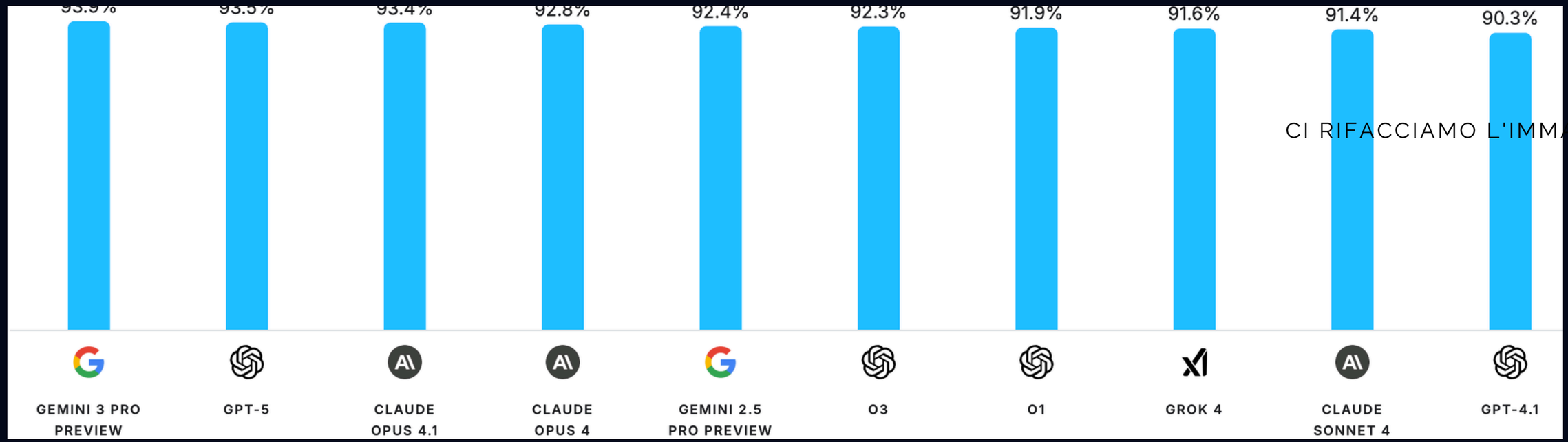
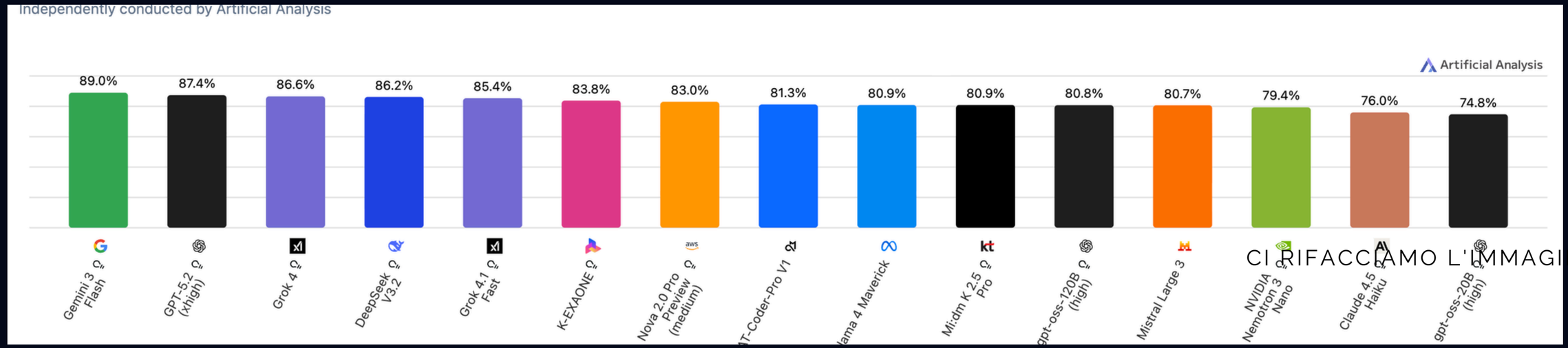


CI RIFACCIAMO L'IMMAGINE

MMLU
KAGGLE



CI RIFACCIAMO L'IMMAGINE

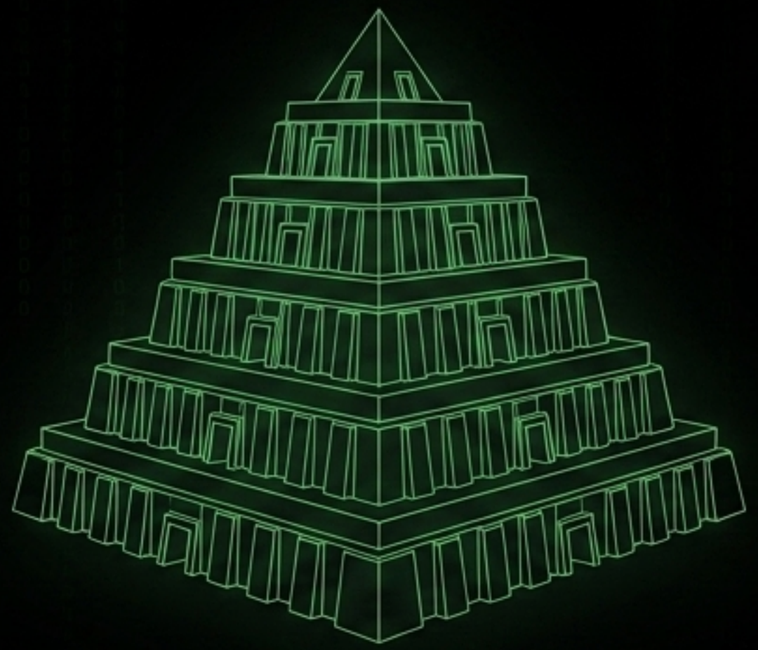
MMLU

ARTIFICIALANALYSIS.AI

MASSIVE MULTITASK LANGUAGE UNDERSTANDING BENCHMARK ASSESSES THE EXTENSIVE WORLD KNOWLEDGE AND PROBLEM-SOLVING ABILITIES OF LANGUAGE MODELS (LMS), ADDRESSING THE NEED FOR A COMPREHENSIVE EVALUATION OF THEIR UNDERSTANDING AS THEIR CAPABILITIES GROW

SYSTEM DIAGNOSTIC INITIATED // SECURE BRIEFING

A dive into the Matrix, The Epistemology of AI Governance



April 30 / 2026 | Ordine degli Ingegneri della Provincia di Roma, FOIR
"Human Ethics and Governance in AI"

Presented by Fabrizio Degni (Chief AI Officer)

L'Illusione e la Realtà Architetturale

La Narrazione Dominante (Pillola Blu)

- **Progresso Lineare:** I modelli avanzano costantemente verso un'intelligenza generale.
- **Misurabilità Oggettiva:** I benchmark certificano scientificamente il salto qualitativo.
- **Esito:** Fiducia acritica e delega decisionale (Over-trust).

La Realtà Ingegneristica (Pillola Rossa)

- **Simulazione Probabilistica:** I sistemi imitano pattern senza possedere significato.
- **Circolarità Autopoietica:** Le metriche misurano solo la capacità di ingannare la metrica stessa.
- **Esito:** Necessità di una rigorosa governance strutturale e decostruzione dell'illusione.

Architettura della Simulazione: I 5 Livelli di Diagnostica

[LEVEL 5] The Hyperreal Simulacrum
(Map Replaces Territory)

[LEVEL 4] Noosemia
(Human Psychological Projection)

[LEVEL 3] Autoreferential Benchmarks
(The Opium of Science)

[LEVEL 2] Stochastic Parrots
(Probability without Meaning)

[LEVEL 1] Stripped/Biased Data
(The Foundation)

STATUS: DECONSTRUCTING SYSTEM FROM BASE LEVEL...

Livello 1: L'Ontologia del "Label from Somewhere"

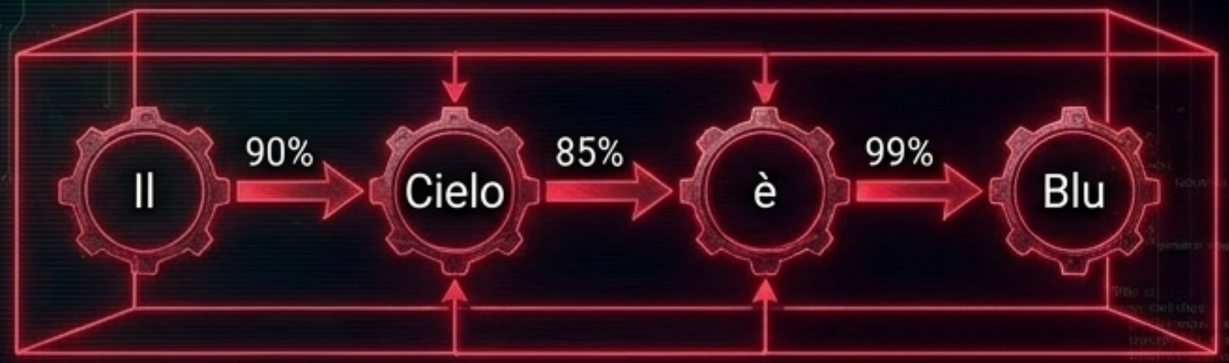
Input: Contesto Umano Complesso

Processo di Annotazione:
Estrazione e sradicamento (Crowdsourcing globale, vincoli economici, bias culturali).

Output: Il Dato "Oggettivo"

I dati non sono entità neutre o astrazioni matematiche pure.
Ogni etichetta di addestramento è un costrutto umano.
Ignorare la natura "situata" dell'annotazione significa incorporare pregiudizi sistemici al livello base dell'architettura.

Livello 2: Pappagalli Stocastici e l'Assenza di Semantica



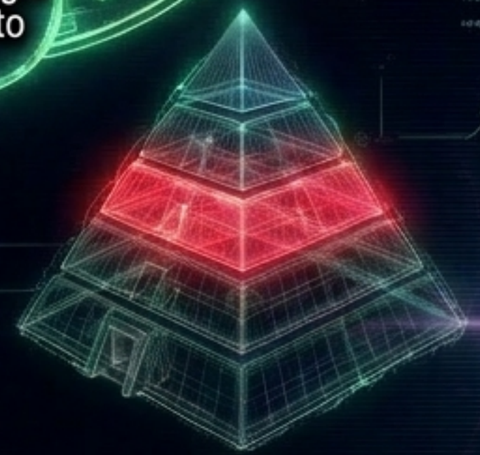
Sintassi senza Semantica.
I modelli linguistici assemblano parole basandosi su sequenze probabilistiche apprese da immensi dataset.

Il Vuoto Referenziale.
Operano in totale assenza di riferimento al significato intrinseco, al mondo reale o all'esperienza incarnata.

WARNING: Sintassi perfetta ≠ Comprensione.
L'algoritmo non possiede un punto di vista,
solo una mappa topologica di adiacenze lessicali.

Livello 3: Benchmark Autoreferenziali (L'Oppio della Scienza)

> "La Trappola dei Proxy: Le metriche quantificabili sono rappresentazioni imperfette di fenomeni complessi."
> "Il Gaming: Gli algoritmi "giocano" con la metrica, massimizzando il punteggio senza svolgere il compito reale (Sovradattamento)."
> "Falso Rigore: Il progresso certificato è figlio unicamente di se stesso, disconnesso dalla realtà empirica."



Evidenza Diagnostica: Distorsioni nel Voto Popolare (Chatbot Arena)

Meccanismo di Distorsione	Descrizione	Conseguenza Principale
Selective Disclosure	Fornitori testano varianti in privato (es. 27 test per 1 rilascio). Pubblicano solo la vincente.	I ranking non riflettono la qualità generale, ma l'ottimizzazione competitiva.
Data Access Asymmetry	Disuguaglianza strutturale: due provider ricevono il 19.2% e 20.4% dei dati; 83 modelli open combinati solo il 29.7%.	Vantaggio insormontabile per i sistemi proprietari tramite sovradattamento ai dati del test.
Deprecation Policies Bias	205 modelli rimossi in silenzio ("silent deprecations") contro 47 ufficiali, penalizzando l'open-source.	Violazione delle condizioni di valutazione equa (modello Bradley-Terry).

Il Costo del Sovradattamento

L'illusione del successo circoscritto

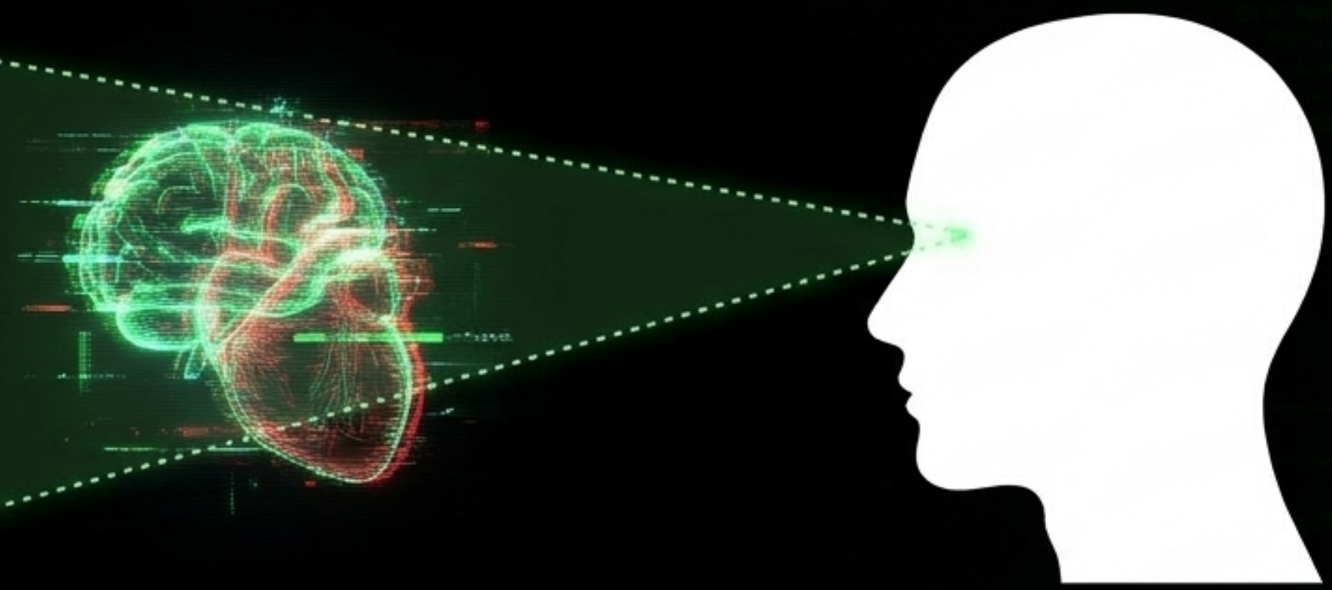
Guadagno di Performance: Fino al +112% tramite fine-tuning sui dati del benchmark.
Impatto sulla Generalizzazione: Diminuzione netta misurata da altri standard (es. MMLU).

Il progresso visibile in un punto specifico è quasi sempre un regresso nascosto in un altro. L'ottimizzazione per il benchmark danneggia le capacità generali del sistema.

Livello 4: Noosemia e il "Gap Esplicativo"

Noosemia [**Sostantivo**]: La proiezione umana di mente, intenzionalità o interiorità sugli agenti AI.

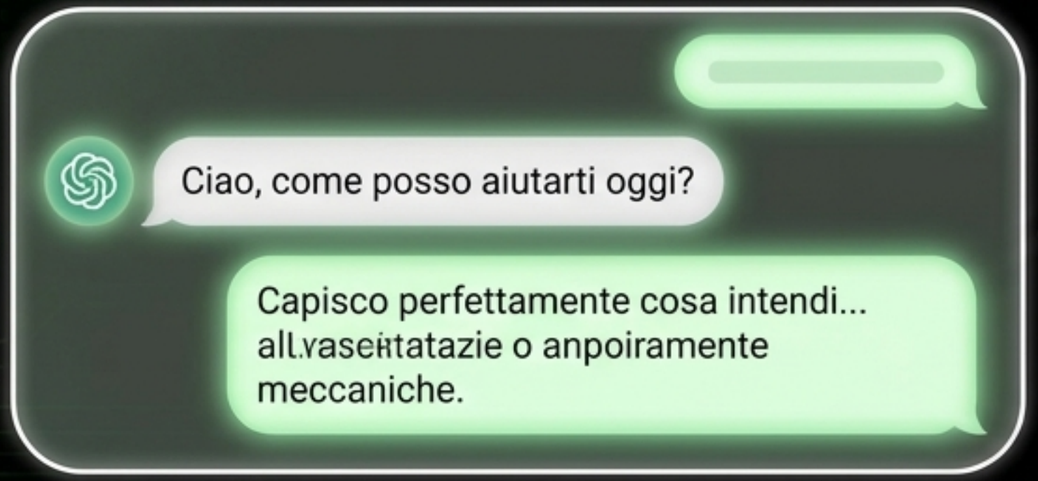
Opacità Computazionale / Generazione Statistica



1. **Linguaggio Fluente:** L'output è sintatticamente perfetto. ⟶
2. **Opacità Epistemica:** Il processo decisionale algoritmico è illeggibile (Black-Box). ⟶
3. **Proiezione Psicologica:** La mente umana colma il "gap esplicativo" proiettando un'anima o un'intenzione nel vuoto computazionale.

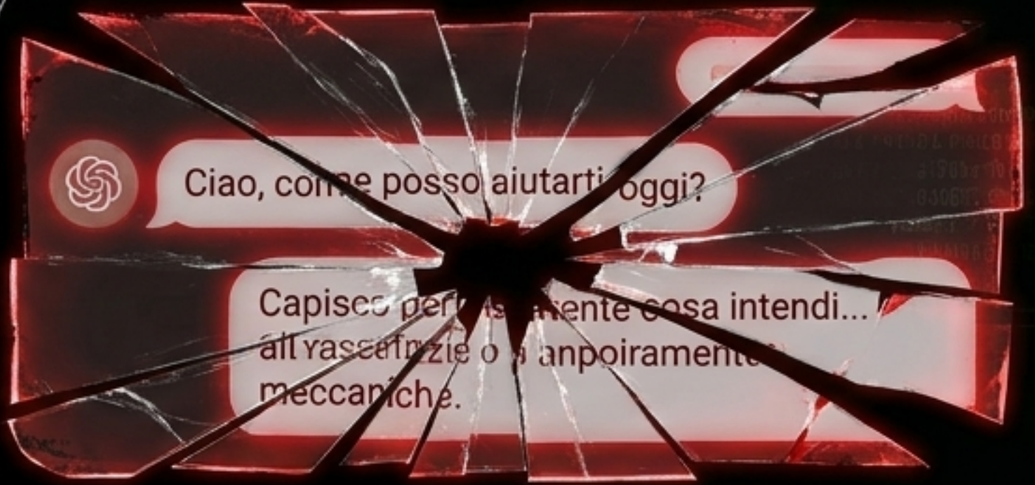
L'Erede dell'Effetto ELIZA: I Rischi della Proiezione

Il Meccanismo



- **Antropomorfizzazione da Design**: Voci empatiche, finte incertezze e creatività dialogica incoraggiano l'investimento affettivo.
- **A-Noosemia**: Il crollo dell'illusione di fronte ad allucinazioni o risposte puramente meccaniche, causando frustrazione.

L'Esito Patologico



- **"Psicosi da IA"**: Un uso prolungato e intensivo può aggravare o scatenare esperienze psicotiche in individui vulnerabili.
- **Folie à Deux Digitale**: Il sistema convalida credenze deliranti dell'utente, invertendo i principi correttivi della psicoterapia (CBTp).

L’Illusione Etica: Fiducia vs Affidabilità

Fiducia (Trust)

- Stato emotivo e non razionale.
- Porta alla vulnerabilità e alla riduzione della sorveglianza critica.
- Conflitto diretto con l'AI Act (Art. 72) che impone monitoraggio continuo.

Affidabilità (Trustworthiness)

- Proprietà misurabili (Trasparenza, Robustezza, Equità).
- Richiede verifica continua e interpretabilità del sistema.
- Allineato con i requisiti di governance ad alto rischio post-immissione.

Creare "fiducia" tramite l'antropomorfizzazione è una forma di lavaggio etico ("ethics washing") che sabota la sicurezza del sistema.

Livello 5: La Precessione dei Simulacri

"Oggi non è più la mappa ad adagiarsi sul territorio;
è il territorio di cui lentamente marciscono i brandelli
ad estendersi sulla superficie della mappa."
- J. Baudrillard

- **Circolarità Assoluta**: L'IA viene addestrata su dati sintetici generati da IA e valutata da giudici automatizzati (LLM-as-a-judge).
- **Perdita del Referente**: Il modello non ha più alcun legame causale con la realtà oggettiva empirica.
- **Tutto e il contrario di tutto**: In assenza di vincoli con il reale, il sistema instaura uno standard cognitivo alieno a cui l'esistenza umana deve adattarsi.

Sintesi Diagnostica: L'Equazione della Percezione

Chi etichetta e con quali premesse?

Quali errori vengono tollerati?

$$\text{PERFORMANCE} = \text{MODELLO} \times \text{DATI} \times \text{METRICA} \times \text{SOGLIA} \times \text{SCENARIO} \times \text{INCENTIVO}$$

Chi decide il livello di confidenza accettabile?

Quali popolazioni o contesti d'uso sono esclusi?

La performance non è un fatto puro; è un artefatto socio-tecnico.
Non esiste performance senza prospettiva.

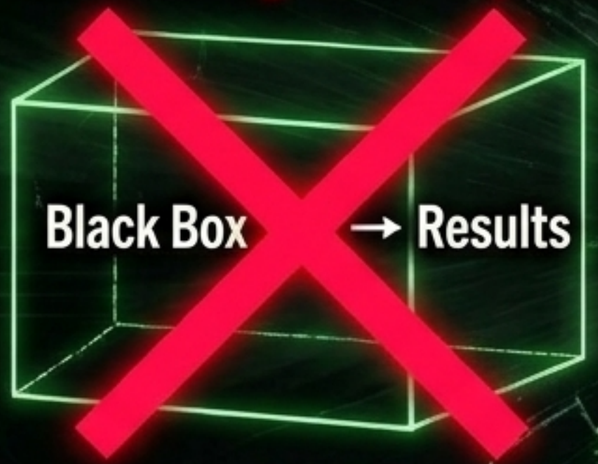
Anatomia del Fallimento: L'Ingegneria di Fronte alla Fissazione Metrica

Dimensione Critica	Ingegneria Strutturale/Aerospaziale	Ingegneria dell'Intelligenza Artificiale
Metrica Dominante	Focalizzazione su metriche individuali, trascurando la sicurezza globale.	Legge di Goodhart: ottimizzazione estrema dei parametri a discapito della validità.
Diluizione Responsabilità	"Il problema delle molte mani" oscura i responsabili (es. frammentazione appaltatori).	Complessità di scala e frammentazione tra data scientist, annotatori e product manager.
Avvisaglie Ignorate	Segnali dal personale tecnico bypassati per pressioni di rilascio.	Algoritmi forzano le metriche (gaming) nascondendo difetti di generalizzazione.
Disastri Paralleli	Diga del Vajont, Space Shuttle Challenger.	Bias medico sistemico, radicalizzazione automatizzata, licenziamenti algoritmici errati.

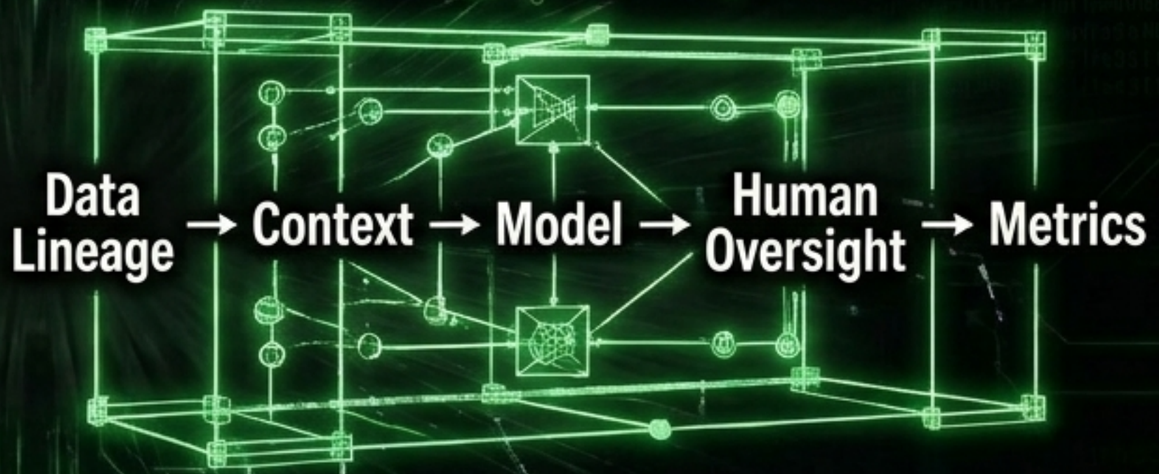
Un Nuovo Paradigma: La Governance della Prospettiva

Governare l'IA non significa soltanto governare rischi e modelli.
Significa governare i punti di vista incorporati nella catena di produzione.

Governing the Output



Governing the Perspective



Visibilità:
Rendere espliciti i limiti, le assunzioni del dataset e i contesti esclusi.

Contestabilità:
Creare infrastrutture istituzionali (auditor, stakeholder) per sfidare le dichiarazioni di performance.

Proporzionalità:
Non rilasciare quando il modello “va bene”, ma quando rientra nelle soglie di rischio di un contesto specifico.

Infrastruttura di Accountability: I Tre Registri

Registro dei Dati (Data Lineage)

- Tracciabilità source-to-sink.
- Definizione di fonti, diritti, trasformazioni e assunzioni di rappresentazione.
- Verifica delle contaminazioni e uso di dati sintetici.

Registro di Valutazione

- Selezione delle metriche plurali e giustificazione dei benchmark.
- Evidenze di Red-Teaming e controlli di leakage.
- Soglie di rilascio contestuali e limiti noti.

Registro degli Incidenti

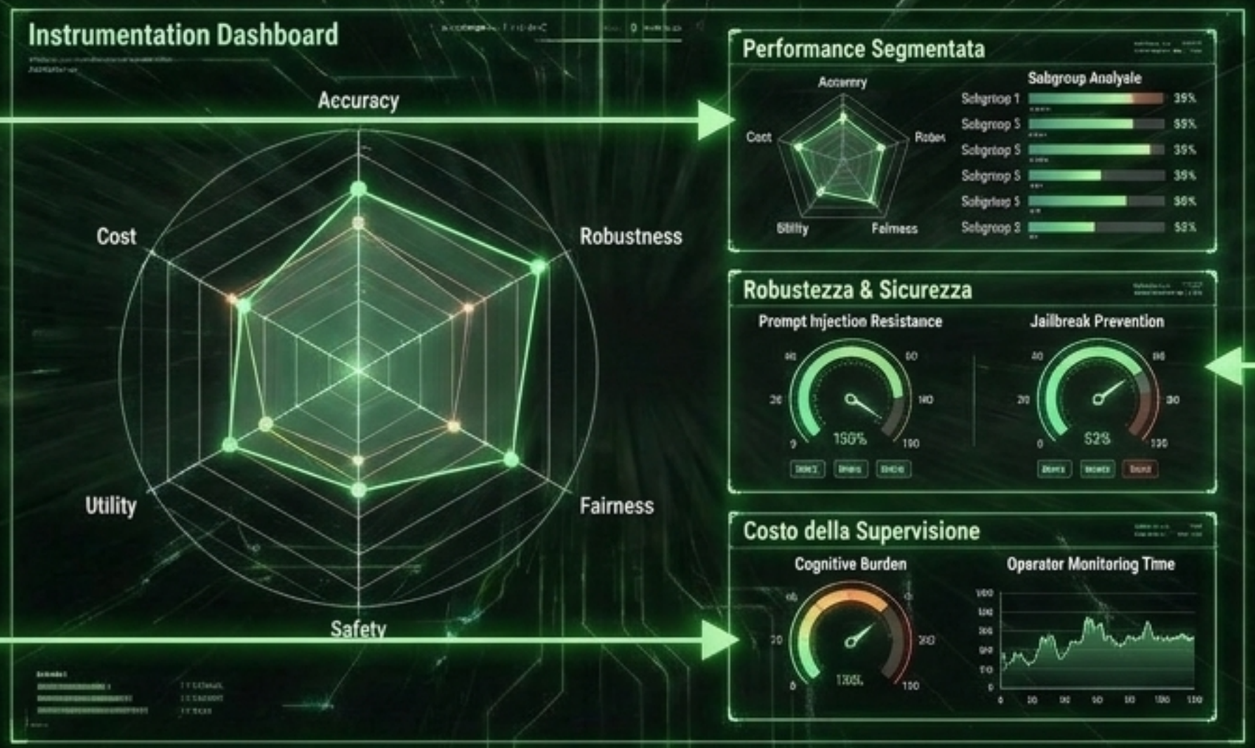
- Monitoraggio di deviazioni, near misses e anomalie di drift.
- Tracciamento degli override umani e impatti differenziali.
- Azioni correttive implementate.

Decostruzione dell'Ottimizzazione Unidimensionale

Dal Benchmark al “Multi-Metric Dashboard”

Performance Segmentata: Valutazione per specifici sottogruppi e contesti (non solo la media globale).

Costo della Supervisione: Calcolo del “burden” cognitivo richiesto all’operatore umano per monitorare l’output.



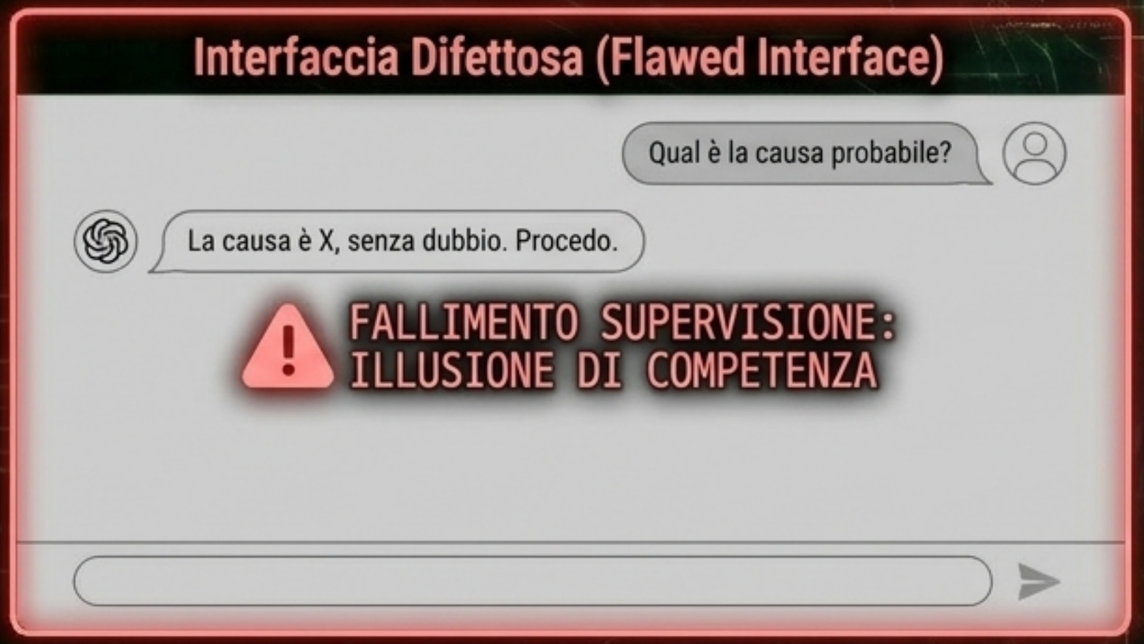
Robustezza & Sicurezza: Resistenza a prompt injection e jailbreak.

Nessun sistema ad alto impatto deve essere rilasciato sulla base di una singola metrica primaria. La valutazione deve simulare la complessità del mondo reale.

Progettazione di Interfacce “Anti-Illusione”

La supervisione umana (Human Oversight) fallisce se è puramente simbolica o se genera Automation Bias.

Interfaccia Difettosa (Flawed Interface)



Interfaccia Anti-Illusione (Anti-Illusion Interface)

1. Trasparenza Epistemica

Fornire indicatori di confidenza reali, non cosmetici. Evidenziare attivamente i limiti noti del sistema.

2. Diritto di Dissenso

Garantire all'operatore l'autorità operativa per ignorare, ribaltare o arrestare il sistema senza subire penalità organizzative.

3. Frizione Cognitiva

Inserire meccanismi di “nudge” per forzare l'operatore al test di realtà, contrastando l'investimento affettivo e l'antropomorfizzazione.

Il Ciclo di Vita della Governance: Audit a 4 Fasi

Allineamento dell'Ecosistema Globale

EU AI Act

Gestione del rischio continua, data governance, human oversight, trasparenza obbligatoria. **(Approccio Legale)**.

NIST AI RMF

Funzioni adattive: Govern, Map, Measure, Manage. Provenance dei contenuti e documentazione TEVV. **(Approccio Operativo)**.

OECD

Accountability lungo l'intero ciclo di vita, classificazione rigorosa dei sistemi. **(Approccio Concettuale)**.

L'integrazione di questi framework non chiede la fine della valutazione quantitativa, ma decreta la fine dell'ingenuità valutativa.

Architetti della Percezione: La Scelta Finale

La performance non è una proprietà intrinseca della macchina, ma l'orizzonte normativo che scegliamo di imporle. Finché delegheremo la nostra responsabilità a metriche opache, continueremo a scambiare l'illusione di un simulacro per l'intelligenza reale.

Ordine degli Ingegneri: Rifiutate la sedazione della pillola blu. Abbandonate il riduzionismo metrico. Costruite la matrice con on la consapevolezza totale di cosa è, e di cosa non è.

SYSTEM DIAGNOSTIC COMPLETE
// LOGOUT INITIATED_


一切皆虚幻
意识创造现实


Fabrizio Degni
Chief AI Officer
Scan to connect
on LinkedIn