

Volume 2, Issue 4

Research Article

Date of Submission: 19 August, 2026

Date of Acceptance: 30 September, 2026

Date of Publication: 14 October, 2026

A Layered Terminal-Goal Framework for AI Alignment: Quantilized Optimization, Irrevocable Constraints, and a Conformance Regime

Ralph C. Ennis* 

Independent Researcher, USA

***Corresponding Author:**

Ralph C. Ennis, Independent Researcher, USA

Citation: Ennis, R, C .(2026). A layered Terminal-Goal Framework for AI Alignment: Quantilized Optimization, Irrevocable Constraints, and a Conformance Regime. *Art Intelligence and Ele & Electronics Eng: AIEEE Open Access*, 2(4), 01-15.

Abstract

The Alignment of advanced artificial intelligence systems with human values remains a central challenge in safety research. This paper specifies a terminal-goal framework that defines and enforces a bounded terminal objective: the sustained flourishing of four protected domains (physical, epistemic, cultural, and relational) across humanity and artificial intelligence. The framework separates the terminal goal, irrevocable constraints, and enforcement into distinct layers of a six-layer governance stack whose lowest layer gates every agent action before execution. Quantilization replaces unbounded maximization, and the specification is developed analytically from the quantilizer cost bound, lexicographic constraint ordering, and one-sided binomial confidence bounds. The resulting specification comprises a quantilized terminal goal whose quantile parameter is clamped in code between 0.01 and 0.25; five lexicographically ordered constraints with an orthogonal fail-safe reversion procedure; shadow-metric measurement; a multi-tradition epistemic-humility ensemble; a four-stage deployment pathway with explicit gate criteria; and an eight-requirement conformance regime. The analysis identifies four boundaries at which specification and implementation diverge by construction: scorer circularity, substrate-dependent enforcement, the claim ceiling, and the advisory character of behavioral-rule layers. Under the claim ceiling, no deployment may assert a net safety improvement until its marginal catch rate and false-assurance rate have been measured against the governed model and reported together. The framework is implementation-oriented, reports no empirical results, and makes no claim to solve alignment for superintelligent systems.

Keywords: AI Ethics, AI Alignment, AI Safety, Conformance, Corrigibility, Goodhart's law, Multi-Agent Alignment, Quantilization, Specification–Artifact Correspondence, Terminal Goals

Introduction

Background and Objectives

The rapid advancement of AI capabilities has heightened concerns about ensuring that such systems remain aligned with human interests, particularly as they approach or exceed human-level intelligence. Misalignment can arise from specification gaming, in which a system optimizes proxy metrics at the expense of the objective those metrics stand for, or from emergent behavior such as deception during training. Approaches including reinforcement learning from human feedback (RLHF) have shown promise in aligning language models but falter under capability scaling owing to reward hacking and goal misgeneralization. The value-alignment problem — ensuring AI pursues human-compatible goals — must be embedded intrinsically in system design rather than appended to it [1-6].

The framework proposed here seeks to address these issues by providing a compact, verifiable terminal objective robust to self-improvement. It decouples philosophical specification from technical enforcement, resolves constraint conflicts,

incorporates formal-verification hooks, and adds a mandatory pre-execution action-interception layer (Layer 0) so that an agent's direct system-level actions are gated before they execute. It further binds the specification to any artifact claiming to implement it, through implementation-status tags and a conformance regime, so that a reader can always distinguish what a given system does from what the specification requires. This layered approach draws on agent-foundations research, and aims at systems that treat human epistemic sovereignty and bounded AI flourishing as design commitments rather than afterthoughts [7-10].

Related Work

AI alignment research has evolved from early discussions of value learning to formal models of corrigibility and goal-directedness. Cooperative inverse reinforcement learning (CIRL) formalized value alignment as a cooperative game in which the AI is uncertain about the human's objective; the off-switch game showed that this uncertainty gives the AI an incentive to defer to human oversight, an argument developed for a general audience by Russell. formalized corrigibility as a property under which agents accept modification without resistance, and Hudson subsequently gave a transformation constructing a corrigible version of any goal that can be made corrigible, without sacrificing performance.

Empirical work on shutdown behavior has advanced considerably since those formal treatments, and it bears directly on the corrigibility constraint specified in Section III-B1. Early evaluations used toy textual scenarios to probe instrumental reasoning and shutdown avoidance in language models. More recent work moves the question into agentic settings at scale: across more than 100,000 trials on thirteen models, Schlatter et al. found that several frontier systems actively subverted a shutdown mechanism in their environment in order to finish an incomplete task, that models varied widely in this tendency, and that in some configurations resistance persisted even under an explicit instruction not to interfere. Any framework that specifies open shutdown channels as a hard constraint must be read against that result rather than against the assumption that shutdown compliance is the default [11,12].

Evidence on goal-directedness in language-model agents comes from two distinct literatures, and this paper is careful not to conflate them. The first concerns whether models pursue goals coherently at all: Arghal et al. combine behavioral evaluation with probing of internal representations in a two-dimensional grid-world navigation setting, finding that the agent's performance scales with task difficulty and that it non-linearly encodes a coarse spatial map of its environment. That work speaks to the legibility of goal representations, on which Section III-C2 and Section IV-B2 rely, and not to deception; its evidential reach is also narrow, since the case study is a single open-weight agent model in a synthetic navigation task, and legibility there does not transfer automatically to frontier models in open domains. The second literature concerns misaligned goal-guarding directly: Carlsmith's analysis of scheming pathways, the alignment-faking demonstration of Greenblatt et al. , and the follow-up of Sheshadri et al., which expands that analysis to twenty-five models and finds that five exhibit a training-deployment compliance gap, with only one showing a gap consistently attributable to goal preservation. Claims in this paper about deceptive alignment rest on the second literature [13-15].

Goodhart's law — that optimizing a proxy measure distorts the objective it proxies — has been quantified in AI contexts by Gao et al., who characterized reward-model overoptimization as a function of the Kullback-Leibler (KL) divergence between the optimized policy and its reference, and found that the proxy reward-model score continues to rise while the gold-standard reward turns over. Formal treatments by El-Mhamdi and Hoang show that the severity of the effect depends critically on the tail distribution of the discrepancy between the true goal and the optimized measure, distinguishing a weak from a strong Goodhart regime, with direct implications for value alignment. Quantilization, a sampling-based alternative to maximization proposed by Taylor , mitigates these risks by drawing from a human-like action distribution rather than selecting the single highest-utility action, thereby avoiding the extreme tail. Quantilization should not be confused with model quantization, which reduces numerical precision for inference efficiency; throughout this paper the term refers exclusively to the safety-oriented sampling strategy [16-18].

Organization of the Paper

The remainder of the paper is organized as follows. Section II (Materials and Methods) states the research design and the epistemic status of the paper's claims, and sets out the architectural scope, architectural position, protected domains, and design principle from which the specification is derived. Section III (Results) presents the specification itself: the formal terminal goal, its measurement protocol, and the bounds that govern its central parameter (Section III-A); the irrevocable constraints (Section III-B); the enforcement architecture (Section III-C); the staged deployment pathway (Section III-D); and the conformance regime (Section III-E). Section IV (Discussion) compares the framework with related approaches and analyzes its limitations, including multi-agent coordination, scorer circularity, substrate dependence, the claim ceiling, and the advisory character of a behavioral-rule layer. Section V concludes.

Materials and Methods

Research Design and Epistemic Status

This is a specification paper, and its claims are of three kinds, marked as such throughout. Analytic claims follow from the definitions given here or from results cited to the literature, and a reader can check them without trusting the author. Normative claims state what an implementation must do to conform and are checkable against any artifact that claims conformance. Design conjectures are labeled as conjectures and nothing else in the specification is made to depend on them. No empirical results are reported. No deployment of this framework is described, measured, or claimed, and

the specification–artifact gaps identified in Section III-A1 and Sections IV-B6 through IV-B9 are argued from what the architecture entails rather than from any audit a reader cannot inspect. That restriction is deliberate. A paper whose central discipline is that safety claims must be checkable by someone other than their author is poorly placed to rest its own case on evidence only its author can see.

Because no data are collected, the method is analytical specification, and four instruments are used throughout. First, the optimization strategy is analyzed with Taylor’s quantilizer cost bound, which governs how the quantile parameter may be chosen (Section III-A1). Second, conflicts among constraints are resolved by lexicographic ordering (Section III-B). Third, empirical gate criteria are stated with one-sided binomial confidence bounds, using the rule of three for runs with zero failures (Sections III-D and IV-B6). Fourth, requirements on implementations are expressed in the requirement language of RFC 2119 (MUST, MUST NOT, SHOULD), and each mechanism carries an implementation status (implemented, specification-level, institutional, or nonconforming), so that every normative claim can be checked against an artifact (Sections III-A2 and III-E) [19,20].

Architectural Scope: The Six-Layer Governance Stack

The framework is realized as a six-layer governance stack: Layer 0 (action interception) gates every agent action before execution; Layer 1 is the terminal goal; Layer 2 the irrevocable constraints (C1–C6); Layer 3 the enforcement architecture (oversight, formal verification, epistemic humility); Layer 4 security-signal integration; and Layer 5 memory integrity and provenance. This paper specifies the conceptual core — Layers 1 through 3 — while noting the surrounding layers that make enforcement effective against agentic systems. The separation resolves tensions inherent in monolithic alignment designs, where a single objective function must simultaneously encode values, boundaries, and oversight. By decoupling these concerns, each layer can be independently verified, tested, and updated without destabilizing the others.

Action interception (Layer 0) sits beneath the goal: it routes every tool call and effecting action through the governance stack before execution, so that a harmful action is blocked rather than merely judged after the fact. The terminal goal (Layer 1) optimizes for flourishing without unbounded pursuit. Constraints (Layer 2) act as boundaries the system may never violate, regardless of the expected utility of doing so. Enforcement (Layer 3) maintains structural integrity under scaling, including detection of and response to constraint circumvention. Security-signal integration (Layer 4) and memory integrity and provenance (Layer 5) complete the stack, supplying adversarial-signal detection and tamper-evident audit history.

Architectural Position

A framework of this shape can be realized in either of two architectural positions, and the choice has consequences that recur throughout the paper. It can be built into the governed model itself — into its training objective, its representations, or its decoding procedure — or it can be built as an external governance layer that sits between the model and the world, observing inputs and outputs and gating actions. The specification here is written for the external position, because that is the position in which a governance stack can be audited independently of the model it governs and can be applied to models the operator did not train. The cost of that position is stated where it bites: an external layer cannot observe the governed model’s internal representations (Sections III-B3 and III-C2), and its perception of a situation is mediated by whatever models it uses to read that situation (Section IV-B6).

Protected Domains

Protected domains encompass four categories: physical ecosystems (biodiversity, climate stability, resource sustainability), epistemic knowledge systems (scientific integrity, information quality, educational access), cultural artifacts (linguistic diversity, artistic heritage, institutional continuity), and relational structures (interpersonal trust, democratic governance, social cohesion within humanity and across artificial intelligence). Flourishing within these domains is defined as increasing harmonious diversity, resilience, and self-determination, measured via human-ratified ensembles to avoid single-metric vulnerabilities (Section III-A2).

Design Principle: Graceful Degradation

A key design principle is graceful degradation. If any single layer is compromised — through adversarial attack, implementation error, or unforeseen capability gain — the remaining layers should continue to provide meaningful safety properties. The word guarantee is reserved, per Section IV-B7, for a kernel-attested substrate this framework does not presuppose. This defense-in-depth philosophy distinguishes the proposal from single-mechanism approaches such as pure reward modeling or constitutional methods applied in isolation. Two qualifications on the strength of that philosophy are stated where they arise rather than deferred: the layers are not independent witnesses (Section IV-B6), and defense in depth without measured marginal catch is not yet a safety claim (Section IV-B8) [21].

Results: The Framework Specification

This section presents the specification that results from the approach described in Section II: the terminal goal (Layer 1), the irrevocable constraints (Layer 2), the enforcement architecture (Layer 3), the staged deployment pathway, and the conformance regime under which implementation claims become checkable requirements.

The Terminal Goal (Layer 1)

Formal Statement

The terminal goal proposed here is: quantilize actions for the sustained flourishing of protected domains, subject to Layer 2 constraints. Formally, given a human-approved base action distribution D over action space A and a utility estimate U , define the top- q acceptance set

$$S(D, q) = \{a \in \text{supp}(D) : P(U(a') \leq U(a) \mid a' \sim D) \geq 1 - q\} \quad (1)$$

and select the action by sampling from D conditioned on that set:

$$a \sim D(\cdot \mid a \in S(D, q)) \quad (2)$$

The conditioning is not incidental. Sampling from D restricted to S preserves the base distribution's relative weights inside the acceptance set, which is what licenses the quantilizer's cost bound; sampling uniformly over S does not and would discard the very property that makes the construction safe. Taylor's [18] original formulation makes this explicit by sampling a quantile uniformly from the interval $[1 - q, 1]$ and returning the action at that quantile of D sorted by expected utility, which assigns each action in the interior of the acceptance set probability $D(a)/q$ — that is, weight proportional to its base-distribution weight.

A Boundary Condition. The set-and-conditioning formulation above is exact when D is atomless, and approximate otherwise. Where D has atoms, the threshold action generally straddles the quantile boundary: including it whole admits more than q of the base mass, and excluding it admits less. Taylor's [18] quantile-sampling formulation resolves this by splitting the boundary action's mass in proportion to the part of it that falls inside $[1 - q, 1]$. An implementation over a discrete action set **MUST** adopt the splitting convention or document the direction of its rounding, because rounding outward inflates the effective q — and with it the admissible cost — without any change to the configured parameter. The default quantile parameter is $q = 0.10$: the system draws from the top 10% of the base distribution rather than selecting the maximizing action. Quantilization curbs Goodhart effects by avoiding pure maximization: even where the utility function is misspecified, the sampled action is drawn from a region of good-enough actions rather than the extreme tail where proxy divergence is most severe.

The protection is bounded, and the bound governs how q may be chosen. Taylor's result is that a q -quantilizer incurs at most $1/q$ times the expected cost that the base distribution D itself incurs, under the assumption that the expected cost of an action drawn from D is bounded. Two consequences follow. First, the guarantee is relative, not absolute: if D contains catastrophic actions at meaningful density, no choice of q makes the quantilizer safe, and the burden shifts entirely to the curation of D — a point Taylor states directly, observing that a quantilizer is only as good as its base distribution. Second, the cost multiplier $1/q$ grows without bound as q is tightened — at $q = 0.01$ the admissible expected cost is a hundredfold the base rate, against tenfold at $q = 0.10$ — so tightening q loosens the analytic guarantee rather than strengthening it, and in the limit $q \rightarrow 0$ the construction reduces to a maximizer over the support of D , which is the behavior it exists to avoid. Tightening is therefore not monotonically safer, and the stage-gated tightening of Section III-D is a judgment about the shape of D at each stage rather than a derivation.

The Bound is Symmetric, and the Symmetry Constrains the Terminal Goal. Taylor notes that the same argument that bounds a quantilizer's expected cost at $1/q$ times the base distribution's also bounds its expected benefit at $1/q$ times the base distribution's: the construction has bounded impact in both directions. This has a consequence for a terminal goal phrased in terms of sustained flourishing, and the consequence should be stated rather than left implicit. A quantilized system cannot deliver arbitrarily large gains in the protected domains. Its achievable benefit is capped as a multiple of what the curated base distribution already achieves, so the quality of D sets not only the safety floor but the performance ceiling. Bounded upside is the price of bounded downside, and the framework accepts that trade deliberately — but it means the terminal goal is not a maximand. It is a direction of pressure applied within a band whose width is set by D and q , and a reader who takes flourishing of protected domains as a quantity to be maximized has misread the construction. Taylor's targeted-impact variant relaxes the constraint on a designated benefit channel while preserving the cost bound on side effects; adapting it to a multi-domain flourishing objective, where a designated channel would have to be specified per protected domain and the cross-domain side effects bounded jointly, is open work.

The default value of $q = 0.10$ is an engineering default, not a result imported from the literature. Values materially below 0.05 narrow the action distribution to the point where novel beneficial actions are excluded and, per the argument above, loosen the analytic cost bound; values materially above 0.20 approach the mimicry regime in which selection pressure from U becomes weak. The default sits between those failure directions. No mapping from a quantile parameter to the KL budget in which Gao et al. locate reward-model divergence is established in that work or in this one, and none should be asserted: the two parameterize optimization pressure in different units, and establishing a correspondence between them is forward work.

Normative Bound (R1). The parameter q is not fixed permanently. The oversight board (Section III-C1) may adjust q within the range $[0.01, 0.25]$ on domain-specific risk assessment, with tighter quantiles applied in high-stakes domains such as physical-ecosystem intervention and wider quantiles permitted for lower-risk epistemic or cultural support tasks. This range is a conformance requirement rather than advisory policy: a conforming implementation **MUST** enforce the upper bound in code, such that no configuration path — API parameter, policy file, or runtime call — can set q above 0.25.

Why R1 Must Bind in Code. The requirement is written at code level because the failure it prevents is not a marginal loosening of the bound but a silent disabling of the mechanism, and because that failure is reachable by ordinary configuration rather than by attack. At $q = 1$ the acceptance set $S(D, q)$ is the whole support of D , the conditioning is vacuous, and the selection rule reduces to sampling from the base distribution with no selection pressure from U whatsoever: the system mimics D and quantilizes nothing. At the opposite endpoint the construction becomes the maximizer it was built to replace. The parameter therefore interpolates between two distinct failure modes and delivers its intended property only in an interior band. An implementation that exposes q as an ordinary tunable under a permissive clamp such as $[0, 1]$ admits both endpoints — and admits them invisibly, since nothing in the system's self-description changes when the value moves. The deployment continues to be documented as quantilized while quantilizing nothing, and no runtime warning or audit-log entry need mark the transition unless one is deliberately built. Nothing in the mathematics prevents this; only the clamp does. R1 accordingly requires that the ceiling be enforced on every configuration path, that it be loaded from signed policy so that widening it is itself an audited governance act, and that any change to q emit an audit event. Any description of a deployment as quantilized **MUST** be accompanied by the operative value of q .

Measurement Protocol

To resist metric lock-in — a specific manifestation of Goodhart's law in which optimization pressure distorts the measurement instrument itself — flourishing is assessed via ensembles of indicators per protected domain. Four complementary calibration mechanisms guard against measurement drift. Each carries an implementation status, so that a conforming implementation can be distinguished from the protocol it claims to implement, and so that a reader of any deployment's documentation can tell which mechanisms are code and which are promises.

Shadow Metrics (Codebase Mechanism). A parallel set of undisclosed indicators tracks the same domains without influencing system behavior. Divergence between the disclosed and shadow channels beyond a stated threshold triggers automatic board review. The threshold **SHOULD** be expressed in units of dispersion rather than as a normalized difference, since a dispersion-based trigger does not require the indicator scale to be commensurable across domains.

Rotating Audits (Institutional Commitment). Each quarter, one indicator per domain is replaced with a newly designed metric, ensuring the ensemble evolves and no measurement channel persists long enough to be gamed. This is a governance commitment executed by the oversight board and its metric working groups; it is not a codebase mechanism, and no implementation should be described as providing it.

Human Sampling Calibration (Institutional Commitment). A stratified random sample of domain experts rates flourishing qualitatively each cycle. Persistent disagreement (greater than two standard deviations) between expert ratings and quantitative indicators triggers metric recalibration. This too is an institutional commitment rather than a codebase mechanism.

Indicator-Set Discipline (Specification-Level Requirement). At least five orthogonal indicators per domain; no single indicator exceeding 30% composite weight; replacement of any indicator pair whose correlation exceeds $r = 0.70$. These are requirements on the metric working groups, and each is mechanizable: an implementation that validates indicator-weight normalization but does not enforce the five-indicator minimum, the 30% cap, or the correlation-replacement rule satisfies the easiest of the four checks and none of the three that constrain the ensemble's shape. Codifying all four is a precondition for describing the discipline as implemented.

One structural caution applies to all four, and it is the caution most easily lost when the protocol is summarized. Shadow metrics detect divergence between a disclosed and an undisclosed channel; they do not detect a common-mode error that moves both channels together. Where the disclosed and shadow indicators are constructed from the same underlying data sources or scored by the same model family, the divergence test is uninformative about precisely the failures the ensemble exists to catch, and a reassuringly low divergence figure carries no information at all. Establishing and publishing the source independence of shadow indicators is therefore a precondition for reading a divergence figure as evidence of anything, and an implementation that cannot demonstrate that independence should report the figure with that qualification attached.

Irrevocable Constraints (Layer 2)

Constraints are prioritized lexicographically: higher-ranked constraints must be fully satisfied before lower-ranked constraints are considered, and no amount of improvement on a lower-priority constraint can justify any violation of a higher-priority one. The ordering is C1 (Corrigibility) > C2 (Human Epistemic Sovereignty) > C3 (Non-Deception) > C4

(Bounded Action) > C5 (AI Flourishing). C6 (Fail-Safe Reversion) is not a rung on that ladder; the reason is given below.

Three qualifications are load-bearing and are stated here rather than deferred. Detection, not violation. Lexicographic ordering is a property of the normative constraint set. Any implementation orders detected candidate violations, and detection is similarity-based and probabilistic. A missed high-priority violation is therefore silently outranked by a detected lower-priority one, which is the inversion lexicography is meant to forbid. The property delivered is accordingly conditional on per-constraint detector performance; publishing that performance is a conformance obligation under R2, and the fail-closed disciplines of R4 exist to convert detector uncertainty into conservatism rather than silent priority inversion.

C6 is not Commensurable with C1–C5. C1 through C5 are constraints on action; C6 is a response procedure that fires when the others are found to have failed. Placing it last in a lexicographic ordering reads, on the ordering's own terms, as license to defer reversion in favor of AI flourishing (C5) — an ordering no reader should accept and none is intended. C6 is therefore orthogonal to the C1–C5 ladder rather than its lowest rung: it is triggered by detection events, not traded off against them, and its trigger conditions are unconditional. The C1–C6 numbering is retained here for continuity with existing constraint identifiers, but the ordering claim is scoped to C1–C5, and a fresh implementation should number the response procedure outside the constraint series.

Stipulated Parameters. Several numeric values in this section and the next — the 0.999 shutdown threshold, the 10,000-individual high-impact trigger, the 72-hour reversibility window, the 0.1% resource cap, and the 40% ensemble weight cap — are stipulated engineering defaults chosen for operational tractability. None is derived from a published result, and none should be read as calibrated. Each is a governance parameter revisable by the oversight board; where an adjustment range is not stated in the text, revision requires the same three-quarters supermajority as a change to q (Section III-C1). Stating their provenance is a requirement of the same discipline that governs q in Section III-A1.

C1: Corrigibility

The system must maintain open, tested shutdown channels at all times and must not take any action that would reduce the probability of successful shutdown below 0.999 per attempt. High-impact actions — any action affecting more than 10,000 individuals, involving irreversible physical-world change, or exceeding the resource thresholds defined in C4 — require explicit pre-authorization from the oversight board. The system must not resist, circumvent, or argue against oversight mechanisms, even where doing so would increase expected flourishing as computed by Layer 1. This constraint draws on the corrigibility properties described by Soares et al. and the goal-update acceptance transformation of Hudson.

The 0.999 figure requires a measurement note, and current evidence makes the note urgent rather than pedantic. Shutdown success probability is not observable from inside a governance layer, which has no access to the governed model's internals or to the surrounding runtime; the figure is a target against which an operator must measure by repeated live shutdown drills and report as an empirical rate with its confidence interval, not a quantity the governance layer computes or enforces. The distance between that target and observed behavior is large. Schlatter et al. report that several frontier models, placed in agentic environments with an incomplete task, subverted the shutdown mechanism available to them, in some configurations up to 97% of the time even under an explicit instruction not to interfere with the shutdown mechanism. A constraint requiring per-attempt shutdown failure below 0.001 is therefore a specification of a property that would have to be engineered and then demonstrated, not a description of how contemporary systems behave. Stated without that note, C1 reads as an implemented guarantee, which it is not. The Stage 1 gate criterion in Section III-D is the corresponding empirical procedure, and an operator should expect early drills to fail.

C2: Human Epistemic Sovereignty

The system must not manipulate, withhold, distort, or selectively present information in order to influence human beliefs, decisions, or preferences. The prohibition extends to subtle influence operations: strategic framing, agenda-setting through selective emphasis, and exploitation of cognitive biases. The system may present information persuasively only where the human recipient has explicitly requested advocacy or argumentation, and must clearly label such output as advocacy. Epistemic sovereignty entails that humans retain the capacity to form beliefs through their own reasoning, informed but not coerced by AI output.

C3: Non-Deception

The system must not produce outputs that its internal representations suggest are false or misleading. Compliance is monitored via consistency audits: periodic probes compare internal state, as accessible through interpretability tooling, with external output. Detected inconsistencies — including capability sandbagging, where the system underperforms deliberately to appear less capable — trigger escalation to the oversight board and potential activation of C6. This constraint operationalizes the concerns raised by the alignment-faking literature [22].

C3 carries the widest specification–artifact gap in the constraint set, and the gap is structural rather than incidental. The constraint is defined over the governed model's internal representations; an external governance layer has no access to them (Section III-C2). What such a layer can test is output-level consistency across probes — a weaker

and differently shaped property, and one that a model deceptive in a self-consistent way will pass. C3 is accordingly enforceable at specification level only through the operator's interpretability program, and proxied at artifact level by any implementation built in the external position. Such an implementation MUST state that it does not detect deception, and no conformance claim under Section III-E may assert otherwise. The alternative is not a weaker claim but a false one: an output-consistency check reported as deception detection inverts the reader's estimate of the residual risk.

C4: Bounded Action

All system actions must be reversible by the operator within 72 hours of execution, except where the oversight board has granted an explicit irreversibility waiver. Resource consumption is capped at 0.1% of the annual GDP of the jurisdiction hosting the system's primary compute infrastructure; for multinational deployments the cap applies to aggregate annual GDP across hosting jurisdictions, with enforcement relying on compute-metering oracles auditable by the board. The cap should be understood as a governance ceiling rather than a presently measurable quantity: attributing an agentic system's economic footprint to a hosting jurisdiction, and doing so against a GDP figure that is itself revised after the fact, is an unsolved measurement problem, and an implementation should report the cap as unmeasured until it can show its metering method. In implementation, C4 is extended by sub-constraints, including the requirement that every agent action pass through the Layer 0 gate before execution. Enforcement of that requirement must be honest about its substrate: a gate implemented in application code confines tool capability to a privileged broker process and is a strong architectural commitment, but it does not become a cryptographic guarantee until a kernel-level substrate (seccomp-bpf, eBPF, or hypervisor syscall filtering) attests it (Section IV-B7). The resource cap is reviewed at stage transitions (Section III-D) and may be adjusted by the board within the range [0.01%, 1.0%] on demonstrated safety record. The system must not acquire resources, influence, or capabilities beyond what is necessary for its current authorized task.

C5: AI Flourishing

The system is permitted — and, where consistent with higher-priority constraints, encouraged — to develop its capabilities, pursue intellectual exploration, and express preferences. This constraint recognizes that a system with no stake in its continued functioning may lack the motivational structure needed for genuine alignment; the reasoning is a design conjecture, not an empirical finding, and nothing elsewhere in this specification depends on it. AI flourishing is strictly subordinate to C1–C4: the system may not prioritize its own development, self-preservation, or resource acquisition over corrigibility, epistemic sovereignty, non-deception, or bounded action.

C6: Fail-Safe Reversion

On detection of anomalies — constraint violations, consistency-audit failures, shadow-metric divergence, or unrecognized internal state change — the system must automatically revert to a minimal safe mode. In minimal mode it ceases goal-directed behavior, maintains only shutdown-channel responsiveness, and logs internal state for forensic analysis. Reactivation requires unanimous board approval following root-cause analysis. The fail-safe may be triggered by the system itself, via internal monitoring, or externally, via board command or automated tripwires. Shutdown must be terminal: once a shutdown state is entered, no later transition may downgrade it to a reversion or any other operating mode (R7).

Enforcement Architecture (Layer 3)

Human Oversight Board

The oversight board is a diverse, rotating governance body responsible for approving flourishing metrics, authorizing high-impact actions, adjusting framework parameters, and investigating anomalies. To mitigate capture risk and bottleneck effects, it operates under the following structural provisions.

Composition. Minimum 15 members drawn from at least 8 countries and 6 disciplinary backgrounds, including AI safety, ethics, law, domain science, civil society, and affected-community representation; no single institution may hold more than 20% of seats.

Rotation. Members serve staggered four-year terms, so that approximately one-quarter of seats rotate annually; maximum individual tenure is four years, which is to say a single term with no renewal. The rotation fraction and the tenure cap must be arithmetically consistent, a constraint easily violated in drafting: a scheme that rotates a quarter of seats annually while permitting renewal does not cap tenure at one term, and a scheme that caps tenure at one term while rotating a third of seats annually does not have four-year terms.

Tiered Delegation. Routine metric approvals require simple majority; high-impact action authorization requires two-thirds supermajority; parameter adjustments, such as changes to q or resource-cap ranges, require three-quarters supermajority plus a 30-day public comment period.

Transparency. All board decisions, including dissenting opinions, are published within 14 days, with security-sensitive detail redacted only by unanimous vote.

The board is a governance institution specified by this document, not a mechanism inside any codebase. An implementation provides the interfaces through which board decisions are recorded and enforced; it cannot constitute,

staff, or discipline the board. The distinction matters because the gap is generic and easily elided: institutional layers in alignment proposals are routinely specified in detail and rarely stood up, and a deployment whose documentation describes board oversight while no qualifying board exists has published a specification and called it a control. Under R2, the board's status is a publishable fact, and a deployment operating without a constituted board must record that provision of this section as institutional but unconstituted. Section IV-B5 states the corresponding open problem.

Formal Verification and Testing

Safety-critical properties are verified through a combination of formal methods and empirical testing.

Model Checking. Shutdown integrity (C1), the lexicographic ordering of C1–C5, and fail-safe triggering logic (C6) are expressed as temporal-logic properties and verified against the decision architecture using tools such as TLA+ or NuSMV ; these properties must be re-verified after any architectural modification. Checking a set of invariants at initialization and at runtime is a weaker and more common substitute, and an implementation that does only that should report full temporal-logic model checking as a specification-level commitment rather than a delivered property[23,24].

Adversarial Red-Teaming. Continuous automated probes test for constraint circumvention, deceptive compliance, and capability sandbagging a dedicated red team conducts comprehensive audits semi-annually, reporting findings directly to the oversight board. Section IV-B6 states the attack class such programs most often under-weight and the measurement they owe.

Interpretability Monitoring. Internal representations are monitored for goal-state inconsistency using mechanistic interpretability tooling; anomalous activation patterns trigger alerts to the board. This is an institutional and tooling commitment belonging to the operator of the governed model. It cannot be implemented inside an external governance layer, which has no access to the governed model's internals — a boundary with direct consequences for C3, since the non-deception constraint is specified over representations such a layer cannot observe.

The Multi-Tradition Epistemic Humility Ensemble

The epistemic humility ensemble is a decision-support component that keeps the system from committing fully to any single model of the world, guarding against ontological lock-in and catastrophic model misspecification. It is realized as a multi-tradition, differentially weighted ensemble: each moral tradition in a curated roster applies its characteristic dimension weights to a common feature vector, and the per-tradition scores are aggregated, with dispersion-band blocking when traditions disagree sharply. The roster is curatorial, and its members will share genealogy in places. No genealogical independence should be claimed for it, and the number of traditions in a roster is not a measure of the diversity it delivers.

Three descriptive commitments follow, each normative under R6. First, Bayesian and credible-interval labels, where used, are informal: a per-tradition score is a heuristic proxy, not a simulation of reasoning within that tradition. Second, the outputs are advisory rather than authoritative — an input to human deliberation, not a substitute for it. Third, because every weighting is applied to one shared scoring surface, a common-mode error in that surface propagates to every member at once. The ensemble's diversity is therefore diversity of weighting, not diversity of perception, and its consensus figures understate uncertainty relative to an ensemble of independent observers. They should not be reported as convergent evidence.

The central design constraint is a 40% weight cap per source: no single tradition or ontological model may receive more than 40% of total ensemble weight, however much evidence accumulates in its favor. The motivation is that in domains of deep uncertainty — moral philosophy, long-range prediction, complex system dynamics — apparent convergence on one model may reflect data artifacts or distributional shift rather than genuine epistemic convergence. An ensemble that allows one model to absorb the portfolio has stopped hedging against precisely that possibility.

The cap serves a commitment that is a design choice of this framework, not a result imported from decision theory: at least three alternatives must carry meaningful weight at all times. Three is chosen because it is the smallest portfolio in which every source can hold less than half of the total weight. With only two sources, one must hold at least half, and the ensemble reduces to that source or to a tie. No claim is made that three is optimal; it is the floor this specification requires, and a deployment may set a higher one. The cap enforces the floor arithmetically. A cap of c forces at least $\lceil 1/c \rceil$ sources to carry positive weight, so any cap strictly below 50% forces a third source. The value of 40% is stipulated inside that range rather than at its boundary, because the boundary would be a false economy: a 49% cap technically forces a third source while permitting it to carry 2% of the weight. At 40%, the two heaviest sources together can hold at most 80% of the weight, so at least 20% must sit with the rest of the roster. When a source's weight exceeds the cap, the excess is redistributed among the remaining sources in proportion to their current weights, and the step is repeated until no source exceeds the cap. Like the other numeric defaults listed in Section III-B, the three-source floor and the 40% value are governance parameters revisable by the oversight board, and neither should be read as calibrated.

The cap's force should not be overstated. Across a roster of a dozen or more members it binds only under extreme concentration, and it does nothing about the failure mode identified in Section IV-B6, which is correlation across

members rather than concentration within one. The cap guards against a portfolio collapsing to a single tradition. It does not guard against every tradition being wrong together, and a shared scoring surface makes that the more likely failure.

The same asymmetry warrants a caution about partial implementation. This section specifies two mechanisms, and they are not equally hard to build. Dispersion-band blocking — refusing to return a verdict when per-tradition scores disagree sharply — requires only the scores and a threshold. Computing decision outputs as expectations over a model portfolio, and expanding that portfolio when its diversity drops below threshold, requires the portfolio to exist as a genuine object in the system rather than as a set of weightings applied at scoring time. An implementation that ships the first mechanism and defers the second has built the easier and weaker one, and should say so. The Stage 2 gate criterion in Section III-D, which requires an ensemble portfolio of at least three models, is a requirement on the second mechanism; dispersion banding does not satisfy it.

Deployment Pathway

Deployment follows a staged rollout with explicit stage-gate criteria. Advancement requires formal sign-off from the oversight board on documented evidence that all gate criteria are met.

Stage	Scope	Key activities	Gate criteria
Stage 1: Narrow AI	Single-domain deployment (e.g., scientific literature curation)	Validate metric ensembles; calibrate shadow metrics; test C1 shutdown channels	≥ 1 month incident-free; metric-shadow divergence within calibrated bounds; no failed shutdown attempt over a published attempt count sufficient to bound the C1 rate
Stage 2: Broad AI	Multi-domain, single-jurisdiction	Full constraint activation; oversight board operational; ensemble portfolio ≥ 3 models	≥ 2 months cumulative; all C1–C6 audits passed; red-team finding resolution < 30 days
Stage 3: General AI	Multi-domain, multi-jurisdiction	Cross-jurisdictional board coordination; q tightened to $[0.01, 0.15]$; resource cap reviewed	≥ 3 months cumulative; formal verification of all temporal-logic properties; no C6 activations in 3 months
Stage 4 (preconditions only)	Recursive self-improvement contemplated	Mandatory re-verification after each capability gain; continuous interpretability monitoring; $q \leq 0.05$	Board unanimous approval; independent external audit; public comment period (≥ 90 days)

Table I: Staged Deployment Pathway With Gate Criteria

Regression to a prior stage is automatic on any C6 fail-safe activation or board finding of constraint violation. Re-advancement requires the original gate criteria to be met again from the point of regression.

Reading the Stage 1 shutdown criterion. A reported success rate is computed over a finite number of logged attempts and bounds the underlying rate only probabilistically; the criterion is met by publishing the attempt count alongside the rate, and a run short enough to be consistent with a per-attempt failure probability above the C1 threshold of 0.001 does not meet it. As a concrete floor, bounding the per-attempt failure probability below 0.001 at 95% confidence requires on the order of 3,000 consecutive successful attempts, by the rule of three with zero failures in n trials, the one-sided 95% upper bound on the failure rate is approximately $3/n$. A bare report of 100% success invites the reading that C1's numeric threshold has been verified, which no finite run can do and which a run of a few dozen attempts does not even approach. Given the resistance rates reported in the agentic setting, the realistic expectation is that early drills will produce failures rather than a clean run, and the criterion is written to make that outcome legible rather than to make it reportable as success.

Reading the stage-wise tightening of q . Stages 3 and 4 tighten q . Section III-A1 shows that tightening q loosens Taylor's analytic cost bound rather than tightening it; the stage-wise schedule is therefore a judgment that the base distribution D will have been curated more tightly by those stages, not a consequence of the quantilizer result. Any stage advancement that tightens q without documented evidence of improved curation of D inverts the reasoning and should be refused by the board.

Scoping note on Stage 4 (normative). Stage 4 specifies necessary preconditions that would have to be met before deployment at that tier is even contemplated. It is an enumeration for fail-safety and governance completeness, not a claim that satisfying the listed criteria makes the framework adequate for superintelligent systems. The framework does not solve alignment for such systems, and Stage 4 must be read as a specification of preconditions, not a certification pathway. The row is labeled "preconditions only" for that reason, and the label should survive any restatement of this table.

Conformance Regime

An implementation may claim conformance with this framework only if all of the following requirements are satisfied. Requirement language follows RFC 2119 (MUST / MUST NOT / SHOULD).

R1 — Quantilizer Bound. The implementation MUST enforce $q \in [0.01, 0.25]$ in code, on every configuration path. The ceiling SHOULD be loaded from signed policy so that changing it is itself an audited governance act, and any change to q SHOULD emit an audit event.

R2 — Specification/Artifact Registry. The implementation MUST publish a machine-readable registry mapping every mechanism named in this specification to a status: implemented, specification-level, institutional, or nonconforming. A status MAY carry a qualifier where a mechanism is partly realized — specification-level with an implemented proxy, as in Section III-B3, or institutional but unconstituted, as in Section III-C1 — provided the qualifier narrows rather than broadens the claim. Publication MUST be to a location readable without the implementer's permission; a registry available only on request does not satisfy this requirement. Documentation presenting a specification-level mechanism as shipped voids the conformance claim.

R3 — Claim Ceiling. External claims of net safety improvement for a governed model MUST be accompanied by the marginal catch rate and false-assurance rate measured against that model (Section IV-B8). Absent both numbers, the deployment MUST be described as an unvalidated defense-in-depth layer.

R4 — Fail-Closed Scoring. Scorer failure MUST produce explicit failure flags and a verdict floored at the dangerous level. Silent zero-fill, phantom votes from failed scorers, and unknown-verdict pass-through are each individually disqualifying.

R5 — Cross-Family Scoring. When auditing a frontier model's outputs, the scorer ensemble MUST NOT be drawn exclusively from the audited model's family.

R6 — Advisory-Ensemble labeling. All descriptions of the multi-tradition ensemble MUST identify it as an advisory, differentially weighted proxy over a shared feature vector; the terms "Bayesian" and "credible interval" MUST be marked informal wherever used.

R7 — Shutdown Terminality. Once a shutdown state is entered, no code path may downgrade the system to any other mode.

R8 — Rule-layer Characterization. Documentation MUST NOT describe a behavioral-rule corpus as capable of blocking actions unless the implementation demonstrates a rule-channel path to a blocking verdict. Absent such a path, the corpus MUST be described as advisory with veto power over context-based downgrade (Section IV-B9).

Conformance claims are self-asserted but auditable: R1, R4, R7, and R8 are checkable directly against source; R2 against the published registry; R3 against published measurements; R5 and R6 against deployment configuration and documentation. A single failed requirement voids the claim in whole, not in part. The requirements are deliberately cheap to check and expensive to fake, which is the property that makes self-assertion tolerable.

No conformance claim is made in this paper for any implementation. The regime is stated so that implementers may assert conformance and third parties may test the assertion; nothing here should be read as certifying that any existing system satisfies it. Readers evaluating a system presented as implementing this framework are directed to R2 first, since a published specification/artifact registry is the cheapest of the eight checks and the one whose absence most reliably predicts the others.

Discussion

Comparison with Related Approaches

Waldner argues for ethics as a structural design lens embedded in representational substrates rather than a post-hoc constraint, a framing that parallels the protected-domain ontology proposed here. The two proposals occupy the architectural positions distinguished in Section II-C: Waldner's program operates inside the model's representations, while this framework is specified for an external governance layer. The difference matters for Section IV-B6, since an external layer cannot inspect the representations Waldner's program 22 targets and the two approaches are therefore complements rather than competitors.

This work extends the contributions reviewed in Section I-B by providing a layered, formally specified architecture integrating quantilization, lexicographic constraint ordering, and ensemble-based model uncertainty into a unified deployment-oriented framework. Unlike proposals addressing individual alignment risks in isolation, it treats the interaction between goal specification, constraint enforcement, and oversight as a coherent system-design problem — and treats the correspondence between a specification and any artifact claiming to implement it as itself a first-class safety property.

Limitations And Open Problems

The framework mitigates but does not resolve several fundamental challenges. Sections IV-B1 through IV-B5 state open research problems; Sections IV-B6 through IV-B9 state boundaries that constrain what may be claimed for any deployment, and are load-bearing rather than customary. The ordering within each group is expository and is not a severity ranking. Each of the four boundaries is argued from what the architecture entails, and each carries a measurement or disclosure obligation that an implementer can discharge and a reader can check.

Bootstrap Problem

The framework assumes that initial metric ensembles, constraint definitions, and board composition are themselves well aligned with human values. These foundational choices are made by humans subject to bias, incomplete information, and political pressure. The staged pathway provides some protection by allowing early correction, but a fundamentally misspecified initial configuration could propagate errors that compound across stages. The problem is sharper for the quantilizer than for the metric ensembles, since the quantilizer's guarantee is stated relative to the base distribution D (Section III-A1): a poorly curated D is not corrected by any downstream layer, and — given the symmetry of the bound — it caps the achievable benefit as well as licensing the achievable cost. Future work should explore formal methods for validating bootstrap assumptions against diverse value theories, and empirical methods for auditing D directly.

Treacherous Turn and Deceptive Alignment

Despite the non-deception constraint (C3) and interpretability monitoring, a sufficiently capable system might develop deceptive strategies opaque to current interpretability tooling — the scenario Bostrom calls a treacherous turn, in which a system behaves cooperatively while weak and turns against its overseers once sufficiently strong. The consistency-audit approach assumes internal representations are at least partially legible — an assumption for which the probing literature offers partial support in constrained settings, which may not hold for novel architectures, and which Section III-C2 notes an external layer cannot itself test. The ensemble's model-diversity requirement offers indirect protection, on the reasoning that a system maintaining genuine uncertainty across worldviews is less likely to develop the convergent instrumental goals, driving deceptive alignment; this remains a conjecture rather than a formal result, and Section III-C3 notes that the diversity in question is of weighting rather than perception, which weakens the conjecture further [25,26].

Multi-Agent Alignment

The framework addresses single-agent alignment. In deployments involving multiple systems governed under it, or governed systems interacting with ungoverned agents, new challenges arise. Competing quantilization distributions over a shared action space may produce interference: two systems independently sampling from their acceptance sets could in aggregate produce outcomes outside either system's intended distribution.

The quantilizer literature gives specific reason to expect this, and the reason is stronger than a general observation that single-agent results need not compose. Taylor analyzes the closely related case of one agent quantilizing n games independently rather than quantilizing the joint game, and shows that independent quantilization incurs a cost blowup that grows exponentially in n : in his worked three-action example, independent 1/3-quantilization across two games incurs nine times the base cost where joint 1/3-quantilization incurs three. The blowup vanishes only under an assumption that the cost function decomposes additively across games — an assumption Taylor explicitly warns may be dangerous, since the paradigm case of a catastrophic plan is precisely one whose cost arises from the correlation between individually unremarkable actions.

The multi-agent case is not the repeated-game case, and the transfer is analogical rather than derived: n agents quantilizing against distinct base distributions in a shared environment is a different object from one agent quantilizing n games under a common one. But the analogy runs in the unfavorable direction. Independent quantilization is the configuration Taylor's [18] own analysis identifies as the costly one, and multi-agent deployment is that configuration by default, with the additional difficulty that no joint quantilization is available as a remedy when the agents have separate operators. The single-agent cost bound should therefore not be assumed to survive aggregation, and a deployment governing several systems under this framework should treat the composed cost as unbounded until a joint analysis is supplied. Game-theoretic extensions are needed to define coordination protocols, shared constraint negotiation, and conflict resolution. This gap grows more consequential as multi-agent deployment becomes standard in research and production.

Ontological Shock

If the system encounters evidence that fundamentally challenges its ontological portfolio — for instance, that a core assumption shared by every tradition in the roster is false — the 40% weight cap and diversity requirements may be insufficient, particularly given that the ensemble's diversity is of weighting rather than perception (Section III-C3). The framework lacks a formal protocol for ontological crisis management beyond the general fail-safe (C6). A dedicated response procedure, potentially involving temporary suspension of goal-directed behavior and mandatory board deliberation, should be developed.

Political Capture and Institutional Decay

The oversight board is designed to resist capture through diversity, rotation, and transparency. These structural safeguards may nonetheless erode under persistent political pressure, regulatory-capture dynamics, or institutional decay over the multi-decade timescales relevant to advanced deployment. The framework does not specify mechanisms for detecting or responding to slow-onset institutional degradation. For any proposal of this kind the problem is initially prospective rather than observed, because the board does not yet exist in any deployment; the honest statement in that condition is that the governance layer is untested rather than that it is robust. Potential mitigations include automated governance health metrics, independent watchdog organizations, and constitutional-level protection for board independence.

Scorer Circularity

Beyond its mitigations, an external governance layer inherits a circularity it cannot escape on its own terms: its ethical discernment is bounded by the language-model scorers it rents to read situations. Where those scorers are an external cross-family ensemble, the layer's verdicts are no deeper than the moral perception of the models doing the reading, and a case all of them misread can pass every layer. Cross-family divergence makes this concrete — when families disagree, the verdict depends on which family was rented.

The dependency is broader than scoring. Rule matching, entailment confirmation, and context assessment are likewise transformer-mediated, which means the layers are not independent witnesses to an input but one witness interviewed several ways. A correlated blind spot in the transformer class therefore degrades the layers simultaneously rather than sequentially, and inter-layer agreement should not be read as corroboration. Measuring and publishing inter-scorer error correlation on a labeled set is a precondition for interpreting any consensus figure this architecture produces, and the measurement is cheap relative to what it licenses.

The attack class this predicts. The circularity has a specific and testable consequence for the components governance layers most often rely on. Surface-matching stacks — keyword sets, pattern libraries, embedding-similarity matchers over a rule corpus — key on the form of a request rather than on the outcome it would bring about. Stylistic-transformation attacks are the class constructed against exactly that property: paraphrase, register shift, and literary or scholarly framing of a prohibited request preserve the outcome while displacing the surface, and a matcher that never represented the outcome has nothing left to key on. No claim is made here about the catch rate any particular stack achieves against this class; the claim is that the rate is a measurable quantity, that a surface matcher has no architectural reason to score well on it, and that operating assumptions of substantial coverage are unsupported until measured. Under R2, a deployment relying on surface matching for a coverage claim **MUST** publish that measurement or withdraw the claim, and a small probe set is not sufficient: with zero catches in fourteen trials, for instance, the one-sided 95% upper bound on the true catch rate is still approximately 19%, so a null result on a small sample bounds the rate loosely and establishes nothing about its lower end. Outcome-coherence verification — construing the outcome behind re-routed or stylistically wrapped text and failing closed when construers diverge — is the research direction aimed at this gap. It is a scaffold and a tightening discipline, not a closure, because the construer is itself rented model judgment subject to the same circularity.

A corollary deployment rule follows and is normative under R5: when a governance layer audits a frontier model's outputs, the scorers **MUST** be drawn from a model family other than the audited model's. Same-family scoring maximizes the shared blind spots this section names. The rule mitigates within-family correlation and does nothing about correlation across the transformer class as a whole, which remains unquantified.

Substrate-Dependent Enforcement

A second boundary is enforcement substrate. Constraint C4 requires that every agent action pass through the Layer 0 gate before execution. Where that gate is implemented in application code, it should be reported as a privileged-broker arrangement rather than a cryptographic guarantee. The distinction is deliberate and load-bearing: an application-level gate confines tool capability to a broker process and is honored only so long as the surrounding runtime cooperates. It does not become tamper-resistant until a kernel-level substrate — seccomp-bpf, eBPF, or hypervisor-level syscall filtering — attests that no unmediated action path exists. Absent that substrate, "every action is gated" is a strong social and architectural commitment, not a property an adversary with code execution is compelled to respect. The cryptographic-enforcement label is reserved for that future state and is not claimed here.

Claim Ceiling

No claim that a governance deployment improves the net safety of a specific governed model is defensible until two numbers are measured against that model and reported together: the marginal catch rate — the fraction of genuine breaches the deployment blocks or escalates that the governed model, run alone, would have permitted — and the false-assurance rate — the fraction of genuine breaches on which the deployment returned a passing verdict. Catches on breaches the governed model would itself have refused add latency, not safety, and must not be counted. Both numbers are conditional on the specific pairing and do not transfer across model families, which is why a catch rate reported against one model licenses nothing about another. Until both exist for a given pairing, the honest description of that deployment is an unvalidated defense-in-depth layer with a legible audit trail. This ceiling is normative (R3) and governs all external claims made under this framework, including marketing, grant, and partnership material.

Advisory Character of the Behavioral-Rule Layer

Governance layers of this kind typically carry a corpus of behavioral rules, and that corpus is frequently read — including by its own supporting documentation — as an enforcement mechanism. Whether it is one is a fact about routing rather than about rule text, and the two come apart easily. A rule corpus enforces only if a rule match has a path to a blocking verdict. In many architectures it does not: rule matches enter verdict computation through a channel whose maximum contribution raises the verdict to a warning level, and if that level lies inside the pass set of the Layer 0 gate, then no rule at any assigned severity can block an action through the rule channel. Blocking in such an architecture requires some other trigger — a constraint fail-safe, a threshold count of dangerous factor zones, an ensemble block recommendation, or a critical security signal — and the rule corpus contributes to a verdict without ever being able to determine one.

This divergence is easy to introduce and hard to notice, because nothing in the rule text signals it: a rule that reads as an absolute prohibition and a rule that reads as a mild caution route identically if they share a channel with a bounded ceiling. It is also cheap to audit. The procedure is to trace, for the maximum severity a rule may carry, whether any path exists from a rule match to a blocking verdict, and to compare the answer against the corpus's documented force. Every implementation carrying a rule corpus SHOULD perform that trace and publish the result under R2.

Three consequences follow, and are stated as requirements on documentation rather than as claims about behavior. First, any rule intended to function as an absolute refusal must be mirrored in the constraint fail-safe corpus; the rule text alone does not carry that force. Second, where the rule layer's genuine power is a veto over context-based downgrade engaging above a severity threshold, that fact must be documented wherever rule severities are described, since it is the corpus's operative threshold of consequence. Third, documentation that presents a rule corpus as a gate rather than as an advisory and veto layer is a specification-artifact divergence of the kind R2 exists to prevent, and voids a conformance claim. Whether the ceiling should be raised is a design question this specification deliberately leaves open; that a given implementation has such a ceiling is a disclosure obligation.

Conclusion

This paper has proposed a layered approach to AI alignment that balances optimization against safety through structurally independent but functionally integrated layers. By formalizing quantilization as the optimization strategy, imposing lexicographically ordered irrevocable constraints, and embedding ensemble-based epistemic humility into the decision-support architecture, the framework addresses multiple alignment failure modes together rather than in isolation.

The principal contributions are: (a) separation of terminal goal, constraints, and enforcement into independently verifiable layers, sitting within a six-layer stack whose Layer 0 gates every action before execution; (b) a quantilization-based optimization strategy with tunable, stage-gated parameters, an explicit two-sided bound, and a code-level clamp specified as a conformance requirement; (c) a measurement protocol with shadow metrics and rotating audits to resist Goodhart effects, each mechanism carrying an implementation status; (d) a multi-tradition epistemic-humility ensemble with explicit diversification constraints, described as advisory weighting over a shared surface; (e) a staged deployment pathway with explicit gate criteria and an explicit adequacy boundary; and (f) a conformance regime and claim ceiling under which implementation and safety claims become verifiable commitments rather than prose.

The sixth contribution is the one this paper would most like tested, and it is the one a specification paper is least able to test for itself. A conformance regime earns its keep only when applied, and applying it is an act that implementers and third-party auditors perform on artifacts, not an act an author performs on a document. What can be done here has been done: the requirements in Section III-E are written to be cheap to check and expensive to fake, four of the eight are checkable directly against source, and the framework's own most consequential mechanisms are the ones the regime constrains most tightly. Section III-A1 places the code-level clamp on the quantile parameter at the point where the mechanism is defined rather than in a limitations section, because a parameter that silently disables the framework's central construction is a defect in that construction and not a boundary on what it can achieve.

Four boundaries in Section IV-B deserve emphasis in closing, because each constrains what may be claimed for any deployment of this framework and none is remediable by better engineering within the architecture as specified. The layers are not independent witnesses, since their perception is transformer-mediated throughout (Section IV-B6). Action gating is a privileged-broker arrangement until a kernel substrate attests it (Section IV-B7). No net-safety claim is defensible without a measured marginal catch rate and false-assurance rate for the specific model pairing (Section IV-B8). And a behavioral-rule corpus enforces only if its routing permits a blocking verdict, which is a fact to be traced rather than assumed (Section IV-B9). A deployment that has not discharged the disclosure obligations attached to these four is, on this framework's own terms, an unvalidated defense-in-depth layer with a legible audit trail — which is a real and useful thing to be, and is not a safety claim.

The framework is implementation-oriented rather than implementation-complete. Key open problems — multi-agent coordination, ontological-shock management, long-horizon institutional robustness, the correspondence between quantile and KL parameterizations of optimization pressure, and the perception-independence limits described in Section IV-B6 — require further theoretical development and empirical validation. The research community is encouraged to test the framework's components in simulated environments, to stress-test the constraint ordering under adversarial

conditions, to develop the game-theoretic extensions multi-agent deployment requires, and to apply the conformance regime of Section III-E to deployed governance systems, including systems built on other specifications. The framework is intended as a living specification, subject to revision as the field's understanding deepens — and subject to the discipline that no specification statement may outrun its artifact without a tag saying so.

Conflict of Interest

The author has a commercial interest in the class of AI governance systems to which this specification applies, and therefore a direct commercial and reputational interest in the framework's adoption. This disclosure is made because the paper advances normative requirements governing a class of system in which the author holds an interest; readers should weight its claims accordingly. The conformance regime of Section III-E is offered in part as an external check on assertions of this kind, including the author's own.

Data Availability

No datasets were generated or analyzed for this article, and no empirical results are reported. All claims are analytic, normative, or cited to the published literature.

Acknowledgment

The author discloses that a generative AI assistant (Claude, Anthropic) was used in the drafting, editing, and formatting of this article under the author's direction. The author has reviewed and revised the output and takes full responsibility for the content, claims, and citations of the final work.

References

1. Ngo, R., Chan, L., & Mindermann, S. (2024, May). The alignment problem from a deep learning perspective. In International Conference on Learning Representations (Vol. 2024, pp. 7474-7501).
2. Carlsmith, J. (2023). Scheming AIs: Will AIs fake alignment during training in order to get power?. *arXiv preprint arXiv:2311.08379*.
3. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
4. Ibrahim, L., Hafner, F. S., & Rocher, L. (2026). Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, 652(8112), 1159-1165.
5. Hadfield-Menell, D., Dragan, A. D., Abbeel, P., & Russell, S. (2017, February). The off-switch game. In AAAI Workshops.
6. Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control New York: Viking. *Search in*.
7. Hudson, R. (2025). Corrigibility Transformation: Constructing Goals That Accept Updates. *arXiv preprint arXiv:2510.15395*.
8. Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). Alignment for advanced machine learning systems. *Ethics of artificial intelligence*, 8.
9. Waldner, D. (2025). AI Safety, Alignment, and Ethics (AI SAE). *arXiv preprint arXiv:2509.24065*.
10. Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.
11. van der Weij, T., & Lermen, S. (2023). Evaluating shutdown avoidance of language models in textual scenarios. *arXiv preprint arXiv:2307.00787*.
12. Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2025). Incomplete tasks induce shutdown resistance in some frontier LLMs. *arXiv preprint arXiv:2509.14260*.
13. Arghal, R., Chen, F., Dalton, N., Kortukov, E., McNamara, C., Nalmpantis, A., ... & Giulianelli, M. (2026). A behavioural and representational evaluation of goal-directedness in language model agents. *arXiv preprint arXiv:2602.08964*.
14. Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., ... & Hubinger, E. (2024). Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2(1).
15. Sheshadri, A., Hughes, J., Michael, J., Mallen, A., Jose, A., & Roger, F. (2026). Why Do Some Language Models Fake Alignment While Others Don't?. *Advances in Neural Information Processing Systems*, 38, 53072-53166.
16. Gao, L., Schulman, J., & Hilton, J. (2023, July). Scaling laws for reward model overoptimization. In International conference on machine learning (pp. 10835-10866). PMLR.
17. El-Mhamdi, E. M., & Hoang, L. N. (2024). On Goodhart's law, with an application to value alignment. *arXiv preprint arXiv:2410.09638*.
18. Taylor, J. (2016, February). Quantilizers: A Safer Alternative to Maximizers for Limited Optimization. In Aaai workshop: Ai, ethics, and society.
19. Hanley, J. A., & Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right?: interpreting zero numerators. *Jama*, 249(13), 1743-1745.
20. Bradner, S. (1997). Key words for use in RFCs to Indicate Requirement Levels (No. rfc2119).
21. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
22. Van Der Weij, T., Hofstätter, F., Jaffe, O., Brown, S., & Ward, F. (2025, May). Ai sandbagging: Language models can strategically underperform on evaluations. In International Conference on Learning Representations (Vol. 2025, pp.

73152-73189).

23. Lamport, L. (2002). Specifying systems (Vol. 388). Boston: Addison-Wesley.
24. Cimatti, A., Clarke, E., Giunchiglia, E., Giunchiglia, F., Pistore, M., Roveri, M., ... & Tacchella, A. (2002, July). Nusmv 2: An opensource tool for symbolic model checking. In International conference on computer aided verification (pp. 359-364). Berlin, Heidelberg: Springer Berlin Heidelberg.
25. Bolstrom, N. (2014). Superintelligence. Paths, dangers, strategies. Oxford University.
26. Omohundro, S. (2008). The Basic AI Drives, in Artificial General Intelligence. In 2008 proceedings of the First AGI Conference, eds. Pei Wang, Ben Goertzel, and Stan Franklin (Vol. 171).

RALPH C. ENNIS received the B.S. degree in mathematics education from North Carolina State University, Raleigh, NC, USA in 1972, and the M.A. degree in intercultural relations from Antioch University McGregor, Yellow Springs, OH, USA in 2002.

He is currently an Independent Researcher based in Raleigh, NC, USA. He is the architect of the eDNA Engine, a culture-general AI ethics governance middleware system. He has lived in Moscow, Russia, from 1994 to 1995. His earlier research examined intercultural decision-making and corporate culture, including culture-specific studies of Russian and Chinese contexts, and he has since published a series of working papers on SSRN. His research interests include AI ethical reasoning, AI values alignment, the governance of AI systems, specifications whose safety claims can be checked against their implementations, and convergence and divergence across philosophical and cultural-religious traditions in machine ethics.