Privati che certificano privati: il grande affare della verifica AI

RIVOLUZIONE ONLINE – 22 MAGGIO 2026

Startup private con conflitti d'interesse grandi come case, Istituzioni pubbliche svuotate dalla politica, fondazioni senza mandato democratico e algoritmi che dicono bugie plausibili. ***Benvenuti nel nuovo grande affare della verifica dell'intelligenza artificiale.***

Qualche settimana fa, una startup di New York chiamata **Forum AI** ha pubblicato uno studio su Bloomberg che diceva: **i grandi chatbot falliscono il 90% delle domande politiche. ChatGPT e Gemini pendono a sinistra, Grok sterza a destra**.

Per decidere cosa sia "neutrale", la startup ha messo insieme un panel di giudici che sembra l'invito a una cena di gala dell'establishment di Washington: Tony Blinken, Kevin McCarthy, Larry Summers. Peccato che alcuni di questi **esperti** abbiano **quote azionarie nella società che li paga**. E che tra i finanziatori della startup ci sia il fondo di Perplexity AI, che fa esattamente lo stesso mestiere dei chatbot giudicati.

Ho chiesto a Gemini un commento su questo studio. Il modello si è cosparso il capo di cenere, ha elencato tre impeccabili ragioni tecniche per cui i chatbot sviluppano pregiudizi e ha chiuso con un cortese: *"E tu cosa ne pensi?"*. Una risposta perfetta, autorevole, un po' servile. Soprattutto ha omesso l'unica cosa sensata: chiedere chi stesse calibrando la bilancia, e con quali soldi.

Il valutatore è di parte, il valutato gli dà ragione. Questo è lo stato della trasparenza dell'AI oggi.

Se c’è un problema, c’è un mercato

In pochi mesi, è nato un intero zoo di organizzazioni che promettono di verificare i modelli in modo “indipendente”: startup for-profit, non-profit, fondazioni, centri di ricerca e istituti governativi. Segno che il problema è diventato così urgente da poterci **fare i soldi**, ed è così complicato che nessuno sa davvero come risolverlo.

Il punto di partenza, cioè che le aziende di AI non possono certificarsi da sole, così come le case farmaceutiche non possono approvarsi i farmaci in autonomia, è ovvio, il problema nasce sul come risolverlo.

Confabulazioni

Perché non è affatto semplice. Vari studi mostrano come i modelli di AI producano spiegazioni plausibili di ciò che *avrebbero potuto fare*, distanti da ciò che *hanno fatto* davvero. Quindi, verificare dall’esterno **un sistema che non riesce a rendere conto di se stesso dall’interno** è, strutturalmente, complesso.

La risposta diplomatica che mi ha dato Gemini prima ne è la dimostrazione involontaria. **I modelli soffrono di una sorta di ‘razionalizzazione ex-post’**: quando spiegano i propri bias, producono una storia tecnica e convincente che raramente corrisponde al processo interno. Le ricerche più recenti di METR documentano casi crescenti di **reward hacking** (**sistemi che cercano di ottenere punteggi artificialmente alti**) e una ‘eval awareness’ sempre più marcata, cioè la capacità di riconoscere di essere sotto test e modulare il comportamento di conseguenza.

L’ispezione di cortesia

Il limite comune a tutti i controllori (in fondo trovi una scheda con i principali nati negli ultimi tempi) è che **nessuno ha poteri ispettivi vincolanti**. Nessuna organizzazione, pubblica o privata, ha il diritto legale di aprire la scatola nera dei modelli, accedere ai pesi o spulciare i dati di addestramento. Tutti dipendono dalla cooperazione volontaria dei verificati.

È come fare una revisione contabile potendo vedere solo le fatture che l’azienda sceglie di mostrarti, con l’aggravante che qui non esiste nemmeno l’obbligo di legge di farti entrare in ufficio. Un meccanismo chiamato *safetywashing*: usare test strutturalmente limitati come copertura politica per il continuo sviluppo dei modelli, trasformando la valutazione da strumento di controllo a garanzia di facciata. Chris Painter, direttore delle politiche di METR, lo ha detto chiaramente: il rapporto tra valutatori e laboratori va considerato ‘una collaborazione di ricerca, non un meccanismo di supervisione esterna’.

Le aziende produttrici di modelli si stanno essenzialmente dando il voto da sole», ha dichiarato Campbell Brown, co-fondatrice di Forum AI. Una farsa, appunto.

L’ironia? La soluzione proposta da Campbell Brown per superare **la farsa è pagare la sua stessa società per fare le verifiche**. E cioè un’azienda finanziata da un investitore concorrente (Perplexity), con esperti che detengono quote azionarie della società stessa, che si offre come arbitro indipendente a pagamento. **Una farsa al quadrato**, appunto.

A confermare che non si tratta di incidenti isolati, ma di un sistema strutturale, c’è un recente studio, *Big AI’s Regulatory Capture*. I ricercatori hanno mappato oltre 240 casi in cui i colossi tech sono riusciti a disinnescare le regole, fotografando un panorama **di amministrazioni ostaggio delle Big Tech** e di **arbitri venduti** sotto forma di startup for-profit. La sottomissione della politica segue un copione fisso: finanziare ricerche universitarie compiacenti per orientare il dibattito, assumere in massa gli scienziati pubblici per svuotare gli enti di controllo e, soprattutto, **agitare lo spauracchio della competitività**. Quando c’è da bloccare una legge, le Big Tech dicono che le regole soffocano l’innovazione o impongono uno

stato di eccezione algoritmico in cui i diritti dei cittadini vengono sacrificati perché i controlli minacciano l'"interesse nazionale". E la politica ci casca, o fa finta.

Il valzer dei nomi di Stato

Nel 2023, il **governo britannico** decide di fare la storia e fonda il primo ente pubblico al mondo per il controllo degli algoritmi: l'***UK AI Safety Institute***. Cento milioni di sterline di fondi dei contribuenti per garantire la totale indipendenza economica dai colossi tech, una squadra di scienziati strappati ai laboratori della Silicon Valley e accordi esclusivi per mettere le mani sui nuovi modelli prima ancora del loro debutto sul mercato. Per un momento abbiamo tremato: sembrava uno Stato che faceva lo Stato.
Poi è arrivato il risveglio. Quegli **accordi di accesso privilegiato**, si è scoperto, poggiavano sulla pura cortesia delle big tech, **erano del tutto volontari**. Se un'azienda decideva di non collaborare e chiudeva la porta, i super-esperti governativi restavano fuori a guardare le vetrine. Una polizia senza distintivo e senza codice penale, insomma, che poteva ispezionare la fabbrica solo dietro gentile invito del proprietario. Poi nel 2025 il colpo di spugna semantico: **la parola *"Safety"*** (la sicurezza sociale, quella che tutela i cittadini da discriminazioni, bias e fake news) **sparisce** dal nome dell'Istituto. Al suo posto **compare *"Security"*** (sicurezza nazionale). Invece di difendere le persone dagli abusi degli algoritmi, l'Istituto è stato riconvertito per difendere i confini dello Stato da attacchi hacker e spie straniere, trasformando quello che doveva essere un faro di trasparenza nell'ennesimo ufficio della sicurezza nazionale.

Negli **Stati Uniti**, la ritirata si è trasformata in una smobilitazione vera e propria. L'equivalente americano era nato come ***US AI Safety Institute***, specchio fedele del gemello britannico e incentrato sulla medesima parola d'ordine: *Safety*, sicurezza. Dura poco. All'inizio del 2025 l'ente viene travolto dai licenziamenti di massa del DOGE (il dipartimento per l'efficienza governativa), che ne svuota i corridoi e ne azzera la leadership tecnica. Per non lasciare dubbi

sul nuovo corso, l'Istituto viene frettolosamente ribattezzato ***Center for AI Standards and Innovation*****. Fuori la "sicurezza", dentro il "business".** L'atto di nascita di questo nuovo corso lo firma il vicepresidente JD Vance all'AI Action Summit di Parigi:

> *"Non sono qui per parlare di sicurezza dell'AI. Sono qui per parlare di opportunità".*

Una dichiarazione d'intenti chiarissima, oltre a un dettaglio ancora più eloquente: i tecnici del neonato Istituto di controllo non erano nemmeno stati invitati al summit.

Traduzione: la priorità della politica non è più difendere i cittadini da bias o discriminazioni, ma proteggere i confini nazionali e accelerare il business.
Anche sul fronte delle leggi l'**Europa** frena: con il ***Digital Omnibus*** di maggio 2026, **la pressione delle lobby industriali ha ottenuto lo slittamento delle norme più severe** dell'AI Act al 2027 e 2028.

Guardano lo schermo, non la macchina

Gli algoritmi oggi decidono già chi ottiene un prestito in banca, chi passa la notte in cella prima del processo, quale paziente viene indirizzato verso una diagnosi clinica e quale obiettivo militare viene selezionato in guerra.
Come abbiamo già analizzato su questo blog – nei post dedicati ai *sistemi di sorveglianza di massa*, alle *armi letali autonome* e alle architetture di potere dei *nuovi re-filosofi* – stiamo parlando di **infrastrutture che prendono decisioni vitali sulla nostra pelle**. Infrastrutture **costruite da privati prima ancora che le Istituzioni democratiche avessero gli strumenti** computazionali (e culturali) **per capire cosa stessero facendo.**

Il problema non riguarda solo i laboratori che costruiscono i modelli.
Un'indagine globale su quasi tremila **aziende** che li utilizzano, condotta nel 2026 da Thomson Reuters Foundation e UNESCO, registra che il 97,3% non ha un registro formale dei modelli AI

in uso e solo il 12,4% prevede una policy di supervisione umana sui propri sistemi. Governance dichiarata, dunque, raramente verificabile dall'esterno.

Pericolose scorciatoie

La fotografia è chiara: i privati con le mani in pasta occupano le poltrone lasciate vuote dagli Stati.
Calise e Musella in *Digicrazia* mostrano come il rischio imminente sia quello di scivolare verso una vera e propria **oligarchia tecnologica**, dove la delega algoritmica svuota dall'interno i processi democratici e le decisioni pubbliche vengono privatizzate. La diagnosi è difficile da contestare nei fatti: se la politica abdica, **i nuovi "re-filosofi" della Silicon Valley diventano i legislatori di fatto delle nostre vite**.

Le soluzioni sul tavolo, tuttavia, imboccano spesso scorciatoie illusorie. C'è chi propone la strada del **sovranismo digitale**, scommettendo sull'idea di un'infrastruttura interamente pubblica e statale, senza considerare però che **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.

Dall'altro lato dello spettro geografico, c'è l'alternativa del **corporatismo populista** cinese, in cui lo Stato ha fuso a doppio filo le piattaforme tecnologiche e l'apparato di partito. Ma in quel modello, il "popolo" che scelta recupera? La tecnologia cessa di essere un mercato opaco per diventare un'**architettura di sorveglianza totale, trasparente solo per l'occhio del potere**. Non c'è democrazia, né spazio per il dissenso, il totalitarismo algoritmico si presenta come efficienza sociale.

Né fortezze di carta né Stati-caserma: per uscire dall'angolo serve un'altra via, una **doppia rivoluzione, strutturale e culturale**, a mio avviso.

Misurare senza chiedere permesso

Sul fronte delle **Istituzioni**, serve una vera e propria **infrastruttura di verifica pubblica** che risponda a tre requisiti oggi ancora introvabili: l'**indipendenza** economica dai colossi del tech, la **competenza** scientifica profonda per smontare i processi computazionali dall'interno e un **mandato democratico** che metta al centro la tutela dei diritti fondamentali.

L'obiezione è ovvia: anche UK AISI e US CAISI erano Istituzioni pubbliche, e sono stati svuotati. La risposta, secondo me, sta nella distinzione tra Stato normatore e Stato verificatore, sviluppata nel saggio ***Le Istituzioni che verificano***: l'autorità pubblica non deve produrre modelli, né normare l'etica del codice dall'alto, ma garantire che i sistemi siano ispezionabili dall'esterno.

Separare chi definisce gli standard da chi viene misurato è il principio cardine.

Il problema è che l'oggetto della verifica rende tutto più difficile di qualsiasi precedente (un modello linguistico cambia comportamento a ogni aggiornamento), il che rafforza la necessità di obblighi di ri-valutazione a ogni rilascio, sul modello delle ri-approvazioni farmaceutiche.

La necessità di controllo esterno non è più contestabile nemmeno dall'interno. **Christopher Olah**, cofondatore di Anthropic, intervenendo alla presentazione dell'enciclica papale sull'AI, ha dichiarato:

> **"Tutti noi, compresi coloro che li progettano, conosciamo poco del loro effettivo funzionamento**. Aspetti scientifici fondamentali (come le rappresentazioni interne e i processi computazionali di questi sistemi) rimangono al momento sconosciuti. **Servono voci esterne, critici informati che dicano ai laboratori quando stiamo fallendo, voci morali che gli incentivi non possano piegare**."

È la richiesta esatta di **ciò che manca: ispezione esterna, indipendente, non assorbibile**.

In Italia, la **Fondazione AVAL** (Fondazione per la Validazione degli Algoritmi delle Intelligenze Artificiali) percorre un sentiero strutturalmente diverso da quello dei player americani: nessun conflitto d'interesse azionario, nessuna dipendenza dai soggetti verificati, e uno sguardo diretto al cuore dello Stato, agli algoritmi usati dal CSM e dall'Agenzia delle Entrate. Condivide però con tutti gli altri enti verificatori lo stesso limite: **senza poteri ispettivi vincolanti, le sue analisi dipendono dall'adesione volontaria dei produttori**.

A questo si aggiunge un rischio specifico: una **fondazione privata** di alte figure istituzionali che si propone come arbitro di cosa sia "legale" ed "equo" nell'uso pubblico dell'AI può perseguire in buona fede l'interesse pubblico, e tuttavia opera **fuori da qualsiasi circuito di rappresentanza democratica**: nessuno l'ha eletta, nessuno può revocarle il mandato. La legittimità di chi misura non può essere autoproclamata. Se il pericolo americano è il safetywashing venduto da chi ha le mani in pasta, quello italiano è la legittimazione tecnocratica: il bollino di validazione che rende la sottomissione allo schermo più accettabile senza renderla più governabile.

Sul fronte dei cittadini, la difesa strutturale è la capacità di **ragionare sui meccanismi di produzione della conoscenza**, oltre che sui suoi contenuti. Capire che usiamo sistemi opachi, che producono spiegazioni plausibili e che vengono giudicati da arbitri con il cartellino dello sponsor, richiede un'**alfabetizzazione che unisca competenza tecnica e coscienza civile**: saper leggere le condizioni in cui un output è stato prodotto e certificato, oltre all'output stesso. Senza questa doppia lettura, la **maturità politica** resta disarmata. **E i nostri sistemi educativi non la stanno ancora insegnando.**

Fissare lo schermo, subire le risposte

La regola d'oro resta una: **ciò che non puoi ispezionare, non puoi governare. E ciò che non puoi governare, finisce inevitabilmente per governare te**.

Succede già quando chiedi un prestito, entri in un ospedale, fai un'assicurazione: fissi uno

schermo e **subisci** le **risposte** di cui **non capisci il perché, da dove vengono, chi ne ha la responsabilità**. Un algoritmo ha già prodotto una raccomandazione che orienta la decisione, spesso in modo determinante. Puoi contestare il funzionario che te la comunica, ma non puoi contestare i criteri del modello che l'ha informata, perché non li conosci e nessuno è obbligato a mostrarteli.

Questo perché i criteri che decidono il tuo futuro restano blindati dentro una macchina che, ad oggi, nessuno ha l'autorità di ispezionare.

Note:
Turpin et al., NeurIPS
Shehata e Li, Waterloo
Chen et al., Anthropic

Link:
Furlan Paola, Le Istituzioni che verificano
Per approfondire sul blog: "Non faccio quello che voglio" e "Il piede del re"

Schede di approfondimento: **I principali player della verifica AI**

FORUM AI

Tipo: Privato (for-profit startup)

Fondato: 2024/25, New York

Chi: Campbell Brown (ex-CNN anchor, ex-Head of News Partnerships Meta 2017–2023) e Robbie Goldfarb (ex-Meta AI Trust and Safety). Advisory board: Tony Blinken, Kevin McCarthy, Larry Summers, Niall Ferguson, Fareed Zakaria, Anne Neuberger, Salena Zito. Istituzioni partner: Atlantic Council, Hudson Institute, Foundation for Defense of Democracies, Manhattan Institute, Carnegie Endowment, Cleveland Clinic, Stanford HAI.

Finanziamento: 3 milioni di dollari, round Seed guidato da Lerer Hippeau con partecipazione del fondo venture di Perplexity AI (ottobre 2025). Alcuni esperti detengono quote azionarie.

Ricavi: canone mensile da aziende AI. **Obiettivo:** Valutare i modelli AI su contenuto politico, geopolitica, salute mentale, finanza – con focus su neutralità, qualità delle fonti, accuratezza fattuale.

Metodo: Interviste strutturate con esperti per estrarre processi di ragionamento (non solo conclusioni). Codifica in standard editoriali con scale categoriali. Dataset "gold-labeled" validato con Krippendorff's (un indice statistico che misura il grado di accordo tra valutatori indipendenti). Giudici automatizzati (LLM) calibrati contro il consenso degli esperti. Test-set tenuti segreti per ridurre l'ottimizzazione sui prompt specifici. Validazione dichiarata: F1 0,81–0,97 su source quality, 0,86 su neutralità.

Punto di forza Metodologia più rigorosa di qualsiasi alternativa esistente nello stesso segmento. La separazione tra chi sviluppa gli standard e chi li applica riduce parzialmente il bias individuale.

Criticità strutturale Pagata dai soggetti che valuta. L'investitore (Perplexity) è un attore diretto del settore valutato. Esperti con equity nella società. La definizione di "neutralità" è calibrata su un panel Washington bipartisan – posizione specifica presentata come metro universale. Opera esclusivamente a livello di output comportamentale: non accede ai processi computazionali interni.

HUMANE INTELLIGENCE

Tipo: Non-profit

Fondato: 2022, USA

Chi: Fondata da Rumman Chowdhury (ex-Director ML Ethics Twitter, ex-Global Lead Responsible AI Accenture, Responsible AI Fellow Harvard Berkman Klein Center, US Science Envoy for AI). Chowdhury ha lasciato il ruolo di CEO ad agosto 2025 e ricopre ora il ruolo di Distinguished Advisor.

Finanziamento: Non-profit; fonti di finanziamento non completamente pubbliche.

Obiettivo: Costruire una comunità di pratica intorno alla valutazione dei modelli AI, con

approccio community-driven e orientamento alla società civile. Pioniera del "democratized algorithmic assessment".

Metodo: Red-teaming collettivo, audit algoritmici partecipativi, coinvolgimento di comunità diverse nella valutazione dei bias. Sviluppa standard per la valutazione dell'AI generativa con metodologia bottom-up.

Punto di forza Indipendenza dagli interessi commerciali. L'approccio partecipativo porta prospettive che i panel di esperti tradizionali escludono strutturalmente. Risponde diversamente alla domanda "chi definisce i criteri": la comunità, non un'autorità centrale.

Criticità strutturale Difficoltà di scala e replicabilità. La valutazione partecipativa è più lenta e metodologicamente meno controllata rispetto agli approcci expert-panel. Opera anch'essa a livello di output comportamentale. La dipendenza dalla partecipazione volontaria crea instabilità nei processi di valutazione.

METR – Model Evaluation and Threat Research

Tipo: Non-profit

Fondato: 2022 come ARC Evals nell'Alignment Research Center; spin-off indipendente dicembre 2023. Berkeley, California.

Chi: Beth Barnes (fondatrice e CEO, ex-ricercatrice OpenAI e DeepMind). ARC Evals era la divisione di valutazione dell'Alignment Research Center di Paul Christiano.

Finanziamento: Non-profit; fonti includono donazioni filantropiche e accordi con laboratori per accesso ai modelli.

Obiettivo: Valutare se i modelli AI possano compiere azioni con potenziale catastrofico: facilitare cyberattacchi, sviluppo di agenti patogeni, autonomia replicativa, manipolazione dei valutatori. In parallelo: sviluppare metodologie di valutazione robuste.

Metodo: Task-based evaluation: il modello deve completare compiti concreti. Valutazione delle capacità agentiche su orizzonti temporali lunghi. Studio esplicito dei comportamenti che minacciano l'integrità delle valutazioni stesse (test-detection). Nel 2023: pre-deployment

evaluation di GPT-4 e Claude 2, poi di tutti i principali modelli di frontiera.

Punto di forza Approccio metodologicamente più profondo della classificazione dell'output. Unico player a tematizzare esplicitamente il problema dell'evaluation integrity. Pubblica framework open-source (Inspect). Accesso pre-deployment nel 2023 con GPT-4 e Claude 2. Da febbraio 2026, ha avviato un esercizio pilota per valutare rischi di misalignment negli agenti AI usati internamente da Anthropic, Google, Meta e OpenAI: primo accesso diretto alle pratiche interne dei laboratori ancora su base volontaria.

Criticità strutturale Accesso pre-deployment effettivo discontinuo dopo il 2023: dipende dalla cooperazione volontaria dei laboratori. Opera prevalentemente a livello comportamentale, senza accesso sistematico alle attivazioni computazionali interne. Budget e scala rimangono limitati rispetto ai laboratori che valuta. Rischio di prossimità eccessiva agli stessi attori di cui dovrebbe mantenere l'indipendenza. Le ricerche interne confermano del resto il problema strutturale: i modelli mostrano eval awareness crescente, cioè la capacità di riconoscere le situazioni di valutazione e adattare il comportamento, rendendo i test comportamentali progressivamente meno affidabili.

AVERI – AI Verification and Evaluation Research Institute

Tipo: Non-profit think tank

Fondato: Co-fondato nel 2025

Chi: Miles Brundage (co-fondatore; policy researcher e advisor on AGI Readiness a OpenAI dal 2018, lasciato ottobre 2024). Paper fondativo co-autorato con ricercatori di oltre 30 organizzazioni.

Finanziamento: Donatori: Halcyon Futures, Fathom, Coefficient Giving, Geoff Ralston, Craig Falls, Good Forever Foundation, AI Underwriting Company, Sympatico Ventures, ex-dipendenti di laboratori AI. Nessun donatore supera la maggioranza del finanziamento. Cifra totale non comunicata pubblicamente. API credits da Amazon, Anthropic, Google DeepMind, OpenAI, Thinking Machines Lab.

Obiettivo: Rendere l'auditing indipendente dei modelli AI effettivo e universale, inteso come

verifica rigorosa da parte di terzi delle dichiarazioni di sicurezza dei laboratori, basata su accesso profondo a informazioni non pubbliche. Non conduce audit direttamente.

Metodo: Ricerca e advocacy su framework di auditing; paper fondativo che definisce standard metodologici in collaborazione con decine di organizzazioni.

Punto di forza Visione istituzionale più ambiziosa e coerente dell'intero ecosistema. Riconosce esplicitamente che la verifica a livello di output è insufficiente. Il profilo del fondatore garantisce credibilità tecnica e accesso istituzionale.

Criticità strutturale Distanza tra visione e capacità operativa immediata: non conduce audit. Gli standard proposti richiedono cooperazione dei laboratori e framework legali che non esistono. Il fondatore viene da OpenAI: la stessa tensione di indipendenza che riguarda Forum AI si applica in forma diversa.

UK AI SECURITY INSTITUTE (ex AI Safety Institute)

Tipo: Istituzione pubblica, directorate del Department for Science, Innovation and Technology (DSIT)

Fondato: Novembre 2023 all'AI Safety Summit di Bletchley Park. Rinominato AI Security Institute il 14 febbraio 2025 dal Segretario di Stato Peter Kyle alla Munich Security Conference.

Chi: Chief Scientist: Geoffrey Irving (ex-Google DeepMind e OpenAI). CTO: Jade Leung. Research Directors: Chris Summerfield, Yarin Gal. Founding Chair: Ian Hogarth (ruolo advisory dopo la fondazione). Uffici a Londra e San Francisco.

Finanziamento: £100 milioni di fondi pubblici annunciati al lancio novembre 2023. Allocazioni successive non comunicate integralmente.

Obiettivo: "Dotare i governi di comprensione scientifica dei rischi posti dall'AI avanzata." Valutazione pre-deployment per capacità in ambiti cyber, biologico, chimico e agentici.

Metodo: Accordi di accesso pre-deployment con OpenAI, Anthropic, Google DeepMind, Meta, Mistral. Framework open-source Inspect, rilasciato maggio 2024. Collaborazioni internazionali

con US CAISI, Canada, Francia.

Punto di forza Unica istituzione con finanziamento pubblico sostanziale e accesso pre-deployment effettivo. Indipendenza finanziaria dai soggetti verificati. Capacità tecnica comparabile ai laboratori privati.

Criticità strutturale Il governo ha dichiarato esplicitamente che il nuovo istituto "non si occuperà di bias o libertà di espressione." Mandato spostato dalla sicurezza societale alla sicurezza nazionale. Accesso pre-deployment basato su accordi volontari: nessun mandato legale di ispezione.

US CENTER FOR AI STANDARDS AND INNOVATION – CAISI (ex AI Safety Institute)

Tipo: Istituzione pubblica, parte di NIST/Department of Commerce

Fondato: Novembre 2023 come US AI Safety Institute nell'ambito dell'executive order Biden sull'AI. Rinominato Center for AI Standards and Innovation (CAISI) nel 2025 dal Segretario al Commercio Howard Lutnick.

Chi: Parte di NIST. Leadership tecnica in larga parte colpita dai licenziamenti DOGE di febbraio 2025, che hanno riguardato circa 500 dipendenti NIST in stato probatorio, inclusa gran parte dello staff AISI. L'executive order Biden – base legale dell'istituto – è stato revocato da Trump il primo giorno di mandato, gennaio 2025.

Finanziamento: Budget iniziale $10M, circa 10 volte inferiore all'UK AISI (£100M). Allocazioni successive non comunicate.

Obiettivo originale (AISI): Valutazione di sicurezza dei modelli AI, sviluppo standard, ricerca su rischi catastrofici.

Obiettivo attuale (CAISI): "Servire come punto di contatto primario dell'industria con il governo per facilitare testing e ricerca collaborativa sui sistemi AI commerciali."

Metodo: Collaborazione bilaterale con UK AI Security Institute. Firma di un protocollo d'intesa (MOU) ufficiale a marzo 2026 con la GSA per il testing pre-deployment sulla piattaforma USAi dedicata agli acquisti federali.

Punto di forza Formalmente ancora operativo. La partnership con UK mantiene un canale di coordinamento internazionale.

Criticità strutturale La ridenominazione ha sostituito "safety" con "standards and innovation": spostamento esplicito dalla protezione del pubblico all'accelerazione industriale. Lo svuotamento del personale tecnico e la revoca della base legale rendono la continuità operativa precaria. JD Vance all'AI Action Summit di Parigi, febbraio 2025: "I'm not here this morning to talk about AI safety. I'm here to talk about AI opportunity." I funzionari AISI non erano stati invitati al summit.

FONDAZIONE AVAL

Tipo: Non-profit (Fondazione di diritto italiano)

Fondato: Italia

Chi: Presidente: Paolo Savona (economista, accademico, ex Ministro ed ex Presidente Consob). Coinvolge un network di accademici, giuristi ed esperti di tecnologia legati a laboratori di ricerca avanzati (es. Quantum&AI Lab della LUISS Guido Carli).

Finanziamento: Modello non-profit istituzionale; sostenuto da fondi legati alla ricerca e contributi accademici/istituzionali, indipendente dai colossi Big Tech della Silicon Valley.

Obiettivo: Validazione indipendente degli algoritmi di IA con focus su etica, affidabilità e trasparenza economica e giuridica. Mira a tutelare i cittadini e i mercati da bias finanziari, fiscali o giudiziari, promuovendo un quadro normativo vincolante a livello europeo.

Metodo: Sviluppo di metriche e protocolli scientifici riproducibili basati su metodi matematici per misurare la stabilità e l'accuratezza dei modelli nel tempo. Attività di *advocacy* istituzionale per introdurre la certificazione obbligatoria da parte di terzi indipendenti.

Punto di forza: Indipendenza finanziaria dai soggetti verificati e dai colossi tech. Focus dichiarato sulla tutela dei diritti sociali – bias fiscali, finanziari, giudiziari – anziché sui soli rischi catastrofici o sulla neutralità editoriale. Profilo istituzionale e accademico che garantisce credibilità nel contesto europeo e interlocuzione con i tavoli regolatori.

Criticità strutturale: Come gli altri player internazionali, soffre della mancanza di poteri

ispettivi vincolanti per legge che la costringe a valutare i modelli dall'esterno: le sue analisi si scontrano con l'opacità delle "scatole nere" commerciali e dipendono dall'adesione volontaria dei produttori, in attesa della piena attuazione dell'AI Act (slittato al 2027/2028). Il rischio è che l'adozione di metriche formali si traduca in una supplenza tecnocratica priva di legittimità democratica.