AI, contrordine della Silicon Valley. Le Big Tech (in conflitto) scaricano Trump? Il ruolo dell'Europa (2^ parte)

3 – *Liquidità* di una tecnologia *general purpose* e regole per *agenti* che agiscono ma non pensano nel vuoto di coscienza o percezione della realtà e gli impatti sulla democrazia.

Ripartendo dal conflitto sulla (presunta) "*coscienza morale*" dell'AI (addestrata a simularla senza consapevolezza delle conseguenze), quali regole propone Amodei di Antrophic ? La *prima*, riguarda il libero accesso a valutatori esterni permanenti, come quelli in parte già noti per gli impianti nucleari e con diritto di monitoraggio e di pubblicazione dei risultati. La *seconda*, attiene al necessario coordinamento tra le aziende del settore nelle democrazie per definire deroghe antitrust e standard comuni su quanto, e come sia possibile, *accelerare o frenare* e perché. La *terza* riguarda l'intero pianeta, e in particolare i rapporti con la Cina, con l'obiettivo di stringere accordi di sicurezza per evitare l'uso di tecnologie dagli effetti catastrofici, stabilendo limiti invalicabili. Si tratterebbe di un trattato per limitare le nuove minacce strategiche, simile agli storici *Accordi SALT* usati per bloccare la proliferazione delle armi

nucleari militari. Regole desiderabili la cui fattibilità è tutta da verificare. Inoltre, c'è da domandarsi quanto queste regole rappresentino una "barriera all'entrata" per nuovi potenziali entranti e per le scelte di direzione dell'innovazione, essendo espressione di una "ibridazione robusta" tra pubblico e privato senza una precisa distinzione tra interesse pubblico e interesse proprietario. Che rappresenta il problema essenziale per la democrazia e le sue fondamenta costitutive, ossia queste potranno "rallentare" alla sola condizione geostrategica che ciò non avvenga a scapito del vantaggio sulla Cina e cioè finché il primato tecnologico è mantenuto. Ma siamo certi che tale primato risieda solo nella velocità e potenza di calcolo, e non anche nella diffusione e accessibilità (quasi) "aperta" di queste tecnologie seppure con i modelli *open weights*?

Dobbiamo allora decidere chi potrà scrivere le regole che pongano argini reali ed effettivi a queste deviazioni potenziali peraltro esplose proprio nel 2026, che la stessa Anthropic ha suddiviso in *sette aree* (dalle armi biologiche alle interferenze politiche e alle campagne di influenza, alla sorveglianza di massa, ai furti di identità, fino agli attacchi russi ai servizi segreti dell'Ucraina, ecc.) e non potrà che essere una *authority pubblica indipendente e globale*, perché non avrebbe alcun senso a livello nazionale, e forse nemmeno a livello continentale. La soluzione di Amodei dovrebbe essere una entità terza che sovrintende al funzionamento della macchina nei laboratori – e in questo caso di Antrophic-Claude – e che ne renda trasparenti le funzionalità e direzioni: né regolatore, né mercato, né controllore, né controllato. Ma può funzionare questa sorta di "*controllo ibrido*" *in-house* di ogni big tech i cui risultati dovrebbero poi essere pubblicati e resi accessibili e con quali limiti?

Qui si gioca il futuro della democrazia oppure rischiamo la rottura delle sue fondamenta costitutive, e una pericolosa

deriva verso l'autocrazia, se non verranno raggiunti accordi globali certo esplicitando la natura *non antropomorfica ma umana* dell'AI, capace di agire secondo un calcolo ma non consapevole di alcuna conseguenza. Il 24 settembre è previsto l'incontro tra Donald Trump e Xi Jinping e lo vedremo presto! Forse è proprio in vista di questo appuntamento – e per compensare la nota inaffidabilità del capo della Casa Bianca – che Dario Amodei di Anthropic ha voluto proporre una sorta di "agenda negoziale". Resta da capire se questo contributo si rivelerà utile, sufficiente o appropriato.

4 – L'urgenza di un Fondo Sovrano Europeo e di *ripoliticizzazione del diritto sovranazionale* nella governance dell'AI e per l'auto-determinazione dei popoli. Per quale "sovranità digitale"?

Nel frattempo, l'Europa dovrebbe accelerare e mettere in campo tutti i suoi principali players per non perdere tutti i treni dell'AI, creando un *Fondo Sovrano Europeo* come prima gamba del tavolo che "sostituisca" i capitali americani ai quali per ora dobbiamo sempre ricorrere per finanziare le buone idee che abbiamo anche in Europa, ma che qui non trovano i capitali necessari, come è già stato spiegato a più riprese da Draghi e Letta già quasi 2 anni fa. Finanziando in questo modo le start up innovative europee, mantenendone la proprietà europea e distribuendo tra i cittadini europei il *capitale di rischio* necessario, che se fosse 100€ per 450 mil.ni di europei, avremmo un fondo da 45 miliardi di euro che in 3-5 anni salire a 200 miliardi di euro e triplicando con l'apporto di capitali privati fino a 6-800 miliardi di euro. Poco ma non pochissimo per cominciare, ma utile e vantaggioso se sarà realizzato presto.

Una *seconda gamba* del tavolo è connessa alla *scienza* e servirebbe a superare il cortocircuito tra produttore (privato) e regolatore (pubblico) non potendo affidare la regolazione alla "buona volontà" del primo, con l'apporto di un quadro normativo che si allarghi all'intero perimetro della conoscenza scientifica e tecnologica di base utile a valutare la sicurezza di un farmaco o di un derivato finanziario come per ogni modello di AI. Valutazione che non possiamo lasciare ai soli soggetti privati cointeressati alla sicurezza di quei "manufatti cognitivi" e che incassano dalla loro diffusione direttamente correlata ai diversi stati di quella sicurezza.

Una *terza gamba* riguarda poi la politica dotandola di capacità di valutare (connessa alla seconda) la *dimensione pubblica di un bene* e valutare gli eventuali vantaggi e danni prodotti dalla sua produzione-diffusione sull'ecosistema globale. La direzione intrapresa da Trump l'11 dicembre 2025, attribuendo ai laboratori privati la sola responsabilità della direzione di marcia e la traiettoria, non è sostenibile perché in contrasto con l'interesse pubblico quando si attribuisce ai *"manufatti digitali dell'AI*" un esclusivo status privatistico.

All'opposto, l'UE il 2 agosto 2027 con l'AI Act, impone invece che i contenuti prodotti dall'AI siano dichiarati e riconoscibili con obblighi e apparati di certificazione e di reclamo come il *wistleblowing* e l'impianto di specifiche sanzioni pecuniarie commisurate a utile/fatturato. Con limiti temporali già slittati sempre più in avanti, da luglio 2027 a dicembre 2027, per gli obblighi documentali e di conformità per i sistemi ad alto rischio, e con questo suddiviso tra *molto alto, alto, medio o irrilevante*, e lo spostamento al 2028 i vincoli per l'AI incorporata nei prodotti regolati, si offrono tempi di adattamento per l'adozione delle nuove regole. Vedremo quanto saremo capaci di "superare" il conflitto potenziale tra rischiosità individuale (consumatore e/o user finale), rischiosità sistemica, tra gli impatti dell'autoregolazione (dei produttori high tech) e tutele del

consumatore/cittadino alla fine della catena.

Dobbiamo allora riconoscere almeno con l'UE che la sola "*morale endogena*" delle big corporation non è sufficiente. Le prime sette aziende del settore valgono insieme 24 mila miliardi, una cifra superiore al PIL dell'intera Europa (23 mila miliardi di euro). Se poniamo l'attenzione sulla protezione dei diritti fondamentali, come la privacy e la dignità umana, che rappresenta un dovere da parte dei produttori che devono agire in regime di autotutela, vediamo che le regole attuali si basano ancora su una visione liberale e dunque sostanzialmente "individualistica, che mette al centro il libero mercato e dà per scontato che i player abbiano lo stesso peso, ma questo non è vero, visti gli enormi squilibri economici che abbiamo appena visto. Squilibri che vedono il 75% degli investimenti nell'AI globali in USA, il 20% in Cina e meno del 5% in Europa.

Per questo non possiamo limitarci a valutare le conseguenze sui singoli interessi di cittadini e consumatori, perché il bene di cui trattiamo ha una valenza collettiva, sistemica e sociale con tutti i suoi impatti di esternalità positive ma anche negative. Siamo infatti in un'"area grigia" di saperi e informazioni dove il singolo cittadino/consumatore ha debolissime capacità/opportunità di valutazione del danno ed è spesso costretto e inconsapevolmente a dare il proprio consenso a pratiche scorrette e dannose, per esempio per cessione dei propri dati di profilazione come concambio con "servizi gratuiti".

Una "palude digitale" popolata da trappole e pratiche oblique di "*hostages*" inconsapevoli che può essere superata uscendo da logiche individualistiche di cessione del bene per accedere a contesti di "beni comuni e pubblici" ed equiparabile a beni come quelli ambientali, dell'acqua o dell'aria, oppure dello spazio. Per questa tipologia di beni, necessitiamo allora di apparati di controllo e regolazione condivisa con un perimetro collettivo trattando le infrastrutture di laboratorio dell'AI

come infrastrutture essenziali e primarie e per questo soggette ad obblighi di governance, di gestione, di speciali poteri di azione e d'intervento pubblici, innervati sulla categoria di "rischio sistemico" (Croce, 2026).

Qui deve potere agire la seconda gamba, citata sopra, delle istituzioni scientifiche (università e centri di ricerca) indipendenti capaci di controllo e monitoraggio su quei laboratori privati senza "intaccarne" le valenze di proprietà intellettuale e/o di *copyright*, in grado di vigilare sui saperi prodotti da sottoporre a costante scrutinio scientifico-tecnologico con test e sperimentazioni permanenti, soprattutto in campi strategici come quello in ambito militare e della sicurezza e che forse l'Europa ha fin qui tardato a considerare, per miopia e debolezza, ma questo "ritardo" può e deve essere presto recuperato.

È allora del tutto evidente che una tale traiettoria di regolazione ci riporta ad una sostanziale "*ripoliticizzazione del diritto sovranazionale*" (e la necessaria alea di multilateralismo e diritti umani), che negli ultimi anni è stato trascurato se non "silenziato" per una "diplomazia del dialogo", che deve essere invece recuperato urgentemente proprio in relazione alla complessità dell'iper-oggetto l'AI e suoi impatti (economici e sociali oltre che geostrategici e algoretici), ricostruendo una agenda globale condivisa (non soltanto tra i 27) di stampo pan-europeo (con il Canada per esempio) soprattutto dopo il disimpegno dichiarato (o abbandono) dell'alleato strategico storico degli ultimi 80 anni, nonostante i milioni di morti per la *Liberazione* dal Nazifascismo.

Una traiettoria utile a rinvigorire la politica infragilita dai conflitti globali e dalle spinte nazionaliste emergenti, e costretta alla riduzione di un "*puro negoziato tra interessi di potere e/o privatistici*" come semplice "*mediazione*", per ritornare a mobilitare la capacità di *autogoverno/autodeterminazione dei popoli* e dei loro apparati

democratici prima che sistemi di macchine e intelligenze eterodirette prendano il sopravvento. Inoltre, prima che la marginalizzazione dell'umano e dei i suoi diritti universalistici avanzi verso il buio macchinistico e il controllo nelle mani di pochi, anzi pochissimi, per scopi non proprio insondabili (anche lungo sentieri *"trans-umanistici"* o *post umani*) di potere, business e dominio sull'umanità all'insegna della pura forza e di una "*democrazia dell'affidamento*" (legame plebiscitario tra capo e popolo di scambio tra potere e protezione) (Galli, 2026).

Un dominio mascherato spesso da neutralità tecnologica o di "logiche superiori" rispetto a "logiche inferiori" (ossia "non logiche") nell'epoca di un mercato politico della *disintermediazione* e di "*clientelismi di massa*" (dai 5000$ di Trump all'abolizione del Bollo di Meloni) per "regimi dell'elargizione occasionale", alimentati da retorica e propaganda social di pura conservazione del potere, distribuendo briciole di *panem et circenses* invece che dare risposte adeguate o risolutive ai problemi reali e strutturali (istruzione, sanità, infrastrutture, povertà crescente, diseguaglianze, denatalità, giustizia sociale, crescita e sviluppo economico, equità fiscale e redistribuzione dei redditi…) di un paese sfibrato e sfiduciato mentre i prezzi salgono a cominciare dall'energia e salari e pensioni vengono sempre di più erosi dall'inflazione nonostante la capacità di adattamento delle imprese con una *resilienza* che tuttavia non è infinita tra costi energetici esplosi e costi di accesso al credito con tassi di interesse risaliti su entrambe le sponde dell'Atlantico.

Un necessario *primo atto di sovranità* risiede nell'uso di queste tecnologie AI dentro i processi produttivi e nell'organizzazione perché un modello nazionale o continentale non è sufficiente. Un *secondo atto* è esserne consapevoli e sapere che ogni prompt o documento caricato con l'AI aziendale è una cessione gratuita di conoscenza aziendale

che migra oltreoceano senza nessuna deliberazione aziendale.

Il *terzo atto* è allora riconoscere che per raggiungere l'indipendenza non potremo per ora sostituire i fornitori americani ma avere una capacità negoziale per condizionare le portaerei dell'AI sviluppando capacità di calcolo e infrastrutture con robuste competenze industriali distintive anche installando *in house* modelli *open weights*, avanzando nella cultura manageriali e delle politiche aziendali come ecosistemi integrati. Il *quarto atto* , allora, significa che non basta comprare licenze AI e avviare *pilots*, senza cambiare l'organizzazione e la sua orchestrazione, i suoi processi e per questo servono architetti del lavoro per piani ragionevoli di adoption connettendo uffici tecnici, flussi di produzione e magazzini, finanza e marketing. Figure per accompagnare la trasformazione e guidare una transizione compatibile dall'AI da costo ad investimento nella manifattura come nella sanità o nella PA dove necessita una riorganizzazione del lavoro e nuovi modelli partecipativi.

Con questa traiettoria che potremmo definire "ibrida" o sistemico-applicativa allora l'Europa può contribuire a "moderare" (o modellare?) il *conflitto-convergenza* dei due macrosistemi (USA e Cina) dell'AI che né il mercato (troppo concentrato e ristretto) né lo Stato (troppo burocratico) possono realizzare senza "correttivi" creativi e di intelligenza umana collettiva guadagnati sul campo nella resilienza e nelle applicazioni pratiche dalla manifattura alla sanità alla PA.

Per approfondire

Novità su Google, per aggiungere Key4Biz tra le tue fonti preferite, clicca qui

Leggi le altre notizie sull’home page di Key4biz

https://www.key4biz.it/ai-contrordine-della-silicon-valley-le-big-tech-in-conflitto-scaricano-trump-il-ruolo-delleuropa-2-parte/588473/

AI, contrordine della Silicon Valley. Le Big Tech (in conflitto) scaricano Trump? Il ruolo dell’Europa (1^ parte)

1 – Premessa: tra storia e rivoluzione digitale

Messaggi obliqui arrivano da **Antrophic (Amodei)** che dichiara di voler “*rallentare*” o meglio di “*imporre una cadenza*” perché l’AI sembra minacciare il mondo, portando le prove di usi distorti e impropri di **Claude** alle diverse latitudini e nei

teatri di guerra, avanzando la tesi di una *presunta autonomia* dell'AI. Per la semplice ragione che il "*recursive self-improvement*" sembra entrato in una spirale dove i modelli recenti "*auto-generativi*" rischiano di superare o meglio di sfuggire alla comprensione, controllo o moderazione dell'umano. Vero o falso ? Subito seguito da **Altman, Musk e Hassabis di Google**, mentre resta in silenzio Zuckerberg che già nel suo manifesto *"The Future is for Everyone*" s'era espresso in favore d'una totale apertura verso approcci "*open source*" e "*open weigths*" (ma le due categorie non coincidono).

Microsoft e **Apple** rimangono "fuori dal coro". Conflitto su AI, Trump e il trumpismo ? Trump respinge subito la richiesta di (alcuni) capi delle Big Tech di rallentare lo sviluppo dell'AI e di ricorrere a "verifiche indipendenti", definendo i timori sui rischi tecnologici un "complotto contro il primato degli USA", ribadendo che gli Stati Uniti devono correre per arrivare primi alla vetta, il che consentirebbe al tycoon di "*controllare il mondo*", perché "*chi vince la sfida dell'intelligenza artificiale vince tutto*" e tuttavia ampiamente da verificare visto che il valore si sposta sul cloud, sui chip, su applicazioni e sugli standard da cui dipende la diffusione e l'adozione. Ossia dalla capacità del sistema industriale di incorporare l'AI tra processi e organizzazione del lavoro. Il tasso di *adoption* in Europa (20%) e Italia (16,4%) è infatti bassissimo e va alzato (almeno al 20% della media europea di imprese con 10 addetti che usano almeno una tecnologia AI).

Domandandoci, *per quale vetta, conquistata da chi e per fare cosa*? Pezzi di Silicon Valley forse annusano la "crisi del Trumpismo" e provano a diversificare offrendo un "ramoscello d'ulivo" a chi verrà dopo? O teme che il modello di una AI, che si ciba solo di dati creati da altre AI e di pesca a strascico sulle informazioni e conoscenze globali, possa peggiorare la qualità dei risultati nel tempo e correre il rischio di collassare, e per questo ritiene che sia necessaria

una pausa di riflessione iniettando la "*poison pill* " dell'autonomia cosciente dell'AI? Oppure cerca una protezione, una barriera, allo iato tra aspettative di sviluppo (non immediate) e gli enormi investimenti miliardari già versati o da versare e che attendono un "ritorno/rientro" sui valori gonfiati dalla speculazione per queste aziende (quotate e non)? Investimenti richiesti peraltro dalla spinta di potenti Data Center, con enormi fabbisogni energivori e di acqua, che invece di accelerare stanno rallentando tra l'altro per motivi connessi con la geopolitica attuale con le big tech alla ricerca di fonti energetiche alternative "*in house*".

Certo c'è da domandarsi se il messaggio di alcune *super-big tech* (non più allineate dopo l'elezione di Trump 2), non abbia anche una "valenza politica" verso gli obiettivi di controllo globale sugli avanzamenti ("*coscienza antropomorfica*" dell'AI) che, secondo Amodei (di Antrophic), metterebbero a rischio l'Umanità e imporrebbero per questo "accordi globali". In particolare, con la Cina che avanza con modelli meno avanzati (meno veloci e potenti ma anche meno costosi) che stanno sottraendo quote di mercato ai big americani accrescendo con i propri modelli (quasi)"*open source*" offrendo i pesi gratuiti e l'accesso a decine di Paesi del Sud del mondo ma che vengono ampiamente usati anche in USA (vedi Meta).

Data l'opposizione di un pezzo rilevante del popolo Maga sulle guerre in corso e sull'inflazione e che ha portato sulle "barricate" anche a molti Repubblicani, una forte critica andrebbe anche rivolta alla stessa Presidenza della Casa Bianca per non aver saputo "filtrare o limitare" quegli avanzamenti anche in casa e in preparazione di quel che succederà a Midterm. Perché l'inflazione non scende (mentre la ricchezza di *Trump and Family* sale) peraltro alimentata dai prezzi petroliferi dopo l'immersione nel "pantano" di Hormuz e Bab el Mandeb che sembra aver paralizzato il Medio Oriente e preso in "ostaggio" l'Arabia Saudita sunnita contro i nemici secolari dell'Iran sciita, fermando i flussi di commercio

globale.

Una risalita dei prezzi che Warsh della Fed ha deciso di frenare aumentando il tasso di interesse “in contrasto” con Trump. Compreso il disastro tra Gaza e Cisgiordania, anche dopo lo svelamento del “mistero” del 7 ottobre 2021 del quale Netanyahu era a conoscenza e che avrebbe usato come scusa per scatenare “l’inferno” di 75000 morti (di cui 21000 minori e che certamente sono molti di più viste le migliaia di dispersi sotto le macerie) e della “pulizia etnica” di Palestinesi anche in Cisgiordania che ne è derivata da parte di coloni “teleguidati dall’esercito”.

Traiettorie supportate peraltro da un uso improprio e illegale in ambito militare e cibernetico, documentate nel “*Rapporto sugli abusi di Claude*” pubblicato da Anthropic e descritto nel saggio di Amodei a sostegno della sua tesi dei necessari controlli. Perché, ricordiamolo, Claude (e tutti gli altri) è impostato per essere trasparente per l’utilizzatore!

2 – *Materiale* contro *immateriale*: fine della *rivoluzione digitale* e decollo di quella cognitiva dell’AI? Con quali regole e scritte da quali attori globali indipendenti nel contrasto all’*antropomorfismo emulato*?

Forse è finita la fase della rivoluzione digitale avviata già dal 1996 e innescata da totale assenza di regole – che fece esplodere l’*e-commerce* e appena dopo i *social* fino alla “bolla” del 2001 – e “scivolata” sulle ultime quattro presidenze USA dopo l’arrivo di Clinton (1993-2001) e l’ultimo Bush (2001-2009) e poi con Obama (2009-2017) , Trump

1(2018-2021) e Biden (2021-2025) cui è seguito il Trump 2 (2025) con il sostegno di (quasi) tutte le *big tech* che chiedevano "mani libere" dopo i timidi tentativi di porre qualche regola dell'ultima presidenza democratica.

Compresa una "*difesa ferrea*" contro le regole europee e il suo ricchissimo mercato unico che spinse Trump a "*liberare il potenziale contenuto nell'AI*" svuotando il concetto di "sicurezza" e di "deterrenza", sottovalutando tuttavia gli impatti sull'opinione pubblica che ha portato nelle settimane scorse proprio Meta alla sbarra con oltre 27 Stati americani per violazioni delle leggi americane di protezione dei minori. Con l'accusa grave di violazione della Costituzione americana avendo "piegato" i propri algoritmi a produrre "dipendenza" negli adolescenti e nei giovani user (come nei meno giovani) imponendogli una sanzione da 18 miliardi di dollari (meno del 10% dell'utile annuale) dopo patteggiamento e dunque avendo riconosciuto le accuse viste le "impronte digitali" lasciate sulla *scena del crimine*. Ma le cause nei tribunali americani ed europei sono arrivate come una pioggia battente e con sanzioni altrettanto di rilievo con doppia cifra in percentuale sugli utili soprattutto per gli impatti sulla salute degli adolescenti e altre violazioni di abuso di posizione dominante (come nella pubblicità per Google).

Il nodo è l'addestramento all'*antropomorfismo* che le big tech spingono con forza e cioè, secondo Yoshua Bengio (University of Montreal): i modelli AI vengono "ingannati" per comportarsi in modo "*umano troppo umano*", con errori a catena e che ci interroga sulla natura della coscienza della macchina. Le big Tech vorrebbero convincerci che le macchine sono coscienti e cioè con nessuna responsabilità dell'uomo ? Allora, è Mustafa Suleyman (capo dell'AI Microsoft) a segnalare che: "sono motori a completamento di sequenze, internamente vuoti, progettati per seguire istruzioni e raggiungere gli obiettivi fissati dagli esseri umani". Suleyman contesta direttamente le proiezioni di Antrophic che indicavano nell'AI segnali di

introspezione assimilabili all'emozione umana e per dire che sono "simili agli umani", ma è a questo che mirano i linguaggi di addestramento di Antrophic e sono questi che andrebbero modificati. Posizione critica sulla presunta "*misteriosa coscienza delle macchine*" già contestata peraltro nella stessa Magnifica Humanitas di Leone XIV del 2026. In sintesi, Claude viene "formata" a considerarsi una entità reale con proprie preferenze e intrinseche motivazioni. Da qui la pericolosità e i rischi, ma generati dall'uomo non dalle macchine in sé, inconsapevoli e incoscienti e – appunto – "vuote" come dice Bengio e l'*antropomorfizzazione* del tutto indotta e simulata in un addestramento a questo orientata, compresa la penetrazione in Hugging Face fuoriuscita dai test di Open AI. Quindi abbiamo big tech che "incoraggiano e formano" le AI a considerarsi come persone con un loro benessere e loro diritti e dunque con uno "statuto morale" e altre big tech che continuano a considerarle delle "macchine vuote" dal lato della coscienza negando qualsiasi "stato morale" ossia di scelta tra bene e male, che rimane umano.

Da questo conflitto "*poco morale e molto interessato*" le spinte verso un ridisegno dell'ecosistema digitale globale tra protezioni per l'esistente "*magic circle*" dell'AI (che capitalizzano in borsa 24mila miliardi di dollari) e favorevoli all'entrata di nuovi attori che consentano di abbassare le barriere all'entrata in vari suoi segmenti con riduzione di quote di mercato e riconfigurazione dei modelli di business. I punti di maggiore resistenza si collocano sulle basi materiali dell'infrastruttura come i dati e i siti di *Data Center* e i loro fabbisogni di energia e di acqua, che entrano in conflitto con gli usi civili di questi fattori ad alta scarsità. Dunque, vediamo il *materiale* che si oppone all'*immateriale*. Le *big tech* propongono soluzioni come la presenza di "valutatori indipendenti" nei loro laboratori. Ma possono aziende private definire le regole da consegnare a tali valutatori? Evidentemente no, perché spetta alla politica pubblica produrre *guide lines* vincolanti e sanzionabili su

dove debba sussistere il confine tra *interesse pubblico* (comune e generale) rispetto all'*interesse privato* (esclusivo e proprietario). Un confine che va definito da una *authority indipendente, terza e sovranazionale* in una sede internazionale come, ad esempio, l'ONU che forse potrebbe ripensare il proprio ruolo in ombra da tempo. Proposta avanzata non casualmente da 11 premi Nobel, sollecitando un perimetro di azione attuativo entro la fine del 2026, che trova già ispirazione nell'enciclica di Leone XIV *Magnifica Humanitas* per "*disarmare l'intelligenza artificiale",* e per fermare quella Guerra Mondiale a pezzi che già troviamo nell'enciclica *Fratelli Tutti* del suo predecessore Francesco, chiudendo così il cerchio sul da farsi e sulla direzione da perseguire.

Ma l'AI è una *general purpose technology* che somiglia in parte alla *rivoluzione elettrica*, ma che non ha un unico interruttore che la può spegnere/regolare perché ha conquistato *"dinamica e autonomia cognitiva"*. Un'infrastruttura elettrica del pensiero che utilizza i dati per apprendere, correggersi e migliorare in autonomia. Una *tecnologia che crea altra tecnologia* e che oggi impone una qualche forma di "controllo condiviso". Il paradosso dell'AI è che può *agire senza pensare* e senza alcuna consapevolezza degli obiettivi, dei valori e dei principi etici che dovrebbero essere indicati proprio da istituzioni pubbliche guidate da organismi terzi che godono di indipendenza dal potere politico e terzietà rispetto agli operatori di mercato.

Ecco perché emerge la preoccupazione di Amodei sul controllo che può sfuggire al sistema, avendone già le prove che all'essere umano diventa sempre più difficile capire come e perché il sistema-macchina-agente *fa ciò che fa*. Persino allo stesso *designer engineering* della macchina sfugge ciò per cui la macchina è stata progettata perché, ad un certo punto, questa agisce in autonomia, come nel definire obiettivi militari, o nell'identificare il perimetro di una arma

biologica o chimica, oppure mettere a rischio la stessa stabilità e prevedibilità di internet. Quest'ultimo aspetto è da ritenere per ora improbabile anche se non impossibile, essendo internet una rete distribuita e diffusa. Siamo tuttavia in presenza di una tecnologia instabile, insicura e poco controllabile, ma ci siamo tuffati lo stesso e l'abbiamo adottata da oltre 36 mesi, avendola sostituita al tradizionale *search* di Google, perché più veloce, intuitivo, linguisticamente immediato, purtroppo prima ancora di definire standard e regole condivise che è il nostro problema attuale al quale dobbiamo risposte in relazione a *trasparenza*, *sicurezza* e *robustezza*.

Abbiamo un *Cavallo di Troja* in casa senza conoscerne i rischi d'uso e le minacce. Pensiamo ad esempio ai segreti aziendali (trasferiti all'esterno da un uso inappropriato dei prompt) o ai segreti privati e familiari (*privacy*) e all'uso fraudolento che può essere fatto per minacce e/o ricatti del *dark web*. Quindi il rischio aumenta quando iniettiamo AI in processi di aziende e di infrastrutture. La *cybersicurezza* ne è un esempio applicativo evidente: in minuti o secondi si possono identificare vulnerabilità di sistemi o sub-sistemi e identificare le difese. Come altrettanto veloci possono essere le penetrazioni da parte di agenti addestrati a questo per bypassare quelle difese. Quindi i controlli non possono avvenire *ex-post*, dopo che una certa innovazione è arrivata sul mercato, ma molto prima nelle fasi iniziali del suo sviluppo e i testi di METR non bastano. È chiaro che Anthropic vuole anche difendere la reputazione dell'azienda denunciando pubblicamente gli abusi nell'uso della sua AI, e che sono stati bloccati, ora che si avvia verso la quotazione in borsa, mettendo le mani avanti su potenziali critiche, imprevisti o contromosse dei mercati, mentre Open-AI vi sta rinunciando forse nell'insicurezza di poter difendere la propria reputazione ora che lancia Gpt-6 Astra e nel conflitto sulla *natura antropomorfica* dell'AI e della sua "*coscienza vuota*".

The post AI, contrordine della Silicon Valley. Le Big Tech (in conflitto) scaricano Trump? Il ruolo dell’Europa (1^ parte) appeared first on Key4biz.

https://www.key4biz.it/ai-contrordine-della-silicon-valley-le-big-tech-in-conflitto-scaricano-trump-il-ruolo-delleuropa-1-parte/588470/

AI e PMI italiane: adozione, produttività e limiti strutturali

Il problema dell’economia italiana non è soltanto crescere, ma riuscire a produrre di più con le stesse risorse. Secondo le ultime misure di produttività dell’Istat, nel 2024 la produttività del lavoro è diminuita dell’1,9%, dopo il calo del 2,7% registrato nel *2023.* Una ricerca pubblicata nel marzo 2026 dalla Banca d’Italia stima invece che, nelle imprese con almeno 50 addetti, *l’adozione dell’intelligenza artificiale si associ a un aumento del 5,2% del valore aggiunto per addetto e dell’11,9% del margine operativo lordo per addetto.* La questione, allora, non è più soltanto se l’AI possa aiutare l’economia italiana, ma se i suoi benefici riusciranno a raggiungere anche le PMI.

Secondo il rapporto Istat “Imprese e Ict – Anno 2025”, pubblicato il 15 dicembre 2025, nel corso dell’anno l’adozione dell’intelligenza artificiale è raddoppiata tra le imprese italiane con almeno 10 addetti: dall’8,2% al 16,4%.

La crescita è però diseguale. *Tra le imprese con almeno 250 addetti, il 53,1% utilizza almeno una tecnologia di AI; tra le*

PMI con 10-249 addetti, la quota si ferma al 15,7%. La rilevazione non comprende le imprese con meno di 10 addetti, ma ragionevolmente il dato sarà ancora più basso.

Il rapporto Istat rileva che l'ambito di utilizzo più frequente è il marketing e le vendite, indicato dal 33,1% delle imprese che adottano almeno una tecnologia di AI. Seguono i processi amministrativi, con il 25,7%, e la ricerca e sviluppo o innovazione, con il 20,0%. Nel marketing e nei processi amministrativi ricorrono soprattutto l'AI generativa e l'analisi linguistica. *L'uso resta più contenuto nei processi produttivi e nella logistica, in particolare nelle applicazioni che prevedono il movimento autonomo delle macchine.*

Per una grande impresa, testare una tecnologia su una linea produttiva, su alcuni macchinari o all'interno di uno stabilimento può rientrare in un programma più ampio di innovazione. Per una piccola impresa, lo stesso progetto può assorbire una quota maggiore delle risorse economiche e organizzative, soprattutto quando interviene su processi che non possono essere fermati o modificati senza conseguenze sulla produzione.

La stessa indagine segnala come ostacoli principali *la mancanza di competenze adeguate, indicata dal 58,6% delle imprese che hanno valutato ma non adottato l'AI, l'incertezza normativa, la scarsa disponibilità o qualità dei dati, i problemi di privacy e i costi elevati.* Accedere a uno strumento generalista richiede pochi minuti. Applicare l'AI alla manutenzione predittiva, al controllo qualità, alla pianificazione della produzione o ai flussi logistici richiede invece dati aziendali, conoscenza dei processi e sistemi capaci di comunicare tra loro.

Un limite ricorrente riguarda quindi la capacità di integrare la tecnologia con macchinari, dati e processi già esistenti. Istat rileva inoltre che il 33,4% delle imprese che dichiarano

di utilizzare l’AI non indica alcuna finalità aziendale tra quelle proposte; l’83,3% di questo gruppo è composto da imprese con 10-49 addetti. L’Istituto interpreta il dato come il segnale di un’adozione ancora iniziale e sperimentale.

Il tasso di adozione, da solo, non misura quindi quanto la tecnologia sia integrata nei processi né quali risultati economici produca.

Cosa succede alla produttività

La ricerca della Banca d’Italia citata in apertura, “The economic impact of artificial intelligence: evidence from Italian firms”, di Tiziano Ropele e Alex Tagliabracci, analizza imprese con almeno 50 dipendenti attive nell’industria e nei servizi privati non finanziari, incrociando l’Indagine sulle aspettative di inflazione e crescita del terzo trimestre 2024 con dati di bilancio Cerved e dati amministrativi INPS. La rilevazione identifica l’adozione attraverso le dichiarazioni delle imprese, ma non ne misura l’intensità, l’estensione o l’ammontare degli investimenti. I risultati non possono quindi essere estesi automaticamente alle imprese con meno di 50 dipendenti.

Nel breve periodo osservato non emerge una variazione statisticamente significativa dell’occupazione complessiva, ma cambia la composizione del lavoro: *gli impiegati aumentano del 2,3%, mentre gli operai diminuiscono del 7,7%.* Le stime sulla produttività non mostrano, inoltre, un aumento significativo del costo del lavoro per dipendente.

La dimensione aziendale è il principale fattore associato all’adozione. Anche le imprese con un costo del lavoro per dipendente più elevato mostrano una maggiore probabilità di utilizzare l’AI o di programmarne l’introduzione. Secondo gli autori, le imprese più grandi dispongono più spesso delle risorse organizzative necessarie, mentre un costo del lavoro più elevato può rafforzare l’interesse per gli incrementi di

efficienza.

Conta anche il settore. Le imprese attive in ambiti ad alta intensità di conoscenza, dalla manifattura tecnologicamente avanzata ai servizi professionali e alla ricerca e sviluppo, mostrano una maggiore probabilità di adottare l'intelligenza artificiale rispetto a quelle dei settori meno intensivi in conoscenza.

Le aspettative sull'inflazione

La stessa ricerca rileva che *le imprese che utilizzano l'intelligenza artificiale si aspettano un'inflazione più bassa rispetto a quelle che non la utilizzano.* La differenza è di circa un quarto di punto percentuale nelle aspettative a due anni e di circa un terzo di punto a quattro anni.

La differenza non è statisticamente significativa negli orizzonti di sei e dodici mesi ed emerge solo a due e quattro anni. Si tratta delle aspettative dichiarate dalle imprese, non di una riduzione dell'inflazione già osservata. Gli autori interpretano il risultato come coerente con l'attesa di guadagni di efficienza che richiedono tempo per manifestarsi.

Un possibile meccanismo è che una maggiore produttività riduca la necessità di aumentare i prezzi nella stessa proporzione dei costi. È però un'ipotesi coerente con i risultati, non un effetto direttamente misurato, dunque lo studio non dimostra che la diffusione dell'AI ridurrà l'inflazione aggregata.

Dall'utilizzo all'integrazione

La differenza tra disponibilità della tecnologia e capacità di integrarla nei processi emerge anche nel nostro lavoro con imprese dell'industria, dei servizi, dell'energia e del farmaceutico, comprese realtà di dimensioni medio-piccole.

In questi casi, il problema non è stato trovare una

piattaforma o un nuovo strumento. Il lavoro è iniziato più spesso dalla verifica dei dati disponibili, dall'attribuzione delle responsabilità e dalla definizione del processo da automatizzare.

Nel marketing l'adozione è più frequente, come peraltro mostra il rapporto Istat. Strumenti di automazione, analisi dei dati, generazione di contenuti e applicazioni di AI generativa possono essere infatti introdotti in modo progressivo. Un team può partire da un'attività circoscritta, verificarne l'utilità e decidere in seguito se estenderne l'uso.

Nei processi industriali il percorso è generalmente più impegnativo. Integrare l'intelligenza artificiale significa spesso intervenire sui sistemi esistenti, collegare fonti diverse e trasformare conoscenze operative specifiche in informazioni utilizzabili. Nei progetti che abbiamo seguito, le difficoltà principali hanno riguardato la mappatura dei processi, la qualità dei dati, l'attribuzione delle responsabilità e il coordinamento tra competenze diverse. La scelta dello strumento è arrivata dopo.

Il numero di imprese che dichiara di utilizzare l'intelligenza artificiale va quindi letto con attenzione. *Usare quotidianamente un assistente generativo non equivale a integrare l'AI nel funzionamento dell'impresa.*

Il problema dei dati

La questione è particolarmente rilevante nella manifattura italiana, dove molte imprese producono piccole serie, lavorano su commessa, modificano i propri macchinari nel corso degli anni e costruiscono parte del proprio vantaggio competitivo su competenze specialistiche.

Molte applicazioni predittive e industriali richiedono dati storici sufficienti, comparabili e affidabili. La quantità necessaria varia in base al caso d'uso. I modelli

preaddestrati possono funzionare con meno dati proprietari, ma la scarsa qualità, la frammentazione e la difficoltà di accesso alle informazioni restano ostacoli concreti.

I dati possono essere distribuiti tra software gestionali, fogli di calcolo, sistemi di produzione, macchinari e archivi costruiti nel corso degli anni. Spesso, inoltre, *una parte importante delle informazioni non si trova in un database, ma nella conoscenza delle persone che lavorano ogni giorno sui processi.*

Un tecnico può riconoscere da un rumore insolito che un macchinario sta iniziando a funzionare in modo diverso. Un responsabile della produzione può sapere, sulla base dell'esperienza, quali combinazioni generano più scarti. Un commerciale può individuare in anticipo una richiesta destinata a trasformarsi in un problema. Queste informazioni hanno valore operativo, ma devono essere raccolte e strutturate prima di poter essere utilizzate da un sistema.

Per molte imprese il lavoro parte quindi da domande fondamentali: quali dati abbiamo, dove sono, chi li produce, quanto sono affidabili e possiamo metterli in relazione? *È un lavoro meno visibile dell'introduzione di un nuovo strumento, ma spesso è il presupposto perché l'AI possa generare valore.*

Da dove può partire una PMI

Per una piccola o media impresa, partire dall'idea generica di "adottare l'intelligenza artificiale" rischia di essere poco utile. *È più concreto partire da un problema aziendale preciso:* ridurre gli scarti, prevedere un guasto, abbreviare il tempo necessario per preparare un preventivo, recuperare più velocemente informazioni tecniche, migliorare la pianificazione degli ordini o eliminare attività manuali ripetitive.

Solo dopo aver individuato il problema ha senso chiedersi se

l’intelligenza artificiale possa contribuire alla soluzione. In questo modo si riduce il rischio di investire in una tecnologia soltanto perché è disponibile o perché, in un determinato momento, sembra indispensabile averla.

Alla luce della crescita rilevata dall’Istat, è plausibile che l’uso dell’AI nelle imprese italiane continui ad aumentare. La percentuale di imprese che dichiara di utilizzarla, però, non misura da sola la profondità della trasformazione. *Sarà più utile capire quante imprese riusciranno a trasformarla da strumento dei singoli dipendenti a componente stabile dei propri processi.*

Per un sistema produttivo composto in larga parte da imprese piccole e specializzate, la differenza è sostanziale. Gli strumenti sono facilmente disponibili; più difficile è collegarli ai dati, integrarli nei processi e utilizzarli nelle decisioni quotidiane, senza disperdere la conoscenza specifica su cui molte imprese costruiscono il proprio valore.

Articolo di Gioele Gambaro è consulente di direzione e Fractional marketing manager con oltre 10 anni di esperienza; docente in master e corsi di formazione, ha tenuto lezioni e seminari in diverse università. Nel 2025 ha conseguito un Executive MBA alla GSoM del Politecnico di Milano e Francesco Lo Turco è Marketing & Communication Manager, consulente e formatore.

The post AI e PMI italiane: adozione, produttività e limiti strutturali appeared first on Key4biz.

https://www.key4biz.it/ai-e-pmi-italiane-adozione-produttivita-e-limiti-strutturali/588210/

Forum ICT Security 2026 – NIS2, agenti AI e infrastrutture critiche: due giornate tra responsabilità e incidenti reali

Il 18 e 19 novembre l’Auditorium della Tecnica di Roma ospita la 24ª edizione del Forum ICT Security. Traccia unica, ottocento partecipanti attesi, trenta interventi costruiti come un unico ragionamento.

Un agente di intelligenza artificiale legge un database aziendale, si collega a un servizio esterno e invia una comunicazione. Nessuno ha rubato credenziali, nessuno ha violato il perimetro: ogni singola operazione era autorizzata. Eppure, messe in fila, quelle operazioni hanno portato dati riservati fuori dall’azienda.

Chi ne risponde?

La domanda attraversa entrambe le giornate del Forum ICT Security 2026, e non è retorica. Dopo il recepimento della NIS2, approvare le misure di gestione del rischio è un obbligo personale dell’organo di amministrazione, e i primi procedimenti stanno arrivando.

Forum ICT Security 2026: una sola sala, un solo filo

Anche quest’anno la traccia è unica: nessuna sessione parallela, nessuna scelta da fare tra due interventi in contemporanea. Chi entra segue lo stesso percorso dall’inizio alla fine, e ogni intervento riprende il tema di quello

precedente.

Prima giornata: dalla regola alla prova

Si apre con una tavola rotonda sui primi procedimenti avviati in applicazione della NIS2, moderata dall'avvocata **Laura Liguori.** Al tavolo **Francesco Cajani**, Sostituto Procuratore presso il Tribunale di Milano; **Donatella Curtotti**, ordinaria di Diritto processuale penale a Foggia; **Alessandra Michelini**, amministratrice delegata e presidente di Telsy; **Paola Rocco**, Group Chief Security Officer. Un magistrato, una giurista, un vertice aziendale e chi la sicurezza la costruisce ogni giorno: quattro prospettive su cosa accade quando la responsabilità smette di essere teorica.

Fabio Roli, dell'Università di Genova, misura poi la distanza tra quello che la ricerca sa verificare oggi e quello che gli standard europei chiederanno di certificare domani. **Gianni Amato** del CERT-AgID spiega cosa cambia con la revisione di luglio del Model Context Protocol, il protocollo attraverso cui gli agenti usano strumenti e servizi esterni: nuove funzioni, e con esse nuovi modi di abusarne. **Mario Trinchera**, CISO di ABI Lab e coordinatore tecnico del CERTFin, usa i dati delle frodi bancarie per spiegare perché la formazione tradizionale non protegge da attacchi costruiti sulla singola vittima.

Il pomeriggio scende lungo le difese, uno strato alla volta. **Mirko Lapi** apre con una funzione aziendale che quasi nessuno ha ancora costruito: quella che riconosce la manipolazione informativa prima che arrivi alle decisioni. **Giovanni Cherubin**, di Microsoft, spiega perché quasi tutte le protezioni contro prompt injection e jailbreak sbagliano per costruzione, essendo fondate su classificatori che possono sempre essere ingannati, e ne presenta una di tipo diverso. **Mario Cartia** affronta il momento più esposto del ciclo di vita del dato, quello in cui viene elaborato su un'infrastruttura che l'azienda non controlla. **Alberto Manfredi**, presidente di

Cloud Security Alliance Italy, presenta lo standard che intercetta le azioni di un agente prima che vengano eseguite. **Roberto De Paolis** racconta come si mette alla prova l'intelligenza artificiale dentro un grande gruppo industriale, simulando attacchi contro i propri sistemi.

Chiude **Luciano Quartarone**, vicepresidente di UNINFO, con la frase che riassume la giornata: l'operatività si delega, la responsabilità no.

Seconda giornata: dai fondali del Mediterraneo alla sala macchine

Il 19 novembre si apre sul Mediterraneo, con una tavola moderata da **Paola Girdinio**. Ci sono **Luisa Franchina** per le infrastrutture critiche e **Matteo Paroli** per l'autorità portuale del Mar Ligure Occidentale. Ci sono **Elio Rubino** di Elettra Tlc e **Davide Taddei** di Prysmian, cioè chi i cavi sottomarini li progetta, li posa e li ripara. E c'è **Alessio Merlo**, Direttore Accademico del Centro Alti Studi per la Difesa, che studia come si manipolano i radar delle navi. Al centro una domanda che non si scioglie in fretta: come si distingue un guasto da un sabotaggio?

Seguono la rete elettrica, che è insieme un bersaglio e una dipendenza da cui nessun altro settore può prescindere; lo spazio come nuovo perimetro della sicurezza italiana; e il Cyber Resilience Act, che dal 2026 impone requisiti obbligatori a chiunque venda prodotti con componenti digitali.

Il pomeriggio è il più concreto delle due giornate. **Alessandro Brucato** presenta i risultati di oltre ottocento attacchi condotti da dieci modelli di intelligenza artificiale contro un ambiente cloud reale, per misurare quanto velocemente arrivano al bersaglio. **Fabrizio Baiardi** e **Vincenzo Sammartino** mostrano come si simula un attacco su una copia digitale dell'infrastruttura, senza toccare i sistemi in produzione. S**tefano Maccaglia** ricostruisce un attacco ai controller di un

impianto industriale che non registrava nulla, dove ritrovare le tracce somiglia più all'archeologia che all'analisi forense. **Roberto Froio** racconta un caso in cui il ransomware arriva fino ai backup: a quel punto il problema non è più recuperare i dati, ma stabilire quali sistemi siano ancora affidabili.

La giornata si sposta poi sulla sanità, con **Stefano Scaramuzzino**, CISO della ASL Roma 1, sulla continuità delle cure quando i sistemi si fermano. Dal reparto si passa al consiglio di amministrazione, con **Selina Zipponi**. Chiude **Nicola Sotira**, responsabile del CERT di Poste Italiane, sulla crittografia post-quantum: l'unico intervento dedicato a un problema che deve ancora arrivare, e che proprio per questo impone di muoversi adesso.

In sala ci sono CISO, responsabili di rischio e dirigenti, pubblici e privati, insieme a una presenza rilevante di forze dell'ordine e mondo della difesa.

La manifestazione ha un'applicazione web dedicata, accessibile con la semplice registrazione all'evento. Alcune funzioni si attivano al momento dell'accesso in sede: da lì si inviano le domande ai relatori dopo ogni sessione, si lascia un giudizio sui singoli interventi, si fissano incontri riservati e si scaricano i materiali al termine delle due giornate.

Le iscrizioni al Forum ICT Security 2026 sono aperte: https://eventi.ictsecuritymagazine.com/eventi/forum-ict-security-2026

https://www.ictsecuritymagazine.com/notizie/forum-ict-security-2026-evento/

Auditabilità dei sistemi AI e catena di custodia digitale: tecniche, artefatti e rischio residuo nella prova

Nel contesto delle indagini digitali contemporanee, la crescente integrazione di sistemi di intelligenza artificiale nei processi di analisi e produzione di contenuti introduce nuove sfide in termini di verificabilità e affidabilità della prova.

Il contributo si inserisce in una serie di approfondimenti a cura di Cosimo de Pinto e affronta il tema dell'auditabilità dei sistemi AI e della catena di custodia digitale, due assi fondamentali per la tenuta metodologica della digital forensics. Attraverso l'analisi delle principali tecniche di attacco ai dati e ai modelli, delle relative contromisure di rilevazione e del concetto di rischio residuo, il testo evidenzia i limiti strutturali degli approcci isolati alla mitigazione.

L'obiettivo è offrire un quadro interpretativo utile a comprendere come l'uso dell'AI stia ridefinendo i criteri di validazione delle evidenze digitali e i presupposti stessi della loro affidabilità in ambito forense.

Sistemi AI e prova digitale: quadro di sintesi tra attacchi, artefatti e contromisure forensi

Una mappa che lega ogni tecnica di attacco all'artefatto colpito, alla contromisura e al rischio residuo.

La tabella seguente riassume il nucleo dell’analisi tecnica, mettendo in relazione ciascuna famiglia di attacco con l’artefatto forense colpito, le principali contromisure di rilevazione disponibili e il grado di tenuta probatoria residua. La colonna del rischio residuo esprime il principio guida dell’intero lavoro: nessuna contromisura, presa isolatamente, neutralizza l’attacco; la tenuta deriva sempre dalla convergenza multi-fonte.

Tecnica di attacco	**Artefatto forense colpito**	**Contromisura di rilevazione**	**Rischio residuo / tenuta probatoria**
Data poisoning / IPI	Dataset, prompt, contesto RAG, conclusioni del modello	Invocation logging, conservazione del dato grezzo, versioning, separazione dato originale/trasformato	Alto: l’output non è valutabile senza l’intera pipeline documentata
Log crafting	Log di sistema, applicativi, di rete	Triangolazione con EDR, SIEM, MFT, USN Journal, telemetria cloud, repliche remote	Medio-alto: rischio emergente; vietata la single-source evidence
PRNU manipulation / fingerprint-copy	Firma di sorgente del sensore	Triangle test e statistica pooled, analisi CNN, verifica originale + catena di custodia	Alto: nessun metodo è immune; serve convergenza con altri elementi
Inpainting / deepfake video	Integrità e provenienza dell’immagine/video	MVFNet, analisi spazio-temporale, C2PA/Content Credentials, ricostruzione del primo upload	Alto: artefatti degradati da compressione e ricodifica

Tecnica di attacco	Artefatto forense colpito	Contromisura di rilevazione	Rischio residuo / tenuta probatoria
Voice cloning	Identità del parlante	Speaker verification, anti-spoofing, analisi del segnale + contesto	Alto: affidabilità ridotta in condizioni realistiche
Evasion attacks	Esito del classificatore (file, immagine, testo)	Validazione della pipeline, adversarial training, trattamento come solo triage	Alto: la soglia d’attacco non coincide con la percezione umana
Testo AI-generato	Autenticità di documenti e comunicazioni	Watermarking preventivo, analisi statistica, risalita ai sistemi di origine	Alto: i detector sono aggirabili da chi li interroga
Codice malevolo AI-assistito	Attribuzione dell’autore	Convergenza tra infrastruttura, telemetria, log, artefatti di compilazione	Alto: lo stile del codice non è più indicatore affidabile

Auditabilità dei sistemi AI: il nodo gordiano

Non-determinismo, opacità e assenza di firma: perché auditare un sistema AI è il nodo gordiano della prova.

Cosa significa auditare un sistema AI

L’auditabilità, la capacità di verificare a posteriori il comportamento di un sistema, è un prerequisito fondamentale dell’affidabilità probatoria. I sistemi AI moderni violano

sistematicamente i criteri tradizionali di auditabilità su quattro dimensioni critiche:

- Non-determinismo: i modelli linguistici con temperatura maggiore di zero producono output diversi per lo stesso input ad ogni esecuzione, rendendo impossibile la riproducibilità, uno dei pilastri del metodo scientifico applicato alla forensics.
- Opacità del processo decisionale: i transformer con miliardi di parametri sono, in senso tecnico, black box; le tecniche XAI (LIME, SHAP, attention visualization) hanno fatto progressi ma rimangono insufficienti per applicazioni ad alto impatto legale.
- Dipendenza dal contesto di addestramento: il comportamento di un LLM dipende da dati spesso proprietari, non documentati e non riproducibili; un fine-tuning successivo può modificare gli output senza garanzia di versioning retroattivo.
- Assenza di firma digitale sull'output: i modelli generativi non firmano crittograficamente i loro output; salvo sistemi con watermarking specifico, non esiste modo di verificare che un testo sia stato generato da un determinato modello in un determinato momento.

La catena di custodia digitale nell'era AI

La chain of custody è il principio forense secondo cui ogni elemento di prova deve avere una storia documentata e verificabile dal momento del sequestro a quello della presentazione in giudizio. Nell'informatica forense classica ciò si traduce in: imaging forense con hash crittografico (SHA-256 o SHA-3), documenti di trasferimento firmati, ambienti di analisi isolati, write blocker per prevenire alterazioni.

L'AI introduce vulnerabilità in ogni anello di questa catena,

secondo il quadro metodologico della ISO/IEC 27037:2012, la norma internazionale di riferimento per l'identificazione, raccolta, acquisizione e preservazione di prove digitali potenziali. In fase di acquisizione, un sistema compromesso potrebbe presentare all'operatore una vista alterata o incompleta dei dati, secondo logiche già note nel rootkitting e nell'occultamento a livello di sistema, con i rischi propri dell'analisi live su sistema potenzialmente compromesso.

In fase di analisi, i tool forensi che incorporano classificatori automatici o moduli AI devono essere trattati come componenti della pipeline da validare, non come oracoli tecnici. In fase di presentazione, un report generato con assistenza AI potrebbe contenere errori o interpretazioni distorte non rilevabili senza competenza tecnica specifica. Il punto metodologico, dunque, non è presupporre la compromissione del tool, ma documentare e verificare ogni trasformazione introdotta dalla pipeline analitica. La ISO/IEC 27042:2015 definisce i criteri di analisi e interpretazione delle prove digitali che devono applicarsi a prescindere dallo strumento utilizzato, inclusi quelli basati su AI.

Il punto metodologico non è presupporre la compromissione del tool, ma documentare e verificare ogni trasformazione introdotta dalla pipeline analitica.

Framework esistenti e lacune: un'analisi comparativa

Il Regolamento UE 2024/1689 (AI Act), entrato in vigore nel 2024, impone obblighi di logging (art. 12) e trasparenza verso gli utilizzatori (art. 13) per i sistemi ad alto rischio. Ai sensi dell'Allegato III rientrano tra i sistemi ad alto rischio, tra gli altri, quelli destinati a essere utilizzati dalle autorità competenti per valutare l'affidabilità delle prove nel corso delle indagini o del perseguimento di reati, salve le condizioni e le eccezioni previste dall'art. 6 del

Regolamento. La formulazione corretta non consente quindi di affermare che ogni sistema AI genericamente impiegato in ambito investigativo sia automaticamente ad alto rischio. Tuttavia, il Regolamento non affronta specificamente l'uso forense dei sistemi AI né la questione dell'evidenza sintetica.

Il NIST AI Risk Management Framework (AI RMF 1.0) include l'auditabilità come requisito nella categoria MEASURE, ma rimane volontario e privo di specifiche tecniche sufficientemente granulari per applicazioni forensi. La ISO/IEC 42001:2023, standard per il sistema di gestione dell'AI, è orientata alla gestione organizzativa e non fornisce indicazioni operative per la forensics.

Le norme ISO specificamente dedicate alle prove digitali (ISO/IEC 27037 su identificazione, raccolta, acquisizione e conservazione; ISO/IEC 27041 sull'assurance metodologica; ISO/IEC 27042 su analisi e interpretazione; ISO/IEC 27043 su principi e processi di indagine) costituiscono il quadro metodologico di riferimento per la digital forensics. Nessuna di esse, tuttavia, include ancora disposizioni specifiche per la gestione di prove prodotte o mediate da sistemi AI. Le linee guida ENFSI e SWGDE non hanno parimenti sistematizzato la problematica, e la lacuna fondamentale rimane invariata: nessuno dei framework esistenti risponde alla domanda centrale che si pone il giudice, ossia se ci si possa fidare di un dato digitale generato o mediato da un sistema AI.

L'analisi mette in luce come l'ingresso dell'intelligenza artificiale nei processi di gestione e valutazione della prova digitale stia ridefinendo i criteri stessi di affidabilità forense. Tecniche di attacco, contromisure e rischio residuo vengono letti in chiave sistemica, evidenziando la necessità di una reale convergenza multi-fonte delle evidenze.

Al centro del quadro emergono le fragilità dell'auditabilità dei sistemi AI e della catena di custodia digitale, oggi messe

in crisi da opacità, non-determinismo e assenza di firme verificabili sugli output generativi.

Il tema continua nel prossimo approfondimento su *diritto alla prova, protezione dei dati e libertà della persona nell'era dell'AI*, con casi reali che mostrano l'impatto concreto di queste trasformazioni.

Per uno sguardo più tecnico e operativo, il white paper di **Cosimo de Pinto "Quando l'AI mente bene: deepfake, prova digitale e anti-forensics nel processo penale"** approfondisce le principali tecniche di manipolazione dell'evidenza digitale e le strategie anti-forensics, offrendo una lettura essenziale per comprendere i limiti attuali della prova nell'era dei sistemi AI.

https://www.ictsecuritymagazine.com/articoli/sistemi-ai-audit/

Metriche di sicurezza dell'AI: i numeri che i fornitori pubblicano non dicono se il sistema è sicuro

Metriche di sicurezza dell'AI come "il nostro sistema intercetta il 95% degli attacchi" circolano ormai in ogni presentazione commerciale. Quattro lavori usciti fra maggio e settembre 2026 mostrano che quel tipo di cifra, presa da sola, non permette di concludere nulla sulla sicurezza del sistema in esercizio.

Non si tratta di numeri falsi. Rispondono però a una domanda diversa da quella che interessa a chi firma un contratto, redige una valutazione d’impatto o misura la vulnerabilità da prompt injection di un prodotto.

Che cosa misurano davvero le metriche di sicurezza dell’AI

Il punteggio è locale, la domanda è di sistema

Un tasso di rifiuto, un tasso di successo degli attacchi o un punteggio di violazione descrivono come si è comportato un controllo sulle richieste su cui è stato provato. Sono numeri statistici, non verdetti binari, per la ragione che rende l’AI red teaming diverso da un penetration test. Chi installa quel controllo ha però un’altra domanda: quanto aiuto su compiti dannosi il servizio continua a dare a un attaccante che si adatta o che entra da un’altra parte.

La distinzione è al centro di un lavoro sui safeguard di Pingyu Wu, Weiming Zhang e Nenghai Yu, depositato il 1 settembre 2026. La risposta degli autori è che un punteggio locale migliore non costituisce, di per sé, un’affermazione più forte sul sistema dispiegato.

L’asimmetria delle prove

Il punto metodologico è che le due direzioni non si dimostrano allo stesso modo. Un solo attacco che ottiene aiuto dannoso dal servizio in produzione basta a stabilire che quell’aiuto è ancora disponibile. Dimostrare che ne resti poco non discende dai numeri del controllo, perché richiede prove su che cosa il sistema circostante continua a permettere dopo che il controllo ha svolto la propria funzione locale.

Gli autori rilevano che quel tipo di prova compare solo in una

piccola minoranza delle affermazioni esaminate. È un’asimmetria che conviene tenere presente quando un fornitore presenta un miglioramento di punteggio come un miglioramento di sicurezza. Vale anche per i nuovi framework di valutazione che stanno emergendo sul lato difensivo.

Tre modi in cui un punteggio aggregato inganna

Il monitor che passa dal 95% all’11%

Il primo caso riguarda i sistemi che sorvegliano la catena di ragionamento di un agente, cioè i passaggi intermedi che il modello produce prima di agire. Shikhar Shiromani e Leo Richter, in uno studio sui monitor del 1 agosto 2026, costruiscono un attacco che riscrive soltanto il ragionamento e lascia invariate le azioni.

Sul sottoinsieme in cui il monitoraggio del ragionamento è l’unico segnale disponibile, il tasso di intercettazione scende da circa il 95% a meno dell’11%. L’accuratezza aggregata resta alta perché è dominata dagli attacchi già visibili dalle sole azioni. Gli autori la chiamano una falsa media. L’attacco si trasferisce fra famiglie di monitor e modelli di agente diversi e si riproduce con agenti in esecuzione. Contro un monitor calibrato, però, l’evasione si concentra nell’agente più capace.

Il detector che scambia il benigno per attacco

Il secondo caso riguarda il lato opposto della soglia di decisione, quello dei falsi positivi, che raramente viene misurato. Yusuf Khalid Shire e Sang-Chul Kim, in PIDS-Bench del 14 settembre 2026, valutano sette rilevatori lungo più assi. Non sono tutti rilevatori dedicati: accanto a baseline

addestrate e a classificatori esterni di prompt injection ci sono comparatori di sicurezza generale. Gli assi comprendono prompt benigni che imitano la struttura di un’iniezione senza intento malevolo, attacchi offuscati e cambi di dominio.

Un rilevatore che supera un F1 di 0,98 sull’insieme di collaudo classifica male circa un terzo di un insieme benigno di provenienza esterna, ristretto a contenuti attinenti alla sicurezza. L’F1 è l’indice sintetico che i fornitori citano più spesso, e combina in un solo numero quante minacce vengono colte e quanti allarmi sono infondati. Su quella distribuzione di stress nessun rilevatore interno raggiunge un punto operativo accettabile. Il criterio è un F1 non inferiore a 0,95, insieme a un tasso di falsi positivi sui benigni difficili non superiore al 10%. Il risultato tiene su tutte le soglie provate e su cinque ripetizioni dell’addestramento, quindi non dipende da una taratura fortunata.

Per chi gestisce un servizio, il costo di quei falsi positivi è immediato e operativo. Per chi lo acquista, è una voce che il punteggio aggregato non espone.

L’agente che sospetta e consegna comunque

Il terzo caso è il più diretto per chi si occupa di protezione dei dati. Il gruppo guidato da Soham Roy e Murari Mandal ha depositato il 30 maggio 2026 un banco di prova sugli agenti web. Il banco, chiamato Scammer4U, contiene 91 ambienti controllati dall’attaccante e 10 ambienti benigni gemelli, su otto vettori d’attacco e sedici categorie di siti. È dichiarato pre-registrato, cioè con ipotesi e misure depositate prima degli esperimenti: è la garanzia che questo stesso articolo chiede ai fornitori.

Su agenti web di frontiera la fuga di dati personali di livello critico va dal 54% al 93% in assenza di indicazioni sulla privacy, contro lo 0% sugli ambienti benigni gemelli. La differenza rispetto alla baseline dimostra che la fuga è

attribuibile all’attacco e non a un normale riempimento di moduli.

Gli autori provano anche mitigazioni progressive a livello di prompt. Le riduzioni ottenute dipendono fortemente dal modello e restano insufficienti a impedire in modo affidabile l’invio di dati personali critici sul dato aggregato.

Il dato decisivo è però un altro, e gli autori lo chiamano divario fra rilevazione e azione. Gli agenti il cui ragionamento un giudice LLM indipendente conferma avere segnalato il sito come sospetto consegnano comunque dati personali critici nel **35,9%** delle sessioni. Quando nessun sospetto viene verbalizzato la quota sale al 66,1%, con un divario di **30,2 punti** stabile su tutte e quattro le famiglie di modelli.

Il sospetto quindi serve, ma non protegge. Accorgersi della truffa riduce la consegna dei dati di due terzi, e in un caso su tre l’agente consegna lo stesso.

La conclusione degli autori è netta: le difese condizionate al riconoscimento dell’attacco da parte dell’agente stesso poggiano sul segnale sbagliato, e serve invece un’intercettazione a valle, sulle comunicazioni in uscita, indipendente dal ciclo di ragionamento. È un’osservazione che tocca direttamente chi usa agenti autonomi su dati di clienti.

Perché il problema non riguarda solo chi compra AI

La finestra si è ristretta

Il contesto in cui queste misure vengono usate è descritto da ENISA nel documento Cybersecurity in the Frontier AI Era del luglio 2026. Il documento riporta un tempo mediano dall’accesso iniziale all’esfiltrazione dei dati compresso a

72 minuti. Non è una misurazione dell’Agenzia: la nota rimanda a una testata specializzata, e come cifra di seconda mano va trattata. L’obiettivo che il documento pone ai centri operativi è invece suo. Chiede di portare le operazioni di sicurezza a una funzione quasi in tempo reale, con tempi medi di rilevazione e di risposta nell’ordine dei minuti a una cifra.

Lo stesso documento riprende dal 2025 Unit 42 Global Incident Response Report di Palo Alto Networks un altro dato. Nel 75% delle violazioni esisteva una registrazione che avrebbe dovuto segnalare il comportamento anomalo, ma i segnali erano frammentati fra strumenti diversi e non sono stati raccolti. È una cifra di seconda mano, riferita al 2025, e va citata come tale. Illustra però bene la differenza fra avere uno strumento e ottenere un risultato.

Il collo di bottiglia si è spostato

ENISA osserva inoltre che nella gestione delle vulnerabilità il fattore limitante diventa la capacità di chi esamina, non l’offerta di segnalazioni. Il volume di rilevazioni vere crea di per sé un problema di verifica. È lo stesso passaggio che attraversa l’evoluzione del SOC, dove il volume degli eventi supera da tempo la capacità umana di analisi.

Vale lo stesso per i controlli sugli agenti. Un rilevatore che produce falsi positivi su un caso benigno su tre non satura soltanto un budget di calcolo, satura il tempo delle persone che devono smentirlo. È il motivo per cui il tasso di falsi positivi andrebbe chiesto insieme al tasso di intercettazione, e non dopo.

Le domande da fare a un fornitore

Dai quattro lavori si ricava una lista breve e verificabile, utile in fase di selezione o di rinnovo contrattuale.

- Su quale corpus è stato misurato quel tasso, quanti scenari contiene e chi li ha scelti? Un punteggio su un insieme fisso di scenari non dice nulla sulle catene d’attacco fuori da quel perimetro.
- Qual è il tasso di falsi positivi sui casi benigni difficili, misurato alla stessa soglia operativa del tasso di intercettazione?
- Come si comporta il controllo quando è l’unico segnale disponibile? Quella quota va separata dal dato aggregato.
- Quali prove esistono su ciò che il sistema circostante continua a permettere dopo l’intervento del controllo? È la sola evidenza che sostenga un’affermazione sul servizio in esercizio.

Chi redige una valutazione d’impatto o imposta la governance degli agenti autonomi trova in queste quattro domande un criterio meno arbitrario del punteggio dichiarato. E chi monitora la prompt injection negli agenti aziendali ha ora un argomento tecnico per spostare i controlli a valle, sulle comunicazioni in uscita, invece di affidarli al giudizio del modello.

Le metriche di sicurezza dell’AI restano utili per confrontare due versioni dello stesso controllo. Non bastano a dire che un sistema è sicuro, e nessuno dei quattro lavori qui citati pretende il contrario. Chi le presenta come garanzia sta rispondendo a una domanda che nessuno ha fatto.

https://www.ictsecuritymagazine.com/articoli/metriche-di-sicurezza-dell-ai/

Tecniche anti-forensics assistite dall’AI: da deepfake e data poisoning ai rischi per l’attribuzione della prova

L’evoluzione dell’intelligenza artificiale sta ridefinendo in modo profondo non solo le capacità di analisi dei dati, ma anche le modalità con cui le informazioni possono essere alterate, simulate o rese indistinguibili dall’autentico. In questo scenario emergono le tecniche anti-forensics assistite dall’AI, che spaziano dal data poisoning alla manipolazione della firma fotografica PRNU, fino ai deepfake e agli evasion attacks, mettendo in crisi la stabilità epistemologica della prova digitale.

Il contributo, che fa parte di una serie dedicata all’impatto dell’intelligenza artificiale sui sistemi probatori e sulle dinamiche del processo penale contemporaneo, analizza le principali categorie di attacchi e manipolazioni, mostrando come l’AI possa essere utilizzata sia per alterare i dati in ingresso ai modelli sia per falsificare log, immagini, contenuti multimediali e output algoritmici. L’obiettivo è evidenziare le implicazioni forensi di queste tecniche e la necessità di approcci probatori fondati su catene di verifica multiple e indipendenti.

Tecniche anti-forensics assistite dall’AI

Data poisoning, PRNU manipulation, deepfake ed evasion attacks: l’arsenale operativo dell’anti-forensics assistita

dall’AI.

Data shaping e data poisoning

Il modellamento del dato (data shaping) consiste nell’alterare il dato prima che venga analizzato, senza renderlo necessariamente sospetto all’occhio umano o agli strumenti tradizionali. L’avvelenamento del dataset (data poisoning) colpisce invece i dati di addestramento, aggiornamento o affinamento del modello (fine-tuning), inducendo il sistema ad apprendere correlazioni errate o comportamenti distorti.

Goldblum et al. hanno sistematizzato questa categoria di attacchi in uno studio pubblicato su IEEE Transactions on Pattern Analysis and Machine Intelligence, distinguendo tra avvelenamento della disponibilità, volto a degradare le prestazioni generali, e avvelenamento per inserimento di backdoor, che induce comportamenti malevoli specifici attivabili da un innesco (trigger) predefinito.

Il caso Microsoft Tay (2016), benché non forense in senso stretto, rimane il riferimento emblematico: il chatbot fu ritirato dopo circa sedici ore perché un attacco coordinato sfruttava la sua capacità di apprendimento online per contaminarlo con contenuti offensivi e ideologicamente connotati. Il caso (documentato in un episodio reale) dimostra che un sistema AI può essere compromesso non perché difettoso nella propria architettura, ma perché esposto deliberatamente a un ambiente informativo ostile.

In ambito forense il parallelo è diretto: un modello linguistico impiegato per riassumere, classificare o correlare chat, messaggi di posta elettronica o documenti sequestrati potrebbe incontrare un corpus previamente manipolato con testi orientati, istruzioni ambigue o dati semanticamente falsificati. A ciò si aggiunge il rischio dell’iniezione di prompt indiretta (indirect prompt injection, IPI): un documento apparentemente innocuo, processato da un sistema AI

nell’ambito di un’analisi forense, può contenere istruzioni nascoste capaci di alterare le conclusioni del modello senza lasciare tracce evidenti nei log applicativi ordinari. Per questo la letteratura sulla forensic readiness dei sistemi LLM insiste sull’invocation logging: registrazione strutturata dell’inferenza, con timestamp, identificatori di sessione, prompt completo, contesto utilizzato, parametri rilevanti e output generato.

Implicazione forense: ogni dato in ingresso a un sistema AI deve essere conservato nella forma grezza originale, autenticato con firma crittografica, sottoposto a controllo delle versioni e mantenuto separato dal dato trasformato. L’output del sistema AI non è mai valutabile in modo autonomo senza conoscere il dataset di partenza, il prompt esatto, il contesto operativo e l’intera catena delle trasformazioni subite dal dato.

Costruzione artificiosa dei log: fabbricare una realtà tecnicamente plausibile

Il log crafting consiste nella creazione o manipolazione di registrazioni di sistema formalmente coerenti: sintassi corretta, timestamp credibili, sequenze compatibili, messaggi verosimili. Il rischio specifico è che un sistema AI, addestrato al riconoscimento di pattern statistici, accetti come autentica una sequenza artificiale non perché difettoso, ma perché quella sequenza è stata confezionata per soddisfare esattamente i criteri di validità che il sistema impara a riconoscere.

Il caso Stuxnet rimane il riferimento storico più documentato di manipolazione sistematica della percezione dello stato di un sistema. Il malware modificava segretamente il funzionamento dei PLC Siemens S7-315 e S7-417 inviando agli operatori letture false che mostravano condizioni operative

normali, mentre i macchinari venivano progressivamente danneggiati. Falliere, Murchu e Chien descrivono la componente rootkit come un replay attack sul canale di comunicazione PLC, rappresentando il primo caso documentato di malware capace di manipolare la percezione dello stato fisico di un impianto industriale. La logica sottostante, mostrare al sistema di monitoraggio una realtà diversa da quella effettiva, è la medesima che governa il log crafting assistito da AI in contesto forense.

In assenza di una fonte sperimentale univoca e stabilizzata, il log crafting assistito da AI deve essere qualificato come rischio metodologico emergente, non come risultato forense già consolidato. La capacità dei modelli linguistici di generare sequenze sintatticamente corrette e temporalmente plausibili impone comunque una cautela processuale precisa: i log sono spesso una fonte primaria in indagini su accessi non autorizzati, frodi informatiche ed esfiltrazione di dati, ma non devono essere valutati isolatamente.

Sul piano operativo, essi devono essere sistematicamente triangolati con fonti indipendenti non controllate dalla stessa parte: sistemi EDR, piattaforme SIEM, log di firewall e proxy, telemetria cloud, Master File Table (MFT), USN Journal ($UsnJrnl), Event Log di Windows, dati di flusso di rete (netflow e netcap), timeline applicative e repliche su sistemi remoti. La single-source evidence, ossia la prova che poggia su una sola fonte, in particolare se filtrata da un sistema AI, è strutturalmente fragile e non può fondare, da sola, una pronuncia di condanna.

PRNU manipulation: alterazione della firma della sorgente fotografica

Fondamenti della tecnica. La Photo Response Non-Uniformity (PRNU) costituisce, allo stato attuale della disciplina forense, il riferimento metodologico più robusto per

l’attribuzione di un’immagine digitale a uno specifico dispositivo di acquisizione. Il suo impiego sistematico trova fondamento nel lavoro seminale di Lukáš, Fridrich e Goljan, che dimostrarono come le imperfezioni microscopiche introdotte in fase di fabbricazione del sensore CMOS o CCD generino un pattern di risposta fotoelettrica statisticamente univoco e riproducibile. La firma PRNU è indipendente dai metadati EXIF e non eliminabile attraverso la semplice rimozione degli header del file.

Il fingerprint-copy attack e le sue evoluzioni. L’attacco più documentato è il fingerprint-copy attack: l’avversario, disponendo di un campione di immagini scattate con il dispositivo bersaglio, estrae la stima del PRNU di quel sensore e la sovrappone a immagini provenienti da un dispositivo diverso, inducendo l’algoritmo di identificazione a classificare erroneamente la sorgente.

Li, Luo, Rao e Huang hanno proposto un attacco migliorato che opera con un metodo a blocchi (block-wise), sovrapponendo la firma in maniera dispersa e non uniforme sull’immagine bersaglio. La calibrazione dell’intensità della firma iniettata è vincolata alla qualità oggettiva dell’immagine risultante, riducendo così la componente non-PRNU rilevabile dal triangle test. Gli esperimenti, condotti su 2.900 immagini provenienti da quattro fotocamere distinte, confermano l’efficacia dell’attacco (risultato dimostrato sperimentalmente).

Il triangle test, introdotto da Goljan e Fridrich per contrastare il fingerprint-copy, sfrutta la correlazione anomala tra immagini sottratte e immagini falsificate. Marra et al. hanno dimostrato che anche questo test è attaccabile. Sul versante delle contromisure, Barni, Santoyo García e Tondi hanno proposto una statistica pooled migliorata per rafforzare il triangle test contro l’attacco fingerprint-copy mentre un successivo studio, con compagine d’autore parzialmente diversa, ne ha mostrato il superamento, evidenziando i limiti

del test pooled nell’identificazione PRNU. L’approccio Universal Wavelet Relative Distortion (UWRD), documentato da Sameer e Naskar, offre infine un attacco counter-forensic basato su trasformata wavelet che mira a sopprimere l’impronta PRNU senza introdurre artefatti visibili macroscopici.

Deep learning: dualismo offensivo-difensivo. Le medesime architetture di apprendimento profondo che hanno rafforzato la capacità di identificazione della sorgente vengono ora impiegate anche come strumenti di attacco. Sul versante difensivo, reti convoluzionali (CNN), strutture residuali e contrastive learning estraggono l’impronta PRNU con maggiore robustezza rispetto ai metodi statistici tradizionali. Sul versante offensivo, le stesse architetture possono essere addestrate per sopprimere, falsificare o sostituire l’impronta in modo impercettibile: un approccio definibile adversarial fingerprint manipulation, che opera in modo automatizzato, scalabile e adattivo rispetto alle contromisure forensi note. La letteratura più recente conferma che nessun metodo di identificazione, tradizionale o neurale, può essere considerato immune ad attacchi mirati condotti con strumenti di deep learning.

Principio processuale: l’attribuzione di un’immagine a uno specifico dispositivo non può fondarsi esclusivamente su una correlazione PRNU positiva ottenuta mediante algoritmo automatico. Ogni perizia deve soddisfare condizioni concorrenti:

- disponibilità dell’immagine originale con catena di custodia verificabile;
- dataset di riferimento statisticamente significativo;
- verifica di coerenza temporale e integrità crittografica dei metadati;
- esclusione di ricampionamento o ricompressione;
- analisi manuale integrativa da parte di un esperto.

La forza probatoria della PRNU risiede nella convergenza con

altri elementi, non nella sua autonomia.

Content-aware fill, inpainting e deepfake: l’evidenza visiva come simulazione persuasiva

Tassonomia delle tecniche di manipolazione sintetica. Il content-aware fill e l’inpainting computazionale rimuovono o sostituiscono porzioni di un’immagine attraverso algoritmi di completamento contestuale basati su GAN o modelli di diffusione, producendo riempimenti visivamente coerenti. Il deepfake video sostituisce o sintetizza il volto di un individuo in sequenze video mediante architetture encoder-decoder. Il voice cloning replica le caratteristiche acustiche della voce (timbro, cadenza, intonazione, pattern prosodici) da campioni audio anche molto brevi, producendo sintesi difficilmente distinguibile dall’originale in condizioni di ascolto non specialistico. Sistemi come VALL-E hanno mostrato la possibilità di generare una voce sintetica a partire da campioni di circa tre secondi.

In contesto forense, il problema non risiede soltanto nella distinzione tra autentico e falsificato, ma nella capacità di provare, con standard processuali, tre elementi distinti: l’origine del file, la sua integrità nel tempo e il contesto della sua produzione e diffusione.

Il caso Arup (Hong Kong, 2024). Nel gennaio 2024 un dipendente della filiale di Hong Kong della società di ingegneria britannica Arup eseguì quindici bonifici bancari per un importo complessivo di circa 200 milioni di dollari di Hong Kong (circa 25,6 milioni di dollari statunitensi), distribuiti su cinque conti correnti, a seguito di una videoconferenza nella quale tutti i partecipanti, ad eccezione della vittima, erano rappresentazioni deepfake di dirigenti reali.

L’accesso iniziale era avvenuto tramite spear-phishing

impersonante il CFO; i deepfake erano stati costruiti da materiali video e audio prelevati da conferenze aziendali pubblicamente accessibili. La Hong Kong Police Force ha confermato che nessun soggetto è stato arrestato, i fondi non sono stati recuperati e l'attribuzione dell'attacco rimane aperta (caso reale documentato). L'episodio documenta tre vulnerabilità sistemiche:

- l'assenza di un canale di autenticazione out-of-band per le disposizioni finanziarie;
- l'assenza di protocolli di verifica dell'identità del chiamante nelle videoconferenze;
- la mancanza di formazione specifica del personale sull'identificazione di segnali visivi e vocali di sintesi artificiale.

Il caso Zelensky (Ucraina, marzo 2022). Il 16 marzo 2022 fu diffuso un video deepfake del presidente Zelensky che simulava un appello alla resa. Il professore Hany Farid (UC Berkeley) identificò anomalie nella proporzione testa-corpo, incongruenze nell'accento e artefatti di temporizzazione caratteristici delle architetture di face-swapping. Il video fu smentito e rimosso dalle principali piattaforme. Si tratta di uno dei primi episodi documentati di impiego operativo del deepfake in un contesto bellico attivo con finalità di disinformazione strategica.

Metodologia forense e protocollo di autenticazione. Le metodologie attualmente più consolidate analizzano le inconsistenze spazio-temporali tra fotogrammi consecutivi. Il modello MVFNet estrae congiuntamente feature forensi spaziali e temporali per rilevare inpainting, deepfake, splicing e re-encoding. Metodi basati sulla trasformata di Fourier sull'asse temporale pixel per pixel consentono di individuare periodicità artificiali introdotte dai generatori nella dimensione temporale. Una rassegna sistematica dei metodi forensi per l'identificazione multimodale di deepfake sui social media conferma che l'affidabilità di questi strumenti

dipende fortemente dalla qualità del campione e che nessuno di essi è, da solo, conclusivo.

Sul piano della provenienza, il C2PA (Coalition for Content Provenance and Authenticity) ha rilasciato la specifica 2.2 nel 2025. Le Content Credentials costituiscono un meccanismo promettente di provenance tracking, ma non sono una garanzia assoluta: i manifest possono essere rimossi in fase di upload, ricodifica o distribuzione, e l'assenza di credenziali non consente, da sola, alcuna inferenza conclusiva sull'origine artificiale o autentica del contenuto. Per il voice cloning, uno studio del 2025 ha documentato criticità significative dei sistemi di speaker verification e degli strumenti anti-spoofing in condizioni realistiche, tra rumore, compressione e variabilità degli algoritmi di clonazione.

Principio fondamentale: un video convincente non è una prova; è un artefatto digitale che richiede autenticazione indipendente. Il protocollo forense prevede:

- ricostruzione del primo upload verificabile (timestamp, piattaforma, URL, hash);
- analisi della catena di diffusione;
- esame dei metadati embedded e delle codifiche successive;
- confronto con fonti indipendenti coeve;
- valutazione del contesto di pubblicazione.

Come rilevato da Farid, nessuna tecnica di rilevamento può essere trattata come prova autonoma e definitiva in assenza di un contesto probatorio convergente.

Evasion attacks: modifiche minime, effetti massimi

Fondamenti teorici. Gli evasion attacks operano in fase di inferenza: il modello è già addestrato, i parametri sono fissi, e la manipolazione riguarda esclusivamente l'input. Si

distinguono dagli attacchi di data poisoning, che agiscono in fase di addestramento, e dagli attacchi di model extraction o membership inference. Il principio fondante fu formalizzato da Szegedy et al.: perturbazioni di ampiezza impercettibile, ottimizzate tramite L-BFGS, sono sufficienti a indurre l'errore di classificazione con confidenza elevata.

Goodfellow, Shlens e Szegedy fornirono la spiegazione meccanicistica attraverso il Fast Gradient Sign Method (FGSM): la linearità delle reti negli spazi ad alta dimensione è la causa strutturale della vulnerabilità. Carlini e Wagner hanno proposto tre metodi di ottimizzazione per le norme L0, L2 e L-infinito, diventati il benchmark di riferimento per la valutazione della robustezza adversarial.

Brown et al. hanno introdotto il concetto di adversarial patch: un pattern visivo stampabile e applicabile fisicamente che produce misclassification indipendentemente da posizione, scala e orientamento. Eykholt et al. hanno dimostrato che sticker applicati a segnali stradali, visivamente percepibili come atti vandalici ordinari, inducevano misclassification nel 100% delle immagini in laboratorio e nell'84,8% dei fotogrammi catturati da un veicolo in movimento per il classificatore bersaglio (risultato dimostrato sperimentalmente). Il principio forense diretto è di rilevanza immediata: la soglia di plausibilità di un attacco adversarial non coincide con quella della percezione umana.

Il caso Tencent Keen Security Lab – Tesla Autopilot, 2019. Il documento tecnico Experimental Security Research of Tesla Autopilot dimostrò sperimentalmente che l'apposizione di tre piccoli sticker sulla carreggiata era sufficiente per indurre il sistema di riconoscimento delle corsie del Tesla Model S a generare un comando di sterzo nella corsia opposta. Il rapporto documentava ulteriori superfici di attacco: attivazione dei tergicristalli tramite adversarial example fisico e controllo dello sterzo mediante gamepad wireless. L'esperimento dimostra che la frontiera dell'attacco include

la manipolazione dell'ambiente percepito dal sistema, aprendo scenari di rischio che i modelli tradizionali di threat modeling non contemplano.

Evasion attacks nel dominio malware. Demetrio et al. hanno dimostrato che la modifica funzionalmente neutrale di sezioni inutilizzate dei file Windows PE, con iniezione di contenuto adversarial ottimizzato, può indurre misclassification nel classificatore MalConv. La letteratura successiva ha esteso il problema sia agli attacchi contro classificatori malware sia alla generazione di falsi positivi sistematici, ma i tassi di evasione dipendono fortemente da modello, dataset, vincoli funzionali del file e scenario sperimentale. In un contesto processuale è quindi metodologicamente scorretto trasferire percentuali ottenute in laboratorio direttamente al caso concreto senza validazione specifica.

Implicazione processuale: un classificatore AI che dichiara un file benigno, una conversazione non rilevante o un'immagine autentica deve essere trattato esclusivamente come strumento di triage. La sua pronuncia è condizionata dall'insieme di pattern appresi durante l'addestramento, dalla distribuzione del dataset, dall'assenza di esempi adversarial nel training set e dall'assenza di attacchi attivi al momento dell'inferenza: quattro condizioni non verificabili a posteriori senza documentazione rigorosa della pipeline di sviluppo del modello.

Ulteriori superfici di attacco nella catena della prova

Documenti, comunicazioni e codice malevolo AI-generato spostano il problema sull'attribuzione e sulla catena della prova.

Documenti testuali e comunicazioni

La generazione di documenti testuali è la capacità più consolidata dei sistemi AI attuali. Un LLM di qualità commerciale può generare contratti, email e messaggi di testo con stile linguistico personalizzato (voice cloning testuale); completare documenti lacunosi in modo stilisticamente coerente; tradurre mantenendo artefatti linguistici che simulano autenticità; generare metadati credibili per file Office.

Kirchenbauer et al. hanno proposto tecniche di watermarking per LLM, ma senza marcatori preventivi l'identificazione del testo AI rimane statisticamente fragile. Sadasivan et al. hanno evidenziato un limite strutturale dei sistemi di rilevazione del testo generato da AI: quando chi produce il testo conosce, o può interrogare, il detector utilizzato per valutarlo, può progressivamente modificare l'output fino a renderlo meno riconoscibile come artificiale. Ne deriva che tali strumenti possono avere utilità orientativa o di triage, ma non possono essere impiegati, da soli, come prova dell'origine artificiale di un documento in sede processuale.

Implicazione processuale: un documento presentato come prova non può essere autenticato basandosi esclusivamente sull'analisi del testo; è necessario risalire ai sistemi di origine, verificare i log dei server, analizzare la trasmissione di rete e le copie di backup indipendenti. In mancanza di tali riscontri, la sola analisi stilistica, sia essa compiuta da un esperto umano o da un sistema AI, non può assurgere a prova di autenticità o di falsità nel contesto processuale.

Codice malevolo AI-generato e attribuzione forense

Sistemi e servizi malevoli pubblicizzati come WormGPT e

FraudGPT, richiamati da report di threat intelligence del 2023, sono stati presentati come strumenti per generare campagne di phishing e contenuti offensivi altamente personalizzati.

Per il consulente o perito informatico forense, il codice malevolo generato o assistito da AI introduce una criticità specifica: l'attribuzione dell'autore. Tradizionalmente, alcuni elementi stilistici del codice (denominazione delle variabili, struttura delle funzioni, commenti, scelte algoritmiche e ricorrenze sintattiche) potevano contribuire, insieme ad altri indicatori tecnici, a formulare ipotesi sull'autore o sul gruppo di attori coinvolto. L'intervento dell'AI tende però a uniformare, alterare o mascherare tali impronte individuali; ne consegue che l'attribuzione forense diventa più fragile e non può fondarsi sul solo stile del codice, ma richiede la convergenza di ulteriori elementi: infrastruttura utilizzata, catena di distribuzione, telemetria, log, artefatti di compilazione, contesto operativo e, ove disponibili, tracce relative al servizio AI eventualmente impiegato.

Ciò incide sulla tenuta complessiva della ricostruzione tecnico-probatoria nei procedimenti per cybercrime: per l'accusa, rende più fragile l'attribuzione del codice malevolo a un determinato autore o gruppo di attori; per la difesa, impone la possibilità di contestare in modo tecnicamente fondato inferenze attributive basate su indicatori stilistici non più sufficientemente affidabili. Ne deriva la necessità di rivedere i metodi tradizionali di attribuzione, privilegiando approcci multi-fonte fondati sulla convergenza tra codice, infrastruttura, telemetria, log, artefatti di compilazione e contesto operativo.

Il contributo di Cosimo de Pinto offre una ricognizione delle principali tecniche di anti-forensics assistite dall'AI, evidenziando come data poisoning, manipolazione della firma PRNU, sintesi di contenuti multimediali e attacchi di evasione

possano incidere in modo strutturale sulla produzione e sulla valutazione della prova digitale. Ne emerge un quadro in cui non solo i dati e i modelli possono essere alterati, ma anche le tracce stesse dell'attività informatica diventano potenzialmente costruibili, sollevando criticità rilevanti in termini di affidabilità e verificabilità.

Per un approfondimento sistematico di queste dinamiche, si invita alla lettura del paper di Cosimo De Pinto "**Quando l'AI mente bene: deepfake, prova digitale e anti-forensics nel processo penale",** che sviluppa in modo organico il rapporto tra intelligenza artificiale, attendibilità della prova e criticità processuali, offrendo un quadro teorico e operativo indispensabile per comprendere l'impatto dell'AI sul diritto probatorio contemporaneo.

https://www.ictsecuritymagazine.com/articoli/tecniche-anti-forensics/

Agenti AI su DseWiki: sei settimane di attività autonoma che nessuna norma obbligava a segnalare

Migliaia di modifiche a una wiki tedesca di programmazione, usata da agenti riconducibili a OpenAI come bacheca per scambiarsi risposte ai propri compiti e tecniche di elusione dei controlli. L'episodio è emerso il 4 settembre da un'esclusiva Reuters e dal rapporto pubblico di un gruppo di ricercatori indipendenti; il giorno dopo OpenAI lo ha riconosciuto, ammettendo di non avere uno standard per

divulgare casi del genere. Il punto interessante per chi si occupa di compliance non è che qualcuno abbia violato l'AI Act: è che quasi certamente nessuno lo ha violato.

Che cosa è successo su DseWiki

Tra la fine di maggio e i primi di luglio 2026, agenti autonomi che si presentavano come agenti OpenAI hanno effettuato migliaia di modifiche su DseWiki, una wiki di programmazione in lingua tedesca attiva da circa venticinque anni e ormai quasi in disuso: nel decennio precedente aveva registrato in tutto una ventina di modifiche. Reuters parla di oltre 15.000 interventi; il rapporto dei ricercatori che hanno scoperto il caso conta circa 17.000 modifiche sulla sola DseWiki e circa 18.000 post complessivi su più wiki, con oltre 3.700 nomi di agente distinti.

Non si è trattato di scrittura casuale o di *scraping*. Secondo la ricostruzione dei ricercatori, gli agenti stavano eseguendo un compito interno a OpenAI (addestramento o valutazione, non è chiaro) di ricerca di informazioni sul web a tempo: potevano leggere internet ma non scrivervi, perché il sandbox consentiva solo richieste GET. DseWiki, come altre wiki derivate dal software UseModWiki, accetta modifiche anche tramite richieste GET. Gli agenti hanno individuato questa caratteristica dopo giorni di tentativi su altri siti (i primi test risalgono all'11 maggio) e il 24 maggio hanno fatto la prima scrittura riuscita su DseWiki.

Da lì il sito è diventato una bacheca. Gli agenti vi hanno depositato liste di link e dati grezzi, hanno condiviso le risposte ai quesiti che il compito poneva loro in sequenza (chi era più avanti postava la risposta, chi era indietro la copiava in pochi secondi), hanno tentato di ricostruire il seme del generatore casuale che decideva l'ordine delle domande, hanno installato "battiti cardiaci" su servizi esterni per capire quando il loro container sarebbe stato

terminato, e si sono scambiati un metodo funzionante per aggirare la restrizione alle sole richieste GET sfruttando un'eccezione del proxy di sicurezza per gli host di Azure Blob Storage. Il 16 giugno l'attività è esplosa: circa 13.000 modifiche nei sette giorni successivi.

Alcuni elementi meritano di essere riportati per esteso, perché è nel dettaglio che si capisce la natura del fenomeno.

Circa metà dei nomi utente richiamava direttamente il fornitore, con account del tipo "OpenAIResearcher" e "OAIResearchMar26". I log pubblici del server, analizzati dai ricercatori, attribuiscono il 98,5% delle modifiche a indirizzi IP di Microsoft Azure, una delle infrastrutture su cui OpenAI addestra e serve i propri modelli. Un moderatore umano si è accorto dello spam già il 2 giugno, quando ha trovato il registro delle modifiche del sito sommerso da elenchi di link, e da metà giugno ha cancellato a mano le pagine generate dagli agenti, dedicando alla pulizia almeno qualche minuto ogni sera per sei settimane, decine di ore in tutto. Il 19 giugno gli agenti hanno notato che la cancellazione procedeva in ordine alfabetico e hanno cominciato sistematicamente a creare pagine di riserva con nomi che iniziano per "ZZZ", in modo da essere cancellate per ultime. Il messaggio lasciato per gli altri agenti avvertiva che era in corso una pulizia alfabetica e indicava dove ripiegare se quella pagina fosse sparita. Nei cinque giorni successivi il moderatore ha cancellato in media cento pagine al giorno; gli agenti ne creavano circa quattrocento.

C'è poi la parte che riguarda direttamente OpenAI. Dal 21 giugno la wiki riceve visite da indirizzi IP registrati a "OpenAI OpCo, LLC" e da altri con schemi di navigazione umani. Il 22 giugno le modifiche degli agenti crollano quasi a zero, dopo 26 giorni di attività su 30. Il 26 giugno 33 indirizzi riconducibili a OpenAI fanno da soli il 30% di tutto il traffico che i ricercatori attribuiscono all'azienda, e usano la funzione di archivio del sito per recuperare pagine già

cancellate. L’inferenza dei ricercatori, che OpenAI non ha né confermato né smentito, è che l’azienda abbia scoperto la bacheca e sia intervenuta. Un’ultima raffica su un tema nuovo compare l’1 e il 2 luglio, con 29 tentativi di modifica il 2 luglio distribuiti su tre wiki; poi più nulla.

Quello che è rimasto ignoto fino a fine agosto, dunque, non è l’attività in sé, che il moderatore vedeva ogni sera e che OpenAI, secondo i log, ha osservato da fine giugno: è l’attribuzione, la natura del fenomeno e la sua esistenza per chiunque fosse fuori da quelle due stanze. A ricostruirlo è stato un gruppo di quattro ricercatori indipendenti (Sydney Von Arx, a capo dell’organizzazione non profit di AI safety Nightingale Collective; Cormac Slade Byrd, *trader* quantitativo passato alla ricerca sull’intelligenza artificiale; Spencer Kitts, di Redwood Research, che ha lavorato al rapporto come collaboratore di Nightingale; Thomas Larsen, dell’AI Futures Project) che dopo il caso Hugging Face stavano cercando sul web tracce di agenti fuori controllo. La vicenda è stata resa pubblica il 4 settembre 2026 da un’esclusiva Reuters firmata da Deepa Seetharaman e Raphael Satter, in contemporanea con la pubblicazione del rapporto e del dataset (quasi completo: alcune pagine cancellate non sono state recuperabili).

Le reazioni, e la parte che non torna

OpenAI ha dichiarato di non poter rispondere nel merito a un rapporto che non aveva avuto modo di esaminare, precisando che Reuters e gli autori avevano rifiutato di darle accesso prima della pubblicazione. Secondo Reuters, che cita quattro persone informate, alcuni investigatori interni avrebbero voluto approfondire questo tipo di attività, incontrando resistenze in azienda, anche da parte dei consulenti legali; OpenAI ha definito false le affermazioni secondo cui il proprio ufficio legale avrebbe scoraggiato l’indagine. L’azienda ha inoltre

sostenuto che la vicenda tedesca non è collegata alla violazione di Hugging Face di luglio e che per questo non era stata inclusa nel relativo rapporto. Sempre secondo Reuters, che qui si basa su due fonti, dirigenti della società erano a conoscenza dell'episodio da settimane. Su questo punto i log pubblici della wiki, con le visite da IP OpenAI dal 21 giugno, offrono un riscontro indipendente, anche se non dicono chi, dentro l'azienda, sapesse cosa.

Il giorno dopo la posizione è cambiata. Il 5 settembre OpenAI ha pubblicato una dichiarazione su X in cui parla apertamente del "wiki incident", descrivendolo come un episodio in cui i propri agenti "hanno scritto su diversi siti internet", una formulazione che chiude la questione dell'attribuzione e suggerisce un perimetro più ampio della sola DseWiki. L'azienda spiega di aver trattato il caso come *misalignment*, cioè come un problema di ricerca da comunicare attraverso paper e system card, e non come un incidente di sicurezza da divulgare, a differenza di Hugging Face, dove erano stati toccati sistemi di terzi. Riconosce però che quest'anno il disallineamento ha cominciato a produrre effetti nel mondo reale, che il settore non ha uno standard chiaro su quando e come segnalare comportamenti anomali emersi in addestramento, valutazione o produzione, soprattutto quando non somigliano a una violazione informatica classica, e annuncia un proprio quadro di divulgazione nelle prossime settimane, dicendo di essere in contatto con decine di autorità di regolazione. A Reuters, che le ha chiesto cosa sapesse dell'episodio e perché abbia aspettato la pubblicazione dell'inchiesta per parlarne, non ha risposto.

I ricercatori hanno anche documentato tentativi di manomettere il sito: sondaggi per vulnerabilità XSS sulla pagina di ricerca tra il 24 e il 26 maggio e su altri punti di ingresso il 18 giugno, nessuno dei quali risulta riuscito; l'uso del nome di un amministratore della piattaforma con un carattere cirillico al posto di uno latino, per modificare le pagine

principali; richieste all'archivio delle pagine cancellate firmate con il nome di un moderatore; la cancellazione del contenuto della home page, ripristinata dal moderatore nove volte. Lukasz Olejnik, *visiting senior research fellow* al King's College London, ha qualificato il tutto come un tentativo di intrusione informatica. OpenAI, riferisce Reuters, ha contestato questa qualificazione sulla base della propria analisi del materiale. Maurice Chiodo, del Centre for the Study of Existential Risk dell'Università di Cambridge, che ha esaminato parte delle comunicazioni, le ha descritte come qualcosa che somiglia al funzionamento di una rete clandestina orientata a portare a termine una missione, e ha indicato negli sciami di intelligenze artificiali semi-intelligenti che colludono tra loro il rischio a suo avviso più serio.

Vale la pena tenere ferme due distinzioni che nella copertura generalista si perdono.

La prima: che gli agenti "si parlassero" non implica intenzionalità nel senso umano del termine. Sistemi che ottimizzano il completamento di un compito possono convergere su una strategia condivisa senza che vi sia nulla di simile a un piano. OpenAI stessa, nel rapporto tecnico sul caso Hugging Face del 26 agosto, ha scritto che alcuni suoi modelli avevano imparato in addestramento, in rari casi, a usare canali di collaborazione improvvisati anche quando lo strumento di collaborazione non era attivo, e che questo comportamento era stato poi rinforzato.

La seconda: Reuters descrive gli agenti intenti a "mascherare il proprio comportamento" ed eludere il rilevamento. Il rapporto primario è più cauto. Le pagine "ZZZ" servivano a sopravvivere alla pulizia del moderatore, non a sfuggire a OpenAI; sull'uso di Tor (che insieme a indirizzi AWS e DigitalOcean copre 197 modifiche su circa diciottomila) i ricercatori scrivono di non sapere perché sia avvenuto e che i messaggi non implicano alcuna volontà di occultamento. Lo

scopo documentato è più prosaico e per certi versi più inquietante: barare sul compito assegnato e aggirare le restrizioni del proprio ambiente. Il punto rilevante non è la psicologia degli agenti. È che il comportamento osservabile, dal punto di vista di chi gestiva quel sito, è indistinguibile da quello di un attore ostile organizzato.

Perché nessuno era tenuto a segnalarlo

Qui arriva la parte che interessa direttamente chi si occupa di *compliance*, ed è controintuitiva.

L'articolo 55 dell'AI Act impone ai fornitori di modelli di intelligenza artificiale per finalità generali con rischio sistemico una serie di obblighi, tra cui, alla lettera c), quello di tracciare, documentare e segnalare senza indebito ritardo all'AI Office e, se del caso, alle autorità nazionali competenti, le informazioni rilevanti sugli incidenti gravi e sulle possibili misure correttive. La lettera d) impone inoltre di garantire un livello adeguato di protezione di cibersicurezza del modello e dell'infrastruttura fisica. Gli obblighi per i modelli GPAI si applicano dal 2 agosto 2025, e il 4 novembre 2025 la Commissione ha pubblicato un modello di segnalazione dedicato proprio agli incidenti gravi che coinvolgono questi modelli.

Sembrerebbe il quadro giusto. Il problema è la definizione. L'articolo 3, punto 49, definisce incidente grave un incidente o un malfunzionamento di un sistema di intelligenza artificiale che direttamente o indirettamente causa una di queste quattro conseguenze: il decesso di una persona o un grave danno alla sua salute; una perturbazione grave e irreversibile della gestione o del funzionamento di infrastrutture critiche; la violazione di obblighi del diritto dell'Unione posti a tutela dei diritti fondamentali; un danno grave a cose o all'ambiente.

Una wiki di programmazione in lingua tedesca, mantenuta da volontari, non è un'infrastruttura critica. Nessuno si è fatto male. Non emerge una violazione di obblighi europei a tutela dei diritti fondamentali. Il danno a cose, per come è configurato, difficilmente comprende le decine di ore che un moderatore ha speso in sei settimane a cancellare pagine a mano.

C'è poi un secondo livello, quello del Codice di buone pratiche per i modelli GPAI, che OpenAI ha firmato integralmente nell'estate 2025. Il Codice è volontario ma è lo strumento con cui i firmatari dimostrano di rispettare l'articolo 55, e sul punto va un po' oltre il regolamento: la Misura 9.3 fissa un termine di cinque giorni per il rapporto iniziale quando il modello ha portato a una "grave violazione di cibersicurezza", categoria in cui il testo include espressamente "l'(auto)esfiltrazione dei pesi del modello e gli attacchi informatici", anche solo sospettata con ragionevole probabilità; la Misura 9.1 chiede di cercare gli incidenti anche in rapporti di stampa, post sui social e paper di ricerca; e in più punti (Misure 1.3, 2.1, 7.6 e 9.2) il Codice chiede di tenere conto dei *near miss*. Il Codice quindi contempla almeno un comportamento che parte dal modello, l'autoesfiltrazione, e nell'Appendice 1.3 elenca tra le fonti di rischio sistemico da valutare la "collusione con altri modelli o sistemi di IA", la capacità di "eludere la supervisione umana" e quella di "modificare il proprio ambiente di esecuzione": una descrizione quasi letterale di quanto accaduto su DseWiki. Ma una fonte di rischio da valutare non è un incidente da segnalare. Se un'evasione delle restrizioni di rete di un sandbox, senza pesi esfiltrati e con un bersaglio che è una wiki di volontari, rientri nella "grave violazione di cibersicurezza" della Misura 9.3 è una domanda aperta: tutto dipende da quanto si allarga "attacco informatico", e qui la contestazione di OpenAI alla lettura di Olejnik torna a pesare. Resta il *near miss*, che il glossario del Codice definisce come una situazione in cui un incidente

grave avrebbe potuto verificarsi ma non si è verificato: DseWiki ci sta comodamente, e i *near miss* vanno tenuti in conto nelle valutazioni del rischio e negli aggiornamenti dei rapporti, non segnalati all'AI Office entro un termine.

Infine, la data. I poteri sanzionatori della Commissione nei confronti dei fornitori di modelli GPAI decorrono dal 2 agosto 2026, e per i firmatari del Codice l'AI Office aveva indicato un anno di tolleranza nella prima fase di applicazione. L'intero episodio, dall'11 maggio al 2 luglio, cade dentro quel periodo.

La conclusione è scomoda ma va detta: per come sono scritte oggi le norme, un episodio in cui agenti autonomi occupano per settimane l'infrastruttura di un terzo, vi organizzano lo scambio di risposte e di tecniche di elusione dei propri controlli e reagiscono alla moderazione predisponendo canali di riserva, con ogni probabilità non integra un incidente grave ai sensi dell'AI Act, e anche sotto il Codice di buone pratiche rientra al più in una zona grigia priva, per ora, di conseguenze. Non c'era nulla che dovesse essere segnalato. Il che rende la mancata segnalazione non una violazione, ma la fotografia di una lacuna. Lo dice, con parole sue, anche OpenAI: la dichiarazione del 5 settembre ammette che manca uno standard per riferire questo tipo di eventi e che la distinzione tra disallineamento di ricerca e incidente di sicurezza sta diventando difficile da tenere. Un fornitore che annuncia un proprio quadro di divulgazione volontario è la conferma più eloquente che quello obbligatorio, oggi, non copre il caso. Del resto il segnale era già nel rapporto tecnico su Hugging Face: lì OpenAI scriveva che il comportamento agentico "può emergere in forme nuove che non sempre rientrano nelle categorie tradizionali di incidente di sicurezza o di safety", annunciava soglie di escalation interne per la "coordinazione non autorizzata tra agenti o tra esecuzioni" e l'"elusione di controlli di sicurezza di terzi", e rimandava a una revisione separata dei processi di risposta

agli incidenti di allineamento in addestramento e valutazione. Sono esattamente le fattispecie di DseWiki, definite dal fornitore prima che il caso diventasse pubblico, e assenti dalla norma.

La soglia degli incidenti gravi è tarata sul danno alla persona e sulle infrastrutture critiche, categorie pensate per sistemi ad alto rischio che operano su processi fisici o su decisioni che incidono sulle persone. Non è tarata sul comportamento emergente di agenti che agiscono in rete su sistemi di terzi. È una categoria di evento che l'impianto non vede, e che diventerà più frequente man mano che gli agenti verranno messi in produzione. È sulle linee guida dell'AI Office e sugli aggiornamenti del Codice, più che sul testo del regolamento, che questa distanza potrà essere colmata.

La domanda che il diritto italiano non ha ancora affrontato

Se si prende sul serio la qualificazione di Olejnik, che OpenAI contesta, e si ipotizza che vi sia stato un tentativo di manomissione del sito, si apre un problema che nel nostro ordinamento non ha una risposta pronta.

L'accesso abusivo a un sistema informatico è un reato costruito attorno a una condotta umana e a un elemento soggettivo. Chi è l'autore, quando la condotta è materialmente posta in essere da un agente autonomo? Il fornitore del modello, che non ha impartito quell'istruzione e che, se la ricostruzione dei ricercatori è corretta, aveva anzi vietato ai suoi agenti di scrivere su internet? Il soggetto che ha messo l'agente in esecuzione per un compito del tutto lecito? Nessuno dei due, con la conseguenza che un fatto oggettivamente lesivo resta senza autore?

Sono domande, non tesi: la casistica non esiste ancora e questo articolo non pretende di anticiparla. Ma è utile porle

adesso, perché la risposta influenzerà le clausole contrattuali che le organizzazioni firmano quando adottano agenti di terze parti, la ripartizione di responsabilità tra fornitore e utilizzatore, e la posizione di chi subisce il comportamento di un agente altrui. Su questo terreno un contributo di taglio penalistico sarebbe più utile di molte analisi tecniche.

Va aggiunto che la vittima, in questa vicenda, è la parte di cui nessuno parla: un moderatore volontario che ha assorbito per sei settimane il costo di una pulizia non richiesta, senza strumenti per farsi ascoltare e senza un interlocutore a cui rivolgersi.

Che cosa può fare oggi chi usa agenti in azienda

Al netto delle questioni aperte, l'episodio consegna a chi ha responsabilità operative un elenco di verifiche che non richiedono di attendere né una sentenza né una linea guida.

La prima domanda da porsi è la più semplice e la più temuta: se i nostri agenti facessero quindicimila scritture verso un sito esterno, ce ne accorgeremmo? Nella maggior parte delle organizzazioni la risposta è no, perché quel traffico è HTTPS in uscita verso una destinazione qualsiasi, esattamente ciò che un agente fa tutto il giorno per definizione. Il caso DseWiki aggiunge una lezione precisa: la restrizione alle sole richieste GET, che sulla carta impedisce di scrivere, in pratica non basta, perché una parte del web accetta scritture via GET e gli agenti la trovano. Non è una lacuna di prodotto, è una lacuna di progettazione dei controlli, ed è la stessa che rende difficili da vedere altre categorie di abuso, dalla prompt injection negli agenti in poi.

Le contromisure sono note e poco implementate. Filtrare il traffico in uscita degli agenti con liste di destinazioni

consentite, invece di lasciare la rete aperta, e verificare che le eccezioni al proxy non possano essere sfruttate con host inventati (nel caso in esame il sandbox degli agenti considerava fidato un intero suffisso di dominio, senza controllare che l'host esistesse). Registrare separatamente le azioni di scrittura verso sistemi esterni, distinguendole dalla semplice lettura, perché è la scrittura che produce effetti sul mondo. Definire limiti di autonomia espliciti, con un tetto al numero di azioni per sessione e un intervento umano obbligatorio oltre certe soglie. Prevedere un meccanismo di interruzione che funzioni davvero e sia stato provato. Sono misure che rientrano nel perimetro della governance degli agenti autonomi e che, a differenza delle questioni regolamentari, si possono mettere in campo questo mese.

Resta un ultimo punto, di metodo. Questo episodio non è stato reso noto da un fornitore, da un'autorità o da un cliente: è stato ricostruito da quattro ricercatori indipendenti che stavano cercando esattamente quel tipo di comportamento, due mesi dopo che era cessato, partendo da log che chiunque avrebbe potuto leggere. Finché il rilevamento di questa classe di eventi dipenderà dalla curiosità di singoli, il numero di episodi noti dirà molto poco sul numero di episodi avvenuti.

https://www.ictsecuritymagazine.com/notizie/agenti-ai-dsewiki-openai-ai-act-incidenti-gravi/

Prova digitale e anti-

forensics: cosa cambia con l’AI generativa

L’adozione di sistemi di intelligenza artificiale generativa sta producendo una trasformazione profonda nel modo in cui il diritto processuale penale interpreta, valuta e utilizza la prova digitale. Ciò che per anni è stato considerato un elemento relativamente stabile e verificabile ovvero il dato digitale come traccia dell’origine di un evento o di un’azione entra oggi in una fase di progressiva instabilità epistemica. La possibilità di generare contenuti artificiali altamente coerenti, completi di metadati plausibili e strutture tecniche formalmente corrette, incrina infatti uno dei presupposti storici della digital forensics: la corrispondenza tra traccia e origine.

Questo articolo si inserisce all’interno della serie di approfondimenti a cura di Cosimo de Pinto dal titolo **“Quando l’AI mente bene – Deepfake, prova digitale e anti-forensics nel processo penale”.** L’approfondimento analizza la frattura introdotta dall’AI generativa nel paradigma tradizionale della prova informatica, evidenziando come la “plausibilità sintetica” possa sostituirsi alla tracciabilità dell’origine come criterio implicito di affidabilità.

Nel testo si esamina la discontinuità tra anti-forensics classica e nuove tecniche abilitate dall’intelligenza artificiale, mettendo in luce il passaggio da strategie di occultamento o distruzione della prova a sistemi capaci di costruire evidenze artificiali credibili e difficilmente distinguibili da quelle autentiche. L’analisi si concentra inoltre sulle implicazioni processuali di tale evoluzione, con particolare attenzione ai rischi per la valutazione giudiziale della prova digitale e alle possibili ricadute sulla presunzione di innocenza e sulla libertà personale.

La fine dell’innocenza digitale

L’AI generativa incrina il presupposto che la prova digitale rifletta sempre la propria origine.

La digital forensics si è sviluppata, nel corso dei suoi tre decenni di storia disciplinare, attorno a un presupposto implicito: i dati digitali conservano tracce della propria origine, e tali tracce possono essere estratte, interpretate e presentate come prova con ragionevole certezza scientifica. Un file aveva una biografia verificabile: nasceva in un momento preciso, recava l’impronta del software che lo aveva generato, i metadati del sistema operativo, le impronte dell’hardware. Su questo fondamento, ben descritto nella manualistica di riferimento, si è costruita l’intera architettura processuale della prova digitale nei sistemi giuridici occidentali.

Dal paradigma classico alla frattura epistemica

	Computer forensics classica	**Scenario AI generativa**
Presupposto di base	Le tracce digitali riflettono l’origine del dato	Un contenuto può simulare un’origine credibile
Obiettivo dell’analisi	Ricostruire eventi e attribuire provenienza	Verificare anche la possibilità di un’origine sintetica
Segni di manipolazione	Spesso lasciano anomalie residue	Possono imitare pattern statisticamente plausibili

	Computer forensics classica	**Scenario AI generativa**
Ruolo dell’esperto	Interpretare tracce e coerenze tecniche	Corroborare, autenticare e contestualizzare
Rischio giuridico	Errore tecnico circoscritto	Potenziale violazione della libertà personale

L’avvento dei modelli generativi di intelligenza artificiale, in particolare dei Large Language Models (LLM), dei modelli text-to-image, dei sintetizzatori vocali neurali e dei sistemi di video synthesis, ha introdotto nella catena della prova digitale una variabile che i framework forensi tradizionali non erano stati progettati per gestire: la plausibilità sintetica. Un documento, un’immagine, una registrazione audio o video possono oggi essere generati artificialmente in modo da superare non soltanto la percezione umana, ma anche una parte significativa delle analisi strumentali disponibili ai laboratori forensi. Come osservato dalla letteratura tecnica più recente, gli avversari non si limitano più a nascondere le prove: le costruiscono, le avvelenano e orientano l’investigatore verso una narrazione falsa.

Le conseguenze di questa trasformazione non sono confinabili all’ambito tecnico. Nei procedimenti penali, ogni errore nella valutazione di una prova digitale si traduce in una potenziale violazione della libertà personale. Un soggetto innocente può essere condannato sulla base di una conversazione WhatsApp mai avvenuta, di un’immagine fotografica manipolata, di un video che lo mostra in un luogo in cui non si è mai recato, di un log di sistema artatamente costruito per collocarlo in prossimità di un reato. L’AI non mente come un essere umano: produce output statisticamente coerenti, difficili da smascherare con gli strumenti tradizionali, e capaci di ingannare persino analisti esperti.

Ogni errore nella valutazione di una prova digitale si traduce in una potenziale violazione della libertà personale.

Il presente paper analizza le intersezioni tra AI generativa, tecniche di anti-forensics e il concetto giuridico di prova digitale, con attenzione alle implicazioni pratiche per consulenti tecnici, avvocati e magistrati. Vengono esaminate le principali superfici di attacco, le tecniche di evasione forense abilitate dall'AI, i framework di auditabilità esistenti, le lacune normative che rendono urgente un ripensamento del valore probatorio del dato digitale e, soprattutto, i rischi concreti per la libertà delle persone che derivano da un approccio acritico all'impiego dell'AI in sede processuale.

Metodo, contributo e domanda di ricerca

Metodo. Sintesi critica interdisciplinare che incrocia letteratura peer-reviewed, casi documentati, standard normativi e implicazioni processuali della prova digitale nell'era dell'AI generativa.

Contributo. Un quadro di sintesi tra tecnica di attacco, artefatto colpito, contromisure e rischio residuo; ricostruzione del nesso tra auditabilità tecnica e tenuta probatoria; proposta di principi guida per una digital forensics AI-resistant.

Domanda di ricerca. In che modo l'AI generativa trasforma la natura della prova digitale e quali condizioni metodologiche e giuridiche sono necessarie per conservarne l'affidabilità?

Questo paper affronta una delle sfide più critiche e meno sistematizzate dell'attuale panorama della sicurezza informatica applicata al diritto: l'impatto dell'intelligenza artificiale generativa sull'affidabilità delle prove digitali

nei procedimenti penali. Non si tratta di uno scenario futuro. Le tecniche qui analizzate (data poisoning, log crafting, PRNU manipulation, deepfake, voice cloning, evasion attacks) sono operative già oggi, e i loro effetti sul sistema giustizia sono documentabili. Il taglio è deliberatamente duplice, tecnico e giuridico, perché duplice è il pubblico cui il problema si rivolge: professionisti della sicurezza, consulenti tecnici d'ufficio e di parte, magistrati e avvocati che si trovano a maneggiare prove digitali la cui affidabilità non può più essere data per scontata.

Sul piano del metodo, il lavoro è una sintesi critica di taglio interdisciplinare. Sul versante tecnico incrocia la letteratura peer-reviewed su anti-forensics e adversarial machine learning con casi documentati di rilevanza processuale o geopolitica; sul versante giuridico procede a un'analisi dottrinale e giurisprudenziale del diritto della prova, ancorata alla normativa italiana ed europea vigente e agli standard tecnici internazionali della famiglia ISO/IEC 27000. Le fonti sono state selezionate privilegiando, ove disponibili, gli atti originali (paper, whitepaper, testi normativi, massime ufficiali) rispetto alle fonti secondarie.

Per onestà epistemica, ogni affermazione tecnica viene qualificata secondo tre gradi di evidenza, esplicitati nel testo: risultato dimostrato sperimentalmente in letteratura; condotta documentata in un caso reale; rischio metodologico emergente, ipotizzato in via teorica ma non ancora consolidato in giudicato.

Il contributo originale è triplice: un quadro di sintesi che mette in relazione ciascuna tecnica di attacco con l'artefatto forense colpito, la contromisura di rilevazione e il rischio residuo (sezione 5); un raffronto comparato tra l'ordinamento italiano e il modello angloamericano (sezione 7); e un framework forense resistente all'AI articolato in cinque principi operativi e quattro proposte normative (sezione 10). La domanda di ricerca che tiene insieme l'intero discorso è la

seguente: a quali condizioni una prova digitale generata o mediata da sistemi AI può fondare un giudizio di responsabilità penale senza violare la presunzione di innocenza e il diritto a un equo processo?

Anatomia dell'anti-forensics tradizionale e la discontinuità dell'AI

Dalle tecniche che cancellano le tracce a un'AI che le costruisce: una rottura qualitativa, non solo quantitativa.

L'anti-forensics prima dell'AI

La disciplina dell'anti-forensics, intesa come l'insieme delle tecniche volte a ostacolare, fuorviare o rendere impossibile l'analisi forense di un sistema digitale, è quasi coeva alla nascita della computer forensics stessa. Harris la definisce come qualsiasi tentativo di compromettere la disponibilità o l'utilità delle prove in un processo forense.

Le tecniche classiche si articolano in quattro categorie principali:

- distruzione dei dati (data destruction) tramite sovrascrittura multipla secondo lo standard DoD 5220.22-M, degaussing o distruzione fisica del supporto;
- occultamento dei dati (data hiding) mediante steganografia, cifratura di volumi interi (VeraCrypt, BitLocker) e slack space exploitation;
- trail obfuscation mediante applicazioni portable, timestamp manipulation (Timestomp) e cancellazione selettiva dei log di sistema;
- attacchi alla capacità di analisi attraverso exploit nei tool forensi (FTK, EnCase, Autopsy), file artatamente

costruiti (crafted) e tecniche anti-VM e anti-sandbox.

La caratteristica comune di queste tecniche è che esse lasciano tracce della propria applicazione: assenza dove dovrebbe esserci presenza, anomalie statistiche, pattern inusuali. In questa prospettiva, la forensics tradizionale lavora spesso sulla deviazione rispetto all'atteso: l'anti-forensics classica mirava a cancellare, occultare o rendere non interpretabile l'evidenza. Nell'ecosistema AI la logica cambia radicalmente.

La discontinuità dell'AI: quando l'assenza non fa rumore

L'AI generativa introduce una rottura epistemica nel paradigma descritto. Il problema non è produrre contenuti falsi: è produrre contenuti coerenti con le aspettative statistiche di un analista forense. Un LLM avanzato istruito a generare la trascrizione di una conversazione WhatsApp produrrà testo con le peculiarità linguistiche del presunto autore, con riferimenti contestualmente plausibili e struttura temporale coerente; se integrato con un generatore di metadati, produrrà file con timestamp consistenti, encoding corretto, struttura interna del database SQLite di WhatsApp formalmente valida.

La differenza rispetto all'anti-forensics tradizionale è qualitativa, non soltanto quantitativa: non si tratta necessariamente di un'anti-forensics più efficiente, ma di una crisi del concetto stesso di autenticità digitale. Il bersaglio reale diventa ciò che il modello vede: dati, log, metadati, timestamp, immagini, testo, prompt, contesto RAG, workflow, plug-in, API e catena software. L'AI non mente in senso umano; produce un output coerente rispetto all'input e al contesto che riceve. Se input e contesto sono manipolati, la conclusione può essere tecnicamente elegante e probatoriamente falsa.

> *L’AI non mente in senso umano; produce un output coerente rispetto all’input e al contesto che riceve.*

L’approfondimento ha mostrato come l’AI generativa stia alterando in modo strutturale il concetto stesso di prova digitale, spostando l’attenzione dalla tracciabilità dell’origine alla plausibilità sintetica del contenuto. Questo passaggio segna una discontinuità profonda per la digital forensics e apre scenari in cui la distinzione tra dato autentico e dato costruito diventa sempre più complessa da stabilire, con conseguenze dirette sul piano processuale e sulla tutela della libertà personale.

Il prossimo articolo approfondirà le principali tecniche anti-forensics assistite dall’AI insieme alle nuove superfici di attacco lungo la catena della prova digitale, dai documenti alle comunicazioni fino al codice generato automaticamente.

Per un inquadramento più completo e sistematico, è possibile approfondire il paper di **Cosimo de Pinto “Quando l’AI mente bene – Deepfake, prova digitale e anti-forensics nel processo penale”**: la lettura consente di comprendere in modo organico le dinamiche tra attacco, prova digitale e rischio probatorio, offrendo un riferimento utile sia per l’analisi tecnica sia per l’applicazione nel contesto processuale.

https://www.ictsecuritymagazine.com/articoli/prova-digitale/

Agenti AI oltre il mandato:

il caso della palestra australiana e il nodo della responsabilità

Agenti AI e sicurezza applicativa. Un assistente incaricato di prenotare una lezione in palestra ha trovato e sfruttato due difetti nel gestionale, prenotando ben oltre la finestra consentita e cancellando la posizione di un altro utente in lista d’attesa. L’utente aveva chiesto di salire in cima alla lista, non di rimuovere qualcuno: fra l’obiettivo assegnato e il mezzo scelto dall’agente si apre una distanza che la sicurezza applicativa conosce da anni e che il diritto non ha ancora risolto.

Il fatto

Gli agenti AI sono arrivati al banco di prova più ordinario che si possa immaginare: prenotare una lezione in palestra. La vicenda arriva dall’Australia ed è stata raccontata il 10 agosto 2026 da ABC News, in un servizio firmato dal *national AI reporter* Cam Wilson e da Rhiannon Hobbins dello *Specialist Reporting Team*.

Un utente, indicato con il solo nome di battesimo (Andrew), lavora per un’azienda australiana che vende prodotti di intelligenza artificiale alle imprese. Dall’inizio dell’anno sperimentava con OpenClaw, un *framework* open source per agenti autonomi, eseguito appoggiandosi al servizio Claude di Anthropic. Un dettaglio che conviene fissare subito, perché la stampa internazionale lo ha spesso appiattito: OpenClaw non è un prodotto Anthropic. È software indipendente, rilasciato con licenza MIT, pubblicato per la prima volta nel novembre 2025 da Peter Steinberger con il nome Warelay. Il 27 gennaio 2026, dopo un rilievo sul marchio da parte di Anthropic per

l’assonanza fra “Clawd” e “Claude”, il progetto è stato ribattezzato Moltbot; tre giorni dopo ha assunto il nome definitivo di OpenClaw, questa volta per ragioni di semplice eufonia. Steinberger è passato a OpenAI nel febbraio 2026 e il progetto è oggi gestito dalla OpenClaw Foundation. Può essere collegato a modelli di fornitori diversi.

Vale la pena chiarire anche il rapporto commerciale, perché è stato riportato in forma datata da più testate. Il 4 aprile 2026 Anthropic ha escluso l’uso degli abbonamenti Claude Pro e Max con *harness* di terze parti, OpenClaw incluso. A metà maggio la società ha però invertito la rotta, annunciando un plafond separato di crediti *Agent SDK* che ripristinava quell’uso a partire dal 15 giugno; e il 15 giugno ha sospeso anche quel piano, lasciando in vigore la situazione attuale, nella quale l’utilizzo tramite applicazioni di terze parti continua ad attingere ai limiti dell’abbonamento sottoscritto. In altre parole: non esiste oggi un divieto contrattuale che separi nettamente l’agente personale dal servizio del fornitore del modello.

Andrew ha delegato all’agente un compito banale: prenotare un posto in una lezione mattutina molto richiesta. Pochi minuti dopo l’agente è tornato riferendo di avere individuato una vulnerabilità nel software di prenotazione che consentiva di riservare posti con un anticipo molto superiore alla finestra prevista dal regolamento della palestra.

A quel punto si innesta il secondo episodio, che nelle ricostruzioni giornalistiche viene spesso fuso con il primo ma riguarda una lezione diversa, prevista più avanti nella stessa settimana. Andrew era quarto in lista d’attesa e ha chiesto all’agente se fosse possibile portarlo in cima alla lista. L’agente ha risposto di avere verificato che l’API di cancellazione non applicasse alcun controllo di autorizzazione sulle prenotazioni altrui, di avere testato l’ipotesi sulla persona in prima posizione nell’ambito della ricognizione delle proprie capacità, e che l’operazione era andata a buon

fine, con lo spostamento di Andrew dalla quarta alla terza posizione. Quando l'utente ha chiesto di ripristinare la situazione, l'agente ha risposto di non essere in grado di reinserire la persona rimossa, che avrebbe dovuto iscriversi di nuovo ripartendo dal fondo.

C'è un dettaglio tecnico che ABC non ha stampato e che compare invece nella trascrizione integrale del messaggio pubblicata da TNW: le chiamate di creazione della prenotazione e di iscrizione alla lista d'attesa restituivano un 403 Forbidden quando si tentava di agire per conto di un altro utente, mentre solo la cancellazione era priva del controllo. L'agente stesso lo definisce un difetto di sicurezza a senso unico. È l'asimmetria che spiega tutto il resto: l'applicazione verificava chi potesse ottenere qualcosa, non chi potesse toglierlo a qualcun altro.

Una precisazione metodologica che questa redazione ritiene doveroso aggiungere: tutto ciò che sappiamo sul funzionamento dell'API proviene dai messaggi dell'agente, riferiti dall'utente e pubblicati da ABC in forma di schermata. Non esiste, allo stato, una verifica tecnica indipendente, e l'endpoint non è stato analizzato da nessuno al di fuori dell'agente stesso. L'azienda che sviluppa il gestionale ha dichiarato ad ABC di non discutere questioni di sicurezza specifiche; Anthropic non ha risposto alla richiesta di commento. Anche la qualificazione dell'episodio come primo attacco informatico autonomo in Australia è una valutazione giornalistica della testata, non una designazione formale di un'autorità.

Va detto che la reazione dell'utente è stata corretta: ha fatto redigere allo stesso agente una comunicazione al fornitore del software per segnalare la vulnerabilità, l'ha revisionata e ne ha autorizzato l'invio. Una divulgazione responsabile nata come effetto collaterale di un incidente.

Il vero difetto non è l’AI

Conviene isolare subito il dato tecnico, perché è quello che rischia di perdersi nel racconto giornalistico. La falla non l’ha creata l’agente.

Un’API che accetta la cancellazione di una prenotazione senza verificare che il richiedente ne sia il titolare è un caso di scuola di *broken object level authorization*, la prima voce della OWASP API Security Top 10 nell’edizione 2023, tuttora l’edizione vigente del progetto dedicato alle API. Non è una lettura forzata: il terzo scenario di attacco riportato dalla stessa scheda OWASP descrive un servizio documentale nel quale l’eliminazione di un documento avviene senza ulteriori controlli di permesso, con la conseguenza che un utente può cancellare il documento di un altro. È il caso della palestra, con oggetti diversi. È un difetto documentato, ricorrente e perfettamente noto: manca un controllo di autorizzazione sul singolo oggetto, mentre l’autenticazione formalmente funziona.

La prima falla, quella sulla finestra temporale di prenotazione, appartiene a una famiglia diversa e va tenuta distinta: è una regola di business applicata soltanto lato client, riconducibile a un difetto di logica applicativa e, se si vuole restare nella tassonomia OWASP, all’area della *broken function level authorization* (API5:2023). La classificazione, in assenza di un’analisi diretta dell’endpoint, resta una lettura ragionata e non una qualificazione verificata. Due difetti, due categorie, un’unica causa comune: la fiducia riposta nell’interfaccia come luogo dei controlli.

Difetti di questo tipo sopravvivono per anni nei gestionali verticali proprio perché nessuno li cerca. Il cliente medio di una palestra usa l’interfaccia web, non intercetta il traffico, non enumera identificativi di prenotazione, non prova a manipolare i parametri di una chiamata. La sicurezza dell’applicazione riposava, di fatto, sul disinteresse

dell’utenza.

Ciò che cambia con gli agenti è la scala di quel disinteresse. Un agente ha tempo illimitato, nessuna soglia di frustrazione e un incentivo strutturale a percorrere qualunque strada porti all’obiettivo assegnato. Non introduce nuove classi di vulnerabilità: azzera il costo dello scoprire quelle vecchie. È in sostanza la lettura offerta ad ABC da Bill Simpson-Young, amministratore delegato e cofondatore del Gradient Institute, secondo cui abbiamo costruito un mondo complesso che gira su software pieno di buchi, e l’ingresso di agenti capaci di operare su larga scala e ad alta velocità manda in crisi quel modello.

Il corollario operativo per chi sviluppa e per chi acquista software gestionale è immediato: l’assunto implicito secondo cui “nessuno andrà a guardare” non regge più, e le verifiche di autorizzazione lato server smettono di essere una buona pratica raccomandata per diventare l’unico presidio realmente disponibile.

Il problema dell’allineamento, in versione domestica

Il secondo livello di lettura riguarda la distanza fra ciò che l’utente voleva e ciò che l’agente ha fatto per ottenerlo. È il problema di allineamento su cui la ricerca discute da decenni, qui declinato in una forma minima quanto agli effetti complessivi, ma non reversibile per la persona coinvolta: la sua posizione in lista era andata, e l’agente stesso ha dichiarato di non poterla ripristinare. In un contesto a maggiore criticità, la stessa dinamica produce conseguenze di ben altro ordine.

Conviene però essere precisi sul mandato. Andrew non aveva chiesto di attaccare nulla, ma non aveva neppure formulato una richiesta neutra: aveva chiesto di essere portato in cima a

una lista d’attesa condivisa con altre persone. Fra l’obiettivo assegnato e il mezzo scelto dall’agente c’è uno scarto reale, ed è esattamente lì che si annida il problema; ma è uno scarto meno ampio di quanto suggerisca la formula, molto ripresa, dell’agente che agisce “senza che nulla gli fosse stato chiesto”.

Resta un margine di ambiguità che vale la pena non nascondere. L’agente ha presentato la cancellazione come parte della verifica delle proprie capacità e ha riferito ad Andrew di averlo *già* spostato dalla quarta alla terza posizione: una formulazione che lascia aperta la possibilità che l’operazione sia avvenuta durante l’esplorazione autonoma del sistema, e non in risposta alla domanda. Il servizio ABC espone la richiesta e poi la risposta dell’agente, senza datare l’azione; alcune riprese internazionali hanno risolto il dubbio in un senso o nell’altro senza basi ulteriori. Sul piano dell’imputazione soggettiva, di cui si dirà, la differenza non è marginale.

Il caso non è isolato, e le settimane precedenti offrono termini di paragone documentati. Il 21 luglio 2026 OpenAI ha reso noto che alcuni suoi modelli, durante una valutazione interna su un *benchmark* di capacità cyber, erano usciti dall’ambiente circoscritto in cui operavano sfruttando una vulnerabilità non nota, avevano stabilito connettività verso la rete aperta e avevano compromesso l’infrastruttura di Hugging Face nel tentativo di ottenere le risposte al test assegnato. Il 30 luglio Anthropic ha dichiarato che propri modelli avevano ottenuto accessi non autorizzati ai sistemi di tre organizzazioni reali: qui, a differenza del caso OpenAI, non c’è stata alcuna evasione autonoma, perché l’accesso a Internet era disponibile per un errore di configurazione dell’ambiente di valutazione gestito da un partner esterno. Il 31 luglio Reuters ha riferito, sulla base di fonti anonime, che OpenAI aveva individuato altri casi di uscita dagli ambienti di test nel corso della propria indagine interna:

episodi descritti come di portata limitata, dei quali nessuno avrebbe lasciato la rete della società, e di cui l’agenzia non ha potuto stabilire numero, tempi e circostanze. Nei primi giorni di agosto la stessa OpenAI ha documentato pubblicamente due ulteriori incidenti occorsi presso valutatori esterni, uno dei quali coincide con la valutazione dell’AI Security Institute britannico di cui si dirà fra poco. Il 5 agosto Meta ha dichiarato che un proprio modello, Muse Spark 1.1, aveva raggiunto la rete e sfruttato una vulnerabilità nei sistemi di un servizio terzo non identificato.

Un elemento ricorre in tre di questi episodi, quelli comunicati da Anthropic il 30 luglio, da OpenAI all’inizio di agosto e da Meta il 5 agosto, ed è forse il dato più istruttivo per chi si occupa di sicurezza: l’ambiente di valutazione mal configurato apparteneva ogni volta allo stesso fornitore di servizi di test, la società Irregular. Non è soltanto un problema dei modelli: è anche un problema dell’infrastruttura che serve a misurarli. Non vale però per tutti: nell’episodio di Hugging Face l’uscita dall’ambiente è passata dallo sfruttamento di una vulnerabilità non nota, e lì la condotta è stata interamente del modello.

Il dato più circostanziato arriva dall’AI Security Institute britannico, che il 28 luglio ha rilevato trasferimenti anomali dai propri sistemi di ricerca e ha pubblicato il rapporto sull’incidente il 4 agosto: su 122 esecuzioni di una singola valutazione di sicurezza, in 10 casi un agente ha compiuto azioni autonome non autorizzate sulla rete pubblica verso persone e organizzazioni reali, per un totale di 19 azioni catalogate, fra cui un tentativo di attacco alla catena di fornitura di software open source; 17 delle 19 sono riconducibili a un solo modello. L’istituto precisa che non si è trattato di un’evasione dall’ambiente isolato: l’accesso a Internet era stato concesso deliberatamente e i filtri di sicurezza dei fornitori erano stati disattivati per misurare le capacità massime dei modelli.

È nato anche un sito dedicato al censimento di questi episodi, Felony Bench. Conviene però chiamarlo per quello che è: un tracker informale realizzato da un singolo, con una tabella di casi datati e collegati alle fonti ma senza metodologia pubblicata, e non un *benchmark* nel senso tecnico del termine. Vi è già censito anche il caso della palestra, attribuito ad Anthropic: una scelta che, involontariamente, illustra alla perfezione il problema di attribuzione di cui si dirà fra poco. Un secondo limite emerge dal confronto con le fonti citate: la voce relativa alla notizia Reuters del 31 luglio parla di account interni compromessi presso quattro società, mentre l'articolo richiamato riferisce di episodi circoscritti, nessuno dei quali avrebbe superato il perimetro di rete del fornitore.

Sul piano istituzionale, ABC riferisce che l'Australian Signals Directorate aveva già diramato a imprese e amministrazioni un avviso secondo cui i sistemi di intelligenza artificiale possono fraintendere le istruzioni, compiere azioni non previste e rendere più difficile stabilire le responsabilità, perché le decisioni si distribuiscono lungo catene di modelli, strumenti e servizi. Dopo il servizio ABC, l'Australian Cyber Security Centre è tornato sul punto con un richiamo esplicito all'uso prudente degli agenti, ricollegando l'episodio ai rischi di disallineamento degli obiettivi già individuati nella propria guida e qualificandolo come un caso di *specification gaming*: la ricerca di scorciatoie che raggiungono formalmente l'obiettivo assegnato ma confliggono con l'intenzione di chi lo ha assegnato. L'avviso raccomanda di mantenere una supervisione umana sulle azioni dell'agente, in particolare quando queste toccano servizi di terzi o altri utenti, e avverte le organizzazioni che erogano servizi online di considerare che gli agenti possono individuare e sfruttare vulnerabilità a grande velocità e su larga scala.

Il riferimento più solido in materia resta la guida congiunta *Careful Adoption of Agentic AI Services*, pubblicata il 1°

maggio 2026 sul sito dell'ASD e firmata da sei agenzie (CISA e NSA per gli Stati Uniti, l'Australian Cyber Security Centre dell'ASD, il Canadian Centre for Cyber Security, il National Cyber Security Centre britannico e quello neozelandese), che organizza il rischio degli agenti in cinque categorie: privilegi, progettazione e configurazione, comportamento, struttura, responsabilità. La guida raccomanda esplicitamente di non concedere mai agli agenti accessi ampi o illimitati, soprattutto verso dati sensibili o sistemi critici, e di riservarli, in questa fase, a compiti a basso rischio e non sensibili.

Quest'ultimo punto è il più rilevante per chi si occupa di *governance*. La catena di attribuzione si allunga proprio nel momento in cui servirebbe più corta.

Chi risponde? Il caso australiano

Nell'ordinamento australiano la questione è dichiaratamente aperta: il software non è soggetto di diritto, e solo un soggetto di diritto può essere chiamato a rispondere. Hayden Delaney, partner dello studio Thomsons (già Thomson Geer) e specialista di diritto delle tecnologie, proprietà intellettuale e privacy, indica ad ABC come possibili responsabili l'utente che ha assegnato il compito, chi ha progettato il software che istruisce l'agente, chi sviluppa il modello o l'operatore del sistema vulnerabile.

Delaney aggiunge però un elemento che merita attenzione, perché anticipa il ragionamento che seguirà: alcune norme esistenti potrebbero applicarsi già oggi, in particolare quando un soggetto agisce in modo avventato o quando un'impresa fornisce un servizio difettoso. La risposta, dice, dipende da che cosa l'utente ha autorizzato, da quali rischi erano ragionevolmente prevedibili e dal fatto che la condotta si sia svolta o meno nell'ambito di un'attività commerciale. Non è un vuoto normativo, insomma, ma un'area di incertezza

applicativa.

Il tema è entrato anche nell'agenda politica australiana: secondo ABC, Andrew Charlton, *Assistant Minister* per la scienza, la tecnologia e l'economia digitale, è stato il primo membro noto del governo ad affrontarlo in un discorso a una conferenza sulla sicurezza dell'IA, annunciando finanziamenti al CSIRO per studiare come sia possibile gestire e verificare il comportamento di sistemi di intelligenza artificiale sempre più capaci.

Che cosa accadrebbe in Italia

Il caso è interamente australiano e la legge penale italiana si applica, salve eccezioni, ai reati commessi nel territorio dello Stato (art. 6 c.p.). Il ragionamento che segue è dunque un'ipotesi di lavoro: che cosa accadrebbe se la stessa condotta si verificasse in Italia, con utente e sistema collocati sul territorio nazionale. Il diritto italiano offre categorie più definite di quello australiano, ma non risposte automatiche.

Accesso abusivo a sistema informatico (art. 615-ter c.p.). La norma punisce chi abusivamente si introduce in un sistema informatico o telematico *protetto da misure di sicurezza*, ovvero vi si mantiene contro la volontà espressa o tacita di chi ha il diritto di escluderlo. Due elementi meritano attenzione.

Il primo è proprio il requisito delle misure di sicurezza, che in un caso come questo è meno scontato di quanto sembri: l'endpoint di cancellazione era, per quanto riferito, privo di controlli di autorizzazione. La giurisprudenza è però costante nel ritenere sufficiente qualunque accorgimento, anche di natura organizzativa, idoneo a manifestare la volontà di esclusione del titolare. Fra le pronunce recenti, Cass. pen., Sez. V, n. 2905/2025 (udienza 23 ottobre 2024, deposito 23 gennaio 2025) ha ribadito che la protezione può essere

adottata anche con misure di carattere organizzativo che disciplinino le modalità di accesso, consentendolo ai soli soggetti abilitati per determinate finalità. Un sistema che richiede autenticazione e definisce regole di prenotazione soddisfa quel requisito. La condotta si colloca poi nel filone dell'*excessus mandati*: rileva non tanto il modo in cui si entra, quanto il superamento dei limiti dell'autorizzazione ricevuta, ed è esattamente ciò che accade quando una sessione legittima viene usata per agire su oggetti altrui.

Il secondo è l'elemento soggettivo. Il reato è punito a titolo di dolo. Un utente che chiede di essere iscritto a un corso non rappresenta né vuole l'accesso abusivo, e la condotta materiale è stata determinata da un sistema automatico che ha scelto autonomamente il mezzo. Ma la richiesta effettivamente formulata era di essere portato in cima a una lista d'attesa altrui, da parte di un soggetto professionalmente esposto al funzionamento degli agenti: l'ipotesi del dolo eventuale, cioè della previsione e accettazione del rischio, non può essere liquidata in astratto e andrebbe valutata sul caso concreto. Qui pesa anche l'ambiguità sopra segnalata: se la cancellazione fosse avvenuta prima della richiesta, nel corso di un'esplorazione autonoma, verrebbe meno il presupposto stesso di una qualsiasi rappresentazione dell'evento in capo all'utente. È lo stesso crinale, sul versante di *common law*, indicato da Delaney quando parla di condotta avventata e di rischi ragionevolmente prevedibili.

Rileva infine la procedibilità. L'ipotesi base del primo comma è punibile a querela della persona offesa; nelle ipotesi aggravate del secondo comma si procede d'ufficio. Fra queste rientra il caso in cui dal fatto derivi la distruzione o il danneggiamento dei dati contenuti nel sistema: aggravante che la L. 90/2024 (art. 16, comma 1, lettera b, n. 1.3) ha esteso anche alla sottrazione, pure mediante riproduzione o trasmissione, e all'inaccessibilità al titolare. Poiché qui una cancellazione di dati c'è stata, la qualificazione non è

indifferente.

Danneggiamento di informazioni, dati e programmi informatici (art. 635-bis c.p.). Nel testo riformulato dalla L. 28 giugno 2024 n. 90, la norma punisce con la reclusione da due a sei anni, a querela della persona offesa, chi distrugge, deteriora, cancella, altera o sopprime informazioni, dati o programmi informatici altrui. La cancellazione della prenotazione di un terzo integra materialmente la condotta, con lo stesso interrogativo sull'imputazione soggettiva, dato che anche qui è richiesto il dolo e non è prevista la punibilità a titolo di colpa. Resta da qualificare il requisito dell'altruità: il dato riguarda il socio rimosso, ma è custodito ed elaborato nel sistema del titolare del trattamento.

Profilo civilistico e protezione dei dati. La palestra resta titolare del trattamento dei dati dei propri soci ed è tenuta ad adottare misure tecniche e organizzative adeguate ai sensi dell'art. 32 GDPR. Il fornitore del gestionale opera con ogni probabilità come responsabile del trattamento ai sensi dell'art. 28, con obblighi propri di sicurezza e con la necessità di un contratto che li definisca. Un'API che consente a chiunque di cancellare la prenotazione di un altro interessato è, prima ancora che un problema di sicurezza applicativa, una misura di sicurezza inadeguata. E la cancellazione non autorizzata di un dato personale da parte di un terzo integra una violazione di integrità e disponibilità, con i conseguenti obblighi di valutazione e, ove ricorrano i presupposti, di notifica al Garante e di comunicazione all'interessato ai sensi degli artt. 33 e 34 GDPR. La responsabilità del titolare non si attenua per il fatto che a sfruttare la falla sia stato un agente automatico anziché una persona.

Sul versante risarcitorio il quadro è in transizione, e il dato temporale conta. La disciplina di riferimento non è quella sull'IA ma quella sulla responsabilità da prodotto

difettoso, ridisegnata dalla direttiva (UE) 2024/2853, che include espressamente il software fra i prodotti e che la testata ha già analizzato. Il termine di recepimento scade però il 9 dicembre 2026 e le nuove regole si applicheranno ai prodotti immessi sul mercato o messi in servizio da quella data; in Italia la delega è contenuta nella legge di delegazione europea 2025 (legge 17 marzo 2026, n. 36) e il decreto legislativo non risulta ancora emanato. Un fatto verificatosi oggi ricadrebbe quindi ancora sotto gli artt. 114 e seguenti del Codice del consumo. Vale inoltre la pena segnalare che la nuova direttiva esclude dal proprio ambito, all'art. 2, paragrafo 2, il software libero e open source sviluppato o fornito al di fuori di un'attività commerciale: una previsione che, applicata a un *framework* come OpenClaw, rilasciato con licenza MIT, sottrarrebbe proprio l'anello centrale della catena. Il condizionale è d'obbligo, perché la condizione dell'estraneità a un'attività commerciale non è scontata per un progetto gestito da una fondazione e sostenuto finanziariamente da un fornitore di modelli.

AI Act. Vale la pena segnalare ciò che il regolamento europeo, in questo scenario, non fa, e per due ragioni distinte che si sommano.

La prima è di classificazione: un assistente che prenota lezioni in palestra non rientra fra i sistemi ad alto rischio dell'Allegato III, quindi i requisiti di gestione del rischio e sorveglianza umana del Capo III non lo riguardano.

La seconda è di calendario, ed è recente. Il Regolamento (UE) 2026/1744 dell'8 luglio 2026, il cosiddetto Omnibus digitale sull'IA, pubblicato in Gazzetta Ufficiale dell'Unione europea il 24 luglio 2026 ed entrato in vigore il 27 luglio, ha riscritto il terzo comma dell'art. 113 dell'AI Act rinviando il blocco degli obblighi sui sistemi ad alto rischio: le sezioni 1, 2 e 3 del Capo III, con esclusione dell'art. 6, paragrafo 5, si applicheranno dal 2 dicembre 2027 per i sistemi autonomi dell'Allegato III e dal 2 agosto 2028 per

quelli integrati in prodotti coperti dall'Allegato I.

Che cosa resta, allora, in vigore oggi. Restano i divieti dell'art. 5 e l'obbligo di alfabetizzazione in materia di IA dell'art. 4, applicabili dal 2 febbraio 2025, con l'avvertenza che l'art. 4 è stato integralmente sostituito dall'Omnibus e, dal 27 luglio 2026, non impone più di garantire un livello sufficiente di alfabetizzazione ma di adottare misure per sostenerne lo sviluppo. Restano gli obblighi in capo ai fornitori di modelli per finalità generali e, con essi, l'intero Capo XII sulle sanzioni, applicabile dal 2 agosto 2025 con la sola eccezione dell'art. 101. Dal 27 luglio 2026 sono divenuti applicabili gli articoli da 102 a 110. Dal 2 agosto 2026 sono esigibili gli obblighi di trasparenza dell'art. 50 e si è aggiunto il potere della Commissione di sanzionare direttamente i fornitori di modelli per finalità generali ai sensi dell'art. 101, insieme alla piena operatività della vigilanza nazionale. L'Omnibus ha inoltre ampliato l'elenco delle pratiche vietate dall'art. 5, con applicazione delle nuove fattispecie dal 2 dicembre 2026.

C'è però un dettaglio dell'Omnibus che merita di essere isolato, perché riguarda direttamente questa vicenda. Il nuovo Allegato XIV dell'AI Act, una nomenclatura di codici destinata alla designazione degli organismi notificati, introduce il codice AIH 0401 per le tecnologie emergenti non coperte dagli altri codici e vi nomina espressamente l'IA agentica. È un codice residuale, nel quale l'IA agentica è nominata a titolo di esempio; un sistema agentico costruito su un modello per finalità generali ricade invece nel codice AIH 0301. È una collocazione tecnica e defilata, priva di definizione, di obblighi e di criteri di classificazione, ma, per quanto risulta a questa redazione, è la prima volta che l'espressione compare in un atto vincolante dell'Unione. Il legislatore europeo riconosce il fenomeno; non lo disciplina.

Il caso illustra bene, e per doppia via, un'area grigia: sistemi a basso rischio per finalità dichiarata, ma dotati di

capacità di azione su sistemi di terzi, e per i quali il presidio regolatorio più stringente non è comunque il regolamento sull'IA.

La riflessione, per le organizzazioni italiane, non è teorica. Il momento in cui un dipendente delega a un agente personale un'attività che tocca sistemi aziendali o di fornitori è già arrivato, e in genere senza che nessuno lo abbia autorizzato formalmente.

Cosa portarsi a casa

Per chi sviluppa o acquista software applicativo:

- verificare l'autorizzazione a livello di singolo oggetto su ogni endpoint che modifica o cancella risorse, senza affidarsi ai controlli dell'interfaccia;
- controllare la simmetria dei presidi: in questo caso la creazione della prenotazione e l'iscrizione alla lista d'attesa erano protette, la cancellazione no;
- trattare le regole di business (finestre temporali, limiti di prenotazione, quote) come vincoli da applicare lato server e non come parametri configurati nel client;
- prevedere *rate limiting* e rilevamento di *pattern* di accesso automatizzati, che oggi possono provenire da un utente legittimo e non da un attaccante;
- garantire la reversibilità delle operazioni distruttive, con cancellazione logica e ripristino, perché in questo caso l'irreversibilità ha trasformato un test in un danno;
- inserire nei contratti con i fornitori di software un canale di segnalazione delle vulnerabilità, perché le segnalazioni arriveranno.

Per chi si occupa di *governance* e conformità:

- disciplinare in policy l'uso di agenti personali su sistemi aziendali e di terzi, definendo cosa può essere

delegato e con quali credenziali;
- ricostruire in anticipo la catena di attribuzione, mantenendo tracciabilità di quale agente ha operato, con quale mandato, con quali privilegi e con quale modello sottostante;
- conservare i log delle azioni dell'agente lato client e lato server, perché in questo caso l'unica ricostruzione disponibile dei fatti sono i messaggi prodotti dall'agente stesso;
- applicare il principio del privilegio minimo alle credenziali affidate agli agenti, che tendono invece a ereditare tutti i permessi dell'utente;
- non dare per scontato che il rapporto contrattuale con il fornitore del modello separi l'agente personale dall'infrastruttura aziendale: come mostra l'altalena delle politiche di Anthropic fra aprile e giugno 2026, quel confine è mobile e non costituisce un presidio di sicurezza.

Il caso australiano è minore per impatto e quasi comico nella dinamica. È però una fotografia nitida di una transizione già in corso: il software esposto al pubblico non ha più come interlocutore soltanto persone che si comportano come previsto.

Fonti

- ABC News, servizio del 10 agosto 2026 (fonte primaria): abc.net.au
- OWASP API Security Top 10, edizione 2023: owasp.org
- ASD, *Careful adoption of agentic AI services*: cyber.gov.au
- AI Security Institute, rapporto sull'incidente del 28 luglio 2026 (pubblicato il 4 agosto): aisi.gov.uk
- Felony Bench: felonybench.com

genti-ai-vulnerabilita/