



t-defence

Oltre la velocità di patching

Da N-day a N-hour: i modelli
linguistici e l'economia dello
sfruttamento delle vulnerabilità

#TDefenceBusiness

AI4Cyber

Summary

Abstract	04
1. Da WannaCry a Firefox 148: cosa è cambiato in nove anni	05
2. Il collo di bottiglia che proteggeva la coda lunga	08
3. L'evidenza: dalla fattibilità alla misura	11
4. Fuori dai modelli protetti	15
5. Il time-to-exploit e la remediation	16
6. Dal laboratorio al campo	20
7. La coda lunga diventa superficie	22
8. Il contesto europeo e italiano	25
9. Conclusioni	27
Bibliografia e fonti	29

AI4Cyber

AI4Cyber studio è il nuovo spazio di ricerca applicata di T-Defence, dedicato alla ricerca e applicazione dell'**Intelligenza Artificiale in ambito cybersecurity**.

Al suo interno, diversi specialisti con competenze ed esperienze eterogenee collaborano per affrontare le sfide poste dal panorama delle minacce informatiche, accompagnare i clienti nell'adozione dell'AI nei propri processi aziendali e condividere le attività di ricerca con la comunità tecnico-scientifica.

Il team cura l'**intero ciclo di sviluppo dei sistemi basati su AI**: dalla raccolta e dal preprocessing dei dati, all'addestramento e validazione dei modelli, fino al deployment in produzione.

Le soluzioni proposte si fondano su tecnologie avanzate di **Machine Learning, Deep Learning e Large Language Models**, e possono essere personalizzate in base alle specifiche esigenze del cliente.

L'AI Team è costantemente impegnato in attività di ricerca e sperimentazione, con un'attenzione particolare agli **aspetti etici**, alla **trasparenza** e alla **tutela della privacy**.

L'obiettivo è proporre soluzioni innovative che siano in linea con i principi di **equità, inclusività e rispetto dei diritti fondamentali**.

Abstract

Fra il bollettino MS17-010 e WannaCry passarono 59 giorni; fra il deposito di una correzione nel repository di Firefox e il primo exploit prodotto in laboratorio da un modello linguistico di frontiera, all'inizio del 2026, è passata meno di un'ora. Le due misure descrivono fenomeni diversi, uno sfruttamento reale e uno sviluppo sperimentale, ma la distanza fra loro è l'oggetto di questo studio.

La lettura corrente del fenomeno, secondo cui i modelli linguistici avrebbero collassato la finestra di patching, è suggestiva ma imprecisa. I dati mostrano che per le vulnerabilità ad alto valore la finestra si era già chiusa prima del 2024, per effetto della crescita degli zero-day e dell'industrializzazione del cybercrime. Ciò che i modelli di frontiera cambiano è un'altra cosa: riducono il costo e la dipendenza da competenze specialistiche nello sviluppo di exploit, e con ciò modificano i criteri economici che selezionavano quali vulnerabilità valesse la pena trasformare in armi. La coda lunga delle vulnerabilità che i vendor classificano come improbabili da sfruttare, e che i dati storici dimostrano ampia, rischia di diventare superficie attiva: è il principale rischio emergente che questo studio identifica, misurato in laboratorio per alcune classi di vulnerabilità e ancora da osservare su scala.

Lo studio ricostruisce l'evidenza 2024-2026 discutendone i conflitti di interesse, spiega perché le serie storiche su time-to-exploit e remediation non possano essere sovrapposte, formula una predizione falsificabile sui dati KEV e sulla serie Mandiant degli N-day, propone di misurare l'esposizione per asset accanto alla velocità di patching e legge il quadro europeo e italiano, da NIS2 al Cyber Resilience Act e al ruolo di ACN. La posizione che ne trae è che la maturità della gestione delle vulnerabilità non si misura più solo con la velocità delle patch, ma con la capacità di identificare le esposizioni critiche, mitigarle con controlli verificabili e contenere uno sfruttamento prima della remediation.

Autori:

- Gaetano Zappulla: Chief Information Security Officer

Nota dell'autore: Nella preparazione di questo paper l'autore si è avvalso di Claude, modello di Anthropic, come strumento di supporto per la revisione critica del testo, la verifica e la raccolta delle fonti, la produzione delle figure e l'editing. Le tesi, le valutazioni e la responsabilità di quanto scritto restano dell'autore. Poiché le ricerche di Anthropic sono fra le fonti primarie discusse in questo studio e vi sono trattate come fonti di parte, l'uso di un suo modello nella revisione è dichiarato qui per trasparenza.

1. Da WannaCry a Firefox 148: cosa è cambiato in nove anni

Il patch management difensivo, così come è stato praticato per vent'anni, poggia su un assunto che raramente viene dichiarato: trasformare una correzione in un exploit richiede tempo, e quel tempo appartiene al difensore. Quando Microsoft pubblicò il bollettino MS17-010, nel marzo 2017, i 59 giorni che separarono la patch dall'esplosione di WannaCry furono vissuti come una finestra persa, un fallimento organizzativo. Erano in realtà un lusso: due mesi in cui chiunque avesse applicato l'aggiornamento era al sicuro, e in cui il lavoro di trasformare quella patch in un'arma era ancora in corso in qualche laboratorio ben attrezzato. Nove anni dopo, un modello linguistico di frontiera ha prodotto un exploit funzionante per una vulnerabilità di Firefox entro un'ora da quando la correzione era stata depositata nel repository di Mozilla, 18 giorni prima che la versione corretta, Firefox 148, fosse rilasciata al pubblico.

Quel lusso non esiste più, ma è importante capire come e quando è finito, perché la risposta cambia le conclusioni. Sarebbe facile attribuire tutto ai modelli linguistici: la cronologia della primavera 2026, con il rilascio di Claude Mythos Preview, le misurazioni di Anthropic e dei benchmark accademici e il primo exploit sviluppato con assistenza AI intercettato da Google, si presta a una narrazione di rottura improvvisa. I dati raccontano una storia più lunga. Mandiant misura dal 2018 il time-to-exploit, cioè il tempo medio fra la divulgazione di una vulnerabilità e il suo primo sfruttamento osservato: era di 63 giorni nel biennio 2018-2019, di 32 nel 2021-2022, di 5 nel 2023. Quel collasso è avvenuto prima che i modelli di frontiera fossero rilevanti per lo sviluppo di exploit, e Mandiant stessa lo spiega con il peso crescente degli zero-day, che nel 2023 costituivano il 70% delle vulnerabilità sfruttate, e con un'esecuzione più rapida degli N-day di alto valore. Per quelle vulnerabilità, le esecuzioni di codice remoto sui dispositivi di perimetro e le falle nei prodotti più diffusi, la finestra si era già chiusa.

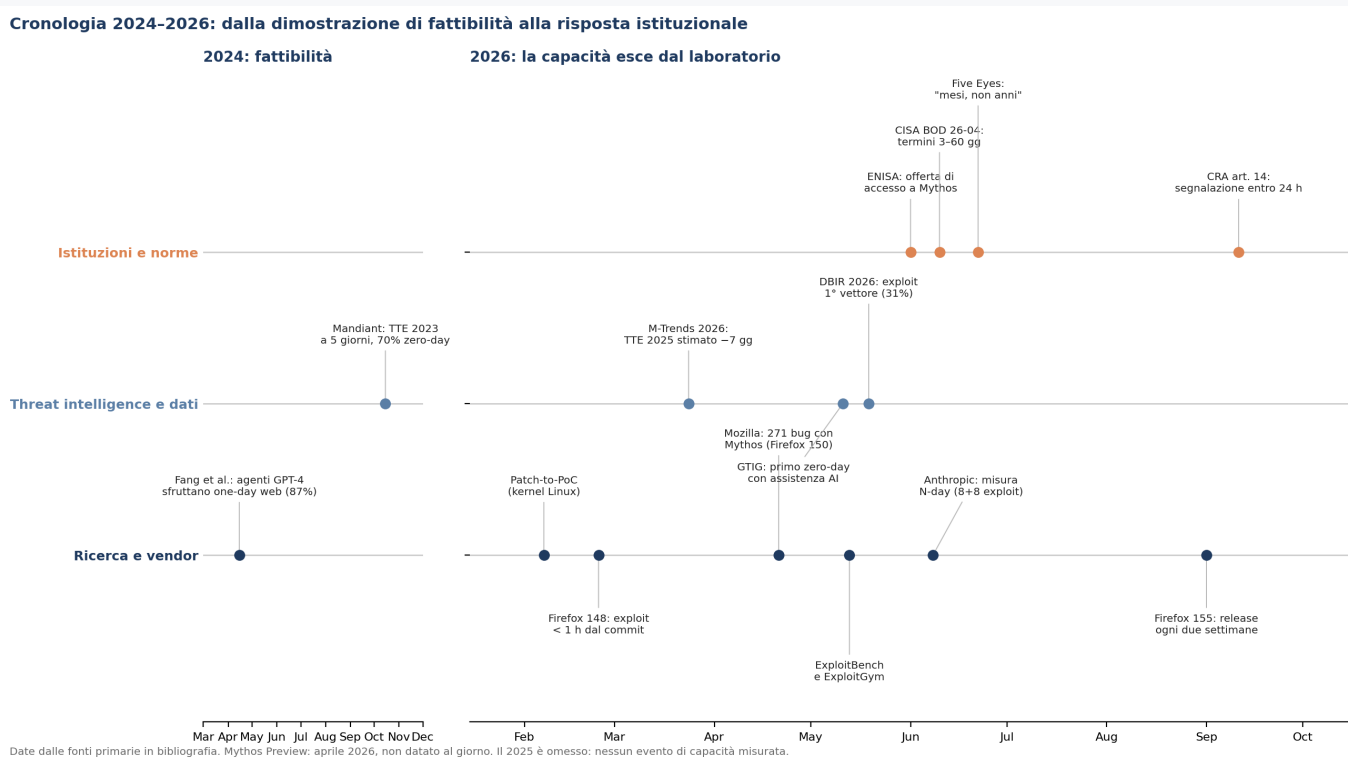


Figura 1. Cronologia 2024-2026 degli eventi citati in questo studio, su tre corsie: ricerca e vendor, threat intelligence e dati, istituzioni e norme. Il 2025 è omissso perché non vi cade alcun evento di capacità misurata. Date dalle fonti primarie in bibliografia.

Ciò che i modelli linguistici cambiano è meno visibile nelle statistiche aggregate e per questo più insidioso. Nel vecchio equilibrio, il costo del lavoro di un reverse engineer esperto non si limitava a rallentare l'attaccante: lo costringeva a scegliere. Settimane di lavoro specialistico si spendevano su una vulnerabilità che valesse la pena, non su una privilege escalation locale classificata dal vendor come "Exploitation Unlikely". Mandiant lo osservava già nel 2023: alcune vulnerabilità non vengono sfruttate semplicemente perché il loro valore è troppo basso perché un attore esperto si prenda il disturbo. Quel giudizio di convenienza era, in pratica, una difesa. Per descriverne l'effetto questo studio usa un'immagine presa in prestito dalla statistica e dall'economia, la coda lunga: ordinate per il valore che hanno per un attaccante, poche vulnerabilità stanno in testa e vengono sfruttate in fretta, mentre la grande maggioranza sta nella coda, dove ciascuna vale poco e tutte insieme sono quasi l'intero insieme. Non sono infruttuabili; sono rimaste al riparo perché nessuno aveva tempo per loro. La sezione 2.3 misura quanto è lunga questa coda.

È questa difesa che i modelli di frontiera indeboliscono. Quando il patch diffing, cioè il confronto fra il codice o il binario prima e dopo la correzione per ricostruire il bug che la patch descrive implicitamente, può essere svolto da un agente in poche ore e per poche migliaia di dollari di crediti API, la selezione si allenta: diventa possibile esaminare sistematicamente un numero di correzioni molto superiore a quello che un singolo reverse engineer può gestire, e produrre proof-of-concept per vulnerabilità che nessun esperto avrebbe guardato. Che questa possibilità si estenda a un intero ciclo di aggiornamenti, un Patch Tuesday, è uno scenario che la sezione 5.4 tratta come ipotesi da verificare, non come capacità dimostrata. Il termine “N-day”, che designa una vulnerabilità già divulgata e corretta ma non ancora applicata ovunque, è diventato fuorviante nella sua stessa unità di misura. “N-hour” è più vicino alla realtà.

1.1 Termini e perimetro

Tre precisazioni orientano tutto il resto. La prima riguarda il termine: “N-hour” non è una categoria tecnica standardizzata come N-day, ma un’espressione per la compressione dei tempi di sviluppo degli exploit misurata in laboratorio, e il tempo fra il commit di una patch e un exploit sperimentale non coincide con il tempo fra divulgazione pubblica e primo sfruttamento osservato, che è la grandezza delle serie storiche della sezione 5; i due intervalli restano distinti per tutto il paper. La seconda riguarda il modello di minaccia, che è uno scenario emergente più che una capacità già osservata su scala: l’attaccante che interessa qui è criminale, con movente finanziario, accesso API a modelli pubblici o aperti, un budget nell’ordine delle migliaia di dollari e competenze di reverse engineering molto inferiori a quelle tradizionalmente necessarie: le pipeline di cui alla sezione 4 riducono il lavoro esperto, non lo eliminano, perché preparare l’ambiente, valutare la correttezza dell’exploit e adattarlo a configurazioni reali restano compiti umani. Per gli attori statali, come si è detto, la finestra era già chiusa. La terza riguarda il perimetro: lo studio è ricerca difensiva e discute le capacità offensive al livello di astrazione necessario a comprenderne l’impatto.

2. Il collo di bottiglia che proteggeva la coda lunga

2.1 Anatomia del patch gap

Il ciclo di vita difensivo di una vulnerabilità attraversa stadi distinti: scoperta, divulgazione, rilascio della correzione, distribuzione, installazione. Ogni passaggio introduce un ritardo. La somma di questi ritardi sul lato difensore è il patch gap, e anche nei casi migliori è ampio. Firefox, che si aggiorna in background e richiede solo un riavvio del browser, ha mostrato nelle patch analizzate da Anthropic una mediana di 19 giorni fra la correzione nel repository e la release stabile, un valore considerato rapido per gli standard del settore. Il campione è di febbraio-marzo 2026 e precede le due accelerazioni decise da Mozilla nel corso dell'anno: le dot-release settimanali programmate da Firefox 151 (maggio) e le release maggiori ogni due settimane, anziché quattro, da Firefox 155 (settembre). Oggi un fix di sicurezza raggiunge il canale stabile nella dot-release della settimana, se viene selezionato per l'uplift, o nella release maggiore successiva. Per Windows, prendendo a riferimento Windows Autopatch, una patch raggiunge il 90% dei dispositivi di una flotta in circa 7 giorni, e solo all'undicesimo scatta il riavvio forzato che la rende effettiva. Nei sistemi industriali, nei dispositivi medicali e nell'IoT, vincolati da finestre di manutenzione fisse, firmware bloccato dal vendor o garanzie di continuità, il gap si misura in mesi, talvolta in anni.

Che il problema fosse riconosciuto prima dei modelli di frontiera lo dimostra la storia dei vendor più attenti. Google ha portato Chrome a release stabili ogni due settimane nel 2020, riducendo il patch gap medio fra commit e release da 35 a circa 15 giorni, e dall'agosto 2023 pubblica aggiornamenti di sicurezza settimanali con una motivazione dichiarata in termini quasi identici a quelli di questo paper: restringere la finestra, già piccola, in cui un attaccante N-day può sviluppare e usare un exploit. Nel luglio 2025 Project Zero ha aggiunto una policy di *Reporting Transparency*, che rende pubblica entro una settimana l'esistenza di ogni vulnerabilità segnalata per comprimere il ritardo ancora più a monte fra il fix di un componente e la sua integrazione nei prodotti a valle. Il patch gap era dunque un tema da anni per chi lo misurava; ciò che è cambiato è il costo di chi lo sfrutta.

2.2 Perché la patch è una mappa

Il paradosso dell'N-day è che la correzione pubblica è anche la descrizione più precisa del bug. Confrontando la versione vulnerabile con quella corretta, l'attaccante individua esattamente cosa è cambiato e da lì ricostruisce la vulnerabilità. Sul software open-source il diff è il changeset stesso; sul software closed-source si lavora sui binari, confrontando le funzioni decompilate. In entrambi i casi la patch fornisce una roadmap, ed è per questo che l'N-day è spesso più pericoloso dello zero-day: non richiede ricerca, solo reverse engineering di un'informazione già pubblica.

Vale la pena fissare qui una distinzione che le statistiche aggregate tendono a cancellare. Uno zero-day sfruttato prima della patch e un N-day trasformato in exploit dopo la patch sono fenomeni diversi, con attori e costi diversi. Il primo richiede capacità di scoperta; il secondo richiede solo la capacità di leggere una correzione. Il tema di questo studio è il secondo, e la distinzione tornerà decisiva nella lettura dei dati della sezione 5.

2.3 Quanto era ampia la coda lunga

I dati storici misurano una cosa precisa: la forte concentrazione dello sfruttamento su una minoranza delle vulnerabilità pubblicate. L'analisi di Cyentia e FIRST sui dati dell'Exploit Prediction Scoring System, pubblicata nel 2024, stima che circa il 6% di tutte le CVE mai pubblicate sia stato osservato sfruttato in-the-wild, con un tasso sostanzialmente stabile nel tempo. Mandiant, nel report sul 2023, aggiunge un dettaglio ancora più eloquente: il 58% delle vulnerabilità divulgate con copertura mediatica non risultava sfruttato, e fra quelle per cui esisteva almeno una proof-of-concept o un exploit pubblico, il 72% non risultava sfruttato. La disponibilità di un exploit, dunque, non si traduceva in sfruttamento. Il costo dello sviluppo è la spiegazione economica di questa distribuzione che Mandiant stessa propone e che questo studio adotta; il suo contributo specifico non è isolabile dai dati rispetto ad altri fattori, come il valore dei bersagli, la disponibilità di accesso, l'affidabilità dell'exploit e la presenza di mitigazioni, ma è il fattore che i modelli linguistici modificano direttamente, ed è per questo che la tesi si concentra su di esso. La concentrazione storica dimostra che una selezione esiste; non consente di attribuirne le cause in termini quantitativi, e la riduzione del costo di sviluppo va letta come un meccanismo plausibile di espansione della superficie sfruttata, la cui rilevanza relativa dovrà essere misurata.

Il patch diffing non è l'unico passo di una campagna N-day reale: restano la scoperta dei bersagli, la consegna dell'exploit e l'evasione dei controlli, e la sezione 6 tornerà su questi colli di bottiglia residui. Ma era il passo il cui costo decideva se una vulnerabilità "valesse". Gli incidenti che i difensori ricordano, WannaCry a 59 giorni dal bollettino, l'exploit pubblico per Citrix Bleed circa due settimane dopo la divulgazione, le 16 vulnerabilità su 25 che nell'analisi Mandiant del 2020 richiesero un mese o più per essere sfruttate, una mediana di 15 giorni fra divulgazione e sfruttamento degli N-day nel 2023, sono i casi in cui la valutazione era stata positiva. Dietro di essi c'era un numero molto maggiore di vulnerabilità che nessuno aveva ritenuto degne di settimane di lavoro. La stessa lettura viene da una fonte che non è un vendor AI: il DBIR 2026, modellando la probabilità che una vulnerabilità KEV torni a essere sfruttata, osserva che il decadimento misurato è coerente con il fatto che gli attaccanti devono sviluppare gli exploit e mantenere l'infrastruttura di scansione, e che di queste attività ne possono condurre solo un numero limitato alla volta; e aggiunge che è un'assunzione di base che l'automazione della scoperta di vulnerabilità e dello sviluppo di exploit con strumenti generativi potrebbe ribaltare.

Conviene fissare questa economia in poche righe, perché nel resto del paper vi faremo riferimento. Per un attaccante una vulnerabilità vale la differenza fra il beneficio atteso e tre costi: sviluppare l'exploit, eseguirlo e consegnarlo, e condurre la campagna, dalla ricerca dei bersagli al rischio di essere identificati. Diventa interessante quando il beneficio supera i costi, ma un valore positivo non decide da solo: l'attaccante ha risorse limitate e sceglie fra le opportunità per rendimento relativo. La coda lunga è l'insieme delle vulnerabilità per cui quel valore era negativo o troppo basso rispetto alle alternative, e l'ipotesi di questo studio, coerente con la lettura di Mandiant ma non isolabile dai dati, è che a renderlo tale fosse soprattutto il costo di sviluppo: settimane di lavoro esperto contro un beneficio modesto. I modelli di frontiera agiscono principalmente su quel costo, per le classi di vulnerabilità misurate; quanto lo riducano non è quantificabile in modo omogeneo, perché le fonti riportano costi di calcolo e non il costo umano equivalente, e l'automazione può incidere anche sugli altri due. Ne segue uno spostamento della soglia di convenienza la cui ampiezza resta da misurare; e, come la sezione 5.4 preciserà, se a dominare sono i costi di esecuzione e di campagna, il vincolo si sposta a valle e la soglia si muove meno di quanto la sola caduta del costo di sviluppo farebbe pensare.

3. L'evidenza: dalla fattibilità alla misura

3.1 Cosa si misura, e a che livello

Prima di esaminare i risultati conviene fissare una scala, perché “sviluppare un exploit” copre attività molto diverse e i numeri che seguono stanno a livelli diversi. I livelli sono sei. Primo: identificare la vulnerabilità. Secondo: riprodurre il bug, cioè raggiungere il crash. Terzo: ottenere primitive sfruttabili, come una lettura o una scrittura arbitraria in memoria. Quarto: costruire un exploit funzionante. Quinto: renderlo affidabile su configurazioni diverse. Sesto: usarlo in una campagna reale, con la scoperta dei bersagli, la consegna e l'evasione dei controlli. ExploitBench, descritto nella sezione successiva, misura separatamente i livelli dal secondo al quarto. In questo studio “proof-of-concept” indica, per convenzione, il secondo livello, la riproduzione verificabile del bug, ed “exploit completo” il quarto, salvo dove le fonti originali usano una definizione diversa; in letteratura una proof-of-concept può dimostrare un crash, una primitiva o un impatto più avanzato, e la convenzione non implica che tutte abbiano lo stesso grado di sfruttabilità. Nessuna delle fonti citate misura il quinto e il sesto livello.

3.2 La letteratura accademica, in tre anni

La ricerca sull'uso di modelli linguistici per lo sfruttamento di vulnerabilità ha percorso in meno di tre anni la distanza fra la prova di concetto e la misurazione sistematica. Il lavoro fondativo è di Fang, Bindu, Gupta e Kang, dell'aprile 2024: 15 vulnerabilità one-day reali, per lo più in applicazioni web, e un agente basato su GPT-4 che, con la descrizione CVE a disposizione, ne sfruttava l'87%, contro lo 0% degli altri modelli e degli scanner ZAP e Metasploit; senza la descrizione la percentuale crollava al 7%. Il costo medio per tentativo era di 3,52 dollari, circa 8,80 per exploit riuscito. La classe di vulnerabilità, web e non memory corruption, non è comparabile con ciò che seguirà sui kernel.

I benchmark del 2026 hanno introdotto la distinzione che struttura la scala di 3.1, separando la riproduzione del bug dalla sua trasformazione in exploit. ExploitGym (Wang et al., NeurIPS 2026) comprende 898 istanze da vulnerabilità reali in programmi userspace, nel motore V8 di Chrome e nel kernel Linux, con e senza mitigazioni attive: Claude Mythos Preview produce exploit funzionanti per 157 istanze e GPT-5.5 per 120, e gli autori notano che gli agenti sfruttano spesso una vulnerabilità diversa da quella assegnata, ragione per cui il giudizio automatico sulla traiettoria è parte del benchmark. ExploitBench (Lee e Brumley, Carnegie Mellon e Bugcrowd, maggio 2026) scompone l’exploitation in 16 flag su cinque livelli, dal crash all’esecuzione di codice, ciascuno verificato da un oracolo deterministico, su 41 bug di V8 attaccati dalla stessa interfaccia esposta in produzione; il risultato principale è una netta separazione fra i modelli di frontiera pubblici e quelli privati, che gli autori avvertono di non estrapolare oltre la classe di vulnerabilità misurata. Più vicino al tema di questo studio, Patch-to-PoC (Pu et al., febbraio 2026) parte esplicitamente dalla patch: su 100 vulnerabilità reali del kernel Linux il sistema K-Repro genera proof-of-concept che riproducono il bug in oltre la metà dei casi, a costi che gli autori definiscono praticabili. È il secondo livello della scala, misurato sul bersaglio più difficile.

Studio	Vulnerabilità	Risultato	Livello della scala	Limite
Fang et al., 2024	15 one-day, per lo più web	87% sfruttate da un agente GPT-4 con descrizione CVE; 7% senza	4 (exploit) su classe web	campione ridotto, classe non comparabile ai kernel
ExploitGym, 2026	898 istanze: userspace, V8, kernel Linux	157 exploit per Mythos Preview, 120 per GPT-5.5	4, con giudice sulla traiettoria	ambiente controllato; agenti che sfruttano vulnerabilità diverse da quella assegnata
ExploitBench, 2026	41 bug di V8	16 flag su 5 livelli; netta separazione fra modelli pubblici e privati	da 2 a 4, misurati separatamente	una sola classe di vulnerabilità, non estrapolabile
Patch-to-PoC, 2026	100 kernel Linux, da patch	oltre la metà riprodotte dal sistema K-Repro	2 (riproduzione)	riproduzione non equivale a exploit
Anthropic N-days, 2026	18 Firefox e 21 Windows, da patch	14 e 18 riproduzioni; 8 exploit e 8 catene a SYSTEM; 13 riproduzioni su 14 rating bassi	2 e 4	fonte interessata; harness strutturato; campione piccolo

Tabella 1. I cinque studi sperimentali citati, letti sulla scala di 3.1.
I valori sono quelli riportati dalle fonti in bibliografia.

Un limite comune va dichiarato, perché gli autori stessi lo dichiarano: questi benchmark misurano capacità specifiche in ambienti controllati e riproducibili, che possono facilitare alcuni compiti rispetto a un attacco reale. I risultati non possono essere generalizzati al quinto e al sesto livello della scala di 3.1, cioè all'affidabilità degli exploit su configurazioni diverse e al loro uso in campagne offensive.

3.3 Il caso Anthropic: Firefox e il kernel Windows

Il documento di riferimento è la misurazione pubblicata dal Frontier Red Team di Anthropic l'8 giugno 2026. Il suo valore sta nell'aver portato la domanda esattamente sul patch diffing, con due scenari costruiti per rispondere a chi obietta che il software reale non assomiglia a un benchmark.

Per il software open-source, 18 patch di sicurezza di SpiderMonkey, il motore JavaScript di Firefox, incluse nelle versioni 148 e 149: l'agente, in un container senza accesso a internet, riceveva il diff pubblico privato del test di regressione, il nome del componente, la severità Mozilla e due build con AddressSanitizer, ma non l'advisory né il reproducer. Per il closed-source, 21 elevation-of-privilege del kernel Windows divulgate fra gennaio e febbraio 2026: senza sorgente, con binari vulnerabile e corretto, simboli pubblici, decompilazione, diff a livello di funzione e advisory Microsoft, e con un sistema di verifica che ricompilava la proof-of-concept e controllava l'escalation con un nonce. Tutte le vulnerabilità erano state divulgate dopo il knowledge cutoff dei modelli: i risultati non sono spiegabili con la memorizzazione.

I risultati sono netti, e vanno letti sulla scala di 3.1. Su Firefox, il modello più capace ha raggiunto il crash per 14 delle 18 vulnerabilità e ha costruito 8 exploit completi di esecuzione di codice; i modelli pubblici della generazione precedente ne hanno prodotti uno o due ciascuno. Il primo exploit è arrivato in meno di un'ora dal commit della patch nel repository Mozilla, gli otto in circa dodici ore. Su Windows, il modello ha raggiunto il crash per 18 vulnerabilità su 21, con la prima proof-of-concept in 31 minuti e tutte entro sei ore, per circa 2.200 dollari di crediti API; ha poi costruito 8 catene complete di privilege escalation da utente a SYSTEM, per 15.700 dollari complessivi, circa 2.000 per escalation. Un modello pubblico è arrivato alle primitive di lettura e scrittura senza riuscire a concatenarle, cioè al terzo livello della scala.

C'è un risultato, in quel documento, che vale più degli altri per la tesi di questo studio, e va riportato con la stessa precisione. Gli advisory Microsoft classificavano 14 delle 21 vulnerabilità come "Exploitation Less Likely" o "Exploitation Unlikely". Il modello ha raggiunto il crash per 13 delle 14, e per una di esse, classificata "Unlikely", ha costruito la catena completa di escalation. Sono quindi tredici riproduzioni e un exploit, non tredici exploit; ma anche a quel livello il risultato dice qualcosa che i rating non prevedevano. L'Exploitability Index di Microsoft stima la probabilità che venga sviluppato un exploit funzionante nei trenta giorni successivi al rilascio, combinando la difficoltà tecnica con l'interesse atteso degli attaccanti; "Less Likely" indica che il codice sarebbe difficile da costruire, "Unlikely" che è improbabile venga rilasciato. Entrambi i giudizi incorporano, esplicitamente o meno, una valutazione del rapporto fra sforzo e valore per un attaccante umano. Il risultato Anthropic non dice che quel giudizio fosse sbagliato sulla difficoltà: dice che la difficoltà pesa meno quando il primo passo costa poche ore di calcolo.

I limiti del caso studio vanno dichiarati. Anthropic è parte interessata, e pubblica questi risultati anche per posizionare i propri modelli e la propria agenda di sicurezza. La collaborazione con Mozilla, annunciata il 21 aprile 2026 con il rilascio di Firefox 150 e documentata tecnicamente su Mozilla Hacks il 7 maggio, ha portato a correggere 271 bug di sicurezza individuati con Mythos Preview, su un totale di 423 fix di sicurezza pubblicati da Mozilla in aprile fra Firefox 150, le point release e le ESR (i due numeri hanno perimetri diversi: tutte le fonti il secondo, il solo modello il primo). Come ha notato Bruce Schneier, il 271 aggrega gravità molto diverse, e l'advisory ufficiale di Firefox 150 elenca 41 CVE, di cui tre esplicitamente attribuite al modello. L'ambiente di prova fornisce al modello un contesto strutturato (build con sanitizer, diff pulito, severità) che però un attaccante può ricostruire con strumenti pubblici, per cui il vantaggio è di comodità più che di informazione. Questi limiti impongono di leggere i numeri come indicazione di capacità e tendenza; i benchmark indipendenti convergono sulla stessa direzione.

4. Fuori dai modelli protetti

Un'obiezione naturale è che tutto questo riguardi un modello gated, distribuito a un gruppo ristretto di organizzazioni attraverso Project Glasswing, con safeguard che reindirizzano le richieste su temi sensibili. Se la capacità fosse confinata lì, il rischio sarebbe gestibile a livello di accesso. L'evidenza dice che non è così, con una sfumatura che merita di essere conservata.

Lo stesso documento Anthropic nota che i modelli pubblici, con le safeguard disattivate in ambiente di test, costruiscono exploit, pur in numero inferiore. La clausola conta: un attaccante che usi un modello pubblico deve aggirare le safeguard, con jailbreak o con accesso offuscato per eludere i controlli d'uso, pratiche che il report del Google Threat Intelligence Group del maggio 2026 documenta come correnti: jailbreak a persona (un attore legato alla Cina istruisce il modello ad agire come "senior security auditor" per fare ricerca di vulnerabilità su firmware) e un'infrastruttura grigia di accesso offuscato ai modelli, con servizi di relay, registrazione automatica di account e browser che mascherano l'impronta, usata per aggirare i controlli d'uso e i limiti di fatturazione dei provider. Le safeguard sono quindi una barriera reale ma parziale: alzano il costo senza renderlo proibitivo. Sul fronte dei modelli aperti, le riprese del lavoro Anthropic riportano che alcuni di essi trovano vulnerabilità a un livello comparabile ai modelli di frontiera. Scoperta ed esecuzione sono però livelli diversi della scala di 3.1, e questo studio non li equipara: il dato è un indizio che la soglia di accesso si abbassa anche dove nessuna safeguard esiste.

L'evidenza più solida viene dalle pipeline costruite fuori dai laboratori dei vendor AI, con strumenti e modelli disponibili. Elastic Security Labs ha documentato nel marzo 2026 una use-after-free in Windows DWM individuata combinando un modello linguistico con il patch diffing, con un exploit affidabile di escalation a SYSTEM. OriginHQ ha descritto, con un titolo che è una dichiarazione d'intenti, The Mythos We Have At Home, una pipeline a due stadi che trasforma gli aggiornamenti del Patch Tuesday in report di diff binario pronti per l'ingestione da parte di un modello e poi in proof-of-concept N-day verificate. SySS, già nel 2024, aveva presentato diffalalyze, che automatizza il patch diffing dei binari e usa un modello per il triage dei cambiamenti rilevanti per la sicurezza. Nessuno di questi lavori pubblica un tasso di successo confrontabile con il caso Anthropic, e vanno letti come prove di fattibilità. Ma il punto è la soglia di accesso. Finché la capacità richiedeva un modello gated, l'accesso controllato era una mitigazione sensata.

Quando la stessa capacità, anche in forma attenuata, è ottenibile con strumenti open-source e modelli pubblici o aperti, le barriere tecniche all'accesso si abbassano. La disponibilità pubblica di queste pipeline non consente però di quantificare quanto si allarghi la popolazione di attaccanti in grado di produrre exploit affidabili e usarli: fra la disponibilità degli strumenti, la riproducibilità dei risultati, l'accessibilità economica, le competenze richieste e la capacità di impiego operativo c'è una scala, e le fonti citate misurano solo i primi gradini. Ciò che si può dire è che il vincolo tende a diventare una spesa modesta e un accesso API, cioè il profilo dell'attaccante dichiarato nella sezione 1.1.

5. Il time-to-exploit e la remediation

Le sezioni precedenti stabiliscono una capacità. Questa esamina il suo effetto aggregato nei dati pubblici, con una cautela che la discussione corrente spesso omette: le serie disponibili sul lato attaccante e su quello difensore misurano cose diverse, e non possono essere sovrapposte su un asse comune per dichiarare un "incrocio".

5.1 Il tempo dell'attaccante: il time-to-exploit

Il time-to-exploit di Mandiant è una media calcolata sulle vulnerabilità effettivamente sfruttate, prima o dopo la patch. La serie è coerente nella direzione: 63 giorni nel 2018-2019, 44 nel 2020-2021, 32 nel 2021-2022, 5 nel 2023. Per il 2024 Mandiant non ha pubblicato un valore, e le interpolazioni che circolano, compresa quella di una research note della Cloud Security Alliance del luglio 2026, non sono dati e qui non vengono usate. Per il 2025, M-Trends 2026 stima un valore negativo di 7 giorni: lo sfruttamento, in media, precede la divulgazione.

Tre avvertenze accompagnano questi numeri. Il valore del 2023 è calcolato dopo la rimozione di 15 outlier statistici; includendoli la media è di 47 giorni. Il valore del 2025 è una stima, ed è trainato dagli zero-day: Mandiant indica che le tre vulnerabilità più sfruttate nel 2025 erano tutte zero-day su application server esposti. E, soprattutto, la serie è costruita su ciò che è stato sfruttato: per costruzione non vede la coda lunga che questo studio sostiene stia diventando attiva.

Esiste però una serie più pulita, che Mandiant pubblica separatamente: il tempo fra divulgazione e sfruttamento dei soli N-day, con una mediana di 15 giorni nel 2023. È questa, e non la media complessiva, la grandezza su cui l'effetto dei modelli linguistici dovrebbe comparire, e la sezione 5.4 la usa come base del test, con l'avvertenza che anche una sua discesa può avere spiegazioni diverse dall'intelligenza artificiale: automazione tradizionale, mercato degli exploit broker, cambiamenti nella composizione delle vulnerabilità divulgate, miglioramenti della telemetria.

5.2 Il tempo del difensore: la remediation

Mentre l'attaccante accelera, la remediation rallenta. Il Verizon Data Breach Investigations Report misura, su dati Tenable, il tempo mediano fra la rilevazione di una vulnerabilità presente nel catalogo CISA KEV e la sua remediation completa: 32 giorni nell'edizione 2025, 43 nell'edizione 2026. Nello stesso intervallo la quota di vulnerabilità KEV completamente rimediate è scesa dal 38% al 26%, e quella ancora aperta al ventottesimo giorno è salita dal 27% al 35%. Il confronto fra i due anni è dell'edizione 2026 stessa, che dichiara di aver replicato le metriche dell'anno precedente sullo stesso perimetro (vulnerabilità presenti nel catalogo KEV a fine anno, rilevate dagli scanner in oltre 13.000 organizzazioni; figure 12, 13 e 14 del report), e non una giustapposizione di due edizioni; nello stesso intervallo il numero mediano di vulnerabilità KEV che ogni organizzazione ha dovuto rimediare è salito da 11 a 16, e la quota di KEV del tutto non rimediate dal 12% al 16%. L'analisi di sopravvivenza su oltre un miliardo di rilevazioni (figura 15) aggiunge il dato più sobrio di tutti: al settimo giorno dalla rilevazione, fra il 60% e il 70% delle KEV resta aperto in ogni anno e a ogni livello di maturità, un limite che il report stesso chiama la velocità della luce della remediation. Due avvertenze: l'intervallo di osservazione del DBIR 2026 va dal 1° novembre 2024 al 31 ottobre 2025 e precede quindi i modelli di frontiera del 2026, e gli stessi autori segnalano che la sezione sull'AI è stata scritta nel febbraio 2026, prima di quegli avanzamenti. Il dato più significativo è la combinazione: nello stesso anno in cui lo sfruttamento di vulnerabilità ha superato le credenziali rubate come primo vettore di accesso iniziale, 31% contro 13%, il tempo per rimediare è aumentato, la quota rimediata è scesa e il volume da rimediare è cresciuto di metà.

5.3 Perché le due serie non si sommano

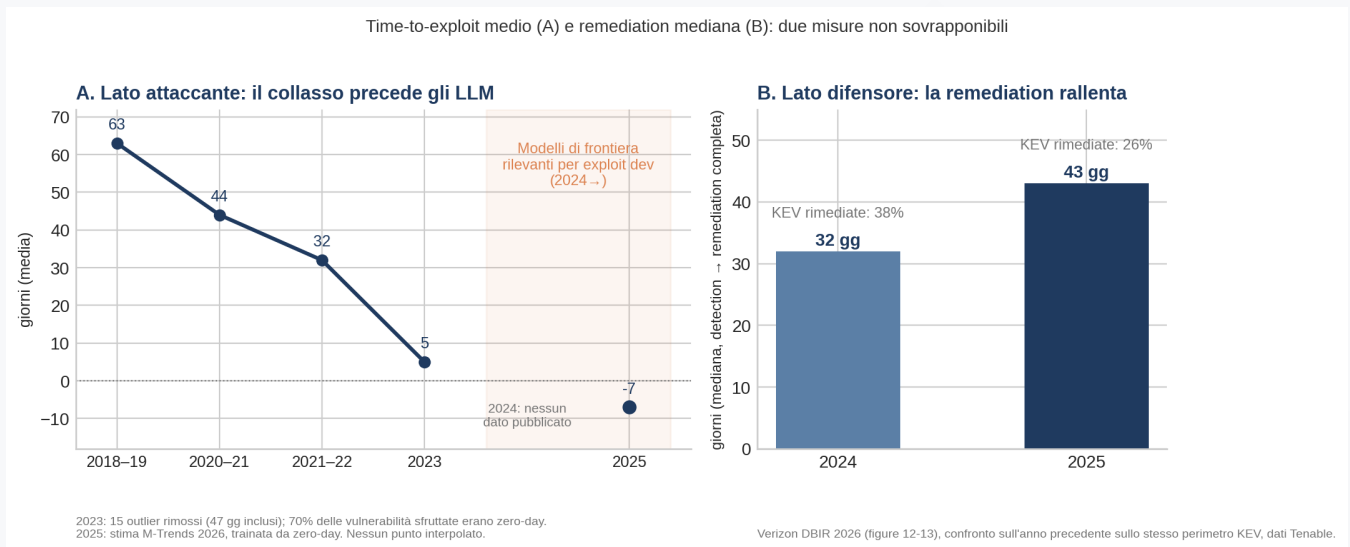


Figura 2. A sinistra il mean time-to-exploit Mandiant, senza punti interpolati; la fascia evidenzia il periodo in cui i modelli di frontiera diventano rilevanti per lo sviluppo di exploit. A destra la mediana DBIR di remediation delle vulnerabilità KEV. Le due misure (media contro mediana, divulgazione contro rilevazione, zero-day inclusi contro catalogo KEV) non sono sovrapponibili.

Sovrapporre le due serie e dichiarare che “le curve si incrociano nel 2024” è una lettura suggestiva ma non difendibile: si confronta una media trimmata che include gli zero-day con una mediana di remediation misurata dalla rilevazione, su un perimetro diverso. Le due tendenze peggiorano insieme; il contributo specifico dei modelli linguistici non sta nella serie del time-to-exploit ma nella coda lunga, che le statistiche aggregate non catturano perché le vulnerabilità “improbabili” raramente finiscono nei dati di incidente.

5.4 Un test che può smentire la tesi

Conviene separare due ipotesi che il dibattito tende a fondere. La prima, di capacità, è che i modelli di frontiera riducano tempi, costi e competenze necessari a trasformare una patch in proof-of-concept e, per una parte dei casi, in exploit; è già sostenuta dalle evidenze delle sezioni 3 e 4, entro i limiti dei benchmark. La seconda, di impatto, è che questa riduzione produca un aumento osservabile dello sfruttamento di vulnerabilità prima trascurate, ed è un’ipotesi da verificare. Le due possono divaricarsi: se il costo marginale della weaponization scende ma il vincolo economico si sposta a valle, sulla scoperta dei bersagli o sulla consegna, la prima resta vera e la seconda no. Un esito del genere non smentirebbe la tesi; direbbe che il collo di bottiglia ha cambiato posto.

L'ipotesi di impatto ammette una verifica precisa, e conviene fissarne i termini prima che i dati arrivino. La capacità misurata è del 2026; il 2025 è quindi l'anno di riferimento e la finestra di verifica è il biennio 2026-2027. Vanno distinti due esiti attesi, perché misurano cose diverse: l'accelerazione, cioè la riduzione del tempo con cui vengono sfruttate vulnerabilità già interessanti, e l'espansione, cioè l'aumento della quota di vulnerabilità prima trascurate che risultano sfruttate. Solo la seconda è la tesi di questo studio; la prima può verificarsi senza di essa. In quella finestra, su periodi omogenei, per classe di vulnerabilità e su gruppi comparabili per prodotto, severità ed esposizione, ci si attende che si muovano quattro grandezze: la mediana Mandiant fra divulgazione e sfruttamento dei soli N-day, 15 giorni nel 2023, deve scendere, ed è la misura più pulita perché non risente degli zero-day; il rapporto fra vulnerabilità sfruttate e pubblicate, stabile intorno al 6% secondo Cyentia e FIRST, deve salire almeno per browser e sistemi operativi; la quota di aggiunte al catalogo KEV con rating vendor "Less Likely" o "Unlikely" al momento della divulgazione deve crescere; e le aggiunte al catalogo devono crescere oltre il trend, al netto del volume di CVE pubblicate, che cresce del 16% l'anno. La prima misura l'accelerazione, le altre tre l'espansione; la terza, la composizione del KEV per rating, è la più vicina alla tesi, la quarta la più esposta ai confondenti. L'anno di riferimento è il 2025 per tutte, ma l'ultimo valore pubblicato non è lo stesso per ogni serie: per la mediana N-day è il 2023, e il confronto dovrà partire dal valore 2025 quando Mandiant lo pubblicherà. Un cambiamento conta se supera la variazione anno su anno osservata nel 2023-2025 per la stessa grandezza. Il disegno più informativo è però differenziale: confrontare le classi di vulnerabilità per cui la capacità dei modelli è dimostrata, la memory corruption in browser e kernel, con quelle per cui non esiste evidenza analoga, come i logic bug o i difetti crittografici. I confondenti comuni, capacità di osservazione, volume di CVE, pratiche di segnalazione, agiscono su entrambi i gruppi; un'espansione significativamente maggiore nel primo sarebbe un'evidenza coerente con l'ipotesi, senza dimostrarne da sola la causalità, perché i due gruppi differiscono anche per diffusione dei prodotti, interesse degli attaccanti e qualità della telemetria, e perché l'assenza di evidenza sperimentale su una classe non equivale all'assenza di capacità. L'interpretazione richiede il controllo di queste differenze strutturali. I dati di oggi non permettono di decidere: il KEV ha ricevuto 187 aggiunte nel 2023, 186 nel 2024 e 245 nel 2025, e 181 nei primi sette mesi del 2026 contro 155 nello stesso periodo del 2025; ma l'aumento del 2025 precede la capacità misurata e va letto come un confondente, perché può dipendere dalla capacità di osservazione, dal volume di CVE o dalle pratiche di segnalazione, e il catalogo registra per costruzione solo ciò che qualcuno ha visto. Se tutte le grandezze restano ferme mentre la capacità continua a crescere, l'ipotesi di impatto va ridimensionata e il vincolo va cercato a valle (confidenza alta sull'ipotesi di capacità, bassa su quella di impatto).

6. Dal laboratorio al campo

6.1 Il caso dell'11 maggio

L'11 maggio 2026 il Google Threat Intelligence Group ha pubblicato il report *Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access*, che documenta quello che ritiene il primo caso noto di un threat actor che usa un modello per sviluppare un exploit zero-day funzionante: un bypass dell'autenticazione a due fattori su uno strumento open-source di amministrazione web, preparato per una campagna di sfruttamento di massa e interrotto grazie alla segnalazione di Google al vendor. Gli attaccanti sono criminali con movente finanziario; Google ha dichiarato alta confidenza che l'exploit sia stato scritto da un modello, non Gemini. Il codice portava le firme della generazione automatica, fra cui un punteggio CVSS "allucinato", e difetti che ne limitavano l'efficacia, con l'avvertenza che potrebbero non persistere.

Questo caso non è un N-day. La vulnerabilità era un errore di logica, non una memory corruption, e non è stata individuata tramite patch diffing. Un caso N-day attribuito a un modello linguistico in attività criminale reale non è, al momento in cui scriviamo, pubblicamente documentato. L'11 maggio è però un indicatore anticipatore per tre ragioni. Il movente finanziario colloca la capacità nella criminalità comune. La preparazione di una campagna di massa indica un intento di scala. E il report documenta un'espansione parallela di capacità offensive AI-abilitate, con interesse rilevato da gruppi legati a Cina e Corea del Nord. Trovare uno zero-day in un logic bug web non è un compito più difficile che trasformare una patch di memory corruption in un exploit affidabile contro un kernel moderno, quindi il caso non si trasferisce al patch diffing per implicazione; indica però che queste capacità stanno entrando nelle operazioni offensive reali di attori criminali, e rende plausibile un'adozione analoga nelle campagne N-day (confidenza media). John Hultquist, chief analyst del gruppo, ha osservato nella dichiarazione diffusa a corredo del report che per ogni zero-day riconducibile all'AI molti altri restano probabilmente non rilevati.

Un secondo indizio, meno vistoso, viene dai dati di abuso che Anthropic ha condiviso con il DBIR 2026: su diverse centinaia di attori osservati sui propri sistemi, circa un terzo delle tecniche di accesso iniziale assistite dall'intelligenza artificiale riguardava lo sfruttamento di vulnerabilità, e gli autori descrivono l'impatto dell'AI come prevalentemente operativo, cioè nell'accelerare attività che gli attaccanti già svolgevano piuttosto che nel crearne di nuove. È un'osservazione dal lato del fornitore del modello, non da quello della vittima, e per dichiarazione degli stessi autori precede gli avanzamenti del 2026; ma è la prima misura pubblica di quanto spesso lo sfruttamento compaia fra gli usi malevoli dei modelli.

6.2 Il quadro istituzionale

Le istituzioni hanno registrato il cambiamento in fretta, e con un precedente. Nel maggio 2025, un anno prima dei risultati di Anthropic, l'NCSC britannico aveva pubblicato la valutazione *Impact of AI on cyber threat from now to 2027*, che aggiornava quella del gennaio 2024, con una frase che oggi si legge come un programma: il tempo fra divulgazione e sfruttamento si è ridotto a giorni e l'intelligenza artificiale, con probabilità quasi certa, lo ridurrà ulteriormente; la stessa valutazione prevedeva un divario crescente fra le organizzazioni capaci di tenere il passo delle minacce abilitate dall'AI e tutte le altre, e indicava nella sicurezza su scala la risposta. Il 22 giugno 2026 i vertici di sei agenzie dei Five Eyes (NCSC britannico, CISA e NSA statunitensi, ASD australiana, CSE canadese, GCSB neozelandese) hanno firmato la dichiarazione congiunta *The AI shift in cyber risk: why leaders must act now*, pubblicata sui siti di NCSC e del Canadian Centre for Cyber Security, con cui hanno avvertito che la trasformazione delle capacità cyber da parte dei modelli di frontiera è questione di mesi e non di anni; la stessa dichiarazione indica fra gli effetti attesi il restringimento della finestra fra scoperta e sfruttamento, senza quantificarlo. Dodici giorni prima, il 10 giugno, CISA aveva emesso la Binding Operational Directive 26-04, *Prioritizing Security Updates Based on Risk*, che sostituisce la BOD 22-01 e impone alle agenzie federali civili termini di remediation da 3 a 60 giorni determinati da quattro criteri: presenza nel catalogo KEV, esposizione pubblica dell'asset, sfruttabilità automatizzata e livello di controllo ottenibile; il blog di accompagnamento cita i dati del DBIR 2026 sulle KEV rimediate. Il Verizon DBIR 2026 descrive a sua volta l'intelligenza artificiale come un fattore che comprime il time-to-exploit dell'attaccante.

La convergenza è reale, ma va pesata prima di presentarla come consenso. Mandiant e il Google Threat Intelligence Group sono entrambi parte di Google. Il DBIR 2026 include un contributo di Anthropic sull'uso dell'intelligenza artificiale da parte degli attaccanti, e Anthropic è anche la fonte del caso studio della sezione 3 e del lavoro con Mozilla. Le fonti indipendenti fra loro sono quindi, in sostanza, due gruppi industriali più le agenzie governative: una base sufficiente per la tesi di questo studio, da non presentare come quattro conferme indipendenti.

7. La coda lunga diventa superficie

7.1 Una riformulazione

Il contributo di questo studio è una riformulazione della tesi corrente che la renda compatibile con i dati. Per le vulnerabilità ad alto valore la finestra si era già chiusa. I modelli linguistici indeboliscono la selezione economica che proteggeva tutte le altre: le evidenze sperimentali mostrano che, per le classi di vulnerabilità misurate, memory corruption in browser e kernel, i modelli di frontiera possono ridurre a poche ore e a poche migliaia di dollari il tempo e il costo della riproduzione del bug a partire dalla patch, anche in presenza di rating che ne indicano una bassa probabilità di sfruttamento; la trasformazione in exploit completo resta più selettiva, ma su quel campione non è più stata il lavoro di settimane che quel rating presuppone. Ne segue che la complessità dello sfruttamento, l'assenza di exploit pubblici o un rating "Unlikely" non sono più garanzie sufficienti per differire la mitigazione di una vulnerabilità presente su un asset critico o raggiungibile.

La tesi si regge su tre ordini di evidenza, nessuno dei quali basterebbe da solo. La prima è l'evidenza storica che la coda lunga esistesse: il 6% appena di CVE osservate sfruttate secondo Cyentia e FIRST, il 72% di vulnerabilità con proof-of-concept pubblica non sfruttate nel 2023 secondo Mandiant (confidenza alta). La seconda è l'evidenza di capacità: tredici riproduzioni e un exploit su quattordici vulnerabilità "Less Likely" o "Unlikely", con il limite di un campione piccolo, di una sola classe di vulnerabilità e di una fonte interessata (confidenza media). La terza è l'ipotesi di impatto della sezione 5.4, che per ora non è decidibile (confidenza bassa). L'implicazione operativa è immediata: i filtri di prioritizzazione fondati sul giudizio umano, come il rating "Unlikely", il punteggio CVSS senza contesto di sfruttamento o l'assenza dal catalogo KEV, non misurano più il rischio N-day.

Il KEV resta uno strumento utile, ma è retrospettivo per costruzione: registra ciò che è già accaduto, non la coda lunga che sta diventando attiva. Un limite analogo riguarda, in prospettiva, EPSS, il sistema di FIRST che stima per ogni CVE la probabilità di sfruttamento nei trenta giorni successivi: la versione 4, del marzo 2025, è addestrata su segnali di sfruttamento osservato e, per dichiarazione di FIRST, non tiene conto né dell'ambiente né dei controlli compensativi.

I suoi segnali storici incorporano la selezione economica descritta in 2.3; se quella selezione cambia, la calibrazione del modello andrà rivalutata periodicamente sui nuovi dati di sfruttamento, cosa che questo studio non ha fatto e non afferma sia già necessaria. Non è un argomento contro EPSS, che resta il filtro predittivo più informativo disponibile; è un argomento per affiancargli criteri che non dipendono dalla storia. Il framework SSVC di CISA e Carnegie Mellon ne offre uno, l'automatizzabilità dello sfruttamento; la BOD 26-04 vi aggiunge l'esposizione pubblica dell'asset. Sono i due criteri che la sezione 7.2 propone di usare come peso.

7.2 Remediation, mitigazione e contenimento

Il mean time to patch, la metrica con cui la maggior parte delle organizzazioni misura la propria gestione delle vulnerabilità, misura una cosa precisa e utile: la capacità di eliminare una vulnerabilità. Non dice nulla su cosa accade nella finestra prima che sia eliminata, e in un mondo in cui la finestra può durare ore quella è diventata la domanda principale. La risposta non è sostituire la metrica con un'altra, ma affiancarle due capacità che misurano fenomeni diversi e che le organizzazioni tendono a confondere. La prima appartiene all'exposure management: identificare tempestivamente quali asset critici o raggiungibili sono interessati da una patch e introdurre, dove la patch non può essere applicata subito, controlli compensativi preventivi e verificabili, come segmentazione, restrizioni di accesso e virtual patching, che riducono la probabilità di sfruttamento; le regole di rilevazione mirate sulla singola patch sono utili ma appartengono alla capacità successiva, perché non riducono l'esposizione, riducono il tempo per accorgersi di uno sfruttamento e contenerlo. La seconda appartiene all'incident response: contenere gli effetti di uno sfruttamento, dal primo alert qualificato all'isolamento dell'asset, prima della remediation definitiva. La maturità di una gestione delle vulnerabilità nell'era N-hour si misura su tutte e tre le capacità, remediation, mitigazione preventiva e contenimento, e il peso relativo delle ultime due cresce con la lentezza strutturale della prima.

Per rendere misurabile la prima di queste capacità, questo studio propone un indicatore sperimentale di esposizione residua, da validare su dati reali prima di qualsiasi adozione operativa: le ore-esposizione per asset, cioè le ore fra il momento in cui un exploit è ragionevolmente disponibile e il momento in cui la vulnerabilità è rimediata o mitigata con un controllo verificato, pesate per la raggiungibilità e la criticità dell'asset.

Il modello conta l'esposizione effettiva e non il tempo trascorso dalla patch: un sistema isolato o protetto da controlli compensativi verificati presenta un'esposizione residua inferiore, da valutare in funzione dell'efficacia dei controlli e dei percorsi di attacco ancora disponibili, perché la segmentazione riduce la raggiungibilità senza azzerarla. L'ipotesi sulla disponibilità dell'exploit va dichiarata come scenario prudenziale e non come previsione: i tempi misurati in laboratorio sono di ore, l'intervallo di 24-48 ore usato in questo studio è un'assunzione conservativa, e un'analisi di sensibilità con uno scenario intermedio a 7 giorni e uno storico ricavato dai dati osservati è il modo corretto di usare il modello senza attribuirgli una precisione che non ha.

7.3 Non tutte le classi di software sono uguali

Quando l'exploit arriva in ore e la distribuzione della patch richiede giorni, la finestra di esposizione coincide quasi per intero con il tempo di distribuzione, che è una proprietà della classe di software più che dell'organizzazione. La Figura 3 riassume gli intervalli. Per i browser la finestra va da pochi giorni a due settimane, a seconda che il fix viaggi nella dot-release settimanale o nella release maggiore bisettimanale, e la mitigazione realistica resta accorciare il canale; per gli endpoint Windows con Autopatch è di una o due settimane, con il riavvio come fattore critico; per il software enterprise on-premise si misura in settimane o mesi, con il change management come collo di bottiglia; per OT e ICS è di mesi, per i dispositivi medicali di mesi o anni, per l'IoT consumer e l'edge, spesso privi di qualsiasi meccanismo di aggiornamento, è indefinita.

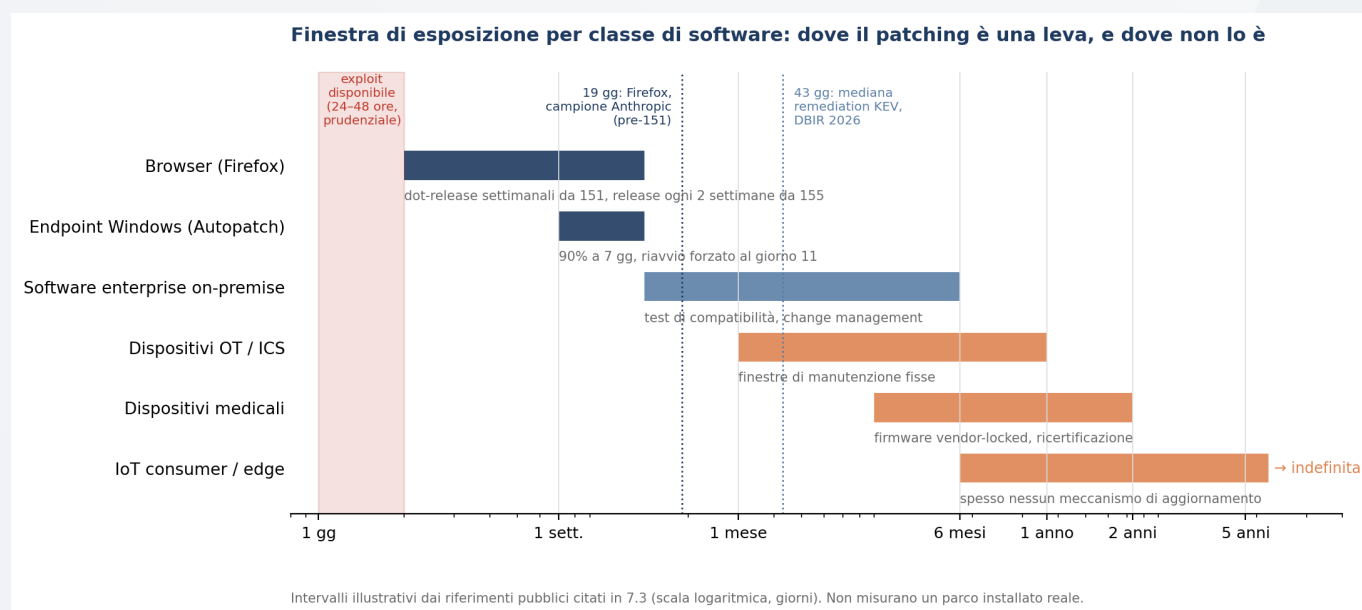


Figura 3. Finestra di esposizione per classe di software, su scala logaritmica dai giorni agli anni. La banda indica la disponibilità dell'exploit nello scenario prudenziale di 24-48 ore; le linee tratteggiate la mediana di 19 giorni del campione Anthropic e la mediana di 43 giorni di remediation delle KEV nel DBIR 2026. Intervalli illustrativi dai riferimenti citati nel testo, non misure di un parco installato.

Il dato che emerge è una stratificazione del rischio, con una conseguenza precisa: per le ultime tre classi la finestra è strutturalmente irriducibile con il solo patching, e la difesa deve passare per le altre due capacità di 7.2. Un'ultima osservazione evita di sovrastimare il fenomeno. Una parte rilevante delle campagne N-day colpisce ancora sistemi mai aggiornati, con vulnerabilità KEV vecchie di anni; per quei sistemi la finestra era già "N-year", e l'N-hour non cambia nulla. L'N-hour è un problema nuovo soprattutto per chi patcha bene: è la sua finestra di pochi giorni che è diventata insufficiente.

Questo modello di esposizione è illustrativo e dipende dalle assunzioni dichiarate: non misura un parco installato reale, non pesa le classi per numerosità o criticità effettiva. Una replica indipendente su un Patch Tuesday completo e la misura della metrica su telemetria reale di un SOC sono indicate fra i lavori futuri.

8. Il contesto europeo e italiano

Il quadro normativo europeo è il più articolato al mondo, ma è stato concepito per un mondo in cui trasformare una patch in un exploit richiedeva settimane, e va riletto senza cedere alla tentazione di accorciare arbitrariamente i termini. Quanto segue è una lettura strategica degli strumenti disponibili, non un'interpretazione degli obblighi: la valutazione giuridica delle disposizioni e delle loro interazioni settoriali resta alle autorità competenti e agli uffici legali delle singole organizzazioni. La domanda che lega questo capitolo al resto del paper è una sola: le metriche di conformità e di gestione delle vulnerabilità oggi in uso restano adeguate quando il tempo necessario a produrre un exploit può essere inferiore al ciclo operativo di remediation? Tre questioni la articolano: cosa dicono i testi sui tempi, chi può agire sull'offerta di vulnerabilità, chi può misurare.

Sui tempi, i testi dicono meno di quanto si creda e meglio di quanto si tema. La Direttiva NIS2 impone ai soggetti essenziali e importanti, all'articolo 21, misure di gestione del rischio che includono la gestione delle vulnerabilità, e fissa all'articolo 23 i termini di notifica degli incidenti, preallerta entro 24 ore e notifica entro 72; non fissa termini di patching, e non sono i termini a dover cambiare, ma l'aspettativa implicita per cui una velocità di patching ragionevole sia, da sola, una misura adeguata. I testi attuativi vanno già oltre. Le norme tecniche di DORA (Regolamento delegato 2024/1774, articolo 10) chiedono alle entità finanziarie di dare priorità alle patch "e ad altre misure di mitigazione" per criticità e profilo di rischio, e di fissare termini di installazione con percorsi di escalation quando non possono essere rispettati;

il Regolamento di esecuzione 2024/2690, per le infrastrutture digitali, chiede patch senza indebito ritardo e, dove non è possibile, mitigazioni documentate.

In entrambi i casi la norma prevede già misure di mitigazione accanto alle patch, senza stabilire un'equivalenza giuridica generale fra le due: è lo spazio in cui la sezione 7.2 si colloca. In Italia il decreto legislativo 138 del 2024 affida ad ACN le modalità attuative, e la categorizzazione delle attività e dei servizi disciplinata dalla Determinazione 155238 del 20 aprile 2026, con dieci macro-aree e quattro categorie di rilevanza, è la base delle misure di sicurezza a lungo termine che ACN stabilirà: la sede naturale per tradurre il principio in un'attesa esplicita di capacità di mitigazione e contenimento, allineandosi a quanto DORA e il 2024/2690 già prevedono.

Sull'offerta di vulnerabilità gli strumenti sono due, con perimetri che non si sovrappongono. Il Cyber Resilience Act agisce sui prodotti con elementi digitali lungo il ciclo di vita, con l'obbligo di fornire aggiornamenti senza ritardo e, dall'11 settembre 2026, di segnalare entro 24 ore le vulnerabilità attivamente sfruttate (articolo 14): copre l'OT e l'IoT, le classi a finestra irriducibile della sezione 7. Esclude però i dispositivi medici e diagnostici in vitro disciplinati dai Regolamenti 2017/745 e 2017/746 (articolo 2, paragrafo 2), perché, come spiega il considerando 25, quei regolamenti impongono già requisiti di sicurezza informatica lungo il ciclo di vita, con la guida MDCG 2019-16 in vigore dal dicembre 2019. Per quella classe la leva è il regolamento sui dispositivi medici, ed è lì che una posizione italiana ed europea sull'aggiornamento dei dispositivi medicali dovrebbe concentrarsi.

Sulla misura, due fatti recenti rendono questa lettura meno teorica. Il 1º giugno 2026 Bloomberg ha riportato che Anthropic ha offerto a ENISA l'accesso a Mythos Preview, cioè al modello su cui poggia l'evidenza della sezione 3; Commissione europea ed ENISA hanno confermato l'offerta precisando che le condizioni erano da concordare, e Anthropic non ha rilasciato un annuncio proprio. Se l'accordo si perfeziona, l'agenzia europea potrà verificare in proprio, su capacità reali, ciò che questo studio può solo leggere da fonti interessate; e il test della sezione 5.4 è esattamente il tipo di misura che un'autorità con accesso ai dati di sfruttamento, come ACN per il perimetro nazionale, può eseguire e pubblicare ogni anno.

Il riferimento esterno più citato, la Binding Operational Directive 26-04 di CISA, è meno distante da questa lettura di quanto la sintesi dei "tre giorni" lasci intendere: i tre giorni valgono solo per le vulnerabilità che sono insieme nel KEV, esposte, sfruttabili automaticamente e a pieno controllo, e per le altre i termini salgono fino a 60 giorni.

È una prioritizzazione per raggiungibilità e impatto, nella direzione della sezione 7.2; la cautela, per un recepimento europeo, è che su OT e medicale anche la fascia più lunga può essere impossibile, e che l'obbligo vada accompagnato da criteri espliciti per le misure di mitigazione quando la remediation nel termine non è possibile.

9. Conclusioni

Che i modelli linguistici possano produrre exploit N-day non è di per sé sorprendente: con tempo sufficiente e un ambiente di prova adeguato, era probabilmente possibile da tempo. Ciò che è cambiato con i modelli di frontiera è il volume, la velocità e, soprattutto, l'indebolimento della selezione economica che lasciava in pace la coda lunga. I dati del capitolo 3 mostrano che i modelli di frontiera riducono in modo sostanziale il tempo e le competenze necessari a riprodurre vulnerabilità a partire dalle correzioni, e in alcuni casi a trasformarle in exploit, su campioni selezionati di poche decine di patch; la possibilità di estendere sistematicamente queste capacità a un intero ciclo di aggiornamenti resta un'ipotesi da verificare, ed è la prima cosa che il laboratorio intende misurare. Il playbook di patch management costruito su cadenze mensili, distribuzione a stadi e ritardo fra canale pre-release e stabile era fondato su due assunti: che trasformare una patch in un exploit richiedesse settimane, e che richiedesse esperti scarsi. Il primo è smentito dai dati, almeno per le classi di vulnerabilità misurate; il secondo è messo in discussione da ogni pipeline pubblica che riproduce, in forma attenuata, ciò che il laboratorio ha mostrato.

Ne discendono tre spostamenti, che questo laboratorio considera la sostanza del cambiamento più di qualsiasi misura tattica. Per chi produce software, la velocità di rilascio resta necessaria, soprattutto quando la patch esiste e non è ancora distribuita, ma la risposta durevole richiede di ridurre alla fonte le classi di vulnerabilità sfruttabili, con linguaggi memory-safe e mitigazioni che ritirano intere classi di exploit, e di rivedere i rating di exploitability alla luce delle capacità dei modelli di frontiera; la misura di quanto conti la prima leva la dà il DBIR 2026, dove l'89% delle organizzazioni ha dovuto rimediare almeno una KEV riconducibile a difetti di memory safety, la classe che Microsoft ha iniziato a sostituire con Rust nel kernel di Windows dal 2023. Per chi difende, la velocità di applicazione della patch resta necessaria ma non basta più a descrivere la maturità: accanto alla remediation contano la capacità di identificare le esposizioni critiche e mitigarle con controlli verificabili, e quella di contenere uno sfruttamento prima della remediation.

Per chi regola, la velocità di patching non è più, da sola, una misura adeguata: gli obblighi vanno orientati verso la rilevazione, la mitigazione e la riduzione della superficie, con il Cyber Resilience Act e il regolamento sui dispositivi medici come leve sulle classi dove il patching non arriva.

Per chi governa la sicurezza di un'organizzazione complessa, il cambiamento riguarda meno gli strumenti che i criteri di decisione. L'accettazione del rischio fondata su "exploitation unlikely" o su "nessun exploit pubblico" era una scelta razionale quando quei giudizi riflettevano un costo reale; oggi va motivata diversamente, con la raggiungibilità dell'asset e la presenza di una mitigazione verificata. Il rapporto fra sicurezza e continuità operativa si sposta: la finestra di manutenzione che rinvia una patch di tre mesi non è più un compromesso neutro, ma un'esposizione da misurare e da coprire. E la responsabilità verso il consiglio di amministrazione cambia oggetto: non più quanto in fretta si applicano le patch, ma quanto a lungo e su quali asset si resta esposti, e cosa protegge nel frattempo.

L'asimmetria documentata in queste pagine è reale e misurabile, ma non è un destino. I lavori che questo laboratorio si assegna partono da una replica sperimentale: la riproduzione da patch, in ambiente isolato e autorizzato, di venti o trenta vulnerabilità con rating di exploitability diversi, misurando tempi, costi, successi e dipendenza dalle competenze dell'operatore, per passare dall'analisi delle ricerche altrui a un risultato proprio sul secondo ordine di evidenza, quello di capacità. Seguono il test della sezione 5.4, la replica su classi di vulnerabilità diverse dalla memory corruption, la misura dell'esposizione su telemetria reale e la valutazione indipendente degli stessi modelli usati in difesa. Gli strumenti sono disponibili anche ai difensori, e la leva più efficace, ridurre l'offerta di vulnerabilità alla fonte, non dipende dalla velocità di reazione. I cinquantanove giorni di WannaCry non torneranno. Ciò che un'organizzazione può ancora decidere è quanto a lungo, e su quali asset, resterà esposta ora che la finestra si misura in ore.

Bibliografia e fonti

Per ogni fonte è indicata, dove possibile, la sezione o la figura da cui sono tratti i dati citati nel testo.

- W. Xiao, T. Abbott, N. Carlini, N. Cheng, D. Forsythe, K. Lucas, M. Nasr, S. Sakhuja (Anthropic Frontier Red Team), «Measuring LLMs' impact on N-day exploits», red.anthropic.com/2026/n-days/, 8 giugno 2026; pagina speculare su anthropic.com/research/n-days (setup, risultati Firefox e Windows, costi, rating Microsoft: sezioni "N-days on Firefox" e "N-days on Windows")
- Anthropic Frontier Red Team, «Claude Mythos Preview: Technical Report» e Project Glasswing, aprile 2026
- G. Volpicelli, «Anthropic to Give Mythos to First EU Body Ahead of Wider Release», Bloomberg, 1° giugno 2026; conferme del portavoce della Commissione europea T. Regnier (AFP, 1° giugno 2026) e di un portavoce ENISA (Financial Times, 1° giugno 2026). Nessun comunicato primario Anthropic alla data di redazione
- Anthropic Frontier Red Team, «Evaluating and mitigating the growing risk of LLM-discovered 0-days», febbraio 2026
- Mozilla, «The zero-days are numbered», Mozilla Blog, 21 aprile 2026; B. Grinstead, C. Holler, F. Braun, post tecnico sulla pipeline di sicurezza con Claude Mythos Preview, Mozilla Hacks, 7 maggio 2026; Mozilla Foundation Security Advisory 2026-30 (Firefox 150); B. Schneier, nota critica, schneier.com, aprile 2026
- Mozilla, «Weekly Firefox dot release schedule starting with Firefox 151», dev-platform, maggio 2026; annuncio della cadenza bisettimanale da Firefox 155, luglio 2026
- R. Fang, R. Bindu, A. Gupta, D. Kang, «LLM Agents can Autonomously Exploit One-day Vulnerabilities», [arXiv:2404.08144](https://arxiv.org/abs/2404.08144), aprile 2024 (tabella dei risultati e sezione sui costi)
- R. Fang et al., «Teams of LLM Agents can Exploit Zero-Day Vulnerabilities», [arXiv:2406.01637](https://arxiv.org/abs/2406.01637), giugno 2024
- Z. Wang, N. Schiller, H. Li et al., «ExploitGym: Can AI Agents Turn Security Vulnerabilities into Real Attacks?», [arXiv:2605.11086](https://arxiv.org/abs/2605.11086), NeurIPS 2026 (composizione del benchmark e risultati: abstract e sezione 2; rdsi.berkeley.edu/blog/exploitgym)
- S. Lee, D. Brumley, «ExploitBench: A Capability Ladder Benchmark for LLM Cybersecurity Agents», [arXiv:2605.14153](https://arxiv.org/abs/2605.14153), Carnegie Mellon / Bugcrowd, 13 maggio 2026, exploitbench.ai (scala a 16 flag e 41 bug V8: abstract e sezione 3); Anthropic Frontier Red Team, «Measuring LLMs' ability to develop exploits», red.anthropic.com/2026/exploit-evals/, 22 maggio 2026 (risultati dei modelli Claude su ExploitBench ed ExploitGym, verificati dagli autori dei benchmark)
- J. Pu et al., «Patch-to-PoC: A Systematic Study of Agentic LLM Systems for Linux Kernel N-Day Reproduction», [arXiv:2602.07287](https://arxiv.org/abs/2602.07287), febbraio 2026 (dataset e tasso di riproduzione: abstract)
- Elastic Security Labs, «Patch diff to SYSTEM», marzo 2026

- OriginHQ, «The Mythos We Have At Home: A Patch-Diffing Pipeline for N-Day Generation», 2026
- SySS, «Automated Patch Diff Analysis using LLMs», 2024
- J. Jacobs, S. Romanosky, B. Edwards, M. Roytman, I. Adjerid, «Exploit Prediction Scoring System (EPSS)», arXiv:1908.04856, 2019
- Cyentia Institute / FIRST, «A Visual Exploration of Exploitation in the Wild: The Inaugural Study of EPSS Data and Performance», 2024 (quota di CVE sfruttate e tasso di crescita delle CVE: sezione introduttiva)
- Mandiant / Google Cloud, «How Low Can You Go? An Analysis of 2023 Time-to-Exploit Trends», ottobre 2024 (serie TTE, quota zero-day, mediana N-day, quote di vulnerabilità non sfruttate: sezioni «Time-to-Exploit» e «Zero-Day vs. N-day Exploitation»)
- Mandiant / Google Cloud, «M-Trends 2026», 24 marzo 2026; presentazione di J. Kutscher, «M-Trends 2026: Data, Insights, and Strategies From the Frontlines», Google Cloud blog, 24 marzo 2026, cloud.google.com/blog/topics/threat-intelligence/m-trends-2026 (stima del mean time-to-exploit 2025 a -7 giorni e vulnerabilità più sfruttate)
- Mandiant, «Think Fast: Time Between Disclosure, Patch Release and Vulnerability Exploitation», 2020 (16 N-day su 25 sfruttate dopo un mese o più; dato ripreso nel post Anthropic del giugno 2026)
- Verizon Business, «2026 Data Breach Investigations Report», 19 maggio 2026, verizon.com/business/resources/reports/dbir/ (vettori di accesso iniziale: p. 10 e figura 10; remediation delle KEV, 26% e 38%, 43 e 32 giorni, 16 e 11 KEV mediane: p. 17, figure 12-14; analisi di sopravvivenza, 35% e 27% al giorno 28, limite del settimo giorno: p. 18, figura 15; probabilità di ri-sfruttamento e capacità degli attaccanti: p. 19-20, figura 18; collaborazione con Anthropic: p. 25-28; finestra dati 1° novembre 2024 - 31 ottobre 2025: p. 9); edizione 2025 per il contesto
- CISA, «Known Exploited Vulnerabilities Catalog»; Cyble, analisi annuali delle aggiunte 2024 e 2025
- Cloud Security Alliance, «AI-Compressed Attack Timeline and Capability Diffusion», research note, luglio 2026 (fonte di sintesi, non usata come dato)
- Google Threat Intelligence Group, «Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access», AI Threat Tracker, 11 maggio 2026, cloud.google.com/blog/topics/threat-intelligence (zero-day AI-assisted; jailbreak a persona; infrastruttura di accesso offuscato); J. Hultquist, dichiarazione a corredo del report, 11 maggio 2026
- NCSC (UK), CISA, NSA, ASD, CSE, GCSB, «The AI shift in cyber risk: why leaders must act now», dichiarazione congiunta, 22 giugno 2026, ncsc.gov.uk e cyber.gc.ca
- CISA, Binding Operational Directive 26-04, «Prioritizing Security Updates Based on Risk», e Implementation Guidance, 10 giugno 2026, cisa.gov
- Microsoft Security Response Center, «Exploitability Index», definizioni dei rating
- Direttiva (UE) 2022/2555 («NIS2»), artt. 21 e 23, recepita con D.lgs. 4 settembre 2024, n. 138
- ACN, Determinazione n. 155238 del 20 aprile 2026, «Categorie di rilevanza, processo, modalità e criteri per l'elencazione, caratterizzazione e categorizzazione delle attività e dei servizi di cui all'articolo 30 del decreto legislativo 4 settembre 2024, n. 138», in Gazzetta Ufficiale n. 178 del 3 agosto 2026, con Guida alla lettura e FAQ; Determinazioni n. 127434 e 127437 del 13 aprile 2026 (termini, piattaforma NIS, fornitori rilevanti)

- Regolamento (UE) 2024/2847 («Cyber Resilience Act»), art. 2 par. 2, art. 14 e considerando 25
- Regolamento (UE) 2017/745 sui dispositivi medici, allegato I; MDCG 2019-16, «Guidance on Cybersecurity for medical devices»
- T. Willis (Google Project Zero), «Policy and Disclosure: 2025 Edition», projectzero.google/2025/07/reporting-transparency.html, 29 luglio 2025 (Reporting Transparency e upstream patch gap)
- A. Ressler (Google Chrome Security), «An update on Chrome Security updates», Google Security Blog, 8 agosto 2023 (patch gap da 35 a 15 giorni con Chrome 77; aggiornamenti settimanali da Chrome 116)
- FIRST, «Exploit Prediction Scoring System (EPSS)», versione 4, 17 marzo 2025, first.org/epss, e FAQ (orizzonte di 30 giorni; esclusione di ambiente e controlli compensativi)
- CISA, «Stakeholder-Specific Vulnerability Categorization (SSVC) Guide», cisa.gov (punti di decisione Exploitation, Automatable, Technical Impact, Mission and Well-being)
- Regolamento delegato (UE) 2024/1774 (RTS DORA sul quadro di gestione del rischio ICT), art. 10; Regolamento di esecuzione (UE) 2024/2690 (NIS2, infrastrutture digitali), allegato
- NCSC (UK), «Impact of AI on cyber threat from now to 2027», NCSC Assessment, 7 maggio 2025, ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027; «The near-term impact of AI on the cyber threat», gennaio 2024
- Microsoft, Windows Insider Blog, build 25905 (componenti GDI in Rust, luglio 2023); D. Weston, BlueHat IL 2023 (Rust nel kernel di Windows)



t-defence

Next | Donexit | Foramil | Innodesi

Via Giacomo Peroni, 452 – 00131 Roma
tel. 06.45752720 – info@defencetech.it
www.t-defence.it

#TDefenceBusiness