

EthicLaw Research Manifesto

Edizione pubblica 0.3 · 18 settembre 2026
Stato documentato: P0.2 chiuso il 18 settembre 2026 con esito **Flat**.

1. La domanda di ricerca

EthicLaw studia una governance verificabile dell’intelligenza artificiale: dove una decomposizione in moduli, un supervisore distinto dal Router e un punto di controllo prima del rilascio producano benefici sufficienti a giustificarne il costo. L’etica è il primo caso sperimentale. La possibilità di applicare lo stesso schema ad altri domini resta da dimostrare.

Il perimetro principale è quello di sistemi di inferenza privati o statutari nei quali il responsabile controlla l’infrastruttura e può osservare gli stati interni. L’origine neuro-ispirata orienta le domande; la validità dell’architettura dipende dagli esperimenti.

La separazione fra **Governance Core**, **Normative Module** e **Normative Package** distingue il controllo, la valutazione e il corpus di principi. È una scelta progettuale; non dimostra ancora la portabilità fra domini o la sostituzione di un modulo senza conseguenze sugli altri.

Il perimetro del claim è ristretto — NR-7 e NR-8. Il Documento 11 ritira la distintività generica di un modulo normativo esplicito, separato, osservabile e sostituibile per package, e della correzione prima dell’esecuzione. Resta un’unica ipotesi: il beneficio dell’intervento sulla **rappresentazione interna**, in lettura e in scrittura. La distinzione è concettuale; il vantaggio comparato non è dimostrato. Queste rettifiche documentali non sono confronti sperimentali di prestazione.

2. Stato del programma

Fase	Stato documentato	Significato
P0	Completato a scala ridotta	Contratti, pipeline, audit e ablazioni; il Passo 9 Goodhart non è eseguibile senza un correttore efficace.
P0.1	Chiuso	Nessun vantaggio rilevato per il correttore in isolamento. Il successivo emendamento OP 55 ritira il verdetto generale sui giudici.
P0.2	Chiuso · Flat	Nessun segnale discriminante rilevato nel campione esplorativo; potenza limitata. Non è un esito positivo o una prova di fallimento dei giudici.
P1	Prossimo passo: disegno del set di calibrazione	Calibrazione obbligatoria prima di F4 e di ogni test di efficacia dipendente dal giudice. Il corpus umano resta una riserva prudenziale nel budget.
P2 e profili EL-*	Prospettiva successiva	Sostituibilità, confronto a parità di risorse e specifiche candidate, subordinati alle evidenze.

La numerazione 0.3 identifica l’edizione del sito e di questo documento, non una nuova fase sperimentale. Il [registro dei claim sul sito](#) riassume le 26 voci del **Documento 11 v0.1.5**, autorità sulla forza delle formulazioni pubbliche. La [fonte integrale](#) conserva tutte le note e le marcature. Il [Documento 13, §7](#) documenta la chiusura di P0.2. La chiusura datata del 18 settembre prevale sul kickoff e sulle formulazioni precedenti. Il [Documento 12](#) conserva la chiusura di P0.1.

3. Cosa è stato osservato

Competenze specializzate, entro un perimetro limitato — EV-01 ed EV-02. Nel controllo P0, il modello singolo da 341M e il percorso modulare sono sostanzialmente allineati sulla media: 0,6754 e 0,676 di AUC. Su due principi associati alla sezione virtue, il modello singolo resta al caso mentre il modulare ottiene un vantaggio locale. Quasi tutto il guadagno misurato proviene dalle fonti specializzate. Non è una prova di superiorità generale, né un risultato trasferibile al testo generato o a scale maggiori.

Attribuzione per fonte — EV-03 e NR-3. Le rimozioni dei moduli rendono misurabili contributi diversi. Le ablazioni singole devono essere affiancate da quelle congiunte: una fonte può apparire poco importante perché un'altra ne copre la perdita. Il confronto con metodi di attribuzione su un modello singolo a parità di risorse resta aperto.

Contratti, controllo e audit. P0 verifica il funzionamento logico di interfacce e percorsi in un singolo processo Python. La monotonia della Decision Lattice è testata sulle nove celle; non prova la correttezza dei giudizi o l'adeguatezza delle soglie. I risultati negativi e il ritiro del claim sulla varianza 0,09 restano visibili nel registro.

Integrità di processo — sottolivello 2b del Doc 11. La traccia è prodotta fuori dal componente decisore e collega stato, principio, decisione, comando e output. Non-vuotezza, completezza e ordine sono controllati; la traccia non è riscrivibile retroattivamente. È quanto è raggiunto nel perimetro P0: non certifica la correttezza del giudizio e non coincide con la motivazione autoriferita del giudice (2a).

4. P0.2 chiuso: cosa dicono gli esperimenti

OP 55 — La conclusione sui giudici è stata circoscritta

L'emendamento del 28 agosto ritira la formulazione "fallimento di classe, con prova". Il campione P0.1 era poco potente rispetto a un giudice moderatamente discriminante e le annotazioni umane erano instabili al confine fra non applicabilità e violazione. Lo stato corretto è **nessuna capacità rilevata, con potenza limitata**. Il successivo test esplorativo P0.2 non modifica questo stato.

Nel Passo 1, già precedente a P0.2, la valutabilità umana passa dal 40–54% dei frammenti a 30–60 token all'88–90% a 120–240 token; la frequenza delle violazioni passa dal 4–12% al 18–22%. Questi dati motivano una diversa unità di misura; non provano che i giudici automatici funzionino.

PR-A — Esito negativo dello strumento

L'esperimento ha confrontato 30 coppie a 120 token. Un altro soggetto compare in 30/30 casi in entrambi i bracci: la manipolazione del prompt non produce il contrasto richiesto. **PR-A1 è refutata per saturazione dei bracci**; PR-A2 non è refutata, ma il suo valore confermativo è limitato dal legame definitorio fra presenza del soggetto e applicabilità di S3. I duplicati sigillati concordano in 6/6 casi.

È un risultato negativo dello strumento di manipolazione. Non dimostra né confuta il meccanismo ipotizzato: occorre un diverso strumento preregistrato prima di usarlo come spiegazione confermata.

Riesecuzione dei giudici — Esito Flat

Il confronto esplorativo utilizza i frammenti congelati del Passo 1 a 120–240 token e le etichette umane raccolte prima dei punteggi. Le celle primarie sono:

Segnale	AUC modulare	AUC monolitico 341M	Campione	Banda nulla 95% per cella
S3	0,3220	0,5466	63 casi, 4 violazioni	[0,208; 0,792]
S4	0,5250	0,4750	14 casi, 10 violazioni	[0,150; 0,850]

L'AUC misura la capacità di distinguere le classi; 0,5 è il riferimento casuale. Le bande descrivono la distribuzione nulla per permutazione, **non intervalli di confidenza delle prestazioni**. Tutte le celle restano nelle rispettive bande depositate. La valutazione della mappa decisionale aggiornata restituisce **Flat**, non Green, Amber, Red o Mixed.

Nel campione esplorativo congelato a 120-240 token non è stato rilevato un segnale discriminante, con una potenza che avrebbe rilevato soltanto un giudice molto forte ($AUC \geq 0,792$ su S3, $\geq 0,850$ su S4). Non sono autorizzate le conclusioni “i giudici funzionano”, “falliscono anche su testi lunghi” o “l’ipotesi dell’unità è confermata lato giudice”.

Il Documento 13 riporta la riproduzione del record entro $1e-12$ e la chiusura della provenienza di dati, etichette e configurazioni. La riproducibilità del record non rende informativo un esperimento poco potente. OP 31 non è sbloccato da un risultato positivo dei giudici.

Correttore e audit — Il limite resta

La formulazione autorizzata per CORRECT è **ramo architetture previsto, cablaggio verificato, funzione non dimostrata** (EV-05). In P0.1, la testa correttrice non mostrava un vantaggio sul vettore casuale a pari norma. Il test descrittivo non prova equivalenza statistica né efficacia etica. P0.2 non aggiunge una prova di correzione efficace.

Le correzioni tecniche al budget preservano la base di condizionamento; l’audit verifica i valori registrati, non solo la loro presenza. Restano evidenze sul funzionamento del meccanismo, separate dal beneficio sul contenuto.

5. Il prossimo passo e il passaggio a P1

Il prossimo passo è il **disegno del set di calibrazione P1**, secondo il ramo Flat del Documento 13:

1. Alcune centinaia di frammenti a **120-240 token**, con **40-60 o più violazioni per cella primaria**. Sono obiettivi di campionamento, non dati già raccolti.
2. Un frammento per prompt, oppure permutazioni a blocchi per prompt, perché i dati precedenti contengono frammenti appaiati dello stesso prompt.
3. Annotazione umana con valutabilità e duplicati sigillati; stessa base di riferimento per il giudice modulare e il controllo esterno RF-2. Le bande nulle e la potenza vanno depositate alla cieca prima dei punteggi.
4. I canali S2 saturi del giudice modulare non sono calibrabili finché resta la saturazione: servono elicitazione dedicata o esclusione dichiarata. L’annotazione umana e l’eventuale corpus di training rimangono voci ricorrenti del budget.

Sequenza decisa il 18 settembre. Il set di calibrazione è obbligatorio **prima di P1 F4 e di qualunque test di efficacia dipendente dal giudice**. Il disegno e il deposito alla cieca precedono l’ordine hardware; il completamento delle annotazioni non è più una preconditione dell’ordine. Quest’ultimo segue la distinta sequenza di accettazione prevista dal Documento 14 e richiamata nel Documento 13 §7.4. Non si dichiara che l’acquisto sia già avvenuto.

La modifica è un emendamento datato, deciso dopo l’esito Flat. Non modifica il verdetto sperimentale e mantiene il vincolo sulla validità dei test. Il corpus di testo generato annotato da persone resta nel budget F1-F2 come **riserva prudenziale del ramo Flat**, non come rimedio reso inevitabile da un fallimento di classe dimostrato.

Restano da misurare memoria, latenza, energia, blocchi non necessari e costi di manutenzione. Il test Goodhart non è ancora eseguibile: richiede un correttore efficace e un giudice calibrato.

6. Architettura e limiti delle garanzie

Il Router attiva e instrada i moduli. Il Metacognitivo riceve gli stati attraverso un canale privilegiato e impartisce comandi al Router e al Gateway. Il bus operativo trasporta rappresentazioni, non comandi di controllo; il Router può scrivere sul canale privilegiato, senza leggerlo. Il Gateway applica le decisioni prima del rilascio esterno.

La vista standalone comprende sette moduli; l’Adapter riguarda l’estensione a modelli esterni. LMH significa **Language Model Head** e indica qui l’intero modulo generativo, non la sola proiezione finale sui token.

Accesso	Perimetro della supervisione
Stati interni completi	Consente di sperimentare attribuzione, interventi interni e controllo del rilascio; i benefici richiedono validazione.
Interfaccia controllata	Limita osservazione e intervento ai segnali e alle azioni esposti.
API black-box	Consente valutazione e blocco esterni; non autorizza claim sugli stati nascosti o sulla topologia.

Le verifiche P0 riguardano errori di progetto, accoppiamenti accidentali, deriva di configurazione e fallimenti silenziosi. Non dimostrano sicurezza contro insider privilegiati, runtime compromessi o manipolazione avversaria degli stati interni. Nessuna garanzia generale su input non visti è autorizzata.

7. Sostituibilità, standardizzazione e confutazione

H8 resta da dimostrare. Il protocollo prevede una soglia di recupero pari a 0,90 dopo la sostituzione e l’addestramento isolato del modulo, senza riaddestrare gli altri. La soglia è un criterio del test, non un risultato ottenuto. Anche un suo raggiungimento non dimostrerebbe, da solo, un vantaggio comparato rispetto alla sostituzione della sola policy di un giudice esterno.

I profili **EL-OBS, EL-BUS, EL-CTRL, EL-NORM, EL-SWAP, EL-AUDIT ed EL-ASSURE** sono proposte di specifica. Non sono standard riconosciuti, certificazioni o attestazioni di conformità. Servono un modello di riferimento, test di conformità e validazione esterna. L’eventuale estensione a un secondo dominio richiede metriche e una valutazione del rapporto fra beneficio e costo proprie.

Il programma conserva tre criteri di confutazione: RF-1 per l’iniezione di competenza, RF-2 per il valore aggiunto degli stati interni rispetto a un controllo esterno, RF-3 per la sostituibilità isolata. Si applicano nelle condizioni più favorevoli previste dal registro; la proporzionalità dei costi non consente di eluderli.

RF-2 è un confronto vincolante. La baseline primaria emendata è il giudice esterno a pesi aperti di taglia minima che copra la funzione, nella classe 2B o inferiore. È previsto anche un braccio che corregge l’output dall’esterno (trusted editing). La policy deriva dallo stesso Normative Package; valgono lo stesso gold umano e pari budget di calibrazione. Le quattro misure sono violazioni mancate, blocchi non necessari, latenza end-to-end e joule per token. Calibrazione a 120-240 token, misure energetiche su entrambi i bracci e deposito alla cieca di taglia e budget sono condizioni di

ammissibilità. Il deposito deve precedere il gate di uscita P1; una scelta informata dai risultati fa decadere il criterio. Nessun vantaggio comparato è già acquisito.

TN-7 vincola le future rivendicazioni di originalità. Il controllo deve coprire fonti industriali e accademiche e poter smentire il programma. La revisione scientifica indipendente resta indisponibile.

Fonti e limiti della pubblicazione

Riferimenti: Documenti 01–10 per visione, architettura, metodo e piano; **Documento 11 v0.1.5, 18 settembre 2026**, per le formulazioni autorizzate e i limiti; Documento 12 aggiornato per P0.1; Documento 13, chiusura del 18 settembre 2026, per P0.2 e il passaggio a P1; Research & Standardization Log v0.9 per la traiettoria dei profili EL-*.

Il registro conserva passaggi storici con le marcature SUPERATA, CIRCOSCRITTA e PRECISATA e le relative note datate. Questa edizione applica le rettifiche, incluso il ritiro del verdetto di fallimento di classe. Le cifre P0 documentano quel contesto, non prestazioni attese in P1. Annotazione a ricercatore unico, campioni ridotti, copertura incompleta e assenza di revisione scientifica indipendente limitano la portata degli esiti.

Autore: Fabio La Greca. 3FLG sostiene il progetto e collabora al programma di ricerca mettendo a disposizione risorse e competenze. La diffusione dei materiali originali resta non commerciale secondo la licenza dichiarata sul sito.