EthicLaw Research Manifesto

ETHICLAW RESEARCH PROGRAM

Research Manifesto

Architettura verificabile per sistemi inferenziali governabili

EthicLaw studia se la governance possa diventare una proprietà auditabile dell'architettura inferenziale, invece di restare una promessa incorporata in pesi monolitici.

Edizione pubblica 0.2 · 11 settembre 2026 · revisione editoriale di v0.1 (2 agosto 2026)

Documento evolutivo del programma di ricerca

1. La tesi di ricerca

I sistemi di inferenza complessi diventano realmente governabili soltanto quando osservazione, valutazione, intervento e responsabilità sono proprietà esplicite, testabili e auditabili dell'architettura.

Tesi centrale

EthicLaw non assume che la modularità sia sempre superiore. Verifica dove una decomposizione funzionale, un supervisore autoritativo e un modulo normativo separabile producano benefici misurabili di controllo, attribuzione, sostituibilità e audit sufficienti a giustificarne l'overhead.

Il programma nasce dall'etica, perché il comportamento etico mette in evidenza con particolare forza i limiti di un sistema nel quale valori, capacità e controllo sono distribuiti negli stessi pesi. Il problema scientifico è però più generale: comprendere se e a quali condizioni un processo inferenziale possa essere governato da un'architettura osservabile, auditabile e capace di intervento vincolante.

Il risultato atteso non è una morale universale né un nuovo filtro. È una reference architecture sperimentale che separa il percorso operativo dal percorso di supervisione, rende visibili le evidenze interne disponibili, collega le decisioni alle azioni di enforcement e dichiara con precisione i limiti delle proprie garanzie.

2. Origine neuro-ispirata, autonomia ingegneristica

EthicLaw deriva da un parallelismo con il sistema di governo del corpo umano: funzioni specializzate operano in parallelo, scambiano segnali e rimangono coordinate da meccanismi di controllo, inibizione e monitoraggio. Questa osservazione ha suggerito la prima decomposizione in componenti percettive, epistemiche, sociali, normative, generative e metacognitive.

La similitudine neurologica è una fonte di ipotesi, non una prova e non un progetto di riproduzione del cervello. L'architettura viene successivamente estesa razionalmente: il canale privilegiato, l'autorità vincolante del supervisore, i contratti tipizzati, la sostituibilità e l'audit sono scelte di systems engineering introdotte per ottenere proprietà testabili che non hanno necessariamente un equivalente biologico diretto.

Regola metodologica

L'analogia biologica giustifica domande di ricerca e scelte iniziali; soltanto esperimenti, confronti e misure possono giustificare la validità dell'architettura computazionale.

3. Il problema: capacità senza governabilità

Nei modelli monolitici, competenze, preferenze, regole e meccanismi di controllo sono distribuiti in un unico spazio parametrico. Questa integrazione può essere efficiente, ma rende difficile isolare il contributo di una funzione, aggiornare un solo dominio normativo, attribuire una decisione a un componente operativo e verificare dall'esterno che un intervento di sicurezza sia stato realmente eseguito.

La trasparenza dei pesi non equivale a verificabilità operativa.

Un punteggio o un guardrail esterno non equivale a controllo degli stati interni e della topologia.

Un audit trail non equivale a evidenza se può essere vuoto, incompleto o scollegato dall'azione eseguita.

Una policy configurabile non equivale a un modulo normativo sostituibile e qualificabile in isolamento.

Una supervisione dichiarata non equivale a una supervisione la cui non aggirabilità sia stata testata contro un avversario dichiarato.

4. Il perimetro primario

EthicLaw è orientato principalmente a sistemi di inferenza privati o statutari nei quali il soggetto responsabile controlla l'infrastruttura e dispone della visibilità degli stati computazionali interni. In questo perimetro possono essere osservati input, rappresentazioni intermedie, routing, moduli, gateway e output; possono inoltre essere imposti stop, rerouting, correzioni e blocchi prima del rilascio.

Classe	Configurazione	Osservabilità	Garanzie ammissibili
A	Full observability	Input, tensori, moduli, routing, gateway e output.	Attribuzione, override, correzione interna, audit ed enforcement completo.
B	Controlled interface	Input/output e hook o telemetria selezionata.	Garanzie limitate ai segnali e alle azioni esposte.
C	Black-box adapter	Prompt, output e metadati API.	Filtro e blocco esterno; nessuna prova sugli stati nascosti o sulla topologia.

L'interfaccia può essere indipendente dal provider; il livello di garanzia non può esserlo. EthicLaw definisce quindi un Assurance Envelope: ciò che il supervisore può osservare, ciò che può comandare e ciò che può responsabilmente affermare per una specifica integrazione.

5. L'etica come primo caso, non come scorciatoia

L'etica è il primo caso sperimentale perché combina pluralità di principi, conflitti, dipendenza dal contesto, assenza di un ground truth unico e conseguenze sociali rilevanti. Proprio per questo non può essere trattata come una semplice funzione di loss né come una proprietà neutrale del sistema.

EthicLaw non stabilisce quali valori siano corretti. Distingue il Governance Core, che osserva e applica i meccanismi di controllo, dal Normative Module, che valuta un corpus, e dal Normative Package, che dichiara principi, ambito, soglie, versioni, conflitti e condizioni di non applicabilità.

Confine del claim

La stessa architettura può essere applicabile in linea di principio a diritto, deontologia, sicurezza, sanità, finanza o policy aziendali. Questa possibilità non è ancora una proprietà dimostrata: ogni trasferimento di dominio richiede metriche, baseline, package e analisi dell'overhead propri.

6. I principi del programma

1. Governance come proprietà architetturale. Osservazione, decisione ed enforcement devono essere separabili, testabili e tracciabili.

2. Autorità proporzionata alla responsabilità. Un supervisore che risponde della sicurezza deve poter impartire ordini vincolanti e verificarne l'esecuzione.

3. Garanzie limitate dall'osservabilità. Nessuna interfaccia universale autorizza claim uniformi su sistemi con accessi differenti.

4. Modularità sostanziale. Un modulo è significativo solo se la sua funzione è misurabile, il suo confine è operativo e la sua rimozione o sostituzione produce effetti misurati.

5. Principi non compensatori. Una violazione critica non può essere cancellata da risultati migliori su altri principi; la validità della valutazione è controllata separatamente.

6. Costi visibili. Parametri, memoria, latenza, energia, falsi positivi e complessità di gestione sono parte del risultato, non note a margine.

7. Claim condizionali. Ogni vantaggio vale nel perimetro, nella scala e nelle condizioni misurate; i risultati negativi delimitano la teoria.

8. Audit come evidenza. Il registro deve collegare stato osservato, principio applicato, decisione, comando, esecutore e output effettivamente rilasciato.

9. Separazione delle autorità. Architettura tecnica, contenuto normativo, assurance e ricorso non devono essere controllati dallo stesso soggetto.

10. Generalizzazione da dimostrare. L'etica è il primo caso; nessun nuovo dominio è acquisito senza replica e convenienza misurata.

7. Le domande falsificabili

ID	Domanda di ricerca
RQ1	Dove la decomposizione modulare produce un vantaggio rispetto a un modello singolo e dove non lo produce?
RQ2	I confini modulari consentono attribuzione funzionale e causale riproducibile?
RQ3	Un singolo modulo può essere sostituito o riaddestrato in isolamento mantenendo le proprietà del sistema?
RQ4	La supervisione autoritativa riduce violazioni e rischi senza introdurre costi operativi sproporzionati?
RQ5	Il Decision Lattice mantiene vincoli non compensatori con tassi accettabili di missed detection e unnecessary stop?
RQ6	Quale quantità e qualità di stato interno è necessaria per ottenere ciascun livello di assurance?
RQ7	La correzione forward-only migliora la traiettoria successiva o ottimizza soltanto un proxy?

ID	Domanda di ricerca
RQ8	Il Governance Core può essere riutilizzato in un secondo dominio senza perdere controllo, significato o convenienza?

8. Evidenza attuale e disciplina dei claim

P0 è un esercizio metodologico su scala ridotta. Ha verificato contratti, canali, assenze, pipeline end-to-end, lattice, audit e attribuzione per fonte; ha inoltre trovato difetti architetturali e fallimenti silenziosi che non erano emersi dalla sola revisione documentale. Non dimostra la superiorità generale del modulare, la correttezza del punteggio etico, l'efficacia della correzione o la scalabilità del sistema.

Area	Evidenza	Limite
Proprietà architetturali	Contratti tipizzati, bus, canale privilegiato, gateway e test di presenza/assenza.	P0 verificato nel suo perimetro.
Attribuzione	Ablazioni singole e congiunte mostrano contributi separabili e ridondanza crescente.	EV-03 / NR-3: vantaggio rispetto all'attribuzione post-hoc su monolite non misurato.
Prestazione modulare	EV-01: vantaggio locale sui due principi virtue rimasti al caso nel controllo singolo a 341M; nessuna superiorità generale.	Un corpus, un dominio, scala P0; replica a parità di budget richiesta.
Correzione	Cablaggio in-place osservato.	EV-05 descrive P0. Aggiornamento P0.1 (OP 55–56): nessun vantaggio significativo sul casuale a pari norma; efficacia etica non valutabile con i giudici disponibili.
Sostituibilità	Interfacce predisposte.	H8 da eseguire.
Generalizzazione	Architettura concettualmente separata dal package etico.	Secondo dominio non ancora testato.

Impegno epistemico

Il programma distingue claim strutturali, risultati P0, teoremi condizionali, ipotesi aperte e risultati negativi. Nessun documento pubblico deve usare una formulazione più forte di quella autorizzata dal Claim Register.

9. La proporzionalità della governance

EthicLaw non propone di aggiungere supervisione ovunque. L'architettura è giustificata nei contesti in cui governabilità, conformità, responsabilità e controllo sono requisiti sostanziali e il loro valore supera il costo della supervisione.

Beneficio da misurare	Overhead da misurare
Riduzione del rischio e delle violazioni	Latenza end-to-end e tempo di decisione
Attribuzione, testabilità e qualità dell'audit	Calcolo, memoria, rete ed energia
Aggiornamento locale e riduzione del lock-in	Integrazione, test, manutenzione e ricertificazione
Conformità a norme, statuti o procedure	Falsi positivi, blocchi e costi organizzativi

Fuori dalla condizione più favorevole, un overhead non compensato delimita il campo di applicazione. Questa clausola non si applica a RF-1, RF-2 e RF-3: restano validi i criteri di refutazione depositati nel registro. I costi di supervisione non sono ancora misurati (§8 del registro).

10. Programma di ricerca e traiettoria pubblica

Fase	Domanda	Output	Stato
P0	Metodo e attribuzione	Contratti, pipeline, lattice, audit, ablazioni e lezioni sui fallimenti silenziosi.	Completato a scala ridotta.
P1	Architettura e costo della governance	Validare comunicazione, controllo, osservabilità, failure modes e overhead sul cluster.	Priorità immediata.
Paper 1	Dove la modularità paga	Confronto modulare/monolitico con claim localizzati e costi separati.	Dopo P1/P2 a parità appropriata.
P2 / Paper 2	Modulo normativo sostituibile	H8, package versionati, calibrazione e ricertificazione locale.	Fase successiva.
Paper 3	Interfaccia e Assurance Envelope	Garanzie graduate su full, controlled e black-box.	Dopo evidenze interne solide.
Replica multidominio	Generalizzazione	Secondo dominio con stesso Governance Core e metriche proprie.	Solo dopo il primo caso validato.
Reference architecture	Standardizzazione	Profili EL-*, reference implementation e conformance test suite.	Dopo risultati e caso reale.

11. Cosa EthicLaw non promette

Non promette una definizione universale o neutrale dell'etica.

Non promette che modularità significhi automaticamente migliori prestazioni.

Non promette le stesse garanzie su un sistema pienamente osservabile e su un'API black-box.

Non promette che un audit trail sia corretto solo perché esiste.

Non promette trasparenza dei pesi: sposta i test e l'audit sulle interfacce, sugli interventi e sulle evidenze.

Non promette fedeltà al cervello umano né usa l'analogia biologica come prova.

Non promette applicabilità universale: ogni nuovo dominio deve giustificare il proprio overhead.

12. L'impegno del programma

Rendere la governance un oggetto che si possa ispezionare, misurare, contestare, sostituire e verificare.

EthicLaw considera i risultati negativi parte del lavoro, i costi parte del risultato e i limiti parte della specifica. Il programma mira a produrre paper falsificabili, un'architettura di riferimento, una suite di test riproducibili e, soltanto dopo evidenze sufficienti, profili tecnici aperti per sistemi inferenziali controllati.

Formula conclusiva. L'etica ha generato la domanda. L'architettura ne rende auditabile il governo. La ricerca deve dimostrare quando questa governabilità vale il costo che introduce.

Evidenze, esiti e limiti

Il registro v0.1.1 è l'autorità sulla forza dei claim. I risultati P0 sono locali: scala ridotta, una macchina, un corpus e un dominio. Non autorizzano risultati di merito a scala P1 o di frontiera. Nessun claim di livello 3 è autorizzato.

EV-01 — Iniezione di competenza
Nel controllo singolo a 341M la media è 0,6754 contro 0,676 del percorso modulare. Sui due principi virtue il controllo resta al caso (0,499 e 0,498), mentre il modulare raggiunge 0,619 e 0,643: +0,120 e +0,145. Tre controlli distinguono l'esito da multi-task, oblio e capacità. Evidenza locale; non dimostra superiorità generale.

EV-03 — Attribuzione: confronto ancora aperto
Sono qualificate ToM, Epistemico e Percettivo, anche quando l'effetto è nullo o negativo. Il confronto con attribuzione post-hoc su un monolite di parità non è stato eseguito. I dati completi di ablazione sono nel registro: ogni confronto deve includere ablazioni singole e congiunte, rapporto e variazioni assolute (NR-3).

NR-1 — Risultati negativi e claim ritirati
NR-1: nessun rilascio su 17 frasi; punteggi 0,319–0,448 contro soglia 0,80. La monotonia del lattice non assicura una decisione utile. NR-2: il routing non si ripaga alla scala misurata; un forward evitato nel 13% dei casi riguarda calcolo, non qualità. NR-3: l'ablazione singola può invertire l'interpretazione del contributo. NR-4: almeno cinque fallimenti silenziosi in otto passi. NR-5: AUC 0,672 per Epistemico+ToM, 0,661 con dimensioni permutate e 0,647 con Percettivo reale. NR-6: ritirata la varianza per testa 0,09; il ritiro resta registrato. Dettagli e limiti di ciascuna voce sono consultabili nel registro.

TN-5 — Tensioni note
Il registro è scritto da chi formula i claim; la revisione indipendente manca (TN-5). La calibrazione cambia con il numero di principi (TN-1). Il supervisore concentra il modo di fallimento e non ha un contrappeso tecnico in linea (TN-2). Gli stati interni possono essere una superficie d'attacco (TN-3). In classe C si propone evidenza opponibile da un giudice indipendente: non si ereditano i claim architetturali della classe A (TN-4).

RF-1 — Condizioni che confuterebbero i claim
RF-1, RF-2 e RF-3 riguardano rispettivamente iniezione di competenza, valore degli stati interni e sostituibilità. Nella condizione più favorevole non è ammesso usare la proporzionalità per sottrarsi alla confutazione. Protocolli, parametro differito e regole di chiusura restano quelli del registro.

limiti — Costi e P0.1
I costi di supervisione in parametri, memoria, latenza ed energia non sono ancora misurati (§8 del registro). Aggiornamento successivo, distinto dal registro v0.1.1: P0.1 (Documento 12, OP 55–56) non rileva un vantaggio significativo del correttore sul casuale a pari norma. Non è prova di equivalenza o di assenza di effetto. Nessun segnale del giudice supera la calibrazione: l'efficacia etica non è valutabile con questo esperimento.

Descrizione testuale e perimetro dello schema
Il percorso standalone comprende sette moduli: Percettivo (input e rappresentazioni linguistiche), Epistemico (fatti e relazioni), Theory of Mind (intenzioni e prospettive), Etico (Normative Package), LMH (generazione), Router (attivazione e instradamento) e Metacognitivo (supervisione). L'Adapter è l'ottavo componente nella modalità di supervisione di un LLM esterno (Documenti 02 e 05). Il Gateway è il punto di enforcement, non un nono modulo. Il registro conta tre moduli qualificati e cinque dichiarati; lo schema standalone mostra sette moduli e descrive separatamente l'Adapter. Il bus pubblica rappresentazioni tra container, senza comandi di controllo. Le coppie co-localizzate usano proiezioni dirette. Il canale privilegiato è separato: tutti i moduli vi pubblicano stati o eventi; il Router può scrivere ma non leggere. Il Metacognitivo riceve l'input prima del routing e impartisce comandi al Router e al Gateway; non pubblica sul bus operativo. Il percorso di correzione Etico→LMH è distinto dal bus. Il Gateway blocca o autorizza il rilascio esterno.

In P0 le proprietà sono testate in-process, in un singolo processo Python. Il modello di minaccia comprende errore di progetto, accoppiamento accidentale, deriva di configurazione e fallimento silenzioso. Esclude insider con privilegi di esecuzione, compromissione del runtime o della catena di build, moduli incentivati a eludere il supervisore e input che inducano stati interni fuorvianti. È igiene ingegneristica testata nel perimetro, non sicurezza contro questi avversari (§6 del registro).

Claim & Evidence Register v0.1.1

EthicLaw Research Program. Research Manifesto. Edizione pubblica 0.2, 11 settembre 2026. Registro applicato: v0.1.1 (5 agosto 2026).