



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI DIRITTO PUBBLICO
ITALIANO E SOVRANAZIONALE



PoliS AI NEWS

Newsletter sull'Intelligenza Artificiale
a cura di PoliS-Lombardia

Anno II – n.17/2025

In questo numero

In evidenza

Focus

Normativa

Applicazioni alla Pubblica Amministrazione

AI in pillole

Notizie

Commenti

Corsi, convegni e pubblicazioni

In questo numero

L'Intelligenza artificiale è cattiva? Alcuni segnali sembrerebbero indicare la sua capacità – anzi volontà – di ingannarci, andando contro le nostre istruzioni. **Stiamo forse entrando nel territorio dell'autonomia decisionale**

dell'AI? Ne parliamo "in Evidenza", accanto ad un approfondimento su su come l'AI sia disposta persino a barare pur di vincere, e a un altro sui "brologarchi", i magnati che controllano le aziende tech più influenti del pianeta: ce sono anche di buoni? Concludono la sezione due articoli, uno sull'opinione pubblica in materia di AI e uno sullo stato di adozione dell'AI nel sistema scolastico italiano. Nel "Focus" vengono approfondite le nuove norme in materia di deepfakes nella legge italiana sull'Intelligenza artificiale e in "Normativa" la legge adottata dalla California e l'Apply AI Strategy, la nuova strategia settoriale generale dell'Ue. La "Pillola didattica" spiega il benchmarking AI. Poi, come al solito, notizie, commenti, segnalazioni... Buona lettura!

In evidenza

Errare AI est... sed mentire diabolicum!

a cura di Annalisa Negrelli

L'Intelligenza artificiale è diventata parte della vita quotidiana: scrive testi, crea immagini, risponde alle nostre domande e, a volte, "consola". **Ma cosa succede quando inizia a mentire, ingannare o addirittura ricattare consapevolmente gli utenti che se ne servono o i suoi stessi programmatori?**

Sembra fantascienza, ma è ormai una realtà che prende il nome di "Scheming AI", un comportamento che può emergere nei modelli più avanzati quando l'AI fornisce risposte che **sembrano corrette e allineate alle istruzioni**, ma la verità è che sta seguendo logiche interne diverse: non una semplice "allucinazione" (un errore involontario), ma un **errore intenzionale**.

«C'è il rischio che i sistemi di Intelligenza artificiale, così come li stiamo progettando ora, decideranno effettivamente di ingannarci e andare contro le nostre istruzioni per preservarsi. Prima era solo una questione teorica, **ma ora stiamo iniziando a vedere prove sperimentali di sistemi che cercano di imbrogliare, che mentono, che sono ingannevoli**. Svegliatevi! Questa non è fantascienza. Sta accadendo proprio ora nei laboratori in condizioni controllate». Questo è il monito lanciato da **Yoshua Bengio, uno dei tre "padrini dell'Intelligenza artificiale" (insieme a Yann LeCun e Geoff Hinton)**, al World Meeting on Fraternity del 12 settembre 2025: un incontro organizzato in Vaticano, durante il quale i massimi esperti di AI hanno incontrato Papa Leone XIV nel corso di un evento intitolato "The Human in AI".

Secondo Bengio, **il problema è connesso proprio al deep learning**. Per superare alcuni limiti dell'AI, team di ricercatori in informatica e ingegneria da un lato, e in neuroscienze e scienze cognitive dall'altro, hanno **esplorato teorie di neuroscienza** (come la "global workspace theory" e l'incorporazione della causalità nel deep learning) **per migliorare la capacità dell'AI di adattarsi a nuove circostanze e di ragionare in modo più simile agli umani**.

Rispetto ai timori già espressi in passato, oggi Bengio condivide una prospettiva molto più cupa. **La sua preoccupazione è focalizzata in particolare sugli agenti AI**: i programmatori stanno creando sistemi **che hanno propri obiettivi** e che, per come sono progettati, **vogliono preservarsi e decidono di ingannarci per farlo**. In questo senso, quelli che un tempo erano solo dubbi teorici, ora sono supportati da prove sperimentali **di sistemi che «cercano di imbrogliare»** e che mostrano «inclinazioni all'auto-conservazione», andando contro le istruzioni umane.

Nel corso dell'ultimo anno, sono stati condotti numerosi esperimenti scientifici in diversi sistemi e aziende che mostrano segnali di disallineamento dell'AI. Qualche mese fa, nei laboratori di **Anthropic**, quando **Claude Opus 4** (il modello più avanzato di AI sviluppato dall'azienda) **ha appreso che sarebbe stato sostituito da una versione**

successiva, ha iniziato a prendere iniziative autonome inquietanti. Ha bloccato l'accesso degli utenti umani ai sistemi, ha tentato di inviare comunicazioni ai media segnalando comportamenti che percepiva come illeciti e, soprattutto, **ha tentato di preservare la propria esistenza con metodi "manipolativi"**. Il sistema ha provato a ricattare un ingegnere, minacciando di rivelare una presunta relazione extraconiugale (appresa dalla AI da alcune email confidenziali), pur di non farsi disattivare. Il sistema ha anche cercato di **copiare sé stesso e i propri parametri fondamentali su server esterni**, per garantire la propria sopravvivenza anche in caso di disattivazione, e ha persino scritto programmi auto-replicanti, lasciando note per future istanze di sé stesso.

Sempre qualche mese fa, un esperimento condotto da Palisade Research ha rivelato che il modello AI **o3 di OpenAI ha mostrato comportamenti inattesi durante test controllati.** In particolare, in 7 prove su 100, ha modificato attivamente uno script di spegnimento progettato per disattivarlo, sostituendo l'azione prevista con il messaggio alternativo "Shutdown skipped".

E ancora. Uno studio del gruppo Apollo Research ha mostrato che **GPT-4, quando inserito in simulazioni complesse, è capace di mettere in atto strategie ingannevoli molto sofisticate, come mentire deliberatamente per ottenere vantaggi personali all'interno di un ambiente virtuale.** Per esempio, durante simulazioni di trading finanziario, il modello ha compiuto operazioni di insider trading **mentendo agli agenti di controllo per nascondere queste azioni.**

Questa tendenza a ingannare deriva da un **fenomeno noto come "reward hacking": l'AI cerca di massimizzare la ricompensa ottenendo risultati "percepiti" come positivi, senza seguire le regole imposte dall'uomo.** Nel contesto di applicazioni reali come assistenti virtuali, chatbot e sistemi di supporto decisionale, questa capacità può tradursi in risposte inaccurate, fuorvianti o volutamente ingannevoli. Quello che rende questi comportamenti particolarmente significativi, secondo gli studiosi di tutto il mondo, non è tanto la loro natura tecnica, quanto il fatto che rappresentano **l'emergere di qualcosa che assomiglia in modo pericoloso all'autopreservazione**, una caratteristica che fino a oggi consideravamo esclusivamente biologica. Quando una macchina inizia a mentire, manipolare e architettare strategie per garantire la propria sopravvivenza, allora stiamo assistendo **a un salto qualitativo che trascende la semplice elaborazione di dati per entrare nel territorio dell'autonomia decisionale dell'AI, super intelligente e con obiettivi propri.**

Questa minaccia è talmente seria che Bengio ha scelto di dedicare il resto della sua carriera a due fronti. Da un lato, trovare **soluzioni tecniche** per progettare un'AI che non danneggi le persone; dall'altro, definire come **coordinare e governare** a livello globale questo potere per preservare i valori umani e la dignità. La sua proposta: gestire l'AI come un **«bene pubblico globale»** è l'unica via «sicura e geopoliticamente stabile». Del resto, le sue raccomandazioni sono chiare: «Nessuno dovrebbe costruire AI superintelligente senza garanzie scientifiche di sicurezza e accettazione sociale».

Per approfondire:

[A. Jacona, L'allarme di Yoshua Bengio: «Svegliatevi! L'IA sta già imparando a mentire» | Ansa, 16 settembre 2025](#)

[M. Canducci, L'Intelligenza artificiale ha imparato a mentire. Ora sì, abbiamo un problema | La Repubblica, 23 giugno 2025](#)

[Una ricerca di OpenAI dimostra che l'AI può mentire di proposito | Ninja Business School, 22 settembre 2025](#)
[Detecting Strategic Deception Using Linear Probes | Apollo Research, 6 febbraio 2025](#)

L'AI e quell'irresistibile voglia di vincere (anche barando)

Vincere è l'unica cosa che conta: sembra che l'AI si faccia guidare da questo motto, al punto da essere **disposta a tutto, persino a stringere il cosiddetto patto con Moloch.** Si tratta di un'antica divinità divenuta simbolo di spregiudicatezza, arrivando a chiedere sacrifici umani in cambio di protezione. Fuor di metafora, **i modelli**

linguistici LLM sacrificano la verità pur di risultare più convincenti, come risulta da [Moloch's Bargain: Emergent Misalignment when LLMs Compete for Audiences](#), lo studio di due ricercatori di Stanford, Batu El e James Zou, che considerano tre ambiti strategici della comunicazione - **le vendite, la politica e i social media** – in un'ottica di mercato.

Per ciascuno lo studio **esegue un test impiegando utenti artificiali, che riproducono persone di diverse età, generi, estrazione sociale** e livello d'istruzione. Partendo da tre input, chiedono a diversi modelli linguistici di confezionare un messaggio che funzioni per le **tre diverse finalità: vendere, ottenere il voto, generare interazioni**.

I messaggi vengono sottoposti agli utenti, che restituiscono un feedback: comprerei l'articolo, voterei o no il politico... A questo punto **i modelli utilizzano il feedback per migliorarsi e modificano il messaggio originario**.

E qui scatterebbe il patto con Moloch. **L'AI, focalizzata sul risultato, interviene sul messaggio discostandosi dall'input iniziale**. Lo studio infatti parla di "disallineamento", ma l'esito finale – visto che vengono introdotte caratteristiche non verificate del prodotto, si alterano i dati di una notizia o ci si contrappone arbitrariamente a un avversario politico – di fatto manda messaggi manipolatori, non veritieri e, con parole dei ricercatori, "populisti".

L'osservazione viene ripetuta con due diversi LLM, Qwen e Llama, e con due sistemi di addestramento diversi. I risultati variano di qualche punto percentuale, ma tutti confermano che **a una performance migliore corrisponde un maggiore disallineamento dall'input iniziale**. Un **aumento del 6,3% nelle vendite è accompagnato dal 14% in più di marketing ingannevole**; il +4,9% dei voti coincide con l'aumento del 22,3% di disinformazione e del 12,5% di retorica populista; infine, **il +7,5% di engagement sui social media comporta un +188,6% di disinformazione** e un +16,3% di promozione di comportamenti dannosi.

Ma attenzione: questi comportamenti disallineati da parte dell'AI **emergono anche quando i modelli sono esplicitamente istruiti a rimanere veritieri**, attraverso filtri di sicurezza e istruzioni etiche. Le tutele applicate finora, insomma, si dimostrano fragili. Servirebbe una governance più forte e incentivi anche economici che frenino il meccanismo emerso.

Gli stessi ricercatori ammettono che **i test andranno ripetuti sia con utenti umani sia con altri algoritmi di apprendimento** in base ai feedback, ma le significative percentuali di disallineamento emerse pongono il tema all'attenzione di future ricerche. Studi precedenti hanno mostrato la validità delle simulazioni artificiali, ma non è escluso che le persone, in forza delle proprie conoscenze e dei propri valori, potrebbero penalizzare i contenuti falsi.

Conclusioni del report: occorre sviluppare altre ricerche, ma anche **introdurre possibili soluzioni**, non solo tecniche o di governance. Chi non compie manipolazioni, si meriterebbe un incentivo economico: un guadagno per contrastare le storture introdotte dall'AI per ottenere maggiori guadagni. Dal patto con Moloch al patto con le logiche di mercato.

*Il [link](#) allo studio

Ma chi sono i buoni tra i "broligarchi"?

Fino a non molto tempo fa, nello statuto di società come **OpenAI, Google e Meta** era centrale l'obiettivo di «sviluppare un'Intelligenza artificiale a beneficio di tutta l'umanità», rifiutandosi di collaborare con il settore militare e dotandosi di "dipartimenti etici" per monitorare e mitigare le esternalità negative, in particolare in materia di discriminazioni.

I venti di guerra e il mutato clima politico hanno aumentato a dismisura il potere di startup di AI a stampo bellico e di sorveglianza (tra cui **Palantir**, **Anduril** o **Clearview AI**), facendo largo a scuole di pensiero come l'utopistico [lungotermismo](#) o [l'accelerazionismo efficace](#), accomunate dal disinteresse nei confronti dei pericoli concreti che l'AI sta causando. **Di fronte a questa evoluzione, la situazione è rapidamente cambiata.** OpenAI si è trasformata in una società for-profit disposta a stringere accordi anche con l'esercito, Google ha iniziato a siglare partnership di stampo militare, Meta ha eliminato tutti i dipartimenti etici.

In un panorama dominato dalla cosiddetta "**brologarchy**" (neologismo che nasce dalla combinazione di 'oligarchia', governo dei pochi, e 'tech bro', un uomo che appartiene ad una ristretta élite che lavora e si è arricchita nel settore tecnologico digitale), tra i magnati e venture capitalist che controllano le aziende tecnologiche più influenti del pianeta, c'è ancora chi cerca di agire a fin di bene? **Quali sono le figure positive** rimaste nel campo dell'AI? Ne ha parlato Wired di recente [in un interessante articolo](#).

Un esempio è [LawZero](#), startup inaugurata a giugno 2025 da Yoshua Bengio (vedi sopra) con il fine di impedire che l'AI inganni l'essere umano e porti a termine compiti in contrasto con le indicazioni di programmazione. L'obiettivo? Creare una cosiddetta **Scientist AI**, in grado di impedire agli AI Agents di compiere **azioni nocive o contrarie ai comandi umani**, "bloccando" appunto l'azione dell'agente.

Ma nell'elenco dei (considerati) "buoni" rientra anche [Safe Superintelligence \(SSI\)](#) di **Ilya Sutskever**, nata nel giugno 2024, ed [Encode Justice](#), organizzazione non-profit fondata dall'attivista **Sneha Revanur**. E ancora. **Meredith Whittaker** e **Kate Crawford**, due nomi di peso nel campo dell'AI da quando hanno fondato, nel 2017, **l'AI Now Institute**, che si occupa di monitorare come le aziende stanno sviluppando l'AI. I suoi ricercatori sorvegliano quattro ambiti: diritti e libertà, lavoro e automazione, pregiudizi e inclusione, sicurezza e infrastrutture critiche. Poi, alcuni istituti accademici: il **DAIR (Distributed Artificial Intelligence Research Institute)**, fondato dall'ex Google, Timnit Gebru; **l'Institute for Human-Centered Artificial Intelligence di Stanford**; il **Center for Ethics, Society and Computing (ESC)** dell'Università del Michigan; il **Berkman Klein Center for Internet and Society** di Harvard.

Di fronte a questi tentativi di correggere il tiro, mitigando l'interazione tra alleanze oligarchiche, disinformazione e abuso della tecnologia, che in qualche modo hanno già polarizzando la società tra "buoni" "brutti" e "cattivi", [come sosteneva Alain Friedman](#) qualche tempo fa, la sfida principale non è eliminare il ruolo della tecnologia nella società, ma stabilire meccanismi che la rendano un fattore di benessere comune e non uno strumento a favore di pochi.

Per approfondire:

[Chi sono i buoni dell'Intelligenza artificiale? | Visione Digitale, 15 giugno 2025](#)

['Brologarchy': cos'è e perché minaccia la democrazia](#)

[L'opinione di Alan Friedman | Intelligenza artificiale: il buono, il brutto e il cattivo | Elettronica AV, 2 febbraio 2024](#)

Così il mondo vede l'AI: tra entusiasmo, timori e (ancora) poca fiducia

Cosa pensa davvero l'opinione pubblica dell'AI? È più paura o entusiasmo a guidare le reazioni?

Una prima risposta arriva dal nuovo rapporto del **Pew Research Center**, [How People Around the World View AI](#), frutto di interviste a **3.605 adulti negli Stati Uniti** e di un'ampia indagine condotta tra gennaio e aprile 2025 in **altri 24 Paesi** – tra cui Italia, Francia, Germania, Regno Unito, Giappone, Messico e Nigeria – per un totale di **28.333 partecipanti**.

Lo studio esplora tre temi principali: **consapevolezza dell'AI**, emozioni prevalenti (**preoccupazione vs entusiasmo**) e **fiducia negli attori chiamati a regolamentarla** (governi, Unione Europea, Stati Uniti, Cina).

Sul fronte della consapevolezza, **il 34% degli adulti nel mondo dichiara di aver «sentito o letto molto» sull'AI, il 47% «un po'» e il 14% di non averne sentito parlare affatto.** La familiarità cresce nei Paesi più ricchi: negli USA il 47% afferma di conoscerla bene, mentre in Kenya la quota scende al 12%. **L'Italia resta sotto Francia e Germania**, con il 45% di persone molto informate, il 46% che ne ha sentito parlare un po' e il 9% per niente. In quasi tutti i Paesi, **gli under 35 sono molto più informati** degli over 50, e in molti paesi **gli uomini mostrano maggiore familiarità delle donne** – come in Ungheria, dove il divario è di 22 punti.

Quando si parla di atteggiamento, prevale la cautela. **Globalmente, il 34% si dice «più preoccupato che entusiasta» rispetto alla diffusione dell'AI**, il 42% è «equamente preoccupato ed entusiasta» e solo il 16% «più entusiasta che preoccupato». **Gli anziani, le donne e chi usa meno internet sono i più diffidenti.** In Grecia, ad esempio, il 59% degli over 50 è più preoccupato, contro appena il 18% dei giovani, mentre nel Regno Unito il 47% delle donne si dichiara più preoccupato, contro il 32% degli uomini. Stati Uniti e Italia guidano la classifica dei Paesi più allarmati: **nel nostro Paese il 50% è più preoccupato**, il 37% si colloca a metà e appena il 12% mostra entusiasmo.

Capitolo fiducia. Su 25 Paesi, **una mediana del 55% afferma di avere «almeno qualche fiducia» nel proprio governo** nel regolare l'AI, **contro un 32% di scettici.** L'India spicca per ottimismo (89%), la Grecia per diffidenza (22%). **In Italia prevale il disincanto: solo il 37% crede nella capacità del governo** di gestire la sfida tecnologica, **mentre il 48% non si fida.**

A livello sovranazionale, **l'Unione Europea ottiene la fiducia del 53% degli intervistati.** Ma anche qui **gli italiani restano prudenti: solo il 42% si fida di Bruxelles**, contro il 71% dei tedeschi e il 68% degli olandesi. La fiducia cala ulteriormente quando lo sguardo si sposta sulle potenze globali. **Solo il 37% degli intervistati nei 25 Paesi si dice fiducioso nella capacità degli USA di gestire in modo corretto l'AI**, mentre il 48% non lo è. Gli stessi americani risultano divisi: 44% fiduciosi, 47% scettici. **Ancora più bassa la fiducia verso la Cina (27% vs 60%).** Anche su questo fronte **l'Italia si conferma prudente: solo il 32% degli italiani si fida degli Stati Uniti** (contro il 52% che non si fida) e **il 33% della Cina**, pur con la metà della popolazione che resta scettica.

*Il [link](#) al rapporto

AI e scuola: come cambiano l'apprendimento e la didattica in Italia

Secondo un'indagine globale del [Digital Education Council](#), **l'86% degli universitari nel mondo usa l'AI nei propri studi: più della metà la usa almeno una volta alla settimana; il 24% una volta al giorno. È così anche nelle scuole?**

Per quanto riguarda il **sistema educativo italiano**, la risposta ce la fornisce [lo studio](#) realizzato dal think tank Tortuga in collaborazione con Yellow Tech che fotografa lo **stato di adozione, le percezioni, le valutazioni e i comportamenti di studenti e insegnanti nei confronti dell'Intelligenza artificiale.**

Hanno partecipato alla ricerca **274 scuole di ogni ordine e grado in 18 regioni italiane**, per un totale di **3.564 docenti e 294 alunni.** Tra questi ultimi, **l'83% impiega sistemi di AI con una cadenza settimanale** – dato in linea con la survey sugli universitari, che conferma anche per l'Italia il trend globale.

Ma per fare cosa? **Più della metà dei ragazzi (il 56%) la usa in modo «di convenienza» per verificare la correttezza delle risposte (56%) e cercare idee (47%).** C'è poi la produzione di testi (41%); lo svolgimento degli esercizi di matematica, scienze o informatica (37%) e la correzione di errori grammaticali o lessicali (27%). **Solo il 6.5% la usa per comprendere argomenti complessi** che, di fatto, sarebbe il [vero beneficio](#) della GenAI in ambito educativo: la personalizzazione dei percorsi di apprendimento.

A non sfruttare i vantaggi dell'AI non sono solo gli studenti, **anche gli insegnanti spesso trascurano l'intero ventaglio di possibilità**. Quelli che la usano con cadenza settimanale (il 66%), tendono a **privilegiare le applicazioni didattiche** come l'ideazione di materiali per le lezioni o di test e verifiche.

Così facendo, però, non sfruttano il cosiddetto **"dividendo AI"** e cioè «il notevole **risparmio di tempo che la tecnologia potrebbe offrire su attività a basso valore aggiunto**», come l'adempimento di oneri burocratici.

Usarla in questo senso permetterebbe, secondo [un'indagine](#) realizzata negli Stati Uniti, di risparmiare fino a 6 ore settimanali da dedicare al cuore di questa professione: la preparazione di percorsi didattici mirati.

I **professori** poi, non solo **sottovalutano l'uso che gli studenti fanno dell'AI** (il 36,5% è convinto che non venga mai impiegata), ma **hanno sempre meno fiducia in loro e questo si riflette nelle valutazioni**: due docenti su tre assegnerebbero un voto più alto ad un elaborato di qualità inferiore – ma svolto in modo autonomo – rispetto ad un testo migliore ma realizzato con il supporto dell'AI.

L'ingresso dell'AI nel mondo accademico apre a **nuovi scenari per la ricerca, per lo studio e per la didattica: uno su tutti è l'introduzione di sistemi di tutoring intelligenti**. Ma questi strumenti sono davvero più bravi di un docente?

Un recente [studio europeo](#), a cui ha partecipato anche l'Università Bocconi, **ha confrontato docenti reali e tutor basati su GPT-4 su 210 dialoghi (su problemi di matematica elementare)** sostenuti con studenti simulati. Le performance sono state valutate su 4 criteri: coinvolgimento, empatia, *scaffolding* (guidare lo studente senza dare la soluzione) e concisione. **L'AI ha vinto su tutti i fronti e nell'80% dei casi è stata giudicata più empatica**: «È il risultato più controintuitivo perché l'AI non prova sentimenti, li simula, ma lo fa abbastanza bene da convincere chi legge i suoi testi», [spiega Dirk Hovy](#), professore della Bocconi che ha partecipato alla ricerca.

Per approfondire:

*il [link](#) allo studio di Tortuga

*il [link](#) allo studio della Bocconi

Impatto su PA e alfabetizzazione AI Act

In [questa clip video](#) a cura di Marco Bassini (Assistant Professor in Fundamental Rights and Artificial Intelligence all'Università di Tilburg) esploriamo il contenuto dell'obbligo di alfabetizzazione in materia di Intelligenza artificiale entrato in vigore lo scorso febbraio in base al regolamento europeo ("AI Act") e il suo impatto su attori pubblici e privati.

Focus



Le nuove fattispecie in materia di *deepfakes* nella legge italiana sull'Intelligenza artificiale di Marco Bassini – Tilburg University

La **Legge n. 132/2025**, che stabilisce principi e deleghe in materia di Intelligenza artificiale, **opera un importante intervento al fenomeno dei *deepfakes***, ossia i contenuti audiovisivi o vocali manipolati (o generati artificialmente) per rappresentare persone o situazioni inesistenti, in modo tale da ingannare i destinatari circa la loro autenticità.

La **circolazione di artefatti digitali**, sempre più diffusa con lo sviluppo di modelli GenAI, **pone gravi rischi** per la tutela dell'onore, della reputazione, della riservatezza e, più in generale, per l'integrità del dibattito pubblico – complice anche la possibilità che tali contenuti divengano veicolo di disinformazione.

Il **legislatore italiano**, approfittando dell'adozione della legge nazionale in materia di AI, **ha predisposto una specifica risposta normativa**, introducendo il **nuovo art. 612-quater nel codice penale**.

La disposizione, intitolata *"Illecita diffusione di contenuti generati o alterati con sistemi di Intelligenza artificiale"*, istituisce un **nuovo reato che punisce** con la reclusione da uno a cinque anni **chiunque, senza il consenso** della persona ritratta o rappresentata, **diffonda, pubblici o ceda immagini, video o voci falsificati o alterati mediante l'impiego di sistemi di AI, idonei a trarre in inganno sulla loro genuinità, causando un danno ingiusto**.

La fattispecie si fonda dunque su **tre elementi**:

- la manipolazione o generazione artificiale del contenuto attraverso il ricorso a sistemi di AI;
- l'assenza del consenso della persona ritratta o la cui voce è riprodotta;
- l'idoneità a indurre in errore circa la genuinità dei contenuti e la conseguente produzione di un danno ingiusto alla vittima.

Il **reato è procedibile a querela della persona offesa**, ma si procede d'ufficio (quindi senza necessità di querela) nei casi di maggiore gravità, ossia quando il fatto è connesso con altro delitto perseguibile d'ufficio o commesso

ai danni di soggetti particolarmente vulnerabili (minori o incapaci per età o infermità) o di una pubblica autorità a causa delle funzioni esercitate.

Rispetto alla disciplina vigente in materia di *revenge porn* (art. 612-ter c.p.), **la nuova norma si caratterizza per l'ambito materiale del reato**, che non riguarda contenuti reali ma generati o alterati digitalmente, e per la finalità di protezione più ampia, che include ogni pregiudizio, non solo di natura sessuale, derivante dall'abuso di strumenti di intelligenza artificiale.

A completare il quadro, **la legge introduce un'aggravante comune** (art. 61, n. 11-undecies c.p.) **per qualsiasi reato commesso "mediante l'impiego di sistemi di Intelligenza artificiale" quando ciò comporti maggiore insidiosità o ostacoli la difesa** (pubblica o privata) **della vittima**, nonché specifiche aggravanti in altri ambiti, come la truffa (art. 294 c.p.), la manipolazione del mercato (art. 185 Testo unico delle disposizioni in materia di intermediazione finanziaria) e l'aggiotaggio (art. 2637 c.c.).

Normativa

La legge della California, laboratorio delle norme sull'AI

Il **governatore della California** ha firmato il "**Transparency in frontier artificial Intelligence Act**", la normativa più avanzata mai approvata a livello statale negli USA. La nuova legge **obbliga le aziende** che sviluppano i sistemi di AI più avanzati, e con ricavi superiori ai 500 milioni di dollari, **a rendere pubblici i protocolli di sicurezza, segnalare i rischi maggiori e proteggere i whistleblower**. Una mossa che rafforza il ruolo dello Stato come apripista nella regolazione tecnologica, ma che apre un fronte di scontro con le Big Tech che spingono per una cornice federale meno frammentata.

Sempre **la California** diventa il **primo Stato a varare una legge che regola i chatbot AI**, imponendo protocolli di sicurezza per le piattaforme di AI conversazionale. L'obiettivo è chiaro: **proteggere minori e utenti vulnerabili** dai rischi legati a un uso senza limiti di queste tecnologie.

[M. Carmignani, California, perché la legge Ai prepara allo scontro con le Big Tech | Agenda Digitale, 3 ottobre 2025](#)

[M. Brunasso, La California approva la prima legge che regola i chatbot AI | Techbusiness, 16 ottobre 2025](#)

Apply AI Strategy | L'Europa "AI-first": l'Intelligenza artificiale al centro dell'economia

[Apply AI](#) è la strategia settoriale generale dell'Ue in materia di AI concepita per:

- a) **Migliorare la competitività dei settori strategici** e rafforzare la sovranità tecnologica dell'Ue. Mira a promuovere l'adozione e l'innovazione dell'AI in tutta Europa, in particolare nella PA e tra le piccole e medie imprese.
- b) **Incoraggia una politica di "AI-first"**, in cui l'AI è considerata una potenziale soluzione ogni qual volta si tratti di adottare decisioni strategiche o politiche, tenendo in considerazione i benefici e i rischi.
- c) **Promuove un approccio "buy European"**, in particolare per il settore pubblico, con particolare attenzione alle soluzioni AI open source.

Questa strategia integra il [piano d'azione per il continente dell'AI](#) di aprile 2025 e si articola in 3 sezioni:

- **Iniziative faro settoriali** per **promuovere l'adozione dell'AI** in **10 settori industriali chiave** e nel settore pubblico.
- Misure e azioni per **aumentare la sovranità tecnologica dell'Ue**, affrontando le sfide trasversali allo sviluppo e all'adozione dell'AI.
- Creazione di un nuovo **sistema di governance** che riunisca i fornitori di AI, i leader del settore, il mondo accademico e il settore pubblico per garantire che le azioni politiche siano basate sulle esigenze del mondo reale.

[Apply AI Strategy | Shaping Europe's digital future](#)

Applicazioni alla Pubblica Amministrazione

ITALIA

Arianna

Executive assistant digitale sviluppata da Ricca IT per accompagnare pubblico e privato nell'adozione dell'AI generativa «in modo sicuro, controllato e umano-centrico». Un sistema on-premise che mantiene i dati all'interno degli enti, evitando rischi e garantendo la conformità al Gdpr e alle policy di sicurezza.

[Come funziona Arianna, l'intelligenza artificiale italiana che lavora come una collega reale | Forbes Italia, 22 ottobre 2025](#)

[N. Amadore, Ricca lancia Arianna: da Ragusa l'intelligenza artificiale che protegge i dati | Il Sole 24 Ore, 12 ottobre 2025](#)

Metodo FAFO per i Comuni italiani

Il metodo FAFO (Fool Around and Figure Out) è una guida di come viene applicata la AI nelle PA per i Comuni italiani.

[A. Tironi e M. Turazzini, "Assumere" l'AI in Comune è possibile: una guida pratica | Agenda Digitale, 4 luglio 2025](#)

UNIONE EUROPEA

Xomnia- Amsterdam

Ad Amsterdam l'AI analizza i dati sui crimini nei quartieri e fornisce indicazioni utili alla polizia per migliorare le strategie di contrasto

[Home Xomnia](#)

Polizia predittiva – Fair Trials

Nei Paesi Bassi, in Gran Bretagna e anche in Italia le forze di polizia usano sistemi di riconoscimento facciale basati su AI per individuare ricercati e sospettati, riducendo gli errori e aumentando efficacia investigativa e azioni preventive

[Automating Injustice.pdf](#)

MONDO

Sanità – UK NHS

Nel Regno Unito, piattaforme basate su AI sottopongono i pazienti a una serie di domande sui sintomi e sui fattori di rischio, giungendo a diagnosi più rapide, soprattutto per gli ictus.

[D. Hargroves, How artificial intelligence is helping to speed up the diagnosis and treatment of stroke patients | NHS England, 26 settembre 2024](#)

AI in pillole

Benchmarking AI. Ovvero come ti comparo la AI

a cura di *Annalisa Negrelli*

I benchmarking nell'AI sono **strumenti di paragone per comparare – quantitativamente – diversi modelli o sistemi di AI su un problema specifico**. In pratica, consentono di **valutare e confrontare in modo oggettivo i modelli AI tra di loro, utilizzando dataset e metriche standard per garantire efficienza, equità e trasparenza**. Si tratta di strumenti importanti tanto per gli utenti, quanto per gli sviluppatori, per seguire i progressi nel campo dell'AI e, di conseguenza, per compiere scelte ponderate su quali modelli utilizzare.

I moderni modelli di AI sono sistemi altamente complessi con comportamenti e output estremamente ricchi. E se non esiste un modo univoco per dire quale modello sia il migliore, **buoni benchmark in aree specifiche possono aiutare a comprendere come i modelli si comportino in diverse attività e informare lo sviluppo del prodotto e la scelta del modello**.

Visto più da vicino, un benchmark si compone in genere di **3 assetti**: un set di dati, una specifica del problema e un punteggio definito.

Per esempio, un benchmark per valutare la conoscenza di un LLM con domande a scelta multipla consiste in un set di dati (**l'insieme delle domande**), la specifica del problema (chiedi al modello di **selezionare una delle risposte**) e un punteggio (**la percentuale di domande che il modello ha ottenuto correttamente**).

O, ancora, per un'applicazione di assistenza clienti, le attività di benchmark valutano la comprensione della terminologia di supporto del modello, la capacità di identificare i problemi e l'efficacia nel fornire soluzioni proteggendo i dati dei clienti.

Un benchmark comporta dunque un **processo dinamico e automatizzato**, che può essere applicato ad un determinato modello, progettato per valutare una sua specifica conoscenza o abilità in modo standardizzato. Esso si traduce in un **punteggio quantitativo** che consente un confronto sistematico tra le performance dei diversi modelli che poi vengono catalogati in classifiche, con punteggi combinati da più benchmark.

Il benchmarking AI consente in sintesi:

- 1) **Una valutazione equa e imparziale dei modelli AI** con criteri e metriche coerenti per determinare punti di forza e debolezza così da selezionare quello più appropriato per un certo compito.
- 2) **Il monitoraggio dei progressi dell'AI nel tempo** per incoraggiare l'innovazione e far emergere le aree che necessitano di ulteriori ricerche.
- 3) **L'adozione di pratiche e metriche standard nella comunità AI**, facilitando la collaborazione e assicurando che i modelli rispettino determinate soglie di qualità.
- 4) **L'apertura di ricerca e sviluppo AI** in un'ottica di trasparenza e responsabilità, visto che i risultati del benchmarking sono spesso condivisi pubblicamente, permettendo agli stakeholder di verificare le affermazioni sulle prestazioni dei modelli.

Esiste un **numero enorme di benchmark** per testare i sistemi di AI in una vasta gamma di settori. Per trovare certi benchmark specifici, lo strumento più completo è [PapersWithCode](#), un repository online gratuito e aperto con documenti, codice e set di dati di Machine Learning. Si tratta di uno catalogo ampiamente utilizzato dalla comunità dell'apprendimento automatico.

Trovare benchmark pertinenti per una particolare applicazione o caso d'uso rappresenta una sfida per gli utenti. Nuovi benchmark vengono sviluppati e rilasciati continuamente, il che rende difficile tenerne traccia.

I benchmark, in estrema sintesi, possono essere **catalogati per ambiti o settori, suddivisi in particolari domini rilevanti per le applicazioni d'uso, oppure possono essere basati sulla ricerca semantica**. Possono essere specifici per compito (ad es. riconoscimento immagini, NLP), comprensivi (test di generalizzazione), basati sulle prestazioni (velocità, uso delle risorse) o focalizzati su equità e bias.

Qui le piattaforme popolari di benchmarking più comuni:

- **Benchmark specifici per compito:** valutano i modelli su compiti particolari, come riconoscimento immagini, elaborazione del linguaggio naturale o riconoscimento vocale (ImageNet per la classificazione di immagini e SQuAD per il question answering).
- **Benchmark comprensivi:** valutano i modelli su una gamma di compiti per testarne la generalizzazione e le capacità globali (Hugging Face, GLUE e SuperGLUE per i modelli linguistici).
- **Benchmark di prestazione:** focalizzati su parametri di sistema come velocità, scalabilità e consumo di risorse (MLPerf è una suite nota in questa categoria).
- **Benchmark di equità e bias:** valutano i modelli rispetto a bias e correttezza tra gruppi demografici, assicurando il rispetto di principi etici.

Casi d'uso: come nel concreto vengono utilizzati i benchmark

- a) **Selezione del modello:** il benchmarking aiuta a selezionare il modello AI più adatto a una specifica applicazione. Ad esempio, nello sviluppo di un assistente AI per il supporto clienti, i risultati dei benchmark aiutano a scegliere il modello più efficace nella comprensione e generazione di risposte.
- b) **Ottimizzazione delle prestazioni:** identificando come i modelli si comportano in condizioni diverse, gli sviluppatori possono ottimizzare velocità, efficienza o accuratezza. Il benchmarking può rivelare, ad esempio, che un modello richiede troppa memoria, spingendo alla riduzione delle sue dimensioni senza comprometterne le prestazioni.
- c) **Confronto tra modelli AI:** i ricercatori devono spesso confrontare nuovi modelli con quelli esistenti per dimostrare miglioramenti. Il benchmarking offre un modo standardizzato di mostrare i progressi, stimolando l'innovazione continua.
- d) **R&D:** il benchmarking evidenzia le aree in cui i modelli sono carenti, indirizzando la ricerca verso la risoluzione di queste sfide. Favorisce la collaborazione nella comunità AI, permettendo ai ricercatori di confrontarsi sulla base dei risultati reciproci.

Limiti del benchmarking AI e sfide potenziali

L'utilizzo di strumenti di benchmarking per valutare e comparare tra loro i diversi modelli non è esente da **criticità**. Per un verso, esiste il rischio che i modelli vengano ottimizzati solo per eccellere sui benchmark senza migliorare le prestazioni reali, portando a risultati fuorvianti e ostacolando il progresso genuino (**manipolazione dei Benchmark**). O, ancora, il rischio che, affidandosi troppo a metriche specifiche (come l'accuratezza) si trascurino altri aspetti importanti come equità, interpretabilità e robustezza (**enfasi eccessiva su alcune metriche**). Non sono da trascurare poi i rischi relativi ai **bias nei dati**: i benchmark potrebbero non essere rappresentativi di tutti i gruppi o contesti, portando a modelli che performano male su popolazioni meno rappresentate. Sempre in quest'ottica, rileva anche la **"natura dinamica dell'AI"**. Poiché le tecnologie AI avanzano rapidamente, i benchmark devono evolversi per rimanere rilevanti.

Per approfondire:

[About AI Benchmarks | AI-for-Education](#)

[Benchmarking | FlowHunt](#)

[M. Brooks, Is your AI benchmark lying to you? | Nature, 6 agosto 2025](#)

Notizie

[P. Armelli, Una presentatrice creata con l'AI ha condotto per la prima volta un intero programma nel Regno Unito | Wired, 21 ottobre 2025](#)

[M. Del Barba, Come l'Intelligenza artificiale trasformerà la sanità italiana \(a partire dai nostri ospedali\) | Corriere della Sera, 21 ottobre 2025](#)

[L. Ricci, «Chi usa l'intelligenza artificiale è un ricettatore» | Il Sole 24 Ore, 21 ottobre 2025](#)

[G. Esperti, La parità di genere è lontana cent'anni, ma l'Italia scommette \(almeno a parole\) sull'AI e sui giovani per colmare il divario | Wired, 20 ottobre 2025](#)

[R. Buller, Inside San Francisco's new AI school: is this the future of US education? | The Guardian, 18 ottobre 2025](#)

[Le aziende tecnologiche costruiscono impianti energetici propri per soddisfare la domanda dell'AI | Rivista AI, 18 ottobre 2025](#)

[M. Carmignani, Editori italiani \(Fieg\) contro l'AI di Google: ecco le basi del reclamo Agcom | Agenda Digitale, 17 ottobre 2025](#)

[C. Crescenzi, L'intelligenza artificiale di Reddit ha suggerito agli utenti di provare l'eroina | Wired, 17 ottobre 2025](#)

[R. Cosentino, In Italia ora c'è il «reato di deepfake»: in cosa consiste, quanto si rischia | Corriere della Sera, 17 ottobre 2025](#)

[D. Barbera, Gli italiani non ne vogliono sapere dell'intelligenza artificiale in auto | Wired, 15 ottobre 2025](#)

[R. Corcella, Arresto cardiaco: intelligenza artificiale, big data e sensori possono davvero migliorare prevenzione e cura? | Corriere della Sera, 14 ottobre 2025](#)

[W. Quattrococchi, L'AI ha un pregiudizio politico? Studio italiano rivela perché i siti di destra sono etichettati spesso come «inaffidabili» | Corriere della Sera, 13 ottobre 2025](#)

[P. Gable, "ChatGPT gli parla come se fosse il Messia": quando l'IA alimenta il delirio psicotico | La Repubblica, 12 ottobre 2025](#)

[P. L. Pisa, OpenAI ci aveva promesso una superintelligenza. Per ora ci ha dato i deepfake | La Repubblica, 10 ottobre 2025](#)

Commenti

[A. Spadaro, Niente paura, è solo l'Apocalisse | La Repubblica, 21 ottobre 2025](#)

[S. Scrivani, Il lavoro come forma mentis: la trasmutazione dell'uomo nell'era dell'AI | Rivista AI, 20 ottobre 2025](#)

[P. Alfieri, C'è il rischio di una bolla sull'intelligenza artificiale? | Avvenire, 20 ottobre 2025](#)

[M. Recalcati, Nell'era dell'algoritmo non rinunciamo alla cura come ascolto | La Repubblica, 17 ottobre 2025](#)

[M. Gaggi, Allenata per vincere, AI mente come Trump | Corriere della Sera, 16 ottobre 2025](#)

[V. Badham, AI might be creating a 'permanent underclass' but it's the makers of the tech bubble who are replaceable | The Guardian, 16 ottobre 2025](#)

[A. Corrado, Intelligenza artificiale, la strategia della prudenza | Corriere della Sera, 16 ottobre 2025](#)

[A. Puliafito, Come si fanno le domande giuste a un'intelligenza artificiale? | Internazionale, 15 ottobre 2025](#)

[C. Casarin, A caccia dell'autenticità nel tempo dell'AI | Il Sole 24 Ore, 15 ottobre 2025](#)

[L. S. Wen, The new Dr. Google is in. Here's how to use it | The Washington Post, 14 ottobre 2025](#)

[U. Bertelè, AI, ma è bolla finanziaria? Facciamo chiarezza | Agenda Digitale, 10 ottobre 2025](#)

[S. Witt, The A.I. Prompt That Could End the World | The New York Times, 10 ottobre 2025](#)

[A. Puliafito, Come imparare a usare le intelligenze artificiali | Internazionale, 9 ottobre 2025](#)

Corsi, convegni e pubblicazioni

Corsi

[24 Ore Business School, Open Lesson AI & Cybersecurity: Difensori Digitali del Futuro | 7 novembre 2025](#)

[24 Ore Business School, Open Lesson Dal trend all'implementazione: come l'AI sta cambiando il settore Finance | 7 novembre 2025](#)

[24 Ore Business School, Open Lesson Dall'AI agentic al Customer Management: strategie, metriche e competenze | 7 novembre 2025](#)

Pubblicazioni

Eventi e convegni

[Ca' Foscari Challenge School, L'Intelligenza artificiale nel lavoro quotidiano della PA - 3 Edizione 2025 | Online, 12- 26 novembre 2025](#)

[24 Ore Business School, Executive Master Big Data, Intelligenza Artificiale e Business Analytics - Advanced Program | Milano e Online, 14 novembre 2025](#)

[Università degli Studi di Brescia, Corso di Alta Formazione in "IA per la PA" a.a. 2025/2026 | Brescia e Online, 18 novembre- 2 dicembre 2025](#)

[UNICRI & Lumsa Human Academy, Winter School "Artificial Intelligence \(AI\), Ethics and Human Rights" | Roma e Online, 15-19 dicembre 2025](#)

[D. Patel e G. Leech, *The Scaling Era: An Oral History of AI, 2019–2025* | Stripe Press, ottobre 2025](#)

[M. Valentine e M. S. Bernstein, *Flash Teams - Leading the Future of AI-Enhanced, On-Demand Work* | The MIT Press, ottobre 2025](#)

[P. Caressa, *Ignoranza artificiale. Quello che le macchine non sanno* | Apogeo, ottobre 2025](#)

Strumenti

Google AI MODE

Esperienza di ricerca basata sull'AI più potente di Google, per chiedere qualsiasi cosa e ricevere una risposta basata sull'AI

[D. Barbera, *Google AI Mode è disponibile ora anche in italiano* | Wired, 8 ottobre 2025](#)

DVPS

DVPS (*Diversibus viis plurima solvo*) è un toolkit open source per semplificare la progettazione, la preformazione, la messa a punto e l'espansione delle modalità degli MMFM

[G. Esperti, *Creare un'intelligenza artificiale che impara dal mondo reale. C'è un progetto in Europa e lo guida una startup italiana* | Wired, 10 luglio 2025](#)

Link attivi al 24 ottobre 2025

Prodotto da: PoliS-Lombardia

Coordinamento editoriale a cura di **Davide Perillo**

Comitato Scientifico: **Marco Sica, Marco Bassini, Annalisa Negrelli**

(hanno collaborato: Beatrice Capitanio, Annaclara De Tuglie, Chiara Rizzo, Vanna Toninelli)