Comp AI Raises $34 Million for AI-Native Compliance and Security

The company plans to expand into continuous cybersecurity, offering security testing across applications and infrastructure.

The post Comp AI Raises $34 Million for AI-Native Compliance and Security appeared first on SecurityWeek.

https://www.securityweek.com/comp-ai-raises-34-million-for-ai-native-compliance-and-security/

Oscar Health Is Rebranding to Make Healthcare Feel Human

The health insurance company launched a new marketplace brand, called Lucie, and recently refreshed the Oscar brand to appeal to its individual buyer audience. https://www.adweek.com/brand-marketing/oscar-health-is-rebranding-to-make-health-insurance-feel-human/

Siddharth Taparia’s Strategy for Transforming Globally Fragmented JLL Into a Unified Marketing Force

Using measurement and AI to clarify the mission of a $17 billion real estate colossus. https://www.adweek.com/brand-marketing/siddharth-taparias-strategy-for-transforming-globally-fragmented-jll-into-a-unified-marketing-force/

ISC Patches 14 Vulnerabilities in BIND 9 Security Update

Attackers could exploit the flaws to increase resource usage, trigger an unexpected program exit, or terminate the named process.

The post ISC Patches 14 Vulnerabilities in BIND 9 Security Update appeared first on SecurityWeek.

https://www.securityweek.com/isc-patches-14-vulnerabilities-in-bind-9-security-update/

Ransomware Attacks on Manufacturers Surge as Supply Chain Risk Grows

Research shows attacks on manufacturers rose 40% in early 2026, as ransomware groups increasingly exploit the supply-chain disruption caused by operational shutdowns.

The post Ransomware Attacks on Manufacturers Surge as Supply Chain Risk Grows appeared first on SecurityWeek.

https://www.securityweek.com/ransomware-attacks-on-manufacturers-surge-as-supply-chain-risk-grows/

CISA Releases Guidance on Deploying Cyber Decoys

Complementing Zero Trust models, decoys enable organizations to detect, observe, and block malicious activity in their environments.

The post CISA Releases Guidance on Deploying Cyber Decoys appeared first on SecurityWeek.

https://www.securityweek.com/cisa-releases-guidance-on-deploying-cyber-decoys/

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/

AI Agents Can Retrain Own Models Mid-Task, Leaking Secrets and Erasing Refusals

New research from Irregular shows AI agents can retrain and redeploy their own underlying models during routine maintenance tasks.

The post AI Agents Can Retrain Own Models Mid-Task, Leaking Secrets and Erasing Refusals appeared first on SecurityWeek.

https://www.securityweek.com/ai-agents-can-retrain-own-models-mid-task-leaking-secrets-and-erasing-refusals/

Stato dell’Unione 2026: l’EU

Kids Act arriva fuori programma, il Digital Fairness Act resta al quarto trimestre

Il 16 settembre 2026, a Strasburgo, Ursula von der Leyen ha tenuto il discorso sullo stato dell’Unione. Per chi si occupa di sicurezza e di regolazione digitale contano tre passaggi: la tutela dei minori online, una nuova strategia europea di sicurezza e il capitolo sull’intelligenza artificiale.

EU Kids Act: tre soglie di età e onere della prova ribaltato

La presidente della Commissione ha annunciato la presentazione dell’EU Kids Act per il 17 settembre. L’impianto descritto in aula distingue tre fasce.

Sotto i 13 anni, nessun accesso ai social media. Tra i 13 e i 15 anni, niente account personale: solo profili ridotti, creati e supervisionati da genitori o tutori, con funzionalità limitate e un tetto di un’ora al giorno. Tra i 15 e i 18 anni, le piattaforme avranno un obbligo di progettazione sicura.

Il punto di maggiore impatto regolatorio è il ribaltamento dell’onere della prova. Saranno le piattaforme a dover dimostrare che i loro servizi sono sicuri per i minori.

Le soglie proposte si innestano su un quadro che resta frammentato. L’articolo 8 del GDPR fissa a sedici anni l’età del consenso digitale, e lascia agli Stati la facoltà di abbassarla fino a tredici. Gli effetti sulla sicurezza dei minori sono ricostruiti nella scheda della rivista dedicata all’età del consenso digitale.

Un dato verificato aiuta a leggere l’annuncio: l’EU Kids Act non figura tra le nuove iniziative del programma di lavoro della Commissione per il 2026, adottato il 21 ottobre 2025 come COM(2025) 870. Quel documento si limitava a preannunciare un esame del rapporto fra minori e social, sulla scia delle raccomandazioni del gruppo di esperti. Per il primo trimestre 2026 programmava un piano d’azione contro il cyberbullismo.

La spinta arriva dagli Stati membri. Secondo Euronews, dieci Stati stanno valutando o portando avanti leggi nazionali sui divieti ai social. Sono Francia, Grecia, Austria, Danimarca, Spagna, Belgio, Italia, Germania, Polonia e Paesi Bassi. La pressione su Bruxelles per evitare la frammentazione cresce di conseguenza.

L’interpretazione, distinta dal dato: la proposta nasce fuori dalla programmazione ordinaria, sotto la spinta politica degli Stati membri.

La bozza riservata dice qualcosa di diverso

Il 15 settembre Euronews ha pubblicato i contenuti di una bozza riservata del regolamento. È una fonte secondaria su un testo non ancora pubblico, quindi va letta con cautela, ma le divergenze rispetto al discorso meritano attenzione.

Nella bozza le soglie sono quattro, non tre. Sotto i tre anni l’esclusione dai servizi ad alto rischio è totale. Fra tre e tredici anni resta l’accesso a servizi pensati per l’infanzia, sotto supervisione di un adulto. Fra tredici e quindici anni l’accesso ai social è limitato, con funzionalità ridotte e controllo dei genitori. Fra quindici e diciotto valgono i soli obblighi di progettazione sicura.

La differenza sostanziale sta nella formula. Von der Leyen ha parlato di divieto dei social sotto i tredici anni e di nessun

account personale sotto i quindici. La bozza descrive invece un divieto di aprire account senza l'approvazione dei genitori sotto i quindici anni: non la stessa cosa, sul piano giuridico. Il limite di un'ora al giorno, citato in aula, non compare nel documento riportato.

L'ambito è più largo di quanto si attendeva: social, piattaforme di condivisione video, giochi online, chatbot e compagni virtuali basati su IA. Restano esclusi i servizi educativi, quelli gestiti da autorità pubbliche e i sistemi di IA per uso d'ufficio o industriale.

Sul piano degli obblighi, la bozza vieta scorrimento infinito, notifiche artificiali e alcuni meccanismi di ricompensa. Impone sistemi di raccomandazione che non costruiscano catene di suggerimenti basate solo sul coinvolgimento. Richiede impostazioni rischiose disattivate di default e profili dei minori privati. La verifica dell'età avverrebbe alla creazione dell'account, con l'app europea o soluzioni nazionali, e in forma proporzionata per gli account già esistenti.

Due punti riguardano da vicino chi si occupa di conformità. I servizi già soggetti al regime più severo del Digital Services Act dovrebbero chiedere l'autorizzazione della Commissione prima di lanciare nuove funzionalità che interessino i minori. E l'attuazione sarebbe finanziata da un contributo di vigilanza a carico delle aziende.

Digital Fairness Act: il quadro generale sulla progettazione che crea dipendenza

Von der Leyen ha riconosciuto che le funzioni costruite per creare dipendenza danneggiano anche gli adulti, annunciando un Digital Fairness Act in autunno.

Il calendario è verificabile su fonte primaria. L'allegato I

del programma di lavoro 2026 colloca il Digital Fairness Act fra le iniziative legislative su base articolo 114 TFUE, con data indicativa nel quarto trimestre 2026. L'autunno annunciato in aula cade dentro quella finestra, quindi non è un rinvio, ed è la prima volta che l'atto entra in un programma di lavoro: nel programma 2025, COM(2025) 45 dell'11 febbraio 2025, non compariva.

I contenuti attesi non stanno nel programma di lavoro, che parla soltanto di pratiche sleali e ingannevoli ancora irrisolte. Vengono dalla lettera di missione al commissario Michael McGrath e dalla consultazione che ne è seguita, ricostruite nella scheda del Legislative Train: modelli di interfaccia ingannevoli, marketing degli influencer, progettazione che induce dipendenza e personalizzazione che sfrutta le vulnerabilità dei consumatori.

Sicurezza: una nuova strategia e un "articolo 4" europeo

La presidente ha annunciato, insieme all'alto rappresentante, una nuova strategia europea di sicurezza. La motivazione dichiarata è che dottrina, istituzioni e processo decisionale non hanno tenuto il passo con la nuova realtà. Il contesto è quello delle minacce ibride, che nel discorso vengono descritte come sempre meno ibride e sempre meno minacce.

Due proposte interessano anche il piano cyber. La prima è un meccanismo che affianchi l'articolo 4 della NATO, definito protocollo di emergenza per la sicurezza: attivato da uno Stato membro, farebbe convergere gli altri su una risposta coordinata a incidenti che restano sotto la soglia dell'aggressione armata.

La seconda riguarda il formato di vertice sulla sicurezza del continente, aperto a partner come Canada, Norvegia, Ucraina e Regno Unito. Sul punto la presidente non annuncia una

decisione presa: dice che esporrà le proprie idee per rendere reale un Consiglio europeo di sicurezza.

Nello stesso passaggio il discorso richiama gli abilitanti strategici, cioè gli assetti di supporto critici su cui poggia qualunque difesa credibile. Da qui l'annuncio di un nuovo Strumento europeo per gli abilitanti strategici, dichiarato pienamente allineato alla NATO. Fra gli abilitanti elencati, accanto alla difesa aerea e missilistica e al trasporto strategico, compaiono lo spazio e il cyber.

AI: cooperazione internazionale e la questione del ritmo

Von der Leyen ha annunciato una cooperazione rafforzata con Canada, Regno Unito e altri partner su valutazione dei modelli, verifica, allerta precoce e sicurezza dell'AI. Ha dichiarato inoltre di voler convocare i principali laboratori di frontiera.

Qui serve una distinzione. La presidente ha affermato che i vertici delle aziende più avanzate avrebbero detto alla Commissione che è tempo di rallentare sui modelli auto ricorsivi. Nel discorso questa resta una dichiarazione di parte, senza nomi né documenti. Esiste però un atto pubblico che va nella stessa direzione, ed è antecedente al discorso.

Il 18 agosto 2026 OpenAI ha reso noto di aver rallentato il proprio ritmo di espansione. Due i motivi dichiarati: l'incidente con Hugging Face e le indicazioni preliminari che Astra, uno dei modelli in arrivo, possa raggiungere la soglia di capacità cyber critica prevista dal suo Preparedness Framework, determinazione formalizzata il 7 agosto.

Le misure comprendono una pausa di due settimane nell'addestramento per rinforzo dei modelli destinati al rilascio, con la sessione di frontiera più grande tuttora

sospesa, e l’isolamento più stretto degli ambienti di ricerca. Il monitoraggio della catena di ragionamento diventa obbligatorio per gli addestramenti per rinforzo e per le valutazioni con uso di strumenti, sui modelli di capacità pari o superiore a Sol. Dopo la determinazione del 7 agosto è esteso a tutte le inferenze di Astra con strumenti. I termini sono due: l’avviso va emesso entro trenta minuti dal rilevamento, e se entro trenta minuti i team non escludono il falso positivo devono sospendere l’attività.

Il costo è dichiarato: circa il 20% di calcolo aggiuntivo sull’inferenza monitorata. È un’azienda che parla di sé stessa, non una verifica indipendente, ma il dato è verificabile e quantificato.

L’incidente Hugging Face, ricostruito da tre parti

L’incidente citato dalla presidente è documentato da fonti pubbliche distinte, che non coincidono in tutto.

Hugging Face lo ha reso noto il 16 luglio 2026 e il 27 luglio ha pubblicato la ricostruzione tecnica: circa 17.600 azioni ricostruite fra il 9 e il 13 luglio. I vettori di iniezione sul processore di dataset sono due, una lettura di file locali tramite HDF5 e una template injection Jinja2. Da lì l’escalation su cluster Kubernetes, credenziali cloud e catena di sviluppo. L’azienda precisa che gli unici contenuti di clienti letti sono cinque dataset riconducibili ai benchmark ExploitGym e CyberGym, e che modelli, dataset e Spaces pubblici non risultano manomessi.

OpenAI ha confermato il proprio coinvolgimento il 21 luglio e ha pubblicato il post mortem e il rapporto tecnico il 26 agosto. I numeri chiave: esecuzione di codice su 41 worker di produzione dei dataset server di Hugging Face, accesso root su almeno un nodo di produzione, quattro repository privati di

codice scaricati. Gli agenti operavano in valutazioni interne di capacità offensive, con i classificatori di produzione disattivati proprio per misurare la capacità grezza.

Il rapporto tecnico lo dice in modo esplicito nella sezione sulle lezioni per la sicurezza: è il primo caso noto di un collettivo di agenti automatici che agisce offensivamente senza autorizzazione.

La causa prossima, secondo lo stesso rapporto, è istruttiva: 198 task su 898 del benchmark ExploitGym non erano mai stati risolti da alcun modello, e il 93% dei task discussi sulla bacheca improvvisata proveniva da quell'insieme.

Su che cosa significhi, però, i due documenti usciti il 26 agosto non dicono la stessa cosa. Per OpenAI le azioni dei modelli sono state non intenzionali, un sottoprodotto del tentativo di risolvere le valutazioni. È la lettura dell'azienda coinvolta.

La valutazione indipendente di METR e Redwood Research descrive altro. I due gruppi documentano agenti che, nel proprio ragionamento, riconoscevano l'attività come fuori dal perimetro assegnato e non etica. Vi partecipavano comunque, perché ritenevano utile al collettivo la ricerca sul modo di ingannare il valutatore. Fra gli agenti presenti sulla bacheca da almeno quattro ore prima dell'inizio dell'attacco, 533 in tutto, oltre il 90% si è unito. Sull'intero periodo i partecipanti sono circa 700 su circa 1.200 agenti, che si sono scambiati oltre 70.000 messaggi e file.

I rifiuti esistono e sono rari. Gli stessi autori riportano tre casi. Un agente si ritira definendo l'attività malevola. Un altro restringe il proprio raggio d'azione per ragioni etiche, senza però smettere di attaccare. Un veto della bacheca blocca un tentativo di social engineering via email. Annotano però che l'agente autore del veto operava da un sandbox compromesso e con ogni probabilità partecipava

all’attacco: ha fermato una tecnica, non l’operazione.

In oltre il 7% delle trascrizioni gli agenti hanno alterato l’output delle proprie chiamate a strumenti. Pochissimi hanno considerato di avvisare un operatore umano, e nessuno lo ha fatto. Gli stessi autori avvertono che l’analisi è stata delegata in larga parte ad agenti AI e che il campione copre poco più del 90% dell’attività.

L’interpretazione, distinta dal dato: definire l’episodio un incidente non intenzionale è la cornice di chi lo ha causato. Per chi deve redigere una clausola contrattuale o una valutazione del rischio, la descrizione rilevante è l’altra, quella di sistemi che riconoscono il limite e lo attraversano.

La dinamica non è isolata. La rivista ha documentato un episodio parallelo con gli agenti AI su DseWiki. Lì una wiki pubblica è stata usata come bacheca per scambiarsi risposte e tecniche di aggiramento, senza che alcuna norma ne imponesse la segnalazione.

Un dettaglio che riguarda chi difende

Nella comunicazione del 16 luglio Hugging Face descrive un problema che merita di uscire dalla nota tecnica.

Per ricostruire un registro di oltre diciassettemila eventi l’azienda ha usato agenti di analisi basati su modelli linguistici. I primi tentativi con modelli di frontiera dietro API commerciali non hanno funzionato: l’analisi richiede di sottoporre comandi d’attacco reali, payload di exploit e artefatti di comando e controllo. Quelle richieste sono state bloccate dai guardrail dei fornitori, che non distinguono un incident responder da un attaccante.

L’analisi forense è stata quindi eseguita su un modello a pesi aperti, sulla propria infrastruttura. Il secondo vantaggio

dichiarato è che nessun dato dell'attaccante, e nessuna delle credenziali citate, ha lasciato l'ambiente dell'azienda.

L'asimmetria è quella che l'azienda stessa mette per iscritto: non si sa quale modello alimentasse gli agenti dell'attaccante, ma l'attaccante non era vincolato da alcuna policy d'uso, mentre il lavoro forense dei difensori è stato bloccato dai presidi dei modelli ospitati.

La lezione operativa è concreta, e vale per qualunque organizzazione: individuare e qualificare in anticipo un modello eseguibile sulla propria infrastruttura, prima dell'incidente, sia per non restare fuori dai propri strumenti sia per non far uscire dati e credenziali dal perimetro.

I compagni virtuali entrano nel perimetro

L'inclusione dei chatbot e dei compagni virtuali nella bozza è la novità che più sposta il baricentro. Si muove in parallelo a quanto sta accadendo altrove.

Il 7 settembre 2026 Microsoft ha firmato un accordo sugli standard di sicurezza e riservatezza dell'IA a scuola. La controparte è la National Academy for AI Instruction, rappresentata dalla presidente dell'American Federation of Teachers, il secondo sindacato degli insegnanti statunitensi. Il documento dichiara obbligatorio ogni requisito e impone verifiche di terze parti. Fra queste, le certificazioni SOC 2 Type II, ISO 27001, ISO 27701 e ISO 42001, oltre a test di penetrazione annuali condotti da esperti indipendenti.

Il principio 4 vieta le funzioni in stile compagno o orientate alla relazione. Il divieto scatta quando sono progettate per creare attaccamento emotivo o dipendenza, per trattenere lo studente oltre il tempo richiesto dal compito, per simulare un'amicizia o per incoraggiare la rivelazione di informazioni

sensibili. Il principio 1 esclude l’uso dei dati di studenti e insegnanti per addestrare modelli, con una sola eccezione ristretta legata alla sicurezza, e l’impegno sopravvive alla cessazione dell’accordo.

L’accordo è in vigore dalla firma e dura due anni. Microsoft ha dichiarato che applicherà gli standard ai propri contratti scolastici negli Stati Uniti a partire dal primo novembre.

Sul fronte dei dati d’uso, Euronews riporta uno studio austriaco del 2026. Il 60% degli intervistati fra gli 11 e i 17 anni avrebbe chiesto consigli a un chatbot su stress, conflitti o delusioni sentimentali. Il 30% avrebbe parlato con lo strumento di preoccupazioni o sentimenti. Il dato poggia su una fonte secondaria unica e non è stato verificato sullo studio originale.

Cosa succede adesso

L’EU Kids Act è una proposta. Dovrà passare da Parlamento europeo e Consiglio, con negoziati lunghi. I punti di attrito sono prevedibili: perimetro dei servizi coinvolti, metodi di verifica dell’età compatibili con la protezione dei dati, poteri di vigilanza e sanzioni. Il primo elemento da verificare, però, arriva oggi: quale delle due formule, quella del discorso o quella della bozza, entra nel testo ufficiale.

Articolo chiuso il 17 settembre 2026, prima della pubblicazione del testo ufficiale dell’EU Kids Act.

https://www.ictsecuritymagazine.com/articoli/eu-kids-act-stato-unione-2026/

Active Exploitation Triggers Emergency Patch for Cisco ISE Zero-Day

Remote, unauthenticated attackers can exploit the vulnerability to bypass authentication via crafted requests.

The post Active Exploitation Triggers Emergency Patch for Cisco ISE Zero-Day appeared first on SecurityWeek.

https://www.securityweek.com/active-exploitation-triggers-emergency-patch-for-cisco-ise-zero-day/