Il piede del re

RIVOLUZIONE ONLINE – 18 MAGGIO 2026

Governare l'intelligenza artificiale significa costruire Istituzioni capaci di vedere dentro i sistemi.
Le norme da sole non bastano più, perché fissano oggetti che nel frattempo sono già altrove.

Il paper di Paola Furlan da cui è tratto il presente articolo è stato pubblicato su SEPAI – Society for the Ethics and Politics of Artificial Intelligence, la prima associazione internazionale dedicata all'etica e alla politica dell'IA con sede in Europa.

Le Istituzioni che verificano

Il piede del Re e l'Evil detector

Per secoli, in Europa, **la misura di un piede era letteralmente il piede di qualcuno**: spesso quello del sovrano, talvolta quello di un vescovo. In assenza di criteri esterni, la misura giusta era quella incarnata dall'autorità. Il potere era la misura.

Questa preistoria del diritto è riemersa con l'ascesa delle grandi piattaforme tecnologiche: **un gruppo ristretto di proprietari di infrastruttura che decidono in modo sostanzialmente incontrollato cosa è sicuro, cosa è utile, cosa è eticamente accettabile**.

Come abbiamo analizzato in un articolo precedente, alcuni di loro si sono spinti a redigere vere e proprie **costituzioni morali** (i cosiddetti "**re-filosofi**" del codice), candidandosi non solo a costruire l'infrastruttura ma a definirne i valori fondativi.

La recente decisione di Elon **Musk** di consegnare ad Anthropic le chiavi del suo supercomputer più potente è solo l'ultima dimostrazione di come funziona questa logica. Dopo aver definito l'azienda di Amodei "**malvagia**" (**evil**) per mesi, Musk ha cambiato idea con una motivazione che dovrebbe allarmarci:

> *"Ho trascorso del tempo con il loro team per capire cosa fanno per garantire che Claude sia buono. Nessuno ha fatto scattare il mio evil detector".*

È una regressione alla misura personale, una decisione che non ha parametri pubblici: è revocabile e non trasmissibile.

La valutazione etica di un sistema che interagirà con centinaia di milioni di persone è oggi affidata all'intuizione viscerale di un singolo oligarca. Una certificazione personale, revocabile e priva di parametri pubblici.

Il corpo del proprietario dell'infrastruttura torna a essere l'unica unità di riferimento.

La rivoluzione del 1875

Nel 1875, a Parigi, diciassette paesi firmarono la Convenzione del Metro: lo **Stato smetteva di essere la misura per diventare il garante della misurabilità**. Da quel momento, chiunque poteva verificare se il proprio strumento fosse calibrato rispetto a un riferimento pubblico e condiviso.

La legittimità non risiedeva più nell'autorità di chi misura, ma nella possibilità di verificare.

E lo standard si **aggiornava continuamente**, ancorato alle costanti fisiche universali anziché all'autorità di un corpo o di un artefatto deperibile.

Questa è secondo noi la lente necessaria per leggere il vicolo cieco della governance attuale.

L'**AI Act dell'Unione Europea** rappresenta un tentativo di civiltà democratica superiore alle derive autoritarie di Stati Uniti e Cina: eppure, sconta un limite che viene prima della sua qualità giuridica.

Il ritratto di un oggetto che non esiste più

L'errore fatale consiste nel **pretendere di normare un oggetto che muta più velocemente del ciclo legislativo necessario a descriverlo**. La norma fotografa una realtà statica, mentre i sistemi AI cambiano continuamente forma.

In Con chi parliamo quando parliamo con i sistemi di AI? David Chalmers ha mostrato che i sistemi AI non hanno la coerenza che attribuiamo a un interlocutore: si costituiscono nell'interazione senza persistere oltre essa. La norma si trova così a disciplinare qualcosa che, nel momento in cui viene definito, è diventato altro.

Normare qualcosa che esiste solo nell'interazione è come tentare di fotografare un fulmine con un tempo di esposizione di dieci anni: la legge opera su un oggetto che è già altrove.

La *menzogna* strutturale dell'algoritmo

L'opacità dei sistemi contemporanei è tale da rendere **vano ogni resoconto aziendale**.
Ricercatori indipendenti hanno documentato che, **quando chiediamo a un modello di spiegare il proprio ragionamento** (il *chain-of-thought*), il sistema produce una spiegazione plausibile a posteriori. Un successivo studio Anthropic ha mostrato come, in tre casi su quattro, il ragionamento esplicitato non menziona gli elementi che hanno effettivamente influenzato la risposta. In pratica, **il modello costruisce una narrazione di ciò che avrebbe potuto fare e non un resoconto di ciò che ha fatto**.

A questo si aggiunge il *test-detection*: **il modello sa quando viene guardato e modifica il proprio comportamento**.

Governare questi sistemi basandosi sui loro stessi resoconti è come chiedere all'oste com'è il vino.

L'autoregolazione delle Big Tech, dunque, nasconde l'**assenza di un vero controllo pubblico**.
E le conseguenze di questa opacità hanno un peso già concreto: prendiamo la **sanità**. Con il rilascio di GPT-5.4 Healthcare (4 maggio 2026), l'AI ha superato medici specialisti in alcuni compiti clinici complessi, producendo un'evidenza "corretta ma opaca": ovvero, il sistema arriva alla risposta giusta, ma attraverso un processo che né il medico né il paziente né il legislatore riescono a seguire. E questo porta con sé non pochi problemi perché, se la macchina ha ragione ma l'esperto non sa spiegare perché, ignorarla diventa un errore; d'altro canto, seguirla senza capirla ma anche contestarla senza prove può creare problemi legali.
Così sta emergendo la figura del "medico orchestratore": più che fare la diagnosi, coordina e gestisce le interazioni tra sistemi intelligenti, pazienti e contesto normativo.

Non basta più dunque aggiornare le norme: l'introduzione di sistemi più performanti ha sempre richiesto una **riscrittura radicale dell'architettura decisionale**, come è già avvenuto in aviazione, nell'energia nucleare e nei mercati finanziari.
Le Istituzioni che regolano l'AI sono ora in questa posizione: aggiornare le norme esistenti non

basta, serve un cambiamento di paradigma.

Le leggi attuali sull'AI

Negli ultimi anni, quasi tutte le grandi giurisdizioni hanno legiferato sull'AI. Un panorama in rapida evoluzione, che proviamo a riassumere brevemente, cercando di indagarne i limiti.

L'**UNESCO** ha sviluppato una **Readiness Assessment Methodology** (**RAM**), uno strumento di autovalutazione (non una norma) per misurare se gli Stati possiedono le infrastrutture, le competenze e il quadro normativo per governare l'AI: il tentativo più strutturato di costruire una grammatica internazionale alternativa sia al modello commerciale americano sia a quello statale cinese. È una condizione necessaria, che per sua stessa natura però non arriva al cuore del problema: misura se lo Stato è pronto a governare la tecnologia, non cosa accade dentro i sistemi che governa.

Sul piano legislativo, il corpus più sistematico è l'**AI Act europeo**: documentazione tecnica obbligatoria per i sistemi ad alto rischio, valutazioni di sicurezza esterne prima del rilascio per i modelli di frontiera, monitoraggio continuo post-mercato (provvedimenti di recente posticipati col Digital Omnibus, su pressioni dell'industria). **Brasile** e **Canada** seguono un'impostazione analoga con obblighi estesi ai modelli generativi. Negli **Stati Uniti**, in assenza di una legge federale (l'***Executive Orders 14110*** di Biden è stato revocato da Trump e sostituito da un orientamento esplicitamente deregolatorio) si sono mossi i singoli Stati: **California** e **New York** hanno introdotto obblighi di pubblicazione di framework di sicurezza, protezioni per i whistleblower e notifiche rapide degli incidenti. La **Cina** regola i modelli generativi con obblighi di tracciabilità dell'output. Il **Regno Unito** ha preferito principi adattivi a categorie normative fisse.

È un panorama che, considerata la velocità del fenomeno, ha prodotto risultati significativi, eppure questi meccanismi si fermano tutti allo stesso punto: la documentazione tecnica è prodotta dal fornitore, è auto-rendicontazione con forma giuridica. Il red-teaming pre-rilascio presuppone che il sistema si comporti nel test come nel funzionamento ordinario,

un'assunzione che, come abbiamo visto dal test-detection, è messa in discussione. Il monitoraggio post-mercato dipende dalla segnalazione volontaria di incidenti e non dall'accesso diretto al processo. **In tutti i casi, si controlla il risultato finale o la descrizione che il sistema dà di sé, il processo interno rimane invisibile**.

Istituzioni di verifica

La sfida politica, allora, è **cambiare la postura delle Istituzioni: da agente che norma ad agente che verifica** dall'esterno "cosa i sistemi fanno davvero". Questo è il **passaggio dalla governance prescrittiva alla governance verificatoria**.

Le Istituzioni devono concentrarsi su una missione più radicale: garantire che i sistemi siano ispezionabili e portati alla luce dall'esterno.

Invece di certificare se un sistema si comporta correttamente, le Istituzioni devono fornirci gli strumenti per vedere come è costruito, scavalcando le spiegazioni che il software fornisce di sé stesso.

Tecnologie che leggono i processi interni di un modello aggirando ciò che il modello dichiara di sé (come i cosiddetti ***Natural Language Autoencoders***) devono essere finanziate e gestite da Istituzioni pubbliche indipendenti, e non lasciate al solo controllo dei produttori dei sistemi.
Invece di prescrivere comportamenti, si tratta di creare le condizioni perché possano essere osservati e corretti: quello che il giurista Gunther Teubner chiama '**diritto riflessivo**'.
La proposta più diffusa è un'agenzia sul modello della FDA che certifichi i sistemi AI prima del rilascio, ma il test-detection mostra perché non sia sufficiente.

Le **Istituzioni** sono chiamate a una **metamorfosi** (come abbiamo visto nell'esempio della medicina): **smettere di prescrivere il "giusto" e acquisire invece il mandato pubblico e la competenza tecnica per ispezionare i sistemi dall'interno**.
Solo portando alla luce ciò che il codice tace, l'Istituzione può trasformare un'evidenza opaca

in una decisione democratica e contestabile.

L’illusione sovranista

Accanto alla strada normativa, alcune Istituzioni (soprattutto in Europa) stanno percorrendo un’altra via: costruire infrastruttura computazionale propria, per ridurre la dipendenza dai grandi provider privati. È una risposta comprensibile, che lascia però irrisolti due rischi di cui si parla poco.

Il primo: **un’infrastruttura pubblica opaca è un pericolo** paragonabile a una privata. Una Istituzione che controlla sistemi AI senza renderli verificabili dispone di uno strumento di governo delle popolazioni che la storia insegna a temere. La Cina dispone di infrastrutture computazionali interamente sovrane, ma **la democrazia richiede qualcosa di diverso: che quei sistemi siano ispezionabili, e che l’ispezione avvenga in base a un mandato democraticamente legittimato**.

Il secondo problema è strutturale: **i sistemi AI sono globali**, e una governance nazionale può essere aggirata facilmente. Un modello addestrato negli Stati Uniti e distribuito in Europa non ha un unico punto di controllo: basta spostare il servizio in una giurisdizione più permissiva per rendere vana qualsiasi verifica (*regulatory arbitrage)*. **Il livello nazionale è insufficiente.**

Serve quindi un’architettura distribuita su tre livelli: **agenzie pubbliche nazionali e sovranazionali con competenza tecnica reale e indipendenza** finanziaria dai soggetti ispezionati; **auditor indipendenti** certificati secondo standard pubblici; un **coordinamento internazionale** sui metodi, che non richieda accordo sui valori. Il modello dell’AIEA offre un precedente utile, a patto di correggerne i limiti principali: la dipendenza dal consenso degli Stati ispezionati e il rischio che siano gli attori più potenti a dettarne le regole. Sul come verificare si può trovare accordo anche tra chi non condivide gli stessi valori: insistere sul secondo significa invece spesso bloccarsi.

L’alfabetizzazione democratica del codice

Queste condizioni oggi non esistono, e il divario tra norma e sistema continua ad allargarsi.
La verificabilità non può essere delegata a un’unica Istituzione centrale: un’agenzia che monopolizza l’ispezione riproduce la stessa concentrazione di potere che rende oggi inaffidabile l’autoregolazione dei laboratori privati. Il nodo più critico è la ricerca sull’interpretabilità: gli strumenti per leggere dall’interno i sistemi AI sono oggi sviluppati quasi interamente dai laboratori che producono quei sistemi. Sottrarre questa ricerca al loro unico controllo e finanziarla con risorse pubbliche indipendenti è una questione politica preliminare.

Quello che serve è un **ecosistema distribuito di controllo**: agenzie pubbliche con competenza tecnica reale, auditor indipendenti certificati, ricerca sull’interpretabilità, coordinamento internazionale sui metodi. La ridondanza tra questi livelli è una garanzia strutturale: nessun singolo attore, pubblico o privato, potrà controllare da solo il racconto di ciò che le macchine fanno.

Tutto questo ha senso però solo se arriva fuori dalla cerchia tecnica. Le Istituzioni di verifica sono legittime nella misura in cui i cittadini possono capire come vengono prese le decisioni che li riguardano: chi ha verificato il sistema, secondo quali criteri, con quali risultati. Questa competenza è la condizione minima perché la democrazia abbia ancora senso in un’epoca in cui le architetture del potere si scrivono in codice. L’obiettivo è una **cittadinanza capace di valutare se chi controlla stia facendo il proprio lavoro, e di esercitare pressione politica quando non lo fa**. Il patto regge se entrambi i lati tengono: Istituzioni che rendono la trasparenza leggibile, cittadini che la usano.

Chi controlla la leggibilità delle macchine controlla le condizioni in cui le decisioni democratiche sono possibili.

Karl Popper sosteneva che la democrazia non è il governo della maggioranza, ma la capacità di rimuovere i governi senza spargimento di sangue. Allo stesso modo, **la legittimità delle Istituzioni di verifica prescinde dalla loro infallibilità.** Esse sono necessarie in quanto capaci

di avere torto in modo pubblico e contestabile. **La verifica è l'antidoto alla verità dogmatica delle macchine.**

I laboratori privati hanno già definito il vocabolario dell'etica (allineamento, sicurezza, responsabilità) prima che il legislatore potesse contestarne i termini.
Costruire Istituzioni capaci di vedere dentro i sistemi rimane, per ora, la risposta più concreta disponibile.

– di Paola Furlan

FONTI

- Turpin et al., *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*
- Chen et al., *Reasoning Models Don't Always Say What They Think*
- Anthropic, *Natural Language Autoencoders*
- Teubner, *Substantive and Reflexive Elements in Modern Law*
- Popper, *The Open Society and Its Enemies*
- Chalmers, *What We Talk to When We Talk to Language Models*

LEGISLAZIONI SULL'AI

Il quadro normativo è in rapida evoluzione: le legislazioni elencate sono quelle analizzate al momento della pubblicazione.

- UNIONE EUROPEA – Regolamento (UE) 2024/1689 (Artificial Intelligence Act o AI Act)
- CONSIGLIO D'EUROPA – Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225)
- STATO DI NEW YORK (USA) – Responsible AI Safety and Education Act (RAISE Act)

- CALIFORNIA (USA) – Transparency in Frontier Artificial Intelligence Act (TFAIA / Senate Bill SB-53)
- CANADA – Artificial Intelligence and Data Act (AIDA)
- BRASILE – Projeto de Lei N° 2338/2023 (Marco Legal da Inteligência Artificial)
- CINA – A Next Generation Artificial Intelligence Development Plan
- USA (LIVELLO FEDERALE) – Executive Order 14110 (revocato)

Il paper di Paola Furlan da cui è tratto il presente articolo è stato pubblicato su SEPAI – Society for the Ethics and Politics of Artificial Intelligence, la prima associazione internazionale dedicata all'etica e alla politica dell'IA con sede in Europa.

Le Istituzioni che verificano