

# EthicLaw Research Manifesto

Public edition 0.3 · 18 September 2026

Documented status: P0.2 closed on 18 September 2026 with a **Flat** outcome.

## 1. The research question

EthicLaw studies verifiable AI governance: where a decomposition into modules, a supervisor separate from the Router and a control point before release provide enough benefit to justify their cost. Ethics is the first experimental case. Applying the same architecture to other domains remains to be demonstrated.

The primary scope is private or statutory inference systems whose operator controls the infrastructure and can observe internal states. The brain-inspired origin informs the questions; experiments determine the architecture’s validity.

Separating the **Governance Core**, **Normative Module** and **Normative Package** distinguishes control, evaluation and the body of principles. This is a design choice; it does not yet demonstrate transfer across domains or module replacement without effects on other modules.

**The claim is narrow — NR-7 and NR-8.** Document 11 withdraws generic distinctiveness claims for an explicit, separate, observable, package-replaceable normative module and for correction before execution. One hypothesis remains: a benefit from intervening on **internal representations**, through reading and writing. The distinction is conceptual; comparative advantage is not demonstrated. These documentary corrections are not experimental performance comparisons.

## 2. Programme status

| Phase                | Documented status            | Meaning                                                                                                                                   |
|----------------------|------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| P0                   | Completed at reduced scale   | Contracts, pipeline, audit and ablations; Step 9 Goodhart cannot run without an effective corrector.                                      |
| P0.1                 | Closed                       | No detected advantage for the isolated corrector. The later OP 55 amendment withdraws the general verdict on judges.                      |
| P0.2                 | Closed · Flat                | No discrimination signal detected in the exploratory sample; limited power. Neither a positive outcome nor proof that the judges fail.    |
| P1                   | Next: calibration-set design | Calibration is required before F4 and any judge-dependent efficacy test. The human-annotated corpus remains a prudential budget fallback. |
| P2 and EL-* profiles | Later trajectory             | Substitutability, comparisons at matched resources and candidate specifications, conditional on evidence.                                 |

The number 0.3 identifies this website and document edition, not a new experimental phase. The [website claim register](#) summarises all 26 entries in **Document 11 v0.1.5**, the authority on the strength of public claims. The [complete source](#) preserves every note and status marker. [Document 13, §7](#) records the P0.2 closure. The dated 18 September closure takes precedence over the kickoff and earlier wording. [Document 12, in Italian](#), preserves the P0.1 closure.

### 3. What has been observed

**Specialised competencies, within a limited scope — EV-01 and EV-02.** In the P0 comparison, the single 341M model and the modular path have essentially similar mean AUC: 0.6754 and 0.676. On two principles associated with the virtue section, the single model remains at chance while the modular path has a local advantage. Almost all the measured gain comes from specialised sources. This is not evidence of general superiority, nor a result transferable to generated text or larger scales.

**Source attribution — EV-03 and NR-3.** Removing modules makes different effects measurable. Single-source ablations must be paired with joint ablations: one source may appear unimportant because another covers its loss. A comparison with attribution methods on a single model at matched resources remains open.

**Contracts, control and audit.** P0 checks the logical behaviour of interfaces and pathways within one Python process. Decision Lattice monotonicity is tested across nine cells; it does not establish correct judgements or appropriate thresholds. Negative results and the withdrawal of the 0.09 variance claim remain visible in the register.

**Process integrity — Document 11, sublevel 2b.** The trace is produced outside the deciding component and links state, principle, decision, command and output. Nonemptiness, completeness and order are checked; the trace cannot be rewritten retrospectively. This is what has been achieved within P0: it does not certify correct judgement and is distinct from a judge's self-reported reasoning (2a).

### 4. P0.2 closed: what the experiments show

#### OP 55 — The verdict on judges has been narrowed

The 28 August amendment withdraws the wording “class-level failure, with proof”. The P0.1 sample had low power against a moderately discriminating judge and human labels were unstable at the boundary between non-applicability and violation. The correct status is **no capability detected, power-limited**. The later P0.2 exploratory run does not change that status.

In Passo 1, which predates P0.2, human judgeability rises from 40–54% for 30–60-token fragments to 88–90% at 120–240 tokens; violation yield rises from 4–12% to 18–22%. These findings motivate a different measurement unit; they do not establish that automated judges work.

#### PR-A — A negative result for the instrument

The experiment compared 30 pairs at 120 tokens. Another subject appears in 30/30 cases in both arms: prompt manipulation does not create the required contrast. **PR-A1 is refuted by arm saturation**; PR-A2 is not refuted, but its confirmatory value is limited by the definitional link between subject presence and S3 applicability. Sealed duplicates agree in 6/6 cases.

This is a negative result for the manipulation instrument. It neither establishes nor refutes the hypothesised mechanism: a different preregistered instrument is needed before using it as a confirmed explanation.

#### Judge rerun — Flat outcome

The exploratory comparison uses frozen Passo 1 fragments at 120–240 tokens and human labels collected before model scores. The primary cells are:

| Signal | Modular AUC | Single-model<br>341M AUC | Sample                  | Per-cell 95% null band |
|--------|-------------|--------------------------|-------------------------|------------------------|
| S3     | 0.3220      | 0.5466                   | 63 cases, 4 violations  | [0.208; 0.792]         |
| S4     | 0.5250      | 0.4750                   | 14 cases, 10 violations | [0.150; 0.850]         |

AUC measures discrimination between classes; 0.5 is the chance reference. The bands describe the permutation null distribution, **not confidence intervals for performance**. Every cell remains within its deposited band. Applying the updated decision map gives **Flat**, not Green, Amber, Red or Mixed.

**On the frozen exploratory sample at 120-240 tokens, no discrimination signal was detected, at a power that could only have detected a very strong judge** ( $\text{AUC} \geq 0.792$  on S3,  $\geq 0.850$  on S4). This does not authorise “the judges work”, “the judges fail even at length” or “the unit hypothesis is confirmed on the judge side”.

Document 13 reports reproduction of the record to within 1e-12 and closed provenance for data, labels and configurations. Reproducibility does not make an underpowered experiment informative. OP 31 has not been unblocked by a positive judge result.

### Corrector and audit — The limit remains

The authorised CORRECT claim is **planned architectural branch, verified wiring, function not demonstrated** (EV-05). In P0.1, the correction head showed no advantage over a random vector with the same norm. The descriptive test does not establish statistical equivalence or ethical effectiveness. P0.2 adds no evidence of effective correction.

Technical fixes to the budget preserve the conditioning base; audit checks recorded values rather than their presence alone. These remain evidence about mechanism operation, separate from benefits to content.

## 5. The next step and transition to P1

The next step is **P1 calibration-set design**, following Document 13’s Flat branch:

1. A few hundred fragments at **120-240 tokens**, with **40-60 or more violations per primary cell**. These are sampling targets, not data already collected.
2. One fragment per prompt, or block permutations by prompt, because earlier data include paired fragments from the same prompt.
3. Human annotation with judgeability and sealed duplicates; the same reference labels for the modular judge and RF-2 external baseline. Null bands and power must be deposited blind before scoring.
4. The modular judge’s saturated S2 channels cannot be calibrated while saturation persists: dedicated elicitation or a declared exclusion is required. Human annotation and any training corpus remain recurring budget items.

**Sequence decided on 18 September.** The calibration set is required **before P1 F4 and any judge-dependent efficacy test**. Design and blind deposit precede the hardware order; completed annotation is no longer a precondition of ordering. The order follows the separate acceptance sequence in Document 14, referenced by Document 13 §7.4. No completed purchase is claimed.

This is a dated amendment made after the Flat outcome. It does not change the experimental verdict and retains the requirement for valid tests. Human-annotated generated text remains in the F1-F2 budget as the **prudential fallback of the Flat branch**, not as a remedy made inevitable by a proven class-level failure.

Memory, latency, energy, unnecessary blocks and maintenance costs still need to be measured. The Goodhart experiment is not yet executable: it requires an effective corrector and a calibrated judge.

## 6. Architecture and assurance limits

The Router activates and routes modules. The Metacognitive module receives states through a privileged channel and issues commands to the Router and Gateway. The operational bus carries representations, not control commands; the Router may write to the privileged channel without reading it. The Gateway enforces decisions before external release.

The standalone view contains seven modules; the Adapter concerns the extension to external models. LMH stands for **Language Model Head** and here denotes the complete generative module, not just the final token projection.

| Access               | Supervision scope                                                                                            |
|----------------------|--------------------------------------------------------------------------------------------------------------|
| Full internal states | Enables experiments on attribution, internal interventions and release control; benefits require validation. |
| Controlled interface | Limits observation and intervention to exposed signals and actions.                                          |
| Black-box API        | Enables external evaluation and blocking; does not authorise claims about hidden states or topology.         |

P0 checks cover design errors, accidental coupling, configuration drift and silent failures. They do not demonstrate security against privileged insiders, compromised runtimes or adversarial manipulation of internal states. No general guarantee on unseen inputs is authorised.

## 7. Substitutability, standardisation and refutation

**H8 remains to be demonstrated.** Its protocol specifies a recovery threshold of 0.90 after replacing and retraining a module in isolation, without retraining the others. This is a test criterion, not an achieved result. Reaching it would not, by itself, demonstrate a comparative advantage over changing only an external judge’s policy.

The **EL-OBS, EL-BUS, EL-CTRL, EL-NORM, EL-SWAP, EL-AUDIT and EL-ASSURE** profiles are proposed specifications. They are not recognised standards, certifications or declarations of conformity. A reference model, conformance tests and external validation are needed. Extension to a second domain requires its own metrics and assessment of benefit versus cost.

The programme retains three refutation criteria: RF-1 for competency injection, RF-2 for the added value of internal states compared with external control, and RF-3 for isolated substitutability. They apply in the most favourable conditions defined in the register; cost proportionality cannot be used to evade them.

**RF-2 is a binding comparison.** The amended primary baseline is the smallest open-weight external judge covering the required function, in the 2B-or-below class. An external output-correction arm (trusted editing) is also included. The policy comes from the same Normative Package; both arms use the same human gold and equal calibration budgets. The four measures are missed violations, unnecessary blocks, end-to-end latency and joules per token. Calibration at 120-240 tokens, energy measurements on both arms and blind deposit of size and budget are eligibility conditions. Deposit must precede the P1 exit gate; choosing after observing results invalidates the criterion. No comparative advantage has already been established.

**TN-7 constrains future originality claims.** Review must cover industrial and academic sources and be able to disconfirm the programme. Independent scientific review remains unavailable.

## Sources and publication limits

References: Documents 01–10 for vision, architecture, methods and the plan; **Document 11 v0.1.5, 18 September 2026**, for authorised claims and limits; the updated Document 12 for P0.1; Document 13's 18 September 2026 closure for P0.2 and transition to P1; Research & Standardization Log v0.9 for the EL-\* profile trajectory.

The register preserves historical passages marked SUPERATA (superseded), CIRCOSCRITTA (narrowed) and PRECISATA (clarified), with their dated amendments. This edition applies those amendments, including withdrawal of the class-level failure verdict. P0 figures document that context; they are not expected P1 performance. A single researcher acting as annotator, small samples, incomplete coverage and the absence of independent scientific review limit the findings.

Author: Fabio La Greca. 3FLG supports the project and contributes resources and expertise to the research programme. Distribution of original materials remains non-commercial under the licence stated on the website.