Zero-day Magento e Adobe Commerce: StyleSmuggler sfruttato attivamente, Adobe non ha ancora rilasciato la patch

Dal 4 settembre 2026 un attacco senza correzione ufficiale colpisce i negozi Magento Open Source aggiornati. Le mitigazioni disponibili sono di terze parti. Per gli e-commerce italiani si aprono due questioni giuridiche: la notifica al Garante e la responsabilità dell'agenzia.

Dal 4 settembre 2026 gli e-commerce basati su Magento Open Source e Adobe Commerce sono esposti a un attacco attivo. Il produttore non ha ancora pubblicato una correzione. La vulnerabilità, chiamata **StyleSmuggler** dalla società olandese Sansec che l'ha scoperta, permette a un attaccante non autenticato di eseguire codice sul server del negozio e di installare una backdoor persistente.

Al 6 settembre Adobe non ha pubblicato avvisi, identificativi CVE, patch o soluzioni temporanee. L'indice dei bollettini di sicurezza di Adobe Commerce si ferma all'aggiornamento dell'11 agosto (APSB26-92). Quel bollettino correggeva sette CVE, tra cui un'escalation di privilegi non autenticata con punteggio CVSS 9.1 (CVE-2026-71362), e dichiarava che Adobe non era a conoscenza di exploit in circolazione. Secondo Sansec, la prossima release di sicurezza è prevista per l'8 settembre, ma non è noto se coprirà questo difetto.

Questo articolo non ripercorre la cronaca. Si concentra sul vuoto tra la scoperta e la patch. Cosa può fare oggi un esercente italiano che tratta dati di pagamento? Quali limiti

hanno le mitigazioni non ufficiali? Quali obblighi giuridici scattano quando la compromissione è probabile ma non accertata?

Cosa è StyleSmuggler

La cronologia dell'avviso di Sansec fissa la prima esecuzione confermata dell'exploit alle 22:20 UTC del 4 settembre. Alle 22:40 Sansec ha individuato la campagna; alle 23:10 il suo scanner eComscan ha segnalato l'impianto su negozi non collegati tra loro. Nelle ore successive Sansec ha riprodotto l'intera catena su installazioni pulite di Magento Open Source 2.4.7, 2.4.8 e 2.4.9. Le regole di blocco del suo prodotto Shield sono entrate in funzione il 5 settembre alle 07:15 UTC. L'avviso è uscito lo stesso giorno, prima del completamento dell'analisi, perché i negozi venivano compromessi in quel momento.

L'attacco abusa del sistema di template di Magento, in particolare delle proprietà styles, per aggirare le protezioni esistenti. Si svolge in due fasi:

1. **Iniezione.** L'attaccante fa scrivere codice PHP in un file che Magento stesso genera. Sansec cita i rapporti di errore in var/report/. Disrex Group, società olandese di hosting e sviluppo Magento, ha gestito la risposta a due negozi compromessi. In entrambi i casi ha osservato l'avvelenamento di var/log/system.log: l'attaccante invia un codice negozio non valido e Magento lo registra così com'è. Secondo Disrex questa fase non è filtrabile a livello di server web, perché è indistinguibile da un'integrazione malfunzionante.
2. **Esecuzione.** L'attaccante provoca l'invio dell'email standard "Payment Transaction Failed Reminder". Il codice iniettato viene eseguito mentre Magento prepara il messaggio. Nessuno deve aprire l'email; l'attacco riesce anche se la consegna fallisce. L'assenza di

messaggi in casella non è quindi una prova di sicurezza.

La ricostruzione del meccanismo, pubblicata da Disrex in un documento dedicato, è un'interpretazione indipendente. Secondo questa lettura, una direttiva {{block}} nel testo iniettato conduce, attraverso una catena di classi native di Magento, fino a codice pensato per il solo compilatore di dependency injection da riga di comando. Quel codice termina con un include su un percorso scelto dall'attaccante: il log avvelenato un istante prima. Disrex ha corretto una prima versione che identificava un punto di ingresso sbagliato. Sansec non ha confermato la ricostruzione né pubblicato la catena completa; Disrex non ha diffuso la richiesta assemblata. Per le stesse ragioni questo articolo non riporta dettagli riproducibili.

Una volta ottenuta l'esecuzione, un dropper PHP prova in ordine sei funzioni di avvio processo (shell_exec, exec, system, passthru, proc_open, popen) e usa la prima disponibile. Poi scarica e avvia l'impianto: un binario Rust statico e privo di simboli di circa 1,9 MB, per x86-64 e arm64. Il file viene installato in ~/.local/share/.gvfsd/gvfsd-user, nella home dell'utente del sito e non nella cartella web. Il processo si maschera con il nome [kworker/u:8:0], tipico di un thread del kernel Linux. Una voce cron lo riavvia ogni cinque minuti; è scritta direttamente nel file di spool, così il registro di sistema non mostra alcuna sostituzione del crontab.

Il perimetro reale: perché le patch di agosto non proteggono

Il dato che più dovrebbe preoccupare i responsabili IT è il profilo della prima vittima documentata da Sansec. Girava su Magento 2.4.6-p15, con le patch di luglio e agosto 2026 applicate e security:patch-status pulito. È il livello più alto che Adobe distribuisce per la linea 2.4.6, quello che il

bollettino di agosto etichetta 2.4.6-2026-aug. Secondo la cronologia delle versioni, il supporto ordinario di quella linea è terminato l'11 agosto 2026, con supporto esteso fino ad agosto 2027. Essere aggiornati non ha protetto.

Il secondo caso documentato è l'opposto. Il negozio 2.4.7-p2 gestito da Disrex girava su un livello del 13 agosto 2024, otto livelli indietro rispetto al 2.4.7-p10 del 12 maggio 2026. L'esito è stato lo stesso. Come scrive Disrex, il livello di patch non dice nulla sull'esposizione.

Il perimetro noto al 6 settembre è il seguente:

- **Vulnerabile per riproduzione diretta:** Magento Open Source 2.4.7, 2.4.8, 2.4.9 (Sansec).
- **Compromesso in attacchi reali:** 2.4.6-p15 (Sansec), 2.4.7-p2 (Disrex).
- **Non verificato:** Adobe Commerce e Adobe Commerce on Cloud. L'avviso Sansec li dichiara interessati, ma riporta riproduzioni solo su Magento Open Source. Adobe non ha confermato quali versioni siano coinvolte. Vanno trattate come esposizione sconosciuta, non come sicure.
- **Fuori perimetro:** Magento 1, fuori supporto dal 2020. Non esistono dati e le mitigazioni descritte più avanti non sono state provate.

La cronologia di Disrex aggiunge un elemento che va oltre il caso specifico. Il suo primo negozio è stato colpito alle 23:10 UTC del 4 settembre, cinquanta minuti dopo la prima esecuzione confermata a livello mondiale e otto ore prima che esistesse qualsiasi regola di blocco. Nessun livello di patch e nessuna firma avrebbero coperto quella finestra. È l'argomento più forte a favore delle misure di sistema indipendenti dalla vulnerabilità, descritte più avanti.

Non esistono statistiche ufficiali sulla diffusione di Magento in Italia. Chi voglia stimare l'esposizione nazionale può usare servizi di rilevamento tecnologico come BuiltWith o

W3Techs con filtro geografico. Questi strumenti contano i domini, non le installazioni, e non distinguono in modo affidabile Open Source da Adobe Commerce. La domanda pratica è un'altra: il mio sito espone l'endpoint /graphql e quale vetrina lo usa? Vale anche ricordare che StyleSmuggler non è un caso isolato. A giugno una falla nel modulo di terze parti Mirasvit Full Page Cache Warmer (CVE-2026-45247), segnalata in questa rassegna, apriva anch'essa a esecuzione di codice remota non autenticata sui negozi Magento e Adobe Commerce, ed è stata inserita da CISA nel catalogo delle vulnerabilità sfruttate. La differenza è che StyleSmuggler colpisce il core della piattaforma, non un'estensione, e non ha una patch.

Indicatori di compromissione e verifica

Il primo passo, prima di qualsiasi mitigazione, è verificare se il negozio è già compromesso. Applicare regole di blocco su un server infetto non serve: l'impianto è dentro e si riavvia da solo.

Gli indicatori pubblicati da Sansec e nell'elenco di Disrex, aggiornato nel corso del 5 settembre, convergono su questi elementi:

- **Processo:** [kworker/u:8:0] di proprietà di un utente non root. Un vero thread del kernel appartiene sempre a root e non ha memoria residente. Un nome tra parentesi quadre sull'utente del sito, con memoria reale, è l'impianto. L'impianto imposta la riga di comando alla stringa letterale, quindi un controllo sul campo comm del processo non trova nulla; va letto args.
- **File:** ~/.local/share/.gvfsd/gvfsd-user, ~/.local/share/.gvfsd/.gvfsd_<8hex>.lock, /tmp/.gvfsd_<8hex>.lock, /tmp/.kw_<random>.
- **Cron:** */5 * * * * exec <home>/.local/share/.gvfsd/gvfsd-user, con una variante che punta a /tmp/.kw_. Su un

negozio la stessa riga compariva 1.728 volte e tornava entro un secondo dalla rimozione.

- **Hash SHA-256:** Sansec pubblica due campioni (e315687a...a26a7 e b79dfdc1...51420). Disrex ne aggiunge uno trovato su disco in entrambi i negozi (8334b434...d06ef) e uno diverso in esecuzione in memoria su uno di essi (251fabd5...bf220). L'operatore aggiorna l'impianto in memoria: va calcolato l'hash anche di /proc/<pid>/exe.
- **Rete:** 247.cdnflare[.]xyz (host di scaricamento), 99.84.67[.]186:443 (comando e controllo via WebSocket su TLS), windwsecurity[.]run:443 (shell remota). Tre domini imitano server NTP con traffico UDP sulla porta 123: ntp.timesysnc[.]net, time.microsft[.]run, pool.microsft[.]studio. Le imitazioni di Microsoft sono evidenti e vanno cercate nei log DNS.
- **Sorgenti:** 88.216.72[.]181 (Sansec, vista su più vittime), 5.181.86[.]133 e 91.238.181[.]19 (Disrex, traffico in massa). La seconda è la sorgente quasi esclusiva della seconda ondata, nel pomeriggio del 5 settembre.
- **Firme di richiesta:** POST /graphql?styles[...]= con codifica percentuale; POST /paypal/transparent/response/ con codice PHP nella stringa di query. C'è poi la richiesta ordinaria GET /customer/section/load/, che nella variante osservata da Disrex portava il payload nell'intestazione User-Agent. Le richieste di exploit arrivano con user agent python-requests (2.15.0 nella prima ondata, 2.32.4 nella seconda), mai con un browser.
- **Prova di esecuzione:** secondo Disrex il payload racchiude il proprio output in un delimitatore MG<20 hex>::...::/MG<20 hex>. La sua presenza nei log prova che il codice è stato eseguito, non solo inviato. Un secondo segnale è un TypeError di array_merge() con argomento intero in system.log, subito dopo l'include: indica che l'exploit è riuscito. La variante più silenziosa restituisce un array vuoto e non lascia

traccia.

Tre avvertenze emerse sul campo meritano attenzione. Primo: il controllo pubblicato da Sansec cerca X_TRACE_ in var/report/. Entrambe le infezioni gestite da Disrex passavano da var/log/system.log e sarebbero sfuggite; vanno ispezionate entrambe le cartelle. Secondo: l’intestazione di innesco è cambiata nella stessa giornata, da X-TRACE-<10 hex> al mattino a X-<12 hex> al pomeriggio, con valore rigenerato a ogni richiesta. Una rilevazione ancorata alla stringa letterale è diventata cieca in un giorno: la ricerca deve intercettare la forma. Terzo: su uno dei due negozi l’impianto non ha aperto alcuna connessione in uscita. Ha letto le sessioni Magento dall’istanza Redis locale con 28 connessioni a 127.0.0.1:6379. Due catture di traffico di oltre 200 MB ciascuna non contenevano un solo pacchetto verso gli indirizzi noti. L’assenza di traffico sospetto non prova nulla.

Sugli scanner c’è un dato discordante, da riportare come tale. Sansec raccomanda il proprio eComscan e ha rilasciato la versione 1.9.7, che termina i processi per i clienti Shield. Disrex riferisce invece che su uno dei negozi infetti lo scanner, con i controlli su processi in background e attività pianificate attivi, ha dato esito pulito con 1.728 righe cron malevole presenti. Disrex non indica versione e orario dell’esecuzione. In ogni caso gli scanner puntati sulla sola cartella web non vedono un impianto installato un livello sopra.

Un segnale indiretto: raffiche inattese di email “Payment Transaction Failed Reminder” sono un motivo per indagare, anche se pagamenti rifiutati legittimamente generano la stessa notifica.

Mitigazioni disponibili ora

In assenza di una correzione del produttore, le opzioni si dispongono su tre livelli. L’ordine riflette la robustezza. La

ragione sta nella struttura della catena come la descrive Disrex. Il punto di ingresso non può essere protetto in modo pulito, perché il codice che elabora il testo iniettato prepara anche ogni email transazionale. Un controllo lì romperebbe le conferme d'ordine. Restano il blocco del punto di uscita e le misure che non dipendono dalla conoscenza della vulnerabilità.

Livello 1: misure di sistema indipendenti dalla catena. Sono le due misure che Disrex mette davanti a ogni altra regola del proprio repository. Su uno dei negozi compromessi le prime quattro delle sei funzioni tentate dal dropper erano disabilitate, ma proc_open no. È stata quella ad avviare l'impianto, e open_basedir non ha contenuto il processo figlio, che una volta avviato gira fuori da PHP. Una lista disable_functions che ferma cinque funzioni su sei non ferma nulla. Le misure sono quindi due: aggiungere proc_open e le altre cinque funzioni a disable_functions in PHP; montare /tmp, /var/tmp e /dev/shm con opzione noexec, così che un binario scaricato non possa essere eseguito. Entrambe vanno provate in staging, perché alcune estensioni legittime dipendono da proc_open.

Livello 2: disattivazione temporanea di GraphQL. È il consiglio interinale di Sansec per i negozi che non usano Shield. Le vetrine headless e PWA richiedono GraphQL; la maggior parte delle vetrine classiche e Hyvä non lo usa. Un conteggio delle richieste POST /graphql nei log di accesso dice se la disattivazione è praticabile. È la strada scelta anche da alcuni fornitori di hosting. Nexcess, marchio di Liquid Web, ha comunicato la sera del 5 settembre un blocco completo delle richieste GraphQL su tutti gli ambienti Magento e Adobe Commerce gestiti. Ha avvertito che le vetrine headless e PWA possono risentirne. Un esercente italiano su hosting gestito potrebbe quindi trovarsi con GraphQL bloccato senza averlo deciso: va verificato con il fornitore.

Livello 3: patch e regole non ufficiali. Tre soggetti hanno

pubblicato mitigazioni il 5 settembre:

- **Disrex Group** ha rilasciato regole nginx e Apache che bloccano le richieste con i parametri dell'exploit nella stringa di query. Ha pubblicato inoltre un guard per tre metodi degli scanner di dependency injection di Magento (ArrayScanner, ClassesScanner, XmlInterceptorScanner), che impedisce loro di girare fuori dalla riga di comando. Il guard è disponibile anche come patch composer che si riapplica a ogni deploy. Secondo Disrex, la patch è stata verificata in prova a secco sui tag 2.4.6, 2.4.7, 2.4.8 e 2.4.9. È stata poi applicata su installazioni attive 2.4.7-p2 e 2.4.8-p4, con setup:di:compile e vetrina funzionanti. Il repository dichiara i propri limiti. È stato scritto con assistenza di intelligenza artificiale durante un incidente e non è stato revisionato. Le regole Apache non sono mai state eseguite su un server Apache reale; nulla è stato provato su hosting condiviso, Docker o pannelli di controllo. Una prova su un negozio attivo ha mostrato che le regole per il server web si aggirano spostando gli stessi parametri nel corpo di una POST. Fermano la campagna come è condotta oggi, non la vulnerabilità. Disrex avverte che il guard non è una correzione completa, perché in Magento esistono altri punti di include non proteggibili allo stesso modo.
- **ProxiBlue** ha pubblicato su GitHub tre patch con lo stesso guard sugli stessi tre metodi, in modo indipendente. Disrex le cita come conferma incrociata e raccomanda di applicare una sola delle due soluzioni.
- **Graycore LLC** ha pubblicato un modulo Magento su GitHub e Packagist che irrobustisce tre punti della catena. La direttiva {{block}} dei template email rifiuta i blocchi di backend. Il generatore di URL delle righe di griglia valida la classe prima di costruirla. I tag di apertura PHP nei rapporti di errore delle Web API vengono spezzati. L'ultima release (2.0.0) aggiunge il rifiuto

di template caricati da stream wrapper o percorsi con byte nullo e rimuove un override di un resolver PayPal presente nella prima versione. L’autore avverte che il pacchetto cambierà mentre prova nuovi livelli di mitigazione e che va letto il changelog prima di ogni aggiornamento. Precisa che si tratta di irrobustimento e non di correzione: la vulnerabilità resta aperta.

Va detto con chiarezza: né Sansec né Adobe hanno confermato che gli scanner DI siano il punto in cui la catena termina. Chi applica queste patch lo fa a proprio rischio, sulla base di una ricostruzione indipendente non convalidata dal produttore.

Se il negozio è già compromesso. La guida di pulizia di Disrex fissa un ordine. Prima confermare e preservare le prove; poi contenere, cercare altre persistenze e ruotare i segreti; quindi decidere se pulire o ricostruire. Le mitigazioni si applicano solo alla fine. Tre regole contano più delle altre. Rimuovere la voce cron prima di terminare il processo, perché il processo la ripristina. Non riavviare la macchina, perché la copia in /proc può essere l’unico binario superstite. Non eseguire composer install come pulizia, perché sovrascrive le date di modifica che documentano cosa è stato toccato.

La ricerca va estesa oltre il crontab: timer systemd anche per utente, at, auto_prepend_file, file PHP in pub/media, chiavi SSH sconosciute. Dentro Magento: account amministrativi non riconosciuti, integrazioni e token API, script iniettati in core_config_data, blocchi CMS modificati. Segue lo svuotamento delle sessioni, che l’impianto leggeva. Vanno poi ruotate la crypt/key in app/etc/env.php, la password del database, tutte le password amministrative, le chiavi API dei fornitori di pagamento e ogni altra credenziale in quel file. Disrex indica anche quando la ricostruzione è preferibile alla pulizia. I casi sono quattro: negozio non sotto controllo di versione; file modificati non spiegabili in vendor/ o app/code/; impianto attivo per giorni; impossibilità di datare l’inizio

della compromissione.

Un ultimo elemento sulla campagna. L’elenco di Disrex, nella revisione più recente, conta 27 indirizzi sorgente distinti su tre negozi e due ondate. La cifra era 26 nella prima versione ed è stata corretta; il README dello stesso repository riporta 28. Tre indirizzi appartengono a infrastrutture di hosting che inviavano in massa; il resto è un insieme di proxy residenziali con due-sei richieste ciascuno. Disrex raccomanda di non bloccare in blocco gli indirizzi residenziali: sono connessioni di consumatori affittate o compromesse, e il blocco costerebbe clienti reali. Bloccare il solo indirizzo citato nell’avviso Sansec avrebbe fermato meno di un quarto del traffico osservato. È un dato di una sola fonte, ma è coerente con l’infrastruttura distribuita tipica delle campagne di web skimming degli ultimi anni.

Riquadro giuridico: cosa scatta quando la compromissione è probabile ma non accertata

Questa sezione ha finalità informativa e non sostituisce il parere di un legale.

Notifica al Garante. L’articolo 33 del GDPR impone al titolare di notificare la violazione all’autorità di controllo senza ingiustificato ritardo e, ove possibile, entro 72 ore da quando ne è venuto a conoscenza. L’obbligo cade se è improbabile che la violazione presenti un rischio per i diritti e le libertà delle persone fisiche. Il nodo, in una situazione come questa, è il momento della “conoscenza”. Le Linee guida 9/2022 dell’EDPB (versione 2.0, adottata il 28 marzo 2023) lo definiscono al paragrafo 31. Il titolare è a conoscenza quando ha un ragionevole grado di certezza che un incidente di sicurezza abbia compromesso dati personali. Il paragrafo 34 ammette un breve periodo di indagine iniziale,

durante il quale il titolare non è ancora “a conoscenza”. Il paragrafo 36 precisa però che le verifiche preliminari vanno completate poco dopo il primo allarme, e possono richiedere più tempo solo in casi eccezionali. Il paragrafo 40 aggiunge che chi non agisce tempestivamente e poi scopre la violazione rischia la contestazione di mancata notifica. L’esempio 3 del paragrafo 33 descrive il caso di specie: un titolare rileva una possibile intrusione, controlla i sistemi, conferma la compromissione dei dati e da quel momento è a conoscenza.

Tradotto in pratica. Un esercente che rileva gli indicatori sopra descritti ha superato la soglia, perché l’impianto leggeva le sessioni e aveva accesso a credenziali e chiavi di pagamento; le 72 ore partono da lì. Un esercente che non rileva indicatori, ma non può escludere l’esposizione tra il 4 settembre e l’applicazione delle mitigazioni, si trova nella fase di indagine. Deve avviarla subito, documentarla con date e attività svolte e chiuderla in fretta. L’articolo 33, paragrafo 5, impone di registrare ogni violazione; il paragrafo 125 delle Linee guida chiede di motivare per iscritto anche la decisione di non notificare. L’articolo 33, paragrafo 4, consente la notifica per fasi. Il paragrafo 57 delle Linee guida indica proprio gli incidenti informatici che richiedono analisi forense come il caso tipico in cui il quadro completo manca nelle 72 ore. Il paragrafo 60 ricorda infine che non c’è sanzione per aver notificato un incidente che poi si rivela non essere una violazione.

Comunicazione agli interessati. L’articolo 34 richiede la comunicazione agli interessati quando la violazione è suscettibile di presentare un rischio elevato. La compromissione di un server che gestisce pagamenti rientra quasi sempre in questa categoria. L’esempio ii dell’allegato B delle Linee guida (attacco a un servizio online con esfiltrazione di dati) prevede la notifica all’autorità e, a seconda della gravità, agli interessati. L’esenzione per dati cifrati (paragrafi 76-78) vale solo se la chiave non è stata

compromessa. È da escludere quando la crypt/key di Magento era leggibile dall'utente del sito con cui girava l'impianto. Le stesse Linee guida avvertono che, se la chiave risulta compromessa in un secondo momento, l'obbligo di notifica si riattiva.

Ripartizione delle responsabilità. Nella filiera tipica di un e-commerce italiano l'esercente è titolare del trattamento. L'agenzia che gestisce l'installazione e il fornitore di hosting sono di regola responsabili ai sensi dell'articolo 28. Due conseguenze pratiche. La prima: l'articolo 33, paragrafo 2, obbliga il responsabile a informare il titolare senza ingiustificato ritardo. Il paragrafo 44 delle Linee guida chiarisce che il responsabile non deve prima valutare il rischio; deve solo accertare che la violazione c'è stata e avvisare. Da quel momento il titolare è considerato a conoscenza. Un'agenzia che ha visto gli indicatori sui server dei clienti e non lo ha comunicato ha quindi una responsabilità autonoma. L'esempio vii dell'allegato B (fornitore di hosting che scopre un difetto nel proprio codice) mostra che l'obbligo scatta anche prima di aver accertato lo sfruttamento effettivo.

La seconda: il contratto di servizio è il luogo in cui si decide chi doveva applicare le mitigazioni e in quanto tempo. Molti contratti prevedono l'applicazione delle patch ufficiali entro un certo termine, ma tacciono sul caso in cui una patch ufficiale non esista. Chi decide se applicare una patch di terze parti non revisionata? Chi risponde se quella patch rompe il negozio, o se la mancata applicazione lo espone? Sono domande che i contratti in essere raramente affrontano e che questa vicenda rende urgenti. L'articolo 82, sul diritto al risarcimento, ripartisce la responsabilità civile tra titolare e responsabile in base al rispettivo contributo al danno.

PCI DSS. Per chi è soggetto allo standard, la versione 4.0.1 richiede al requisito 6.3.3 che le patch per le vulnerabilità critiche siano installate entro un mese dal rilascio. Il

requisito presuppone una patch. In sua assenza operano il requisito 6.3.1 (identificazione delle vulnerabilità tramite fonti riconosciute, compresi gli avvisi dei CERT, e loro classificazione per rischio) e i controlli compensativi. I requisiti 6.4.3 e 11.6.1 sono qualificati dallo standard come buone pratiche fino al 31 marzo 2025 e obbligatori da quella data. Il primo impone autorizzazione, integrità e inventario di tutti gli script caricati nelle pagine di pagamento. Il secondo richiede un meccanismo che rilevi le modifiche non autorizzate alle intestazioni HTTP e agli script così come li riceve il browser del cliente, eseguito almeno ogni settimana. Sono esattamente i controlli che intercetterebbero uno skimmer inserito dopo una compromissione via StyleSmuggler. Il requisito 12.10.1 impone infine un piano di risposta agli incidenti, da attivare al primo indicatore.

NIS2. Il singolo negozio online non rientra tra i soggetti del D.Lgs. 138/2024. L'allegato II del decreto include invece i fornitori di mercati online tra i fornitori di servizi digitali. Un fornitore di hosting che gestisce molti e-commerce può rientrare nel perimetro come fornitore di servizi cloud o di data center (allegato I, settore 8, infrastrutture digitali). Un'agenzia che amministra le installazioni Magento dei clienti può rientrarvi come fornitore di servizi gestiti o di servizi di sicurezza gestiti (allegato I, settore 9, gestione dei servizi TIC business-to-business). Restano fermi i criteri dimensionali e la registrazione presso l'ACN. In quel caso l'articolo 25 prevede tre adempimenti per gli incidenti significativi. Una pre-notifica al CSIRT Italia entro 24 ore dalla conoscenza; una notifica entro 72 ore con valutazione iniziale e indicatori di compromissione; una relazione finale entro un mese. Il comma 3 precisa che la notifica non espone chi la effettua a una responsabilità maggiore di quella derivante dall'incidente. I commi 9 e 10 prevedono la comunicazione ai destinatari dei servizi sia degli incidenti sia delle minacce significative, con le misure di mitigazione adottabili. Per un fornitore in perimetro,

StyleSmuggler è una minaccia significativa da comunicare ai clienti anche prima di qualsiasi compromissione accertata.

Cosa aspettarsi dall’8 settembre

La release di sicurezza di Adobe dell’8 settembre è il primo momento in cui il quadro può cambiare. Tre scenari sono possibili. La release include una correzione con relativo CVE. Oppure non la include, e Adobe pubblica un bollettino fuori ciclo nei giorni successivi. Oppure Adobe pubblica prima una soluzione temporanea ufficiale e poi la patch. In ogni caso la finestra successiva va preparata ora.

Chi ha applicato patch non ufficiali deve pianificarne la rimozione prima di installare quella del produttore. La sovrapposizione di modifiche agli stessi file può produrre conflitti o, peggio, l’illusione di aver aggiornato; la patch composer di Disrex prevede un percorso di ritorno esplicito. Chi ha disattivato GraphQL deve avere pronto il piano di riattivazione, con prove funzionali sul pagamento. Chi ha pulito un negozio compromesso deve ricordare che la patch non rimuove una backdoor già installata né invalida credenziali già sottratte. La rotazione dei segreti resta necessaria anche dopo l’aggiornamento.

Resta il punto più scomodo. Un e-commerce compromesso tra il 4 settembre e l’applicazione delle mitigazioni può aver perso dati di pagamento senza che nulla lo segnali, perché l’impianto può operare senza traffico verso indirizzi noti. Per quegli esercenti l’8 settembre non chiude l’incidente. Apre la fase in cui bisogna decidere, con i log alla mano e con un legale al fianco, se e cosa notificare.

https://www.ictsecuritymagazine.com/notizie/zero-day-magento-stylesmuggler-adobe-commerce/

MikroTik RouterOS: la catena “MikroTrick” è sfruttata attivamente, patch urgente

Bastano un servizio SSH esposto su Internet e nessuna credenziale. Con questi soli requisiti, almeno dal 2 settembre, alcuni attaccanti prendono il controllo completo di dispositivi MikroTik RouterOS e lasciano come firma un account amministrativo chiamato ops. Lo ha confermato il 5 settembre 2026 CERT Polska, il team che ha scoperto le vulnerabilità e ha chiamato la catena “MikroTrick”. MikroTik ha pubblicato le versioni corrette su tutti i canali e raccomanda di aggiornare subito.

Cosa è successo

Il team polacco ha individuato sei vulnerabilità in RouterOS, due delle quali critiche, e ne ha coordinato la divulgazione. Combinandone due, un attaccante ottiene privilegi amministrativi completi senza autenticarsi, purché il dispositivo esponga SSH su rete pubblica. Le falle riguardano il server e il client SSH, il servizio bandwidth-test, la gestione dei certificati X.509 e l’interfaccia WebFig.

Secondo i record NVD, sono interessate tutte le release RouterOS 6.x precedenti alla 6.49.21, le 7.x precedenti alla 7.23.4 e la 7.24 precedente alla 7.24.2. I punteggi che seguono sono i CVSS 4.0 assegnati da CERT Polska, ente assegnatario delle CVE; NVD non ha ancora pubblicato una propria valutazione.

Le tre vulnerabilità principali sono:

- **CVE-2026-67276** (CVSS 9.2): elusione dell’autenticazione SSH. RouterOS non confrontava per intero la chiave pubblica RSA associata a un utente. Chi conosceva il nome utente e il modulo pubblico della chiave poteva costruire una chiave diversa e autenticarsi senza la chiave privata, con i privilegi dell’account bersaglio.
- **CVE-2026-86060** (CVSS 9.2): elevazione dei privilegi di sessione tramite un nome utente costruito ad arte. Il meccanismo di login SSH gestiva male i nomi utente che iniziano con un carattere non ammesso; l’attaccante poteva così modificare la maschera dei permessi e ottenere una sessione con pieni diritti amministrativi.
- **CVE-2026-67277** (CVSS 8.8): il servizio bandwidth-test permetteva a una connessione non autenticata di raggiungere uno stato riservato agli utenti autenticati. Insieme a due ulteriori difetti (esposizione di dati non inizializzati e integer underflow nella verifica delle dimensioni), consentiva di leggere memoria del kernel o di provocare un riavvio remoto del sistema.

L’elenco completo delle sei CVE è disponibile su una pagina dedicata di CERT Polska. Il team spiega di aver anticipato la pubblicazione per una ragione precisa: i pacchetti corretti erano già pubblici e il confronto tra le versioni aveva permesso alla comunità di ricostruire alcune correzioni. Non ha diffuso codice di exploit né dettagli utili ad automatizzare gli attacchi.

Gli attacchi osservati e gli indicatori di compromissione

CERT Polska dichiara di aver ricevuto conferma che la catena MikroTrick viene usata per assumere il controllo di dispositivi con SSH esposto e che le patch bloccano gli attacchi osservati. Gli attacchi riusciti, compresa la creazione dell’account ops, provengono dall’indirizzo IP

82.192.72.4 e risalgono almeno al 2 settembre. Un secondo indirizzo, 103.102.31.18, è stato usato in tentativi di sfruttamento della stessa catena.

Nei log di RouterOS l'attacco lascia due tracce riconoscibili: un accesso fallito per l'utente -2 via SSH, seguito dalla creazione di un nuovo utente da parte di ssh:-2@<ip>. La presenza di uno qualsiasi di questi elementi va indagata subito. La loro assenza, avverte il team, non esclude un'attività non autorizzata.

La risposta del vendor

Il bollettino di sicurezza MikroTik, datato 3 settembre 2026, definisce l'aggiornamento importante e afferma che la maggior parte delle configurazioni non è a rischio. Conferma che la correzione è inclusa in RouterOS 7.25beta3, 7.24.2, 7.23.4 e 6.49.21. Il 4 settembre è uscita anche la 7.23.5 per il ramo long-term: non aggiunge correzioni di sicurezza, ma risolve una regressione del DHCP IPv6 introdotta dalla 7.23.4, ed è quindi la versione a cui puntare. Il vendor non ha ancora pubblicato dettagli tecnici, per dare tempo agli amministratori di aggiornare. Per gli utenti domestici, scrive, il rischio non è immediato, ma consiglia comunque a tutti di aggiornare. CERT Polska riferisce inoltre che MikroTik, per la prima volta, ha inviato una notifica push agli utenti dell'app mobile ufficiale.

Le release corrette utilizzano anche il meccanismo Flagged, presente da tempo nel device-mode di RouterOS. All'avvio il sistema analizza la configurazione alla ricerca di segni noti di modifiche non autorizzate; disabilita le voci sospette, scrive un messaggio critico nel log e imposta lo stato "Flagged", descritto nella pagina dedicata del manuale. CERT Polska precisa però tre cose: il meccanismo rileva solo alcune tracce; l'assenza del marcatore non prova che il dispositivo sia integro; non si può escludere che il vendor abbia corretto

anche vulnerabilità ignote al team e non descritte nelle note di rilascio.

Perché conta per l’Italia

RouterOS è usato, in Italia come altrove, da operatori ISP e WISP, in reti aziendali e di PMI e in contesti edge. La catena colpisce esattamente i dispositivi che espongono SSH su Internet. Un dato di contesto, datato ma indicativo, viene dal rapporto Eclypsium del dicembre 2021: su oltre due milioni di dispositivi MikroTik con interfacce di gestione raggiungibili dalla rete, circa 300.000 risultavano vulnerabili a falle note. L’Italia figurava tra i cinque Paesi più interessati, con Cina, Brasile, Russia e Indonesia.

Al momento non risultano vittime italiane confermate pubblicamente e CSIRT Italia non ha pubblicato un avviso su questa catena. Lo stesso CSIRT era però già intervenuto sull’ecosistema MikroTik nell’ultimo anno. Il 29 ottobre 2025 ha pubblicato un alert sul componente WebFig di RouterOS e SwitchOS; il 17 gennaio 2025 un bollettino su una botnet di dispositivi MikroTik compromessi.

Un router o gateway compromesso offre all’attaccante persistenza a livello di rete, intercettazione o deviazione del traffico, tunnel verso l’esterno e movimento laterale verso i sistemi interni. È il modello, ormai documentato, dei dispositivi di rete usati come infrastruttura offensiva. Va aggiunta una considerazione normativa: gli obblighi di notifica NIS2 al CSIRT Italia sono operativi dal 15 gennaio 2026, quindi un dispositivo di frontiera compromesso può diventare rapidamente un incidente da segnalare per i soggetti in perimetro.

Cosa fare subito

1. **Aggiornare senza attendere** a RouterOS 7.24.2 (stable),

7.23.5 (long-term; la 7.23.4 contiene la stessa correzione ma soffre della regressione DHCP IPv6), 6.49.21 (long-term v6) o 7.25beta3, o a versioni successive.
2. **Dopo l’aggiornamento**, controllare nel log l’eventuale messaggio di compromissione e il valore del marcatore Flagged con /system/device-mode/print. Verificare la presenza di utenti sconosciuti (in particolare ops), script, attività dello scheduler, proxy e tunnel non riconosciuti.
3. **Se la patch non è applicabile subito**, disattivare i servizi esposti o limitarne l’accesso alle sole reti di gestione fidate, in particolare SSH, WWW/WWW-SSL e il server bandwidth-test. Evitare di avviare connessioni TLS dal dispositivo non aggiornato e di usare i client SSH integrati (/system ssh e /system ssh-exec) verso host o reti non fidati. Sono misure temporanee che non sostituiscono la patch.
4. **In caso di compromissione sospetta o confermata**, isolare il dispositivo e mettere al sicuro log e configurazione prima di qualsiasi ripristino. Segnalare l’evento al CSIRT competente. Ripristinare le impostazioni di fabbrica, riconfigurare da una configurazione verificata e cambiare password, chiavi e altri segreti. Non ripristinare alla cieca una copia di sicurezza completa proveniente da un dispositivo potenzialmente compromesso; non cancellare lo stato Flagged prima di aver completato l’analisi.

Vulnerabilità trovate con l’aiuto degli LLM

Un dettaglio di metodo merita attenzione. CERT Polska riferisce di aver individuato le vulnerabilità con il supporto dei modelli GPT-5.5-cyber e GPT-5.6-sol, disponibili al team nell’ambito del programma OpenAI Government and Trust Agency

Collaboration. I modelli hanno lavorato in un ambiente di ricerca agentico su un laboratorio isolato. Hanno automatizzato la creazione delle macchine di prova, il confronto tra versioni e l'analisi di RFC e codice binario. Particolarmente efficace, secondo il team, è stata la modellazione dei protocolli come macchine a stati, per verificare cosa accade quando una fase viene saltata, ripetuta o eseguita nell'ordine sbagliato.

Ogni ipotesi ha però richiesto conferma su sistemi RouterOS reali, test di controllo negativi e valutazione d'impatto da parte dei ricercatori. I modelli, conclude CERT Polska, hanno accelerato l'analisi senza sostituire la verifica umana. È una lezione coerente con quella portata al Forum ICT Security 2025 dal Politecnico di Torino sugli agenti AI applicati all'individuazione delle CVE: l'architettura dell'agente e il controllo umano pesano più del modello.

https://www.ictsecuritymagazine.com/notizie/mikrotik-mikrotrick/

Panzer, il nuovo ransomware-as-a-service che ha già colpito due aziende italiane

Emerso ad agosto 2026, Panzer unisce un programma di affiliazione già maturo, payload per Windows, Linux, ESXi e FreeBSD e doppia estorsione. Tra le prime vittime rivendicate figurano Doimo Cucine e NTE Italia.

Un esordio rapido e strutturato

Nel panorama ransomware la comparsa di una nuova sigla non è, di per sé, una notizia: il ricambio è continuo e molti operatori scompaiono nel giro di poche settimane. Panzer merita attenzione per un motivo diverso: la rapidità con cui ha messo in piedi una piattaforma completa. Secondo il profilo pubblicato il 4 settembre 2026 dalla società di threat intelligence CyberXtron, il leak site del gruppo sulla rete Tor è attivo dal 5 agosto, data della prima osservazione, e in un mese ha pubblicato 16 vittime in almeno 11 Paesi.

La distribuzione geografica non ha un baricentro: tre vittime in Thailandia, due in Italia, due in Indonesia, due in Serbia e una sola in ognuno degli altri Paesi colpiti, cioè Curaçao, Corea del Sud, Spagna, Repubblica Ceca, Germania, Nigeria e Svizzera. Anche i settori sono eterogenei: la tecnologia è in testa con quattro casi, seguita dalla manifattura con tre; poi governo e difesa, agroalimentare, energia e utility, istruzione e retail. È il profilo tipico di una campagna opportunistica, guidata dagli accessi che gli affiliati riescono a procurarsi e non da una scelta mirata di Paesi o settori.

Il conteggio, peraltro, è già superato. Tre tracker indipendenti, SOCRadar, Ransomware.live e l'italiano Ransomfeed, concordano su 19 vittime al 6 settembre. Le ultime due rivendicazioni, del 3 e del 4 settembre, riguardano un ente pubblico indonesiano e un'università tedesca; dopo le 17 rivendicazioni di agosto il ritmo è rallentato, ma il gruppo resta attivo.

Sul numero di Paesi colpiti, invece, i tracker divergono: Ransomware.live ne conta 11, SOCRadar 16, e negli elenchi compaiono Portogallo, Francia e Cipro, assenti dal censimento di CyberXtron. Le differenze dipendono dai criteri di attribuzione geografica, e una stessa vittima può essere

attribuita a Paesi diversi a seconda della piattaforma: il gruppo orologiero Festina, per esempio, è registrato in Spagna da Ransomfeed e in Svizzera da SOCRadar. È un’avvertenza da tenere presente ogni volta che si citano questi numeri. SOCRadar indica inoltre la Russia come Paese di origine del gruppo: un’attribuzione coerente con il divieto di colpire i Paesi della CSI, ma non confermata da altre fonti.

Le vittime italiane

Le due organizzazioni italiane sono censite, con dati coincidenti, sia da Ransomware.live sia da Ransomfeed. La prima è Doimo Cucine, produttore di cucine di design con sede a Nervesa della Battaglia, in provincia di Treviso: la rivendicazione è stata rilevata il 17 agosto e, secondo la scheda di Ransomware.live, il gruppo dichiara di aver esfiltrato 30 GB di dati. La seconda è NTE Italia, società di servizi tecnici e di ingegneria per le telecomunicazioni e la sicurezza dei cantieri, con sede a Catanzaro e uffici a Roma e Catania: è comparsa sul leak site il 21 agosto e il gruppo afferma di aver sottratto 16 GB di dati, descritti come migliaia di documenti sensibili.

Una pubblicazione su un leak site resta una rivendicazione dell’attaccante, non una prova. Nessuna delle due aziende ha confermato pubblicamente l’incidente, e le rivendicazioni di un gruppo appena emerso, privo di uno storico verificabile, vanno trattate con particolare cautela fino a verifica indipendente. La comparsa di due vittime italiane in pochi giorni richiama una dinamica ricorrente nei RaaS di recente formazione: concentrare più bersagli in un breve intervallo per costruire credibilità presso i potenziali affiliati.

Come funziona il programma di

affiliazione

Il tratto più interessante di Panzer non è il malware, di cui si sa ancora poco, ma l'infrastruttura commerciale. L'operazione si presenta come un RaaS semi-aperto: gli aspiranti affiliati presentano domanda tramite un canale Tox e ottengono l'accesso al pannello di controllo solo dopo una selezione. La ripartizione dei proventi è 80/20 a favore dell'affiliato, con la quota della piattaforma trattenuta automaticamente su ogni pagamento. Le regole vietano di colpire organizzazioni nei Paesi della CSI e soggetti coinvolti in abusi su minori; gli account inattivi per più di una settimana vengono disattivati.

Il pannello di controllo descritto da CyberXtron è quello di un servizio già rodato. Comprende il monitoraggio del saldo e delle transazioni, la gestione delle build per le diverse piattaforme, con una chat di negoziazione dedicata a ogni vittima e un generatore integrato di fatture in Bitcoin, un sistema di ticket di assistenza e sotto-account per i team. La pubblicazione dei leak segue un flusso con approvazione preventiva degli operatori, conto alla rovescia e possibilità di mettere in evidenza i casi ritenuti più redditizi. Gli operatori dichiarano inoltre di sottoporre i nuovi affiliati a un monitoraggio automatico durante il primo mese, per individuare ricercatori o forze dell'ordine infiltrati.

Un investimento di questo tipo in strumenti per il reclutamento, i pagamenti e l'autotutela indica un progetto pensato per crescere attraverso operatori esterni, non un gruppo chiuso che conduce in proprio le intrusioni. È la logica del ransomware human-operated portata alle estreme conseguenze: la piattaforma fornisce strumenti e infrastruttura, gli affiliati eseguono le intrusioni. Alcuni osservatori accostano il modello a quello di VanHelsing e di altri RaaS emersi negli ultimi anni.

Cosa si sa, e cosa non si sa, delle tecniche

Sul piano tecnico il quadro è frammentario e il vettore di accesso iniziale non è stato verificato in modo indipendente. Le attività che CyberXtron associa al gruppo, con confidenza medio-bassa, comprendono il dumping delle credenziali del sistema operativo, attacchi di forza bruta, la ricognizione dei servizi di rete, il movimento laterale tramite servizi remoti e account validi, la raccolta di dati dal sistema locale, l'esfiltrazione su protocolli alternativi e la disabilitazione o manomissione degli strumenti di sicurezza. La mappatura di SOCRadar, in parte diversa, aggiunge l'esecuzione tramite interprete di comandi, la persistenza all'avvio e l'inibizione del ripristino di sistema, tecnica che prende di mira le copie shadow e i meccanismi di recupero.

Non esistono al momento indicatori di compromissione confermati, siano essi hash, indirizzi IP, domini o campioni di malware: gli unici indicatori pubblici riguardano l'infrastruttura del leak site e il canale di reclutamento. Ciò che invece dichiarano gli stessi operatori è la copertura dei payload: build per Windows, Linux, VMware ESXi e FreeBSD, con oltre quindici comandi personalizzabili, funzioni anti-rilevamento e un pannello di avvio con monitoraggio in tempo reale dell'operazione. Lo schema è quello della doppia estorsione: i dati vengono esfiltrati prima della cifratura e alcuni campioni finiscono sul leak site come prova.

Perché è rilevante per le organizzazioni italiane

Due caratteristiche rendono Panzer una minaccia concreta per il tessuto produttivo nazionale, al di là dei due casi già noti.

La prima è il supporto a ESXi. Gran parte delle medie e grandi imprese italiane concentra i propri carichi di lavoro su infrastrutture virtualizzate: compromettere un hypervisor significa cifrare in un'unica operazione decine di macchine virtuali e i servizi che vi girano, con tempi di ripristino ben più lunghi rispetto a un attacco limitato agli endpoint Windows. Le build per Linux e FreeBSD allargano ulteriormente la superficie esposta negli ambienti con parchi server misti. Non è un rischio teorico: il CSIRT Italia, nel bollettino dedicato a Qilin, ha già segnalato l'uso di tecniche mirate agli ambienti VMware ESXi da parte di un gruppo il cui impatto sistemico sul Paese è classificato come critico.

La seconda è la doppia estorsione. Backup solidi e testati restano indispensabili per la continuità operativa, ma non neutralizzano il rischio legato alla pubblicazione dei dati sottratti, con le conseguenze reputazionali, contrattuali e regolamentari che ne derivano, comprese le notifiche previste dal GDPR e, per i soggetti che rientrano nel suo ambito di applicazione, dalla NIS2.

Il contesto aggrava il quadro. Secondo Ransomfeed, al 6 settembre le rivendicazioni contro organizzazioni italiane nel 2026 sono già 212, contro le 169 dell'intero 2025, su un totale storico di 850. Nella sola settimana tra il 6 e il 13 agosto, secondo una ricostruzione ripresa anche da Bismark.it, otto gruppi diversi hanno rivendicato 15 vittime italiane. Panzer si inserisce in una fase in cui gli attaccanti puntano soprattutto sulla manifattura, già segnalata dall'ACN come bersaglio privilegiato nei distretti industriali del Nord, e sui fornitori di servizi tecnologici e di telecomunicazione: ne sono esempi NTE Italia e Retelit, quest'ultima colpita da Qilin a fine luglio.

Le priorità difensive

Le raccomandazioni pubblicate da CyberXtron coincidono con le

buone pratiche già consolidate contro il ransomware, ma alcune assumono un peso particolare alla luce delle caratteristiche di Panzer.

Sul fronte delle identità, la priorità è l'autenticazione a più fattori resistente al phishing su tutti gli accessi remoti, le VPN e gli account privilegiati, unita al principio del minimo privilegio e a una revisione periodica degli account amministrativi: le tecniche associate al gruppo ruotano intorno al furto e all'abuso di credenziali valide. Sul fronte architetturale, la segmentazione deve isolare dalle reti utente i controller di dominio, i sistemi di backup e, soprattutto, l'interfaccia di gestione degli hypervisor, confinando i protocolli di amministrazione remota in segmenti dedicati e monitorati.

Contro la doppia estorsione servono controlli sul traffico in uscita e il rilevamento dei trasferimenti anomali di dati: l'esfiltrazione precede la cifratura ed è la finestra più utile per intercettare l'attacco. I backup devono essere immutabili, con almeno una copia offline, isolata dalla rete, e test di ripristino regolari che verifichino l'integrità dei punti di ripristino. Il piano di risposta agli incidenti, infine, dovrebbe prevedere esplicitamente lo scenario di esposizione dei dati, con i flussi legali, regolamentari e di comunicazione pronti prima che una rivendicazione compaia online.

Panzer è ancora in fase di espansione e il suo profilo tecnico potrebbe cambiare rapidamente man mano che nuovi affiliati introducono metodi propri. Proprio per questo va trattato come una minaccia credibile e trasversale, non come un fenomeno di nicchia.

https://www.ictsecuritymagazine.com/notizie/panzer-ransomware-vittime-italiane/

Pagamenti digitali in Italia continuano a crescere in Italia, i dati di Bankitalia

(Adnkronos) – Continuano a crescere in Italia i pagamenti alternativi al contante. Secondo i dati di Bankitalia, le transazioni digitali e senza contante segnano un incremento del 13%. L'88% dei pagamenti al POS è avvenuto avvicinando la carta, senza inserirla nel lettore. Il 24% dei pagamenti nei negozi fisici è stato avviato tramite dispositivi mobili. Le carte restano il principale strumento digitale e coprono circa tre quarti delle transazioni non in contanti.

–

multimedia/news-to-go

webinfo@adnkronos.com (Web Info)

The post Pagamenti digitali in Italia continuano a crescere in Italia, i dati di Bankitalia appeared first on Key4biz.

https://www.key4biz.it/pagamenti-digitali-in-italia-continuano-a-crescere-in-italia-i-dati-di-bankitalia/586322/

Comando e controllo su un

broker MQTT pubblico: le backdoor Toy Ghouls e il punto cieco delle reti industriali

Due impianti Windows analizzati da Kaspersky non usano infrastruttura propria per il comando e controllo: uno si appoggia al broker MQTT pubblico di HiveMQ, l'altro al protocollo Matrix. Le vittime documentate sono russe e non si tratta di un attacco a sistemi industriali, ma la tecnica solleva un problema di rilevamento che riguarda direttamente chi gestisce reti OT.

Due backdoor, nessuna infrastruttura propria

Il 4 settembre 2026 Kaspersky ha pubblicato su Securelist l'analisi di due backdoor che i suoi ricercatori attribuiscono al gruppo a movente finanziario noto come Toy Ghouls, indicato anche con i nomi Bearlyfy, Laboo.boo e Feral Wolf. L'attribuzione è dei team GERT e Security Services di Kaspersky e va riportata come tale.

Il gruppo è attivo dal 2025 contro organizzazioni russe. Ha iniziato appoggiandosi a strumenti disponibili pubblicamente e a *builder* di ransomware trapelati, quelli di Babuk e LockBit, per poi sviluppare codice proprio, tra cui il ransomware GenieLocker, in versioni per Windows, Linux ed ESXi.

I due impianti individuati all'inizio di luglio 2026 si chiamano mqtt-bird-agent 0.1.0 e matrix-bird-agent 0.1.0, e corrispondono a due eseguibili Windows, cplsupport.exe e wtass.exe. Il tratto che li rende interessanti è che nessuno

dei due porta con sé un server di comando e controllo: entrambi si appoggiano a servizi e protocolli legittimi.

Un secondo elemento merita attenzione, e riguarda il modo in cui arrivano sulla macchina. La distribuzione avviene attraverso Windows Remote Management, con utilità di post-sfruttamento *open source* come Evil-WinRM e WinRM-fs. Il report non descrive il vettore di accesso iniziale, ma la scelta di WinRM implica che gli operatori dispongono già di credenziali amministrative quando installano l'impianto: la backdoor non è il punto di ingresso, è ciò che rende l'accesso persistente.

La persistenza si ottiene registrandosi come servizio Windows, con i nomi "cplsupport" (descritto come "Problem Reports Control Panel") e "wtas" ("Windows Telemetry Aggregator Service"), tramite argomenti di installazione e disinstallazione. In entrambe le varianti i campi sensibili della configurazione vengono cifrati al primo avvio con ChaCha20-Poly1305, con una chiave derivata dal valore MachineGuid del registro di sistema: un accorgimento che impedisce di riutilizzare la configurazione su un'altra macchina e complica il lavoro di chi analizza il campione fuori dall'ambiente di origine. La variante MQTT conserva il file cifrato su disco; la variante Matrix lo cancella dopo il primo avvio e sposta i parametri in una chiave di registro dedicata.

Perché un broker MQTT pubblico è un canale quasi ideale

Secondo Kaspersky, la variante mqtt-bird-agent comunica con broker.hivemq.com sulla porta 8883, quella standard di MQTT su TLS. Il dominio è quello del broker pubblico di HiveMQ, uno dei più noti tra quelli messi a disposizione gratuitamente per test e prototipazione: non è un server compromesso né un dominio registrato dagli attaccanti. Il report descrive quattro percorsi di scambio: uno per lo stato del sistema

infetto, uno per le metriche periodiche (CPU, memoria, disco, carico, tempo di attività), uno da cui la backdoor legge i comandi e uno su cui restituisce i risultati. I comandi vengono eseguiti tramite PowerShell in modalità nascosta.

Un'annotazione di precisione: nello stesso passaggio Securelist parla di un "cluster" creato dagli attaccanti e di un piano gratuito con limiti di 100 connessioni concorrenti e 10 GB di traffico mensili. Sono esattamente le caratteristiche del piano Serverless gratuito di HiveMQ Cloud, che prevede credenziali di accesso e TLS, non quelle del broker demo aperto, che non ha né cluster né piani. Il report non scioglie l'ambiguità. La conseguenza pratica è che non si può dare per certo che il canale fosse privo di autenticazione, e che l'indirizzo effettivamente contattato potrebbe essere un sottodominio dedicato del servizio cloud piuttosto che il nome generico del broker pubblico: in entrambi i casi, comunque, la destinazione appartiene a un fornitore legittimo e ampiamente usato.

Vale la pena spiegare perché questa scelta sia efficace, al di là della gratuità.

MQTT è un protocollo *publish and subscribe*: il client non riceve connessioni in ingresso, si limita a stabilire una connessione uscente verso il broker e ad attendere. Non serve alcuna porta aperta sul perimetro, e il traffico ha la forma di una sessione TLS lunga verso un servizio cloud, esattamente come mille integrazioni legittime. I messaggi sono minuscoli, perché il protocollo nasce per la telemetria su reti a banda ridotta, quindi il volume non fa scattare soglie. E poiché il broker è condiviso da chiunque lo usi per test, l'indirizzo di destinazione non è un indicatore di compromissione: bloccarlo a livello di reputazione significa bloccare un servizio legittimo.

C'è però un motivo ulteriore, ed è quello che riguarda direttamente il nostro perimetro editoriale. MQTT è lo

standard di fatto della telemetria industriale e dell'*Internet of Things*. In una rete di stabilimento, in un impianto di trattamento acque, in una flotta di dispositivi connessi, il traffico MQTT in uscita non è un'anomalia: è il funzionamento normale. Un canale di comando e controllo che parla MQTT, in quel contesto, non si nasconde tra il rumore, coincide con il rumore.

Il caso Matrix, e la differenza che quasi nessuno sta facendo

La seconda variante usa il protocollo Matrix. L'agente invia messaggi di stato e metriche con tipi di evento personalizzati, riceve comandi preceduti dal prefisso cmd: da un account chiamato "panel-bot" e li esegue attraverso la riga di comando di Windows, non PowerShell. Accetta inoltre un'istruzione di configurazione che regola l'intervallo di invio delle metriche, in una finestra che va da cinque secondi a un'ora; il valore viene salvato nel registro. Le risposte tornano come eventi di un tipo dedicato.

Qui però la somiglianza con il caso MQTT si ferma, e conviene dirlo chiaramente perché diverse riprese della notizia stanno appiattendo le due cose. Il server Matrix utilizzato è sotto il controllo degli attaccanti, raggiungibile a un dominio che richiama il nome del client Element. Non c'è quindi abuso di un servizio di terzi: c'è l'uso di un protocollo legittimo e federato su infrastruttura propria. È mimetismo di protocollo, non parassitismo di servizio.

La distinzione conta per chi deve difendersi. Nel caso MQTT il dominio di destinazione è inutilizzabile come indicatore, perché legittimo; nel caso Matrix il dominio è un indicatore valido e bloccabile, ed è l'analisi comportamentale a essere più difficile, perché il traffico Matrix cifrato è opaco per definizione.

Un’osservazione di prospettiva, che è nostra e non del report. Matrix sta guadagnando terreno nelle amministrazioni pubbliche europee proprio come alternativa sovrana alle piattaforme di messaggistica statunitensi, e il tema della sovranità digitale ne sta accelerando l’adozione. Nelle organizzazioni che lo hanno introdotto, il traffico Matrix passerà presto dall’essere un’anomalia all’essere ordinario. Vale per Matrix ciò che vale già oggi per MQTT negli impianti industriali: la protezione offerta dall’anomalia statistica ha una scadenza, e coincide con il momento in cui la tecnologia diventa diffusa.

Che cosa questo caso non dice

Poiché l’angolo più interessante riguarda le reti industriali, è necessario essere precisi su ciò che il report documenta e su ciò che non documenta, altrimenti l’analisi si trasforma in allarmismo.

Non si tratta di un attacco a sistemi di controllo industriale. Gli impianti girano su Windows, vengono installati come servizi dopo che gli attaccanti hanno già ottenuto privilegi amministrativi, ed eseguono comandi tramite PowerShell o riga di comando: è una backdoor IT, non un *malware* progettato per interagire con logiche di processo. Le vittime documentate sono organizzazioni russe, e il report non individua settori specifici né fornisce un numero di vittime.

Il motivo di interesse non è quindi ciò che è accaduto, è ciò che la tecnica rende possibile. Un canale di comando e controllo su MQTT esiste ora, è documentato e funziona. Il giorno in cui un impianto simile venisse installato su un sistema di una rete dove MQTT è traffico normale, la difficoltà di rilevamento non sarebbe teorica. È un’osservazione prospettica, e come tale va presa: nessuna delle fonti disponibili afferma che ciò sia già avvenuto.

Detection: dove guardare davvero

Il primo punto è il più controintuitivo, e riguarda la priorità. Poiché la distribuzione avviene via WinRM con strumenti noti, il momento in cui questa catena è più visibile non è quello del comando e controllo, è quello che lo precede. Sul lato Windows i segnali sono ordinari e già disponibili: la registrazione di un nuovo servizio (evento 7045 nel registro System), l'attività nel canale Microsoft-Windows-WinRM/Operational su macchine che non dovrebbero ricevere sessioni remote, connessioni in ingresso sulle porte 5985 e 5986 da postazioni che non appartengono al team di amministrazione, e la scrittura nelle directory cplsupport e SynapseAgent sotto ProgramData. Chi ha un buon rilevamento sulle attività amministrative anomale intercetta la catena prima che il canale MQTT entri in funzione, e per la maggior parte delle organizzazioni questo è l'investimento con il ritorno migliore.

Sul canale, le indicazioni utili sono di natura diversa da quelle a cui si è abituati, e vanno lette con la mappa dell'impianto davanti.

In un'architettura industriale ordinata, i client MQTT stanno in posti prevedibili: gateway e dispositivi edge al livello 2 del modello Purdue, sistemi di raccolta dati e historian al livello 3, un broker in zona demilitarizzata industriale che fa da unico punto di contatto verso l'IT e il cloud. Una connessione MQTT in uscita da una postazione Windows di livello 3 o 4, e a maggior ragione da un server di dominio o da una workstation di ufficio, non ha una funzione operativa che la giustifichi. Il primo strumento non è quindi un sensore di rete: è l'inventario dei client autorizzati, per zona, con l'indicazione del broker a cui ciascuno deve parlare. È lo stesso principio di zone e conduit di IEC 62443-3-2 applicato a un protocollo che spesso resta fuori dal disegno perché considerato "solo telemetria".

Il secondo strumento sono i *topic*. Negli impianti che adottano la specifica Sparkplug B, la struttura è fissa: spBv1.0/<gruppo>/<tipo di messaggio>/<nodo>/<dispositivo>, con un vocabolario chiuso di tipi (NBIRTH, DDATA, NDEATH e pochi altri) e payload binario Protobuf. Anche dove Sparkplug non è in uso, quasi tutti gli impianti hanno una convenzione di denominazione per sito, linea e macchina. Un topic che non appartiene a nessuna di queste gerarchie, con un identificativo opaco al primo livello e percorsi come status, metrics, cmd/req e cmd/res, è riconoscibile a colpo d'occhio da chi conosce la convenzione, molto più di quanto lo sia dal volume o dalla frequenza.

Qui però va detto un limite che il report non discute. Sulla porta 8883 il traffico è cifrato in TLS, e i topic non sono visibili a chi ascolta passivamente sulla rete. Le piattaforme di monitoraggio OT che analizzano il traffico da una porta SPAN riconoscono MQTT in chiaro sulla 1883 e ne estraggono i topic; su TLS vedono l'handshake, il nome del server nell'estensione SNI, il certificato e il ritmo delle sessioni, nient'altro. La visibilità sui topic si recupera in due modi soltanto: sul broker, se è interno e registra le connessioni e le sottoscrizioni, oppure sull'host, con la telemetria di processo che mostra quale eseguibile ha aperto la connessione. Nel caso in esame, un servizio registrato come "Problem Reports Control Panel" che apre una sessione TLS persistente verso un broker MQTT è un'anomalia leggibile sull'endpoint anche senza vedere un solo byte del contenuto.

Resta il ritmo, che è osservabile anche in TLS. Un sensore industriale pubblica su evento o per eccezione, con cadenze legate al processo; un impianto di comando e controllo invia metriche a intervallo fisso e interroga il canale con regolarità. Sulla durata di ore o giorni, la differenza è misurabile.

Sul piano architetturale la misura che chiude il problema alla radice è nota e poco applicata: i dispositivi di una rete

industriale dovrebbero parlare con un broker interno, con autenticazione mutua tramite certificati, e la zona OT non dovrebbe avere alcuna via verso broker pubblici su internet. L'eccezione tipica, e pericolosa, è la manutenzione remota dei fornitori di macchine, che spesso passa proprio da broker cloud del costruttore: va censita e incanalata attraverso la zona demilitarizzata, non tollerata come traffico di fondo. È lo stesso ragionamento di segmentazione tra IT e OT che si applica agli altri protocolli di campo. Il caso polacco della APN privata che nessuno sorvegliava ha mostrato poche settimane fa che cosa succede quando un canale considerato tecnico e innocuo resta fuori dal perimetro di monitoraggio.

Restano infine gli indicatori pubblicati nel report: le *hash* dei due eseguibili, le chiavi di registro sotto HKLM\Software\synapse e HKLM\Software\SynapseAgent, e la richiesta verso un servizio pubblico di geolocalizzazione degli indirizzi IP che entrambe le varianti effettuano all'avvio. Quest'ultima, per un sistema che non ha alcuna ragione di sapere in quale Paese si trova, è un segnale semplice e sorprendentemente efficace.

Dove sta andando il cybercrime

Il dato di fondo di questa vicenda va oltre i due campioni analizzati. Un gruppo criminale che rinuncia a infrastruttura propria in favore di servizi e protocolli legittimi ottiene tre vantaggi in una volta sola: costi di gestione quasi nulli, resistenza ai *takedown*, perché non c'è un server da sequestrare, e una superficie di attribuzione molto più povera, perché non ci sono registrazioni di dominio, hosting o pagamenti da seguire a ritroso.

È la stessa direzione che si osserva nell'abuso di piattaforme cloud, di servizi di *collaboration* e di repository pubblici come canali di comando e controllo. La differenza, qui, è che il servizio scelto appartiene al mondo della telemetria

industriale, un’area in cui la maturità del monitoraggio è mediamente inferiore a quella dell’IT e in cui l’analisi del traffico è ostacolata da vincoli di disponibilità che non ammettono interruzioni.

Per chi difende, la lezione non è aggiungere un dominio a una lista di blocco. È accettare che l’elenco dei canali plausibili per un comando e controllo coincide ormai con l’elenco dei servizi che l’organizzazione usa legittimamente, e costruire il rilevamento su questa premessa invece che sulla reputazione delle destinazioni.

https://www.ictsecuritymagazine.com/notizie/mqtt-matrix-command-control-backdoor-toy-ghouls/

Speciale Adnkronos dal Forum Teha di Cernobbio – Diretta 4

(Adnkronos) – Speciale Adnkronos dal Forum Teha di Cernobbio – Diretta 3, sabato 5 settembre ore 17.

–

economia

webinfo@adnkronos.com (Web Info)

The post Speciale Adnkronos dal Forum Teha di Cernobbio – Diretta 4 appeared first on Key4biz.

https://www.key4biz.it/speciale-adnkronos-dal-forum-teha-di-cernobbio-diretta-4/586324/

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/

Fiere, Vicenzaoro: Cavalieri (CIBJO): "Vicenza fondamentale per gioielleria e bellezza"

(Adnkronos) – "Per l'inizio delle celebrazioni dei 100 anni di CIBJO – Confederatione mondiale della gioielleria siamo partiti da Parigi, dove la Confederazione nacque nel 1926, ma, essendo l'attuale presidenza italiana, abbiamo scelto Vicenza come sede più idonea per continuare a celebrare questi 100 anni anche in Italia. Vicenza non è solo una città di fondamentale importanza nel campo della gioielleria, ma anche della bellezza in generale". È quanto affermato da Gaetano Cavalieri, presidente di CIBJO – Confederatione mondiale della gioielleria, intervenendo al congresso organizzato dalla Confederazione a Vicenzaoro September 2026 in occasione dell'importante anniversario.

–

economia

webinfo@adnkronos.com (Web Info)

The post Fiere, Vicenzaoro: Cavalieri (CIBJO): “Vicenza fondamentale per gioielleria e bellezza” appeared first on Key4biz.

https://www.key4biz.it/fiere-vicenzaoro-cavalieri-cibjo-vicenza-fondamentale-per-gioielleria-e-bellezza/586326/

Fiere, Vicenzaoro: Giordini (Giordini Gioielli): “Padiglione 2 è nuova casa manifattura italiana”

(Adnkronos) – “Siamo qui a Vicenzaoro nella nuova Hall 2, e ne siamo molto felici: finalmente abbiamo una nuova casa, la casa della manifattura italiana. Per questo, siamo entusiasti di presentare le nostre nuove collezioni ai numerosissimi clienti presenti alla fiera. Giardini Gioielli, in particolare, presenta collezioni classiche per il mercato degli Emirati Arabi e per il mercato americano, ma anche per gli europei. Siamo qui anche con la nuova collezione Elastique, in oro nove carati; si tratta di collane elastiche, che vendiamo solo online. Il nostro intento è far indossare i nostri gioielli a una clientela contemporanea, che vede il gioiello non come pezzo iconico da indossare una volta ogni tanto per l’occasione, ma quotidianamente, sentendosi a proprio agio con

le nostre collezioni”. Lo ha riferito Giordana Giordini, Titolare Giordini Gioielli di Arezzo, presidente della sezione Oreficeria di Confindustria Toscana Sud e presidente di Confindustria Toscana Sud, alla seconda giornata di Vicenzaoro September 2026, il salone internazionale dell’oreficeria e della gioielleria firmato IEG – Italian Exhibition Group, in programma dal 4 all’8 settembre nel quartiere fieristico di Vicenza.

–

economia

webinfo@adnkronos.com (Web Info)

The post Fiere, Vicenzaoro: Giordini (Giordini Gioielli): “Padiglione 2 è nuova casa manifattura italiana” appeared first on Key4biz.

https://www.key4biz.it/fiere-vicenzaoro-giordini-giordini-gioielli-padiglione-2-e-nuova-casa-manifattura-italiana/586328/

Fiere, Vicenzaoro: Renai (Cammilli Gioielli): “anche a settembre è vetrina fondamentale”

(Adnkronos) – “Anche a settembre Vicenzaoro presenta una platea fondamentale, anche se la stagione è avanzata. Noi supportiamo la clientela presentando delle collezioni anche in questa seconda parte dell’anno. Abbiamo, come sempre, uno stile, quindi le nostre unità vanno sempre nella direzione di

una precisa espressione dei volumi e dei colori dell’oro, per cui è famosa la nostra azienda”. Ad affermarlo Riccardo Renai, ceo Cammilli Gioielli, in occasione di Vicenzaoro September 2026, il salone internazionale dell’oreficeria e della gioielleria di IEG – Italian Exhibition Group, in programma dal 4 all’8 settembre nel quartiere fieristico di Vicenza.

–

economia

webinfo@adnkronos.com (Web Info)

The post Fiere, Vicenzaoro: Renai (Cammilli Gioielli): “anche a settembre è vetrina fondamentale” appeared first on Key4biz.

https://www.key4biz.it/fiere-vicenzaoro-renai-cammilli-gioielli-anche-a-settembre-e-vetrina-fondamentale/586330/