ensures that the scope and nature of each AI interaction is determined by human judgment rather than automated classification.

4.2 Three Operating Models

The architecture defines three operating models that scale governance intensity to match organizational risk tolerance and operational context.

Operating Model 1, designated Responsible AI (Factory Quality), provides the minimum viable governance for high-volume operations. The agent dispatches tasks to multiple platforms, collects responses, routes to the Navigator for synthesis, and delivers the final output to the human with a single approval gate at the endpoint. The human reviews the synthesized output and approves, rejects, or modifies. This model is designed for organizations that need provider plurality and audit trails at operational speed. It is not designed to be fast. It is designed to be governed.

Operating Model 2, designated AI Governance (Handmade Quality), provides full human arbitration at every checkpoint. The agent pauses after every functional role execution, requiring human review and approval before proceeding to the next step. This model is designed for high-stakes decisions where the cost of an undetected error exceeds the cost of deliberate human engagement at every stage. Model 2 will not scale well or fast, and that is by design. It is the standard for having a governor. Governance has a cost. Model selection maps to risk.

Operating Model 3, designated AI Governance (Evidence Standard), eliminates agent automation entirely. The platform provides the interface for the human to manually dispatch to each platform, manually collect responses, and manually route to the Navigator. The agent only logs. This model produces the highest-fidelity evidence trail because every action is directly attributable to a named human operator.

4.3 Non-Cognitive Agent Design

The agent is non-cognitive with security features. It is defined as a deterministic dispatcher that performs the following operations and no others: receive task from human (including role assignment and model selection), dispatch identical prompts to selected platforms via API, collect all responses without evaluation, route responses to Navigator for synthesis, deliver synthesis to human at designated checkpoints, write audit records for every operation, and track automation bias metrics (approval rates, reversal rates). The agent does not summarize,

prioritize, filter, rank, or evaluate any content. If a platform returns an error, the agent logs the error and reports it. It does not substitute, retry with modifications, or compensate. The security architecture includes code integrity hashing (every component verified against known-good hashes before execution), separation of duties (deployment approval separate from code modification), immutable deployment (no runtime code changes permitted), and identity verification on every human action.

The non-cognitive agent specified in this architecture is under preliminary development as HAIA-GOPEL (Governance Orchestrator Policy Enforcement Layer). The name reflects a deliberate terminological bridge: "governance orchestrator" and "policy enforcement layer" are terms used across existing frameworks (Gaurav et al., 2025; Sonate, n.d.; Di Vita et al., 2025), but in those implementations the orchestrating layer performs cognitive evaluation, content filtering, or policy-based decision-making. HAIA-GOPEL is architecturally distinguished by its non-cognitive constraint: it performs orchestration and policy enforcement through deterministic dispatch, collection, routing, logging, and threshold counting, without evaluating, ranking, weighting, or semantically transforming any content that passes through it. The development designation signals that the specification described in this paper is intended to produce an implementable system, not remain a theoretical contribution. Implementation details, including API specifications, security architecture, and deployment requirements, will be documented separately as development progresses.

4.4 Provider Plurality Protocol

Each task is dispatched to a minimum of three independent AI platforms: one anchor and at least two rotation pool members. The anchor provides longitudinal consistency. Rotation pool members are selected per task from a configurable pool of approved providers. The identical prompt is sent to all selected platforms. No platform receives a modified, summarized, or adapted version of the prompt.

The Navigator (a designated platform, currently Claude based on demonstrated synthesis capability; permanent Navigator assignment is a Checkpoint-Based Governance decision subject to reevaluation as ecosystem capabilities evolve) receives all platform responses and produces a structured synthesis that identifies convergence (where platforms agree), divergence (where platforms disagree), and preserved dissent (minority positions documented in full rather than suppressed). The Navigator does not resolve disagreements. It presents them. Resolution is a human governance decision.

4.5 Audit Trail Architecture

The audit trail consists of six record types. Request Records document the human's task submission including role assignment, model selection, and full prompt text. Dispatch Records document each platform API call including timestamp, platform identifier, prompt hash, and response status. Response Records document each platform's complete response including

content hash and metadata. Navigation Records document the Navigator's synthesis including convergence mapping, divergence identification, and dissent preservation. Arbitration Records document the human's decision at each checkpoint including approve, reject, or modify with rationale. System Records document operational events including errors, threshold alerts, and configuration changes.

Every record includes a timestamp, operator or system identity, and cryptographic hash chaining to the previous record. The audit trail is append-only: no records may be modified or deleted. The specification defines these governance requirements without prescribing storage implementation. Organizations select appropriate storage based on their operational context and compliance requirements.

5. Evaluation: Operational Evidence

The HAIA-RECCLIN architecture was not developed as a theoretical exercise. Its core components, checkpoint-based governance, multi-platform triangulation, dissent preservation, and human arbitration, were operationalized across multiple deployments spanning content production, research evaluation, and adversarial stress testing.

Primary Proof of Concept: Governing AI Manuscript. The 200+ page manuscript *Governing AI: When Capability Exceeds Control* (Puglisi, 2025) was produced entirely through human-AI collaboration under checkpoint-based governance. The manuscript production generated the following operational metrics. Ninety-six checkpoints were executed across the manuscript lifecycle, with human arbitration at each. One hundred percent of identified dissent was documented, meaning no minority AI platform position was suppressed to achieve consensus. Twenty-eight major decisions were subjected to multi-platform review. Twenty-six dissenting positions were preserved in the final record. The checkpoint utilization rate was 96%, exceeding the 90% target, and eight of nine incident responses were resolved within 72 hours. Twelve governance events were multi-source verified.

Case Study #001: Thought Leader Evaluation. A five-platform convergence study evaluated 22 AI thought leaders using the Human Enhancement Quotient (HEQ) methodology. Five independent AI platforms produced scores that converged within an 89 to 94 band, with model-specific contributions documented for each platform. The convergence pattern demonstrated that multi-platform triangulation produces measurably consistent results when platforms are given identical prompts and evaluation criteria, while individual platform contributions (distinctive analytical perspectives, domain-specific strengths) are preserved rather than averaged.

Kimi Outlier Case Study. An adversarial stress test across nine AI platforms evaluated the HAIA-RECCLIN specification itself. Eight of nine platforms recommended publication. One platform (Kimi) sustained an adversarial rejection, mischaracterizing the framework as "theoretical" despite the operational evidence documented in the specification. Human arbitration reviewed Kimi's objections, identified the mischaracterization, and documented the correction with full rationale. The case demonstrates the architecture's designed behavior: a minority position was preserved, reviewed, and resolved through human judgment rather than automated consensus, exactly as the dissent-preservation protocol specifies.

Chapter 11 Override. During the Governing AI manuscript production, the human arbiter overrode a four-of-six AI platform majority on a structural decision. Subsequent review confirmed that the minority position was correct and the majority had introduced citation errors. This case provides direct operational evidence for the necessity of human checkpoint authority: multi-platform consensus is not infallible, and the governance architecture must preserve the human's ability to override even strong AI agreement when judgment warrants it.

Perplexity Hallucination Case Study. During the second round of six-platform structured review of the specification (v2.1 to v2.2), one platform (Perplexity) produced the most methodologically rigorous review in the first batch, including detailed section-by-section analysis with specific citation verification. In the second batch, the same platform fabricated entire specification sections (Sections 7.5, 9.7, 9.8, 10.1, 10.2, 10.3), invented a phantom Appendix B glossary, and generated quoted text attributed to the specification that did not exist in the source document. Cross-validation across five other platforms detected the fabrication immediately: no other platform referenced the nonexistent sections. When confronted with the discrepancy, Perplexity self-corrected and revised its assessment. This case provides live operational evidence for three architecture claims. First, single-platform reliability is not guaranteed across sessions; a platform that performs rigorously on one task can hallucinate on the next. Second, multi-platform triangulation catches fabrication that single-platform review cannot detect. Third, the governance architecture's cross-validation mechanism functions as designed: fabricated content was identified, documented, and corrected through the same checkpoint process that governs all other decisions.

Additional validation evidence includes the HAIA-SMART v1.5 calibration framework (a structured assessment methodology for measuring collaboration quality) and the End-of-Year 2025 Collaboration Audit, which documented framework performance across the full calendar year of operational deployment.

These metrics are self-reported and derived from the author's own operational records. They have not been independently verified by a third-party auditor. The author acknowledges this limitation and notes that the architecture's audit trail design is specifically intended to enable independent verification: all records are append-only, timestamped, and cryptographically chained, meaning a third-party auditor could reconstruct the complete decision history from the audit trail alone.

6. Multi-Platform Review Under Checkpoint-Based Governance

The HAIA-RECCLIN specification underwent structured review across ten AI platforms: Claude (Anthropic), ChatGPT (OpenAI), Gemini (Google), Grok (xAI), Perplexity, Mistral, DeepSeek, Kimi, and additional evaluation instances. The review process followed the framework's own methodology: identical prompts were dispatched to all platforms, responses were collected, convergence and divergence were documented, and human arbitration resolved all disagreements with documented rationale. The review process is framed as structured adversarial review under governance, not peer review in the traditional academic sense. No AI platform possesses the authority, institutional standing, or methodological independence to serve as a peer reviewer. The platforms served as structured critics operating under checkpoint governance, with the human arbiter retaining final decision authority.

The specification evolved through v1.7 (initial structured review across ten platforms), v2.1 (EU compliance integration with prEN 18286 alignment), and v2.2 (six-platform synthesis incorporating two rounds of structured adversarial review). The v1.7 review incorporated ten decision points resolved through human arbitration with documented rationale. The v2.1 to v2.2 cycle added enforcement timeline corrections, prEN 18286 quality management system integration, Navigator Balance Audit procedural updates, and a closing positioning statement. Key findings sustained across all review cycles include the following. All platforms identified no architectural contradictions in the automation bias trigger mechanism (95% approval threshold with configurable reversal rate thresholds, three-cycle escalation), the agent security architecture (code integrity, separation of duties, immutable deployment), the three-tier categorical framework (Ethical AI, Responsible AI, AI Governance), and the evidence hierarchy designating Model 3 as the gold standard.

The most contested design decision was Navigator concentration: reviewing platforms across all cycles identified single-platform Navigator designation as a potential single point of failure, with multiple remedies proposed ranging from rotation to structural liability language. The human arbiter resolved this through documented rationale: permanent Navigator assignment is a Checkpoint-Based Governance decision subject to reevaluation as ecosystem capabilities evolve. Post-project Navigator Balance Audit uses three-platform review with a dissent-loop resolution procedure: any dissent is returned to the Navigator for resolution, and persistent disagreement after the loop is overruled by two-thirds majority. This resolution was sustained across four review cycles, with dissent preserved in documentation and the arbiter's reasoning recorded at each decision point.

The integration gap claim was separately verified through independent search commissions dispatched to six AI platforms across two review cycles. All platforms confirmed the claim at medium to high confidence. No platform identified a published architecture occupying the governance layer between quality management systems and multi-AI platform workflows. The closest candidates identified (Bandara et al., 2025; Vijayaraghavan et al., 2026;

Sonate Trust Protocol) each covered three of five integration components, with non-cognitive agent design and automation bias detection consistently absent. The term "non-cognitive agent" is this architecture's vocabulary; functional equivalents may exist under different terms, and the claim holds for those alternatives as well.

7. Limitations and Future Work

7.1 WEIRD Bias Limitation

As documented in Section 3.6, all major AI platforms share overlapping training data drawn predominantly from WEIRD sources. Multi-platform triangulation can detect divergence between platforms but cannot detect biases shared across platforms. This is a structural limitation of any multi-AI governance approach, not unique to HAIA-RECCLIN, but it must be acknowledged rather than obscured.

During the multi-platform review process documented in Section 6, two platforms with distinct commercial branding returned identical review text on the same task, demonstrating that provider plurality can fail when platforms share overlapping model architectures or training pipelines. This is a concrete instance of the limitation this section describes: triangulation requires genuinely independent platforms, and commercial branding alone is not a reliable indicator of architectural independence.

The long-term structural remedy for WEIRD bias in multi-AI governance is policy-driven: government subsidies for independent AI developers with diverse geographic, cultural, and institutional origins would expand the rotation pool beyond platforms sharing overlapping WEIRD corpora. Without such diversification at the ecosystem level, governance architecture can document the limitation but cannot resolve it.

7.2 Self-Validation

The framework was validated using the framework's own methodology: multi-platform dispatch, synthesis, and human arbitration. This creates a circularity that the author acknowledges. The operational evidence from the *Governing AI* manuscript and subsequent case studies demonstrates that the methodology produces documented, attributable decisions with preserved dissent, but independent third-party validation using alternative methodologies remains a necessary next step.

7.3 Navigator Concentration

Permanent Navigator assignment is a Checkpoint-Based Governance decision subject to reevaluation as ecosystem capabilities evolve. The current designation of Claude as primary Navigator reflects demonstrated synthesis capability across operational deployments. Post-project Navigator Balance Audit uses three-platform review with a dissent-loop resolution procedure: any dissent is returned to the Navigator for resolution, and persistent disagreement after the loop is overruled by two-thirds majority. This is a known governance compromise necessitated by the uneven distribution of synthesis capabilities across current platforms, not a structural flaw. The specification documents backup Navigator protocols for continuity.

7.4 Model 2 Scalability

Operating Model 2 requires human arbitration at every checkpoint. At enterprise scale, this creates a human bottleneck that limits throughput. This is acknowledged as a design feature rather than a flaw: governance has a cost, and Model 2 is intended for contexts where that cost is justified by the risk profile of the decisions being governed. Organizations requiring speed select Model 1. The architecture does not pretend that full governance and maximum throughput are simultaneously achievable.

7.5 Future Work

Priority areas for future research include: independent third-party audit of the operational evidence trail; longitudinal studies of automation bias threshold effectiveness across domains; comparative analysis of Navigator synthesis quality across platforms; formal security verification of the non-cognitive agent design and HAIA-GOPEL implementation; controlled experimental studies of threshold calibration across diverse organizational contexts; pilot implementations across regulated industries (financial services, healthcare, government) to validate regulatory compliance mapping claims; alignment mapping between HAIA-RECCLIN's six directly supported quality management system elements and the full twelve-element structure of prEN 18286:2025 once the final standard is published; and development of a companion implementation guide covering cost analysis, team sizing, frequently asked questions, troubleshooting protocols, and audit checklists for organizations adopting the specification at each of the three operating models.

8. Discussion

HAIA-RECCLIN occupies a specific and previously unoccupied position in the AI governance landscape. It is not an orchestration framework: it does not compete with LangGraph, AutoGen, or CrewAI for task routing. It is not an ethics framework: it does not prescribe what is right or wrong. It is not a responsible AI toolkit: it does not provide model testing or bias mitigation techniques. It is a governance architecture: it provides the structural and operational framework within which ethical judgments are made, responsible practices are enforced, and compliance is documented.

More precisely, HAIA-RECCLIN is an operational governance layer designed to sit between a quality management system and multi-AI platform workflows. Quality management system standards such as prEN 18286:2025 and ISO 42001 tell organizations what governance documentation is required. AI risk frameworks such as NIST AI RMF and the EU AI Act tell organizations what risk management obligations they face. Orchestration tools such as LangGraph, AutoGen, and CrewAI tell engineers how to route tasks between platforms. Alignment techniques such as Constitutional AI and RLHF tell model providers how to make individual models safer. None of these occupies the operational governance layer between regulatory obligation and the AI systems that produce work. That layer, connecting quality management system requirements to operational AI workflows through a single coherent evidence-producing architecture, is the integration gap this specification addresses. The analogy is COBIT, which does not compete with ERP software but tells organizations how to govern IT and sits between regulatory obligation and operational systems. HAIA-RECCLIN occupies the equivalent position for multi-AI workflows.

The non-cognitive agent design represents the most counterintuitive contribution of this work. In a field racing toward greater AI autonomy, HAIA-RECCLIN argues for maximum constraint on the orchestrating agent. The logic follows from the Harold Finch Principle: the component with the most access must have the least intelligence. If the orchestrating agent can evaluate content, it can be manipulated. If it cannot evaluate content, the attack surface reduces to transport and logging, both amenable to formal verification. The preliminary development of HAIA-GOPEL as the implementable form of this specification bridges the gap between architectural theory and operational deployment.

The provider plurality protocol addresses a concern that the antitrust literature has identified (Narechania & Sitaraman, 2024; FAS, 2025) but that no operational framework has previously implemented: the risk of single-vendor capture in AI-dependent workflows. Organizations that deploy on a single AI platform face the same governance risks that telecommunications regulators identified with monopoly providers: no competitive pressure on quality, no alternative when the provider fails, and no ability to detect systematic bias without an external reference point.

Provider plurality as an architectural requirement, however, is necessary but not sufficient. The Harold Finch Principle explains why: if all available platforms are built by corporations with overlapping profit incentives, overlapping training data, and overlapping institutional cultures, rotation produces the appearance of diversity without its substance. The policy argument follows directly. Government subsidies for independent AI developers, free of corporate influence and profit motive, with diverse geographic and cultural origins, expand the pool of genuinely independent platforms from which HAIA-RECCLIN's rotation protocol draws. HAIA-RECCLIN captures the value of diversity when diversity exists in the ecosystem. The ecosystem must first be diverse. That is a policy problem, not an engineering problem.

The regulatory compliance implications are significant and have strengthened since the EU's shift to self-assessment under Annex VI for most high-risk AI systems. The EU AI Act mandates human oversight (Article 14) and logging (Article 12) for high-risk AI systems. The draft harmonised standard prEN 18286:2025 operationalizes Article 17 quality management system requirements, defining twelve core elements; HAIA-RECCLIN directly supports six of those twelve elements (documentation and record-keeping, risk management integration, testing and validation evidence, incident reporting evidence, technical specifications, and accountability framework infrastructure), with the remaining six requiring organizational governance that deployers build around the specification. The EU Commission found ISO 42001 not aligned with the AI Act, making prEN 18286 the relevant compliance target. NIST AI RMF emphasizes traceability and accountability. DORA (Digital Operational Resilience Act) requires ICT risk management for financial institutions; HAIA-RECCLIN does not claim DORA compliance but its audit trail architecture produces documentation relevant to ICT risk management requirements. NYDFS 23 NYCRR 500 mandates cybersecurity controls for financial services; HAIA-RECCLIN's security architecture addresses access control and audit logging but does not constitute a complete NYDFS compliance program. Enforcement of high-risk AI system obligations under the EU AI Act has been extended to December 2027 through the Omnibus Simplification Package, though Article 17 quality management system requirements remain effective August 2026. HAIA-RECCLIN produces the artifacts required across this compliance stack: timestamped decision records, human approval documentation, error logs, provider identity trails, and automation bias metrics. Under the self-assessment model, these artifacts become the deployer's self-certification evidence rather than supporting materials for an external reviewer. The architecture does not claim compliance. It claims that an organization operating the architecture produces the documentation and operational evidence that compliance requires.

The relationship between HAIA-RECCLIN and orchestration frameworks is complementary, not competitive. HAIA-RECCLIN could be implemented on top of LangGraph, AutoGen, or any other orchestration substrate. The orchestration framework provides the plumbing. HAIA-RECCLIN provides the governance. Plumbing without governance is automation. Governance without plumbing is policy.

9. Conclusion

This paper has presented HAIA-RECCLIN, a governance architecture specification for a non-cognitive autonomous agent that enables audit-grade multi-AI collaboration. The architecture addresses a documented integration gap in the published literature, verified through independent multi-platform search across six AI platforms and two rounds of structured adversarial review: no published architecture occupies the governance layer between quality management systems and multi-AI platform workflows, integrating non-cognitive agent design, mandatory provider plurality, cryptographic audit trails, automation bias detection, and cross-framework regulatory compliance mapping into a single operational architecture.

The framework was developed through iterative multi-platform refinement, operationally validated through multiple proof-of-concept deployments including a 96-checkpoint manuscript production, a five-platform thought leader convergence study, and adversarial stress testing across nine platforms, and independently reviewed under checkpoint-based governance across ten AI platforms. It provides three operating models that scale governance intensity to match organizational risk profiles, from factory-speed single-endpoint approval through full human arbitration at every checkpoint.

The core design philosophy is structural humility: the agent that orchestrates the most powerful AI platforms is itself the least intelligent component in the system. It dispatches, collects, routes, logs, and pauses. It performs zero cognitive work. It is a pipe with a logbook. The humans decide. The platforms reason. The agent governs the space between them.

HAIA-RECCLIN is, to the author's knowledge, the first published operational governance architecture designed to sit between a quality management system (prEN 18286 or equivalent) and multi-AI platform workflows. Individual components exist independently in the literature: audit trails, human oversight mechanisms, compliance mapping, multi-model orchestration, and automation bias detection each appear in isolation or partial combination. The contribution is the integration layer connecting quality management system requirements to operational AI workflows through a single coherent evidence-producing architecture, enforcing human oversight at architecturally defined checkpoints and generating documentation required for regulatory self-assessment. The specification occupies the governance layer between regulatory obligation and operational AI systems. That layer, not any individual component, is the integration gap this architecture addresses.

The non-cognitive agent is under preliminary development as HAIA-GOPEL (Governance Orchestrator Policy Enforcement Layer), bridging the gap between the architectural specification documented here and operational implementation. The specification is published as an open document (GitHub: github.com/basilpuglisi/HAIA) and the author invites implementation, critique, extension, and correction from the research community.

References

Agranat Commission. (1974). Israeli state inquiry into the Yom Kippur War. State of Israel Government Report.

Bandara, E., et al. (2025). Towards responsible and explainable AI agents with consensus-driven reasoning. arXiv preprint arXiv:2512.21699.

Banovic, N., Kraut, R. E., Olson, J. S., & Steinfeld, A. (2023). The effects of explanations and algorithmic accuracy on visual recommender systems of artistic images. Behaviour & Information Technology, 43(9), 1–18. https://doi.org/10.1080/0144929X.2023.2184098

Barrett, H. C., et al. (2023). Psychology's WEIRD problems. Cambridge Elements. https://doi.org/10.1017/9781108974288

Cao, J., Deng, Z., Zhao, M., Guo, L., & Zhang, S. (2024). Diagnosing the Western cultural bias of large vision-language models in generative image understanding. arXiv preprint arXiv:2406.11665.

Clayton Antitrust Act, 15 U.S.C. §§ 12–27 (1914).

Di Vita, S., et al. (2025). EAM-SQL: Cryptographic safety envelopes for table-centric LLM pipelines. OpenReview.

European Data Protection Supervisor. (2025, September 22). TechDispatch #2/2025: Human oversight of automated decision-making. EDPS Office.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, L 188. Articles 12, 14. https://artificialintelligenceact.eu/

Federation of American Scientists. (2025, January 9). Antitrust in the AI era. FAS Policy Publication. https://fas.org/publication/antitrust-in-the-ai-era/

Gaurav, S., Heikkonen, J., & Chaudhary, J. (2025). Governance-as-a-Service: A multi-agent framework for AI system compliance and policy enforcement. arXiv preprint arXiv:2508.18765.

Georgia Institute of Technology. (2024, April 2). LLMs generate Western bias even when trained with non-Western languages. Georgia Tech News.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Heim, L., Bernaerts, K., Metzger, C., & Anderljung, M. (2024). Governing through the cloud: The intermediary role of compute providers in AI regulation. Center for the Governance of AI.

Henrich, J., Atari, M., Xue, M. J., Park, P. S., & Blasi, D. E. (2024). Which humans? Measuring and understanding the responses of large language models to cultural diversity. Harvard Department of Human Evolutionary Biology. [Working Paper].

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hoover, A. (2025, June 17). There's a '10% to 20% chance' that AI will displace humans completely, says 'godfather' of the technology. CNBC.

Janis, I. L. (1972). Victims of groupthink: A psychological study of foreign-policy decisions and fiascos. Houghton Mifflin.

Kreps, S., Kriner, D. L., & Schneider, J. (2024). Automation bias in national security contexts: Experiential and attitudinal drivers. International Studies Quarterly, 68(2), sqae020. https://doi.org/10.1093/isq/sqae020

Mallaband, A. (2025, December 29). Deterministic guardrails for nondeterministic agents. LinkedIn.

Mökander, J., & Floridi, L. (2025). Human oversight under Article 14 of the EU AI Act. In Regulation of Artificial Intelligence: International Standards and Governance Mechanisms (pp. 127–154). Edward Elgar Publishing.

Mosier, K. L., & Skitka, L. J. (1996). Does automation bias decision-making? International Journal of Industrial Ergonomics, 18(5), 311–325. https://doi.org/10.1016/0169-8141(95)00070-4

Narechania, T. S., & Sitaraman, G. (2024). An antimonopoly approach to governing artificial intelligence. Yale Law & Policy Review, 43(1), 1–74.

Navarrete, J., Vaca-Barahona, B., Peña, S., & Hernández-Suárez, A. (2024). Human oversight of automated decision-making systems: A multi-method study on automation bias with recidivism risk assessment algorithms. PLOS ONE, 19(1), e0296384.

Nolan, J., & Nolan, L. (Creators). (2011–2016). Person of Interest [Television series]. CBS.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Pierucci, F., Galisai, M., Bracale, M. S., Prandi, M., Bisconti, P., Giarrusso, F., Sorokoletova, O., Suriani, V., & Nardi, D. (2026). Institutional AI: A governance framework for distributional AGI safety. arXiv preprint arXiv:2601.10599.

prEN 18286:2025. Quality management systems for providers of AI systems. European Committee for Standardization (CEN). Public enquiry completed January 22, 2026.

Puglisi, B. (2025). Checkpoint-Based Governance v4.2.1. basilpuglisi.com.

Puglisi, B. (2025). Governing AI: When capability exceeds control. basilpuglisi.com.

Puglisi, B. (2026). HAIA-RECCLIN Agent Architecture Specification v2.2 (EU Compliance Version). basilpuglisi.com. https://github.com/basilpuglisi/HAIA

Puglisi, B. (2026). HAIA-RECCLIN Multi-AI Framework Updated for 2026. basilpuglisi.com.

Schrepel, T. (2025). Decoding the AI Act: Implications for competition law and computational antitrust. Journal of Competition Law & Economics, 21(3), 374–412. https://doi.org/10.1093/jcle/qhae024

Sherman Antitrust Act, 15 U.S.C. §§ 1–7 (1890).

Shur-Ofry, M. (2023). Multiplicity as an AI governance principle. International Journal for Law and Information Policy, 12(3), 374–410.

Sonate. (n.d.). SONATE Trust Protocol. Sonate.

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

Vallance, C. (2024, December 27). 'Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years. The Guardian.

VeritasChain Standards Organization. (2026, January 18). Why the EU AI Act needs cryptographic audit trails—and how VCP v1.1 delivers them. VeritasChain Blog.

Vijayaraghavan, G., et al. (2026). If you want coherence, orchestrate a team of rivals: Multi-agent models of organizational intelligence. arXiv preprint arXiv:2601.14351.

Zhao, R., Shoaib, M., Hoang, V. T., & Hassan, W. U. (2025). Rethinking tamper-evident logging: A high-performance, co-designed auditing system. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25). ACM.

Zheng, K. (2025). Antitrust in artificial intelligence infrastructure. NeuroComputing. https://doi.org/10.1016/j.neucom.2025.127235

Zhou, K., Constantinides, M., & Quercia, D. (2025). Should LLMs be WEIRD? Exploring WEIRDness and human rights in large language models. arXiv preprint arXiv:2508.19269.