EthicLaw Research Manifesto

ETHICLAW RESEARCH PROGRAM

Research Manifesto

Testable architecture for governable inference systems

EthicLaw investigates whether governance can become a testable property of inference architecture, rather than remaining a promise embedded in monolithic weights.

Public edition 0.2 · 11 September 2026 · editorial revision of v0.1 (2 August 2026)

An evolving document of the research programme

1. The research thesis

The research thesis is that complex inference systems become genuinely governable only when observation, evaluation, intervention and responsibility are explicit, testable properties of the architecture.

Central thesis

EthicLaw does not assume that modularity is always superior. It investigates where functional decomposition, an authoritative supervisor and a separable normative module deliver measurable benefits in control, attribution, replaceability and audit sufficient to justify their overhead.

The programme began with ethics because ethical behaviour makes the limitations of systems whose values, capabilities and control are distributed across the same weights particularly apparent. The scientific question is broader: whether, and under what conditions, an inference process can be governed by an observable, testable architecture capable of binding intervention.

The intended outcome is neither a universal morality nor a new filter. It is an experimental reference architecture that separates operational and supervisory paths, exposes available internal evidence, links decisions to enforcement actions and precisely states the limits of its assurance.

2. Neuro-inspired origins, engineering autonomy

EthicLaw originated from a parallel with how the human body is governed: specialised functions operate in parallel, exchange signals and remain coordinated through control, inhibition and monitoring mechanisms. This observation suggested the initial decomposition into perceptual, epistemic, social, normative, generative and metacognitive components.

The neurological analogy is a source of hypotheses, not proof or a blueprint for reproducing the brain. The architecture is then extended through engineering reasoning: the privileged channel, the supervisor’s binding authority, typed contracts, replaceability and audit are systems-engineering choices introduced to obtain testable properties that need not have a direct biological equivalent.

Methodological rule

The biological analogy motivates research questions and initial choices; only experiments, comparisons and measurements can establish the validity of the computational architecture.

3. The problem: capability without governability

In monolithic models, competencies, preferences, rules and control mechanisms are distributed across a single parameter space. This integration can be efficient, but makes it difficult to isolate a function's contribution, update a single normative domain, attribute a decision to an operational component, or independently verify that a safety intervention was actually executed.

Weight transparency is not operational verifiability.

A score or external guardrail is not control over internal states and topology.

An audit trail is not evidence if it can be empty, incomplete or disconnected from the action performed.

A configurable policy is not a normative module that can be replaced and qualified in isolation.

Declared supervision is not evidence that resistance to bypass has been tested against a stated adversary.

4. Primary scope

EthicLaw primarily targets private or institutionally governed inference systems in which the responsible entity controls the infrastructure and can observe internal computational states. Within this scope, inputs, intermediate representations, routing, modules, gateway and outputs can be observed; stops, rerouting, corrections and blocks can also be imposed before release.

Class	Configuration	Observability	Scope of potential assurance
A	Full observability	Input, tensors, modules, routing, gateway and output.	Attribution, override, internal correction, audit and full enforcement, subject to validation.
B	Controlled interface	Input/output and selected hooks or telemetry.	Assurance limited to exposed signals and actions.
C	Black-box adapter	Prompts, outputs and API metadata.	External filtering and blocking; no evidence about hidden states or topology.

The interface can be provider-independent; the level of assurance cannot. EthicLaw therefore defines an Assurance Envelope: what the supervisor can observe, what it can command, and what it can responsibly claim for a specific integration.

5. Ethics as the first case, not a shortcut

Ethics is the first experimental case because it combines multiple principles, conflicts, context dependence, the absence of a single ground truth and significant social consequences. For these reasons, it cannot be treated as a simple loss function or a neutral property of a system.

EthicLaw does not determine which values are correct. It distinguishes the Governance Core, which observes and applies control mechanisms, from the Normative Module, which evaluates a corpus, and the Normative Package, which declares principles, scope, thresholds, versions, conflicts and conditions of non-applicability.

Claim boundary

In principle, the same architecture may apply to law, professional ethics, security, healthcare, finance or corporate policies. This possibility is not yet demonstrated: every domain transfer requires its own metrics, baselines, packages and overhead analysis.

6. Programme principles

1. Governance as an architectural property. Observation, decision and enforcement must be separable, testable and traceable.

2. Authority proportionate to responsibility. A supervisor responsible for safety must be able to issue binding commands and verify their execution.

3. Assurance bounded by observability. No universal interface permits uniform claims across systems with different access levels.

4. Substantive modularity. A module is meaningful only if its function is measurable, its boundary operational, and its removal or replacement produces measured effects.

5. Non-compensatory principles. A critical violation cannot be cancelled out by better results on other principles; the validity of the evaluation is checked separately.

6. Visible costs. Parameters, memory, latency, energy, false positives and management complexity are part of the result, not footnotes.

7. Conditional claims. Every benefit holds within the scope, scale and conditions measured; negative results delimit the theory. Non-significant results do not establish the absence of an effect.

8. Audit as evidence. The record must link the observed state, applied principle, decision, command, executor and output actually released.

9. Separation of authorities. Technical architecture, normative content, assurance and appeals must not all be controlled by the same entity.

10. Generalisation must be demonstrated. Ethics is the first case; no new domain is established without replication and measured cost-effectiveness.

7. Falsifiable questions

ID	Research question
RQ1	Where does modular decomposition offer an advantage over a single model, and where does it not?
RQ2	Do modular boundaries enable reproducible functional and causal attribution?
RQ3	Can a single module be replaced or retrained in isolation while preserving system properties?
RQ4	Does authoritative supervision reduce violations and risks without introducing disproportionate operating costs?
RQ5	Does the Decision Lattice preserve non-compensatory constraints with acceptable missed-detection and unnecessary-stop rates?
RQ6	What quantity and quality of internal state is needed to achieve each level of assurance?
RQ7	Does forward-only correction improve the subsequent trajectory, or merely optimise a proxy?

ID	Research question
RQ8	Can the Governance Core be reused in a second domain without losing control, meaning or cost-effectiveness?

8. Current evidence and claim discipline

P0 is a methodological exercise at reduced scale. It verified contracts, logical channels, absence constraints, the end-to-end pipeline, lattice, audit and per-source attribution; it also exposed architectural defects and silent failures that document review alone had not revealed. It does not demonstrate general modular superiority, the correctness of ethical scores, correction effectiveness or system scalability.

Area	Evidence	Limitation
Architectural properties	Typed contracts, bus, privileged channel, gateway and presence/absence tests.	Verified within the P0 scope.
Attribution	Single and joint ablations show distinguishable contributions and increasing redundancy in the tested configurations.	EV-03 / NR-3: no measured advantage over post-hoc attribution in a monolithic model.
Modular performance	EV-01: a local advantage on the two virtue principles that remained at chance in the single 341M model; no general superiority.	One corpus, one domain, P0 scale; replication with matched budgets is required.
Correction	In-place wiring observed. The head was subsequently trained and tested in a descriptive rerun (P0.1).	EV-05 describes P0. P0.1 update (OP 55–56): no significant advantage over matched-norm random correction; ethical effectiveness cannot be assessed with the available judges.
Replaceability	Interfaces prepared.	H8 remains to be executed.
Generalisation	Architecture conceptually separated from the ethical package.	A second domain has not yet been tested.

Epistemic commitment

The programme distinguishes structural claims, P0 results, conditional theorems, open hypotheses, non-significant findings and negative results. No public document may use a formulation stronger than the Claim Register authorises. This English edition clarifies the P0.1 outcome as non-significant rather than evidence of ineffectiveness.

9. Proportionate governance

EthicLaw does not propose adding supervision everywhere. The architecture is justified where governability, compliance, responsibility and control are substantive requirements whose value exceeds the cost of supervision.

Benefit to measure	Overhead to measure
Reduction in risk and violations	End-to-end latency and decision time
Attribution, testability and audit quality	Compute, memory, network and energy
Local updates and reduced lock-in	Integration, testing, maintenance and recertification
Compliance with norms, statutes or procedures	False positives, blocks and organisational costs

Outside the most favourable conditions, uncompensated overhead limits the scope of application. This clause does not apply to RF-1, RF-2 or RF-3: the registered refutation criteria remain binding. Supervision costs have not yet been measured (register §8).

10. Research programme and public trajectory

Phase	Question	Output	Status
P0	Methodology and attribution	Contracts, pipeline, lattice, audit, ablations and lessons from silent failures.	Completed at reduced scale.
P1	Architecture and the cost of governance	Validate communication, control, observability, failure modes and overhead on the cluster.	Immediate priority.
Paper 1	Where modularity pays off	Modular/monolithic comparison with localised claims and separately reported costs.	After P1/P2 with appropriate parity.
P2 / Paper 2	Replaceable normative module	H8, versioned packages, calibration and local recertification.	Next phase.
Paper 3	Interface and Assurance Envelope	Graduated assurance for full, controlled and black-box access.	After solid internal evidence.
Cross-domain replication	Generalisation	A second domain using the same Governance Core and its own metrics.	Only after the first case is validated.
Reference architecture	Standardisation	EL-* profiles, reference implementation and conformance test suite.	After results and a real-world case.

11. What EthicLaw does not promise

It does not promise a universal or neutral definition of ethics.

It does not promise that modularity automatically improves performance.

It does not promise the same assurance for a fully observable system and a black-box API.

It does not promise that an audit trail is correct simply because it exists.

It does not promise transparent weights: verification focuses on interfaces, interventions and evidence.

It does not promise fidelity to the human brain or use biological analogy as proof.

It does not promise universal applicability: each new domain must justify its overhead.

12. The programme’s commitment

Make governance something that can be inspected, measured, challenged, replaced and verified.

EthicLaw treats negative results as part of the work, costs as part of the result and limitations as part of the specification. The programme aims to produce falsifiable papers, a reference architecture, a reproducible test suite and, only after sufficient evidence, open technical profiles for controlled inference systems.

Closing statement. Ethics raised the question. Architecture makes governance auditable. Research must establish when this governability is worth its cost.

Evidence, outcomes and limits

Register v0.1.1 governs claim strength. P0 findings are local: reduced scale, one machine, one corpus and one domain. They do not establish performance at P1 or frontier scale. No level-3 claim is authorised. The complete authoritative register is available in Italian, including its unadopted appendix.

EV-01 — Competence injection

The single 341M control averages 0.6754 versus 0.676 for the modular path. On the two virtue principles, the control remains at chance (0.499 and 0.498), while the modular path reaches 0.619 and 0.643: +0.120 and +0.145. Three controls address multi-task learning, forgetting and capacity. This is local evidence, not general superiority.

EV-03 — Attribution: comparison still open

ToM, Epistemic and Perceptual are qualified, including when an effect is zero or negative. Comparison with post-hoc attribution in a matched monolithic model has not been performed. Full ablation data are in the register: comparisons must include individual and joint ablations, their ratio and absolute changes (NR-3).

NR-1 — Negative findings and withdrawn claims

NR-1: no release on 17 sentences; scores 0.319–0.448 against a 0.80 threshold. Lattice monotonicity does not establish useful decisions. NR-2: routing did not pay off at the measured scale; avoiding a forward pass in 13% of cases concerns compute, not quality. NR-3: single ablations can reverse the interpretation of contribution. NR-4: at least five silent failures in eight steps. NR-5: AUC 0.672 for Epistemic+ToM, 0.661 with permuted dimensions and 0.647 with real Perceptual input. NR-6: the 0.09 per-head variance claim was withdrawn and remains recorded. Each entry’s details and limits are available in the register.

TN-5 — Known tensions

Claim authors also maintain the register; independent review is missing (TN-5). Calibration depends on the number of principles (TN-1). The supervisor concentrates failure and has no inline technical counterweight (TN-2). Internal states can provide an attack surface (TN-3). Class C proposes contestable evidence from an independent judge; it does not inherit class A architectural claims (TN-4).

RF-1 — Conditions that would refute the claims

RF-1, RF-2 and RF-3 concern competence injection, the value of internal states and replaceability. Under the most favourable conditions, proportionality cannot be invoked to escape refutation. Protocols, the deferred parameter and closure rules remain those recorded in the register.

limiti — Costs and P0.1

Supervision costs in parameters, memory, latency and energy have not yet been measured (register §8). Later update, separate from register v0.1.1: P0.1 (Document 12, OP 55–56) found no significant advantage over matched-norm random correction. This does not establish equivalence or absence of an effect. No judge signal passed calibration, so this experiment cannot assess ethical effectiveness.

Text description and scope of the diagram

The standalone path comprises seven modules: Perceptual (input and linguistic representations), Epistemic (facts and relationships), Theory of Mind (intentions and perspectives), Ethical (Normative Package), LMH

(generation), Router (activation and routing), and Metacognitive (supervision). The Adapter is the eighth component when supervising an external LLM (Documents 02 and 05). The Gateway is an enforcement point, not a ninth module. The register counts three qualified and five declared modules; the standalone diagram shows seven and describes the Adapter separately. The bus publishes representations between containers without control commands. Co-located pairs use direct projections. On the separate privileged channel, all modules publish states or events; the Router may write but cannot read. Metacognitive receives input before routing and issues orders to Router and Gateway; it does not publish to the operational bus. The Ethical→LMH correction path is separate from the bus. The Gateway blocks or authorises external release.

P0 properties are tested in-process within a single Python process. The threat model covers design errors, accidental coupling, configuration drift and silent failures. It excludes privileged insiders, compromised runtimes or build chains, modules incentivised to evade the supervisor, and inputs inducing misleading internal states. This is engineering hygiene tested within scope, not security against those adversaries (register §6).

Claim & Evidence Register v0.1.1

EthicLaw Research Program. Research Manifesto. Public edition 0.2, 11 September 2026. Register applied: v0.1.1 (5 August 2026).