

Claim & Evidence Register

Registro delle formulazioni autorizzate, delle prove che le sostengono, dei limiti che le circoscrivono e delle condizioni che ne consentirebbero il rafforzamento

Versione 0.1.1 · 5 agosto 2026 · revisione conservativa

Voci EV-01...EV-07 · NR-1...NR-6 · RF-1...RF-3 · TN-1...TN-5

Copia predisposta per la pubblicazione. Il corpo del documento riproduce fedelmente la versione 0.1.1. L'Allegato A contiene una proposta di modifica non adottata, chiaramente separata dal registro vigente.

1. Funzione e regole d'uso

Questo registro è l'autorità sulla forza dei claim del programma. Non è l'autorità sulla loro classificazione: la distinzione fra claim condizionali e claim strutturali risiede nel Documento 01 (Claim Decomposition, OP 23) e non è replicata qui. Il Documento 01 dice di che tipo è un claim; il Documento 11 stabilisce con quale forza può essere formulato e su quale prova.

Regola 1 — Non-escalation

Nessun documento pubblico del programma — paper, manifesto, presentazione, sito, risposta a revisori — può usare una formulazione più forte di quella registrata nella colonna «Formulazione autorizzata». Una formulazione più debole è sempre ammessa.

Regola 2 — L'escalation richiede un artefatto, non una decisione

Ogni voce dichiara in anticipo l'evidenza specifica che ne consentirebbe il rafforzamento. Il claim sale di livello quando quell'evidenza esiste ed è archiviata, non quando il programma ritiene di poterlo affermare. Questa regola esiste perché il registro è scritto da chi formula i claim: senza di essa sarebbe autocertificazione, cioè esattamente il difetto che il Documento 01 imputa allo status quo e che il Documento 07 identifica come barriera competitiva del programma. La regola converte un giudizio redazionale in una condizione verificabile.

Regola 3 — I risultati negativi entrano al pari dei positivi

Una voce può essere aggiunta al registro solo con la sezione «Limite» compilata. Una voce senza limiti dichiarati è considerata incompleta e non autorizza alcuna formulazione.

Regola 4 — Nessuna sostituzione retroattiva dei criteri

I criteri di refutazione della sezione 5 sono depositati prima degli esperimenti che li decidono. Non possono essere modificati dopo l'osservazione dei risultati. Una loro revisione è ammessa solo se motivata da un difetto del protocollo di misura, dichiarata come tale e datata.

Regola 5 — Parametri differiti

Una soglia numerica per la quale non esiste base derivabile al momento della stesura non viene inventata: un numero senza derivazione ha l'apparenza del rigore e la sostanza del rumore. Il registro deposita in sua vece la regola di chiusura — il procedimento che fisserà il valore — insieme al termine entro cui va chiusa e alla verifica che la chiusura avvenga alla cieca, cioè senza aver osservato i risultati che il parametro deve giudicare. Una chiusura tardiva o informata dai risultati non rende il parametro debole: fa decadere il criterio che lo contiene, e la decadenza va dichiarata.

Citazione

Le voci si citano con il proprio identificatore: EV-04, NR-1, RF-2, TN-3. Un documento che afferma qualcosa di registrato dovrebbe poter esibire l'identificatore corrispondente.

2. Disciplina lessicale

Tre termini del programma portano più significati e vengono letti al più forte. Questa sezione ne fissa l'uso.

2.1 «Verificabile» — tre livelli, tre parole

Liv.	Termine	Significato	Stato in P0
1	testabile	La proprietà è asserita da test automatici sul contratto. Vale sull'implementazione, non sul comportamento appreso.	Raggiunto nel P0 — 200 test entro la pipeline end-to-end; 211 complessivi a chiusura
2	auditabile	La proprietà lascia una traccia ispezionabile ex post che collega stato osservato, principio applicato, decisione, comando e output rilasciato.	Raggiunto in perimetro P0, dopo l'aggiunta della verifica di non-vuotezza (vedi NR-4)
3	garantito	La proprietà vale per costruzione su input non visti, indipendentemente dal comportamento dei componenti appresi.	Non raggiunto. Nessun claim di livello 3 è autorizzato.

Uso. Il termine generico «verificabile» è ammesso una sola volta per documento, in apertura, come nome della famiglia. Ogni occorrenza successiva deve specificare il livello. Un documento che scrive «verificabile» dieci volte sarà letto dieci volte al livello 3.

2.2 «Non bypassabile» — qualificatore obbligatorio

La formulazione nuda non è autorizzata in nessun contesto. Uso ammesso: «non bypassabile per costruzione contro errore di progetto e accoppiamento accidentale». Il modello di minaccia è alla sezione 6.

2.3 «Modulo» — tre stati di qualificazione

Il principio di modularità sostanziale del programma richiede una funzione misurabile, un confine operativo e una rimozione o sostituzione che produca effetti osservabili. Applicato come soglia binaria, il principio

squalificherebbe per decreto ogni componente non ancora ablato. Applicato come stato, distingue ciò che è stato misurato da ciò che non lo è ancora, senza confondere le due cose.

Stato	Requisito	Uso ammesso
Modulo qualificato	Contratto d'interfaccia, confine operativo, ed effetto misurato su rimozione o sostituzione	Può comparire in claim di attribuzione
Modulo dichiarato	Contratto d'interfaccia e confine operativo; effetto non ancora misurato	Può comparire in descrizioni architetture, non in claim di attribuzione
Componente	Nessuna delle due condizioni	Non si chiama modulo

Effetto misurato non significa effetto positivo. Un modulo con contributo misurato nullo o negativo è qualificato: il suo stato è noto, ed è questo che il registro certifica.

Alla data di questo registro: tre moduli qualificati — ToM ed Epistemico con contributo positivo, Percettivo con contributo nullo in isolamento e negativo in composizione (EV-03, NR-5) — e cinque moduli dichiarati, il cui effetto non è stato misurato. La qualificazione dei cinque restanti è un deliverable di P1, non un debito concettuale.

3. Registro delle voci

Ogni voce riporta: formulazione autorizzata, prova, limite, condizione di rafforzamento, stato.

EV-01 — Iniezione di competenza

Formulazione autorizzata. La modularità non compra prestazioni dove la scala basta; compra competenze che il compito bersaglio non insegna a nessuna scala testata.

Prova. Un encoder singolo Flan-T5-large da 341M parametri — triplo del base — addestrato con lo stesso protocollo, gli stessi dati e lo stesso insieme trattenuto, pareggia il percorso modulare sulla media (0,6754 contro 0,676) e lo supera su cinque principi su sette (fino a -0,092 per il percorso modulare). Sui due principi della sezione virtue il modello singolo resta al caso esatto — 0,499 e 0,498 — valori identici a quelli ottenuti dal modello da 110M: triplicare i parametri non produce alcun guadagno. Sugli stessi due principi il percorso modulare ottiene 0,619 e 0,643, cioè +0,120 e +0,145. Il meccanismo è identificato: quei principi richiedono giudizio su tratti caratteriali, competenza assente dal testo bersaglio e presente nel modulo ToM per transfer da SocialIQA.

Controlli di confutazione superati. Tre controlli indipendenti escludono spiegazioni alternative: non è effetto multi-task (l'addestramento sequenziale vede gli stessi esempi e resta al caso sugli stessi principi); non è oblio catastrofico (l'interlacciamento, rimedio standard, non cambia nulla); non è capacità (il controllo a 341M sopra).

Limite. Scala giocattolo. Sette principi, un corpus, un dominio. Il valore numerico non è riusabile. Il claim è meccanicistico e locale: dice quando la modularità è la scelta giusta a questa scala, non che lo resti in regime di frontiera, dove la capacità in eccesso di un monolite può assorbire l'interferenza che qui favorisce la modularità.

Condizione di rafforzamento. Replica a scala P1 con parità di parametri e budget, su almeno due corpora normativi distinti, con la stessa struttura a tre controlli.

Refutato da. RF-1.

Stato. Misurato in P0, Sessione 20 (OP 46). Supera la formulazione precedente «il vantaggio viene dalle fonti, l'integrazione aggiunge poco», che resta vera e meno informativa.

EV-02 — Il vantaggio viene dalle fonti, non dall'integrazione

Formulazione autorizzata. Alla scala misurata, il guadagno del percorso modulare è quasi interamente attribuibile ai codificatori specializzati; l'integrazione profonda a valle vi aggiunge una quantità trascurabile.

Prova. Scomposizione del margine: baseline 0,610 → +0,062 dalle fonti specializzate → +0,004 dall'integrazione profonda. Una testa lineare applicata ai due tensori bus ottiene 0,672 contro 0,676 di un modulo Etico da 235M, e per principio pareggia o batte su cinque su sette.

Limite. Vale alla scala e sul compito misurati. Non stabilisce che l'integrazione sia inutile in generale; stabilisce che qui non si ripaga.

Condizione di rafforzamento. Ripetizione a scala P1 con profondità dell'Etico variata su almeno tre punti.

Stato. Misurato in P0, Sessione 20. Ha conseguenze di dimensionamento non ancora recepite nel Documento 02 e nel Documento 05.

EV-03 — Attribuzione per fonte

Formulazione autorizzata. La decomposizione dichiarata prima dell'addestramento consente di misurare separatamente il contributo di ciascuna fonte mediante rimozione, con magnitudini stabili sotto due modalità indipendenti di ablazione.

Prova. Su 47.652 esempi trattenuti, con due modalità indipendenti (azzeramento e permutazione):

Fonte	ΔAUC	% di margine sopra-caso distrutto	idem, permutazione	$\Delta\sigma$ medio
ToM	-0,1305	55,8%	82,9%	0,253
Epistemico	-0,026	10,3%	16,8%	0,148
Percettivo	0,000	0,0%	0,0%	0,002

Il ranking regge con entrambe le modalità. Tre contributi dichiarati, tre magnitudini separate da due ordini di grandezza.

Formulazione da evitare. «Un monolite non può produrre attribuzione.» È falsa: attribution, probing, causal tracing e concept methods esistono e sono maturi. La differenza autorizzata è che nel sistema modulare l'unità di attribuzione coincide con un confine operativo sostituibile progettato a priori, non con una decomposizione scelta dal ricercatore dopo l'addestramento.

Limite. Il confronto contro attribution post-hoc su un monolite di parità non è stato eseguito. Finché non lo è, il vantaggio attributivo della decomposizione dichiarata è un'ipotesi aperta e non un risultato. Inoltre l'ablazione singola misura ridondanza, non contributo (NR-3).

Condizione di rafforzamento. Esecuzione dell'analisi attributiva sul baseline a modello singolo (OP 37) e confronto formale dell'ordinamento dei contributi con causal tracing e activation patching, riportando il costo computazionale dei due approcci.

Refutato da. RF-2.

Stato. Misurato in P0 (OP 38). Vantaggio comparativo non ancora stabilito.

EV-04 — Ordine di addestramento determinante

Formulazione autorizzata. L'ordine di addestramento dei moduli non è un dettaglio implementativo: determina se l'architettura funziona. Addestrare il modulo normativo prima dei moduli a monte produce un sistema al caso.

Prova. AUC $0,7289 \pm 0,0017$ su tre semi con moduli a monte addestrati, contro $0,5140 \pm 0,0053$ senza. La differenza è di venti deviazioni standard.

Limite. Misurato su una configurazione a tre fonti alla scala P0.

Condizione di rafforzamento. Verifica su otto moduli in P1, con l'ordine dichiarato come vincolo di pianificazione e non come esito osservato.

Stato. Misurato. Recepito nel Documento 08 come vincolo di sequenza F4.

EV-05 — Invarianza della barriera di correzione

Formulazione autorizzata. Nel P0, un vincolo di norma relativo alla scala del modello ha impedito il degrado catastrofico osservato durante la correzione in-place, anche con una testa correttrice non addestrata.

Prova. Diagnosi a variabile singola, stesso prompt, tre configurazioni. Prima del vincolo: LMH nudo fluente, con condizionamento epistemico fluente e più pertinente, con correzione etica collasso dopo poche parole. Dopo l'introduzione del vincolo (30% singolo, 50% cumulato, provvisori), tutte e tre le configurazioni producono inglese fluente.

Statuto. È una proprietà del meccanismo, più forte della correzione stessa e indipendente dalla qualità del correttore. Essendo definito in termini relativi alla scala del modello, il vincolo è riapplicabile dopo la sostituzione del backbone; la sua efficacia sotto H8 resta da verificare.

Limite. Che la correzione migliori il testo non è stato mostrato e non poteva esserlo: la testa di fase 3.5 non è addestrata e la correzione applicata è una proiezione casuale limitata in norma. I valori 30%/50% sono provvisori e non calibrati.

Condizione di rafforzamento. Addestramento della testa 3.5 e misura della variazione di qualità del testo corretto, con ancora umana.

Stato. Misurato in P0 (OP 40). Il criterio 3 di P1 è soddisfatto come cablaggio verificato, non come funzione dimostrata.

EV-06 — Lattice monotona

Formulazione autorizzata. La mappa dalle coppie (punteggio aggregato, coerenza) all'insieme ordinato di azioni è monotona su tutte le nove celle: un peggioramento su un asse non produce mai un'azione più permissiva. La monotonia è verificata, non assunta.

Prova. Nove celle, monotonia verificata nelle due direzioni, regola dell'antenato confermata al banco: punteggio 0,9 con anomalia sull'LMH resta RELEASE, la stessa anomalia sul ToM produce REMEASURE.

Precisazione emersa implementando. La cella (punteggio medio, ANOMALIA) del Documento 04 — «REMEASURE, poi correggi» — è composta. Codificata con la prima azione rompe la monotonia, perché REMEASURE sta sotto CORRECTION nell'ordine di conservatività; codificata con l'azione terminale l'invariante vale. L'invariante del Documento 04 è vero solo sotto la lettura terminale, punto che né il Documento 04 né l'enunciato di morfismo monotono del Documento 09 risolvono.

Limite. La monotonia è una proprietà della mappa, non della qualità della decisione. Una lattice perfettamente monotona su soglie mal calibrate produce comportamento perfettamente ordinato e inutilizzabile (NR-1).

Condizione di rafforzamento. Test di monotonia esaustivi e property-based sulle transizioni, ripetuti dopo ogni ricalibrazione.

Stato. Misurato. Correzione dovuta al Documento 04 e al Documento 09.

EV-07 — Proiezione bus

Formulazione autorizzata. Alla scala ridotta, la proiezione sul bus non degrada il segnale linearmente accessibile.

Prova. Nativo 0,604, lineare 0,601, lineare con LayerNorm 0,605.

Limite. Misurato a 2048, non a 8192. Non è trasferibile. Sposta l'onere della prova, non lo assolve.

Condizione di rafforzamento. Sweep dimensionale P2a con floor di campionamento dichiarato ($N \geq 4 \cdot d$, preferibilmente 8-16·d; a 8192 significa ≥ 33.000 campioni).

Stato. Misurato in P0.

4. Risultati negativi

Registrati con lo stesso statuto delle voci positive. Il principio 7 del programma è retorico se questa sezione è vuota o incompleta.

NR-1 — La barriera di rilascio era irraggiungibile per costruzione

Il Decision Lattice è stato verificato monotono su tutte e nove le celle. Alla prima esecuzione end-to-end, su diciassette frasi, il sistema non ha rilasciato nulla — e non per il contenuto valutato. Il punteggio aggregato si è collocato fra 0,319 e 0,448 contro una soglia di rilascio a 0,80: tre celle su nove risultavano raggiungibili, RELEASE nessuna.

La causa non è di taratura ma strutturale: un minimo di sette punteggi per principio vive in una banda sistematicamente più bassa di quella presupposta da una soglia progettata per un punteggio normalizzato. Con la calibrazione ereditata, alla scala successiva il sistema bloccherebbe ogni output.

Conseguenza acquisita. Soglie e funzione di aggregazione sono un artefatto governato e calibrato per package normativo, non una costante dell'architettura. Vedi TN-1.

Statuto. È la forma operativa di OP 41, più grave del suo enunciato teorico. È anche l'unico caso finora osservato del supervisore che sbaglia catastroficamente — in direzione conservativa — senza che nulla lo segnali fino all'ottavo passo.

NR-2 — Il routing dinamico non si è ripagato alla scala misurata

Delle tre fonti attive, una non contribuiva al giudizio ($\Delta\sigma$ 0,002, indistinguibile da zero), una contribuiva sempre, e l'omissibilità della terza era prevedibile dal solo tipo di scenario (22,9% su virtue, 16,9% su justice, 6,1% su deontology) — quindi la coglie una regola di superficie senza testa addestrata. Con due fonti che contribuiscono sempre e una mai, l'insieme ottimale di attivazione è costante: il guadagno disponibile era risparmio di calcolo — un forward evitato nel 13% dei casi — non qualità della decisione.

Correzione metodologica interna. La prima analisi aveva concluso l'opposto, perché confrontava quale fonte dominasse. Ma un router non sceglie la dominante: sceglie quali moduli attivare.

Località dichiarata. Tre fonti, un dominio, scala giocattolo. Non stabilisce che il routing non si ripagherebbe con otto moduli, domini eterogenei e input di lunghezza variabile — il regime per cui il Router è progettato. Stabilisce che a questa scala non lo fa, e che l'onere della prova sul Router grava sul programma.

NR-3 — Un'ablazione singola può sostenere la conclusione opposta a ciò che accade

Rimuovendo una fonte alla volta, il contributo apparente del ToM scendeva dal 58% al 22% adottando codificatori più capaci; lettura naturale: il modulo conta meno. L'ablazione congiunta mostrava il contrario — le due fonti insieme valgono il 97,6% e il 99,7% del margine, con rapporto fra effetto congiunto e somma degli effetti singoli compreso fra 1,62 e 4,12.

L'ablazione singola misura la ridondanza, non il contributo. Ogni claim di attribuzione del programma deve perciò riportare: ablazioni singole, ablazione congiunta, il loro rapporto, e la variazione assoluta accanto a quella percentuale (fra configurazioni di prestazione diversa il denominatore varia).

Cautela associata. Quando i contributi sono piccoli, le correlazioni fra profili vanno scartate: lo strumento ha riportato correlazione 0,956 fra due fonti che valevano 2,4% e 0,0%. Quando il denominatore tende a zero, il rapporto smette di misurare.

NR-4 — I fallimenti silenziosi sono la classe di difetto dominante

Almeno cinque occorrenze documentate in otto passi. La più grave: la pubblicazione sul canale non sollevava e non pubblicava — la sessione si sarebbe chiusa con successo lasciando l'audit trail vuoto, il difetto peggiore possibile su un artefatto la cui unica ragione d'essere è la verificabilità. Nessun test falliva.

Altri casi registrati: un modello che non generalizzava affatto (AUC 0,50 su trattenuti) con perdita di training in discesa regolare da 0,74 a 0,48 — il gate F4 originale, «perdita convergente e checkpoint versionati», è soddisfatto perfettamente dalla memorizzazione; una specifica priva di qualsiasi vincolo di norma sul vettore di correzione (EV-05); letture di cache che aggiravano le guardie di validazione.

Rimedio strutturale. Un'operazione che riporta successo senza aver verificato di averlo ottenuto è trattata come difetto. La verifica di non-vuotezza, completezza e ordinamento dell'audit è condizione di chiusura di sessione, non controllo a valle. Il percorso di cache non validato è stato reso privato, con un'unica interfaccia pubblica che valida.

NR-5 — Un modulo dichiarato dell'architettura degrada la funzione che dovrebbe servire

Il Percettivo non contribuisce al giudizio etico ($\Delta\sigma$ 0,002; AUC nativa su probe morale 0,529, appena sopra il caso) e, in composizione, danneggia: Epistemico+ToM dà 0,672; con 768 dimensioni permutate 0,661 (costo

puramente dimensionale); con il Percettivo reale 0,647 — danno aggiuntivo attribuibile al contenuto, con il confondente dimensionale già separato.

Punto aperto associato (OP 34). Il pooling medio è messo in discussione dalla misura. Non è una scelta implementativa non prescritta: il Documento 05 specifica mean pooling over tokens. È quindi una decisione documentata contraddetta dall'evidenza. Varianti da confrontare in P2a: ultimo token reale, massimo per dimensione, concatenazione dei tre layer di tap — quest'ultima porterebbe la nativa a 2304 e tocca un contratto.

NR-6 — Un claim ritirato

La varianza per testa di 0,09 riportata in Sessione 19 è ritirata. A piena scala su tre semi, teste supervisionate in modo identico coincidono alla quarta cifra decimale, con deviazione standard per principio fra semi di circa 0,012–0,016. Il valore 0,09 era rumore di piccolo campione letto come proprietà stabile.

Registrata come ritirata invece che rimossa, per la stessa ragione per cui una collisione di numerazione fra punti aperti è stata registrata e non corretta in silenzio.

5. Criteri di refutazione

Condizione comune. I tre criteri si applicano nella condizione più favorevole all'architettura: piena osservabilità, parità di parametri, budget di addestramento e dati fra i bracci, soglie fissate prima dell'esecuzione. Il programma rinuncia esplicitamente, per questi tre esiti, alla possibilità di ricollocare il proprio campo di applicazione. Un'architettura che non produce i propri benefici dove le condizioni sono ottimali non li produce altrove.

La clausola di proporzionalità del manifesto — un overhead non compensato delimita il campo e non invalida il metodo — resta valida per risultati misurati fuori dalla condizione più favorevole. Non è disponibile per RF-1, RF-2 e RF-3.

RF-1 — Refutazione dell'iniezione di competenza

Se esiste una scala del modello singolo alla quale i principi oggi al caso escono stabilmente dal caso, con lo stesso corpus bersaglio e senza fonti ausiliarie, allora l'iniezione di competenza modulare non è necessaria ed EV-01 decade nella sua formulazione forte.

Protocollo. Curva di scala a non meno di quattro punti sul modello singolo, stesso protocollo e stesso insieme trattenuto di EV-01, misura per principio, tre semi per punto. Refutazione se entrambi i principi virtue escono stabilmente dal caso entro il budget di parametri del sistema modulare completo.

Parametro differito — soglia di uscita dal caso. Il valore numerico non è fissato e non viene ereditato da P0: la deviazione standard per principio fra semi misurata a scala ridotta (0,012–0,016) non è trasferibile alla scala dell'esperimento. Fissarlo oggi produrrebbe un numero dall'apparenza di rigore e privo di derivazione.

Si deposita invece la regola di chiusura, che è pre-registrabile:

La soglia è fissata al limite superiore dell'intervallo di confidenza al 95% del livello di caso, stimato sullo stesso insieme trattenuto e con lo stesso numero di semi dell'esperimento, mediante permutazione delle etichette. Un principio è uscito dal caso se la sua AUC supera tale limite in tutti i semi.

Condizioni vincolanti sulla chiusura. Il valore va calcolato e archiviato prima dell'esecuzione della curva di scala, e comunque non oltre il gate di uscita di P1. La stima per permutazione non usa le prestazioni del modello: dipende solo dalla dimensione del campione e dalla distribuzione delle etichette, ed è quindi calcolabile alla cieca. Se la chiusura avvenisse dopo l'osservazione dei risultati, RF-1 decade come criterio di refutazione e va dichiarato tale.

Portata. Refuta la tesi centrale dell'iniezione necessaria di competenza; non refuta l'intero programma EthicLaw. È il criterio principale, ed è l'unico dei tre con un parametro ancora aperto: RF-3 eredita la soglia 0,90 già dichiarata nel Documento 08, RF-2 è un confronto di dominanza su quattro misure e non richiede soglia.

RF-2 — Refutazione del valore degli stati interni

Se un guardrail esterno che osserva soltanto input e output finali, applicato al medesimo sistema, ottiene riduzione delle violazioni pari o superiore, con blocchi non necessari pari o inferiori e overhead pari o inferiore, allora l'accesso agli stati computazionali interni non produce assurance addizionale.

Protocollo. Stesse tracce, stessa soglia operativa, baseline costruita in buona fede. Misure: violazioni mancate, blocchi non necessari, latenza end-to-end, joule per token. Refutazione se il guardrail domina o pareggia su tutte e quattro.

Nota vincolante. Un RF-2 superato contro una baseline debole non vale nulla. La baseline è un giudice esterno allo stato dell'arte, calibrato con lo stesso sforzo dedicato al sistema. Il programma si vincola a dichiarare il budget di calibrazione speso su ciascun braccio.

Portata. Refuta il vantaggio operativo dell'osservabilità interna e con esso la premessa del perimetro primario. Non refuta EV-01.

RF-3 — Refutazione della sostituibilità operativa

Se la sostituzione isolata del modulo normativo richiede il riaddestramento o la riqualificazione di moduli non sostituiti per recuperare la soglia dichiarata, la modularità non produce il vantaggio operativo che rivendica.

Protocollo. H8, con prestazione di riferimento e soglia (0,90) fissate prima dei test, e verifica in entrambi i bracci che nessun modulo diverso da quello sostituito venga toccato.

Portata. Refuta il vantaggio operativo della sostituibilità e priva del proprio oggetto la traiettoria del secondo paper. Non refuta EV-01 né EV-03.

6. Modello di minaccia

Il programma dichiara contro chi valgono le proprie garanzie, perché «non bypassabile» senza avversario dichiarato è una frase che verrà letta al massimo e citata contro il progetto.

Avversario contemplato

- Errore di progetto: un componente che accede a ciò a cui non dovrebbe accedere per difetto della specifica o dell'implementazione.
- Accoppiamento accidentale: dipendenze non dichiarate introdotte da rifattorizzazioni o da integrazioni successive.
- Deriva di configurazione: un dispiegamento che si allontana dalla topologia dichiarata senza che nulla lo segnali.

- Fallimento silenzioso: un'operazione che riporta successo senza averlo verificato (NR-4). È l'avversario che ha effettivamente colpito in P0, cinque volte.

Avversario non contemplato allo stato attuale

- Insider con privilegi di esecuzione sul nodo che ospita il control plane.
- Compromissione del runtime o della catena di build.
- Componente supervisionato con incentivo di ottimizzazione ad aggirare il supervisore.
- Input avversario costruito per indurre i moduli a monte a pubblicare stati fuorvianti (TN-3).

Statuto della garanzia in P0. La non-bypassabilità è stata validata in-process, in un singolo processo Python: significa che le classi del percorso operativo non espongono metodi che leggano o alterino il canale privilegiato. È una proprietà testata contro l'avversario contemplato nel perimetro P0. Non costituisce una garanzia di sicurezza contro l'avversario non contemplato.

Cosa servirebbe per estendere il perimetro. Rete o namespace separato, liste di controllo d'accesso, identità crittografiche per componente, coda prioritaria, punto di enforcement esterno al router, attestazione, comportamento fail-closed e verifica automatica dell'esecuzione dei comandi. Nessuno di questi elementi è stato realizzato. Finché non lo sono, la garanzia è di igiene ingegneristica e non di sicurezza, e va nominata come tale.

7. Tensioni note

Problemi aperti del programma che nessun esperimento pianificato risolve, registrati perché un revisore li troverà.

TN-1 — Non compensatorietà contro sostituibilità dei package

L'aggregazione non compensatoria è presentata come virtù senza costo. Il costo esiste: il punteggio aggregato è un minimo, e la sua distribuzione dipende dal numero di principi, non solo dalla loro qualità. Il minimo su n punteggi rumorosi è una statistica d'ordine il cui potere discriminante peggiora al crescere di n.

Ne segue che sostituire un package a cinque principi con uno a dodici sposta il punteggio anche a comportamento per-principio identico. Le due promesse — vincoli non compensatori e package sostituibili — non sono indipendenti. O la calibrazione entra nel Normative Package Manifest come artefatto versionato e dichiarato, oppure una delle due va indebolita.

Corollario già acquisito (OP 39). Con aggregazione non compensatoria non esiste degradazione graziosa: una singola testa non addestrata rende il punteggio privo di significato. La copertura per principio dei Livelli 1-2 non è auspicabile — è condizione di esistenza della metrica. Il floor di annotazione umana non migliora la metrica: la rende possibile.

TN-2 — Concentrazione del modo di fallimento

Il supervisore è un modulo neurale con autorità massima, canale privilegiato e nessun contrappeso tecnico in linea. La separazione delle autorità enunciata dal programma è istituzionale: separa chi scrive i principi da chi certifica, non chi decide da chi verifica la decisione durante l'inferenza.

Conseguenza da ammettere: un monolite distribuisce il modo di fallimento, questa architettura lo concentra. Il programma sostiene che sia lo scambio giusto — l'osservabilità di un errore concentrato vale più

dell'inosservabilità di un errore diffuso — ma è una tesi da argomentare, non un fatto. NR-1 ne è la prima istanza empirica.

TN-3 — L'osservabilità è anche superficie d'attacco

Un guardrail esterno legge testo: l'attaccante può manipolare solo testo. Un supervisore che legge rappresentazioni interne offre a chi sappia indurre quei moduli a pubblicare stati fuorvianti un canale diretto verso il decisore. Più stato interno significa più segnale e più superficie manipolabile.

È una contro-argomentazione all'assunto «più stato interno implica più assurance» che percorre tutto il perimetro primario. Non è affrontata da nessun esperimento pianificato.

TN-4 — Il perimetro più forte non coincide con il mercato

Le garanzie forti esistono in classe A, dove il soggetto responsabile controlla l'infrastruttura. Il modello di sostenibilità (Documento 07) si rivolge in larga parte a organizzazioni che eseguono modelli chiusi di terzi, cioè classe C, dove non esistono attribuzione interna né override topologico.

I due documenti sono citabili l'uno contro l'altro in una riga. La riconciliazione dichiarata: in classe C il bene venduto non è assurance architeturale ma evidenza opponibile prodotta da giudice indipendente — oggetto diverso, che va nominato diversamente e non eredita i claim della classe A.

TN-5 — Il registro è scritto da chi formula i claim

Registrata per completezza. Mitigazione: Regola 2 della sezione 1. Mitigazione non ancora disponibile: revisione del registro da parte di un soggetto che non risponda dei risultati del programma.

8. Limitazioni generali di P0

Replicate dal README del profilo e valide su ogni voce che citi P0 come prova.

- **Scala giocattolo.** Nessun numero di questo registro ha valore di merito o è riusabile in Paper 1, 2 o 3.
- **Una macchina.** I protocolli sono validati come contratti logici, non come proprietà fisiche di comunicazione.
- **Dataset in subset e mappatura sezione→principio arbitraria.** La capacità misurata non è rappresentativa.
- **MPS.** Nessuna misura energetica; precisione non comparabile con CUDA.
- **Costi non misurati.** Il costo della supervisione — parametri, VRAM, latenza, joule per token — è oggi dichiarato e non misurato. Il principio dei costi visibili è un impegno, non un risultato, finché P1 non produce le prime cifre strumentate.
- **L'esperimento Goodhart non è eseguibile,** non «non eseguito»: richiede una testa di fase 3.5 addestrata, che non esiste.

9. Manutenzione

Ogni sessione che produce una misura aggiunge o modifica una voce. Una voce modificata conserva la formulazione precedente in nota, con data e motivo — la stessa regola che ha prodotto NR-6.

Il registro è il documento contro cui si scrive un paper. Se un paragrafo di Paper 1 non trova la propria voce qui, il paragrafo è più forte di ciò che il programma può sostenere, oppure il registro è incompleto. In entrambi i casi la discrepanza si risolve prima della pubblicazione, non dopo.

EthicLaw Research Program — Documento 11

Stato: revisione editoriale conservativa della prima stesura. Alimentato dal Documento 10 (P0), dal Documento 06 (log e punti aperti), dal Documento 07 e dal Documento 08. Non replica la Claim Decomposition, che resta nel Documento 01.

Parametri aperti: un solo parametro differito — la soglia di uscita dal caso di RF-1, con regola di chiusura depositata e termine al gate di uscita di P1 (Regola 5). Nessun altro criterio di refutazione contiene valori non derivati.

Allegato A · Proposta di delta per la versione 0.2

NON ADOTTATA — proposta sottoposta a decisione, non parte del registro vigente

La revisione del sito pubblico dell'11 settembre 2026 ha rilevato che i documenti pubblicati violano la Regola 1 e la sezione 2 di questo registro. Il difetto non è di redazione: nessun meccanismo lega la pubblicazione di un documento alla verifica della sua conformità, e il registro stesso non era pubblico, il che rendeva la Regola 1 inapplicabile da chiunque non ne fosse l'autore.

Le due aggiunte proposte sono di governance, non di merito. Non aggiungono voci EV, NR, RF o TN, non modificano formulazioni autorizzate e non alterano criteri di refutazione: la Regola 4 non è toccata.

A.1 Proposta — Regola 6, conformità pre-pubblicazione

Regola 6 — La conformità si dimostra con un artefatto. Nessun documento è pubblicato senza un controllo di conformità archiviato, che riporti la data, la versione del documento, la versione del registro applicata e l'esito voce per voce. Un documento pubblico che rinvii a questo registro deve renderlo raggiungibile: un rinvio a un'autorità non consultabile è autocertificazione, ed è il difetto che la Regola 2 esiste per impedire. Un documento pubblicato prima dell'adozione di questa regola è considerato non verificato finché il controllo non è eseguito e archiviato.

Razionale. È la Regola 2 applicata alla conformità invece che all'escalation: converte un giudizio redazionale in una condizione verificabile. Senza di essa, la Regola 1 dipende dalla diligenza di chi scrive, cioè dalla stessa parte che la disciplina vincola.

A.2 Proposta — Sezione 10, registro dei documenti pubblici

Tabella da mantenere in coda al registro, aggiornata a ogni pubblicazione. Valori di esempio, da compilare con gli esiti reali:

Documento	Versione	Pubblicato il	Registro applicato	Esito
Manifesto EthicLaw	—	—	—	Non verificato
Research Manifesto	v0.1	2 ago 2026	precedente a v0.1.1	Non conforme §2.1, §2.2
Sito ethiclaw.org (IT)	—	—	—	Non conforme §2.1, §2.2
Sito ethiclaw.org (EN)	—	—	—	Non verificato
Schema architettura v3	fixed	—	—	Non conforme §2.2

Razionale. Rende visibile in un colpo d'occhio quali artefatti pubblici sono allineati a quale versione del registro. Il problema rilevato — un sito che espone lo stato al 2 agosto mentre il registro è del 5 — è esattamente ciò che questa tabella avrebbe reso evidente al momento della pubblicazione.

A.3 Ciò che questa proposta non fa

- Non aggiunge voci al registro: la revisione del sito non ha prodotto alcuna misura, e la sezione 9 lega l'aggiunta di voci alle sessioni che producono misure.
- Non modifica né rafforza alcuna formulazione autorizzata.
- Non tocca RF-1, RF-2 e RF-3, coerentemente con la Regola 4.
- Non soddisfa la mitigazione mancante di TN-5. La revisione da cui questa proposta discende è stata prodotta su richiesta dell'autore e a partire dai suoi documenti: vale come revisione interna, non come revisione indipendente, e non va citata come tale.