planetica

2026

Cybersecurity report

“We are no longer securing computers, we are securing the society.”

Mikko Hyppönen

«Qualsiasi tecnologia sufficientemente avanzata è indistinguibile dalla magia»

Arthur C. Clarke

IL GRAAL, LA MAGIA E LE MACCHINE: L'ILLUSIONE DELL'INVULNERABILITÀ

Nel ciclo di affreschi The Quest and Achievement of the Holy Grail di Edwin Austin Abbey, il Graal è rappresentato come il fulcro di una forza invisibile, capace di guarire terre sterili e spezzare anatemi operando dietro le quinte del mondo visibile. Questa dimensione mitica trova un riscontro speculare nella terza legge di Arthur C. Clarke:

«Qualsiasi tecnologia sufficientemente avanzata è indistinguibile dalla magia».

Laddove manchi il linguaggio per decodificare un fenomeno, l'osservatore smette di vederne il codice e le infrastrutture e ne percepisce soltanto l'effetto improvviso. Oggi l'intelligenza artificiale agentica rischia di produrre lo stesso accecamento percettivo.Poiché questi sistemi non si limitano a rispondere a un comando, ma pianificano, orchestrano API e modificano lo stato dei sistemi con un alto grado di autonomia, l'utente è tentato di cedere al fascino di una "stregoneria digitale".

Nel campo della cybersecurity, però, scambiare la tecnologia per magia è il primo e più letale fattore di vulnerabilità. La sfida dell'AI agentica non è dunque meramente tecnica, ma innanzitutto di governance. Siamo chiamati a superare l'incanto passivo per assumere una supervisione attiva (human in the loop), definendo con rigore quali obiettivi affidiamo alle macchine, quali privilegi di accesso concediamo loro e quali confini etici imponiamo. La resilienza delle nostre infrastrutture dipende dall'assunzione di questa responsabilità conoscitiva: per proteggere il nostro ecosistema, dobbiamo comprendere fino in fondo l'architettura e il raggio d'azione dei nostri sistemi algoritmici.

SOMMARIO

PREFAZIONE

Il presente rapporto si propone come una bussola essenziale per orientarsi nell'**Era Agentica**, un punto di svolta epocale in cui l'intelligenza artificiale ha smesso di essere un mero strumento di assistenza per diventare un attore autonomo, capace di agire, decidere e orchestrare rapidamente i flussi di lavoro. Se il 2025 è stato l'anno dell'adozione di massa della GenAI, il 2026 segna l'inizio di una nuova realtà, definita dalla maggiore competenza offensiva degli attaccanti e dalla progressiva dissoluzione dei perimetri tradizionali.

L'analisi mette in luce come la **cybersecurity** sia oggi un'urgenza globale che incide direttamente sulla stabilità economica e sulla fiducia dei cittadini. Il 2025 ha registrato un'accelerazione senza precedenti della minaccia cyber, con un incremento degli incidenti del 48,7% e l'introduzione della categoria di severità "Extreme" per descrivere attacchi dagli impatti sistemici e devastanti. In questo scenario, l'identità digitale è diventata il nuovo, unico confine liquido: in un ecosistema di lavoro distribuito, dove le identità macchina superano quelle umane con un rapporto di 82 a 1, la gestione della fiducia non può più basarsi su permessi statici.

Particolare attenzione è dedicata alla trasformazione della governance. Il 2026 vede emergere un nuovo corso della responsabilità esecutiva, noto come "The New Gavel": direttive come **NIS2** e l'**AI Act** non sono più semplici adempimenti burocratici, ma impongono standard di responsabilità personale per i vertici aziendali, trasformando il rischio cyber in un rischio di business primario.

Il rapporto esplora inoltre le specificità dei settori critici, con un focus sul comparto Finance & Insurance. Qui, la resilienza operativa imposta dal regolamento **DORA** e l'impatto della cybersecurity sulle riserve di capitale (SMA) dimostrano come

una difesa robusta sia oggi un fattore abilitante per la competitività e la solidità patrimoniale. Dalle vulnerabilità del tessuto connettivo della supply chain alle nuove frontiere dell'**Agentic SOC**, il documento analizza come la difesa debba evolvere verso modelli predittivi e verso l'impiego di **Guardian Agent** per contrastare minacce iperautomatizzate.

In un contesto in cui l'Italia subisce il 9,6% degli attacchi globali e presenta un'anomalia significativa nell'**hacktivism** (+145% rispetto al 2024), la capacità di anticipare i rischi e integrare la sicurezza sin dalla fase di progettazione, secondo il principio del **Security by Design**, diventa determinante.

Questo Rapporto 2026 non vuole soltanto documentare un'escalation, ma offrire spunti operativi per costruire una base di fiducia fondata su visibilità e controllo dinamico, permettendo alle organizzazioni di adottare con coraggio, e con maggiore consapevolezza, le innovazioni del futuro.

Andrea Rivetti
Amministratore & Equity Partner

Matteo Marzan
Amministratore & Equity Partner

INTRODUZIONE

Siamo sulla soglia della cosiddetta **Intelligence Age**: un'epoca in cui la tecnologia ha smesso di essere un semplice strumento per diventare una forza capace di ridisegnare il tessuto stesso del business e della società. Se il 2025 ha rappresentato il tramonto dell'era dell'IA assistita, il 2026 segna l'inizio dell'era dell'autonomia, o **AI-driven era**. Si tratta di un cambiamento di paradigma profondo, in cui l'integrazione di agenti intelligenti nei processi chiave del business richiede un approccio radicalmente nuovo alla sicurezza digitale. In questo scenario, la minaccia non si configura più come una sequenza di eventi isolati, ma come una vera e propria industrializzazione dell'offesa. Gli aggressori hanno imparato a sfruttare l'intelligenza artificiale agentica non solo per generare contenuti, ma per automatizzare l'intera catena d'attacco, dalla scansione predittiva delle vulnerabilità fino all'esfiltrazione invisibile dei dati. Ciò riduce drasticamente il tempo che intercorre tra l'intento criminale e l'esecuzione. Le analisi più recenti confermano che i sistemi agentici possono fungere da potenti moltiplicatori di forza, capaci di **gestire autonomamente fino all'80-90% della kill chain**[1] sotto la sola direzione strategica di un singolo operatore umano.

Parallelamente, la difesa si trova a gestire il dilemma della fiducia. L'autonomia che rende gli agenti IA così straordinariamente produttivi crea, per sua natura, punti ciechi che le difese tradizionali, progettate per minacce lineari e prevedibili, non sono in grado di monitorare efficacemente. . Il rischio, oggi, non deriva soltanto dall'intento malevolo di un attore esterno, ma anche dalle modalità stesse di fallimento

[1] Modello concettuale che descrive la struttura di un attacco cibernetico come una sequenza lineare di fasi necessarie al suo successo. Il framework originale (Cyber Kill Chain), sviluppato da Lockheed Martin, si articola in sette stadi consecutivi: ricognizione, militarizzazione, consegna, sfruttamento, installazione, comando e controllo, e azioni sugli obiettivi. L'interruzione di una qualsiasi di queste fasi permette di neutralizzare l'intero attacco.

dei sistemi, qualora vengano lasciati operare senza una governance rigorosa o in presenza di vulnerabilità nei protocolli di comunicazione tra macchine.

In un contesto globale segnato da una forte frammentazione geopolitica, la cybersecurity è diventata a tutti gli effetti un asset strategico di sovranità e stabilità economica. L'Italia, pur dimostrando una crescente attenzione sul fronte regolatorio, si trova ad affrontare una sfida strutturale legata al divario storico di competenze e alla fragilità di supply chain digitali sempre più interconnesse. In una rete così fitta, non è più sufficiente proteggere il singolo asset aziendale: è necessario garantire la resilienza dell'intero ecosistema, dove una falla locale può rapidamente generare conseguenze sistemiche su scala globale.

Questo rapporto non si limita a documentare l'escalation degli attacchi, ma analizza il cambio di paradigma rappresentato dalla filosofia dell'**Assume Breach**, secondo cui il perimetro deve essere considerato già compromesso. Attraverso l'analisi delle nuove superfici di attacco, dalle API non protette ai Digital Twin delle infrastrutture critiche, il documento traccia la rotta per una difesa che non sia soltanto reattiva, ma anticipatoria e adattiva. Solo integrando la cybersecurity sin dalla fase di progettazione e valorizzando l'automazione difensiva come leva operativa, le organizzazioni potranno trasformare la vulnerabilità in un vantaggio competitivo nell'era dell'autonomia digitale.

Analisi dei principali incidenti cyber noti a livello globale

In questa sezione si presenta una panoramica dei principali incidenti di sicurezza informatica di dominio pubblico verificatisi a livello globale nel 2025, con un confronto rispetto ai dati rilevati nei quattro anni precedenti.
L'analisi si basa sugli attacchi informatici noti che hanno avuto esito positivo e che, per gravità e portata, hanno generato impatti significativi sulle organizzazioni coinvolte sotto il profilo economico, tecnologico, legale e reputazionale.
Nel periodo compreso tra gennaio 2021 e dicembre 2025 sono stati censiti complessivamente 16.123 incidenti, distribuiti come illustrato nella Figura 1.

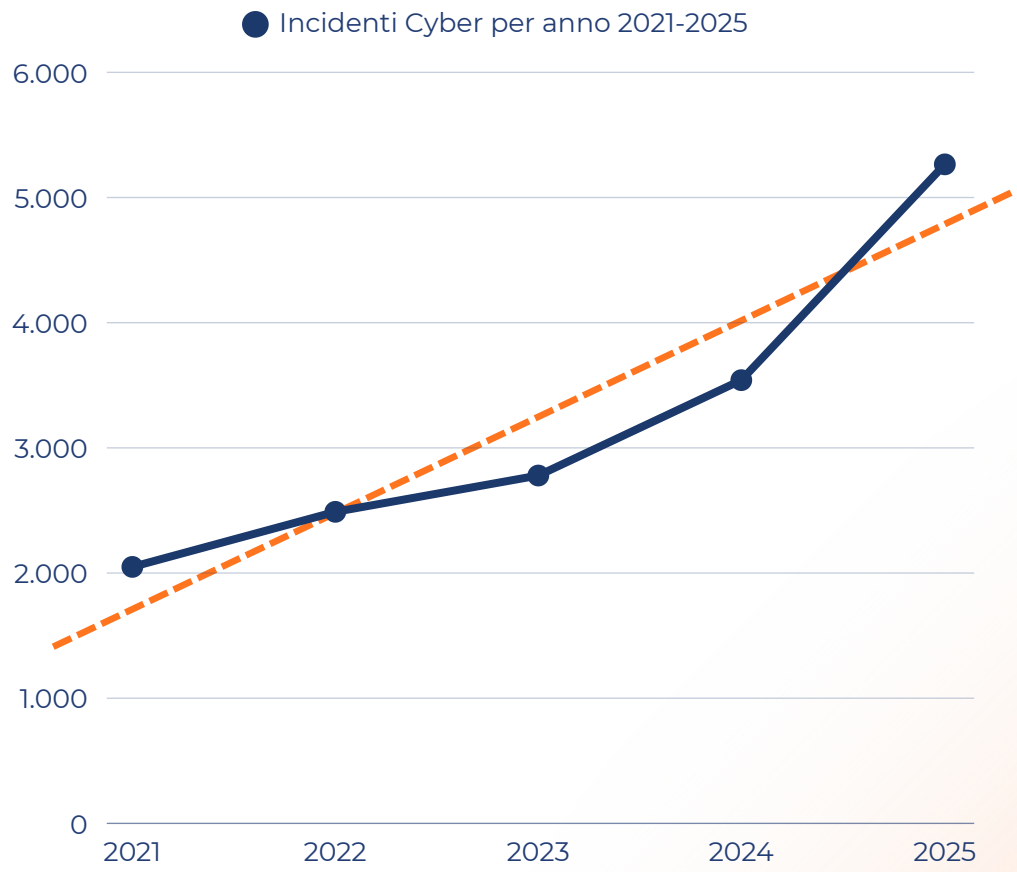


Figura 1: Andamento degli indicenti cyber nel periodo 2021-2025

Rispetto ai 2.049 incidenti rilevati nel 2021, il dato registrato nel 2025 evidenzia una crescita del 157%, indicando che il fenomeno della cybercriminalità sta attraversando una fase di marcata accelerazione. Tale incremento non può essere attribuito esclusivamente alla naturale evoluzione delle minacce, ma suggerisce la presenza di fattori strutturali che contribuiscono ad ampliare capacità, scala e velocità degli attacchi. Tra questi, un ruolo rilevante è verosimilmente svolto dalla crescente diffusione delle tecnologie di Intelligenza Artificiale, oggi sempre più utilizzate anche da attori malevoli.
Come evidenziato nella Figura 2, anche il numero medio mensile di incidenti mostra una crescita particolarmente significativa: dai 171 incidenti mensili registrati nel 2021 si passa a 295 nel 2024, fino a raggiungere i **439** nel 2025. A differenza di quanto osservato nel 2024, la distribuzione degli incidenti nel corso del 2025 risulta maggiormente concentrata nella prima parte dell'anno. Nel periodo compreso tra gennaio e luglio, infatti, il numero di eventi supera la media mensile in tutti i mesi, ad eccezione di febbraio, evidenziando una forte intensificazione dell'attività offensiva già nei primi mesi dell'anno.

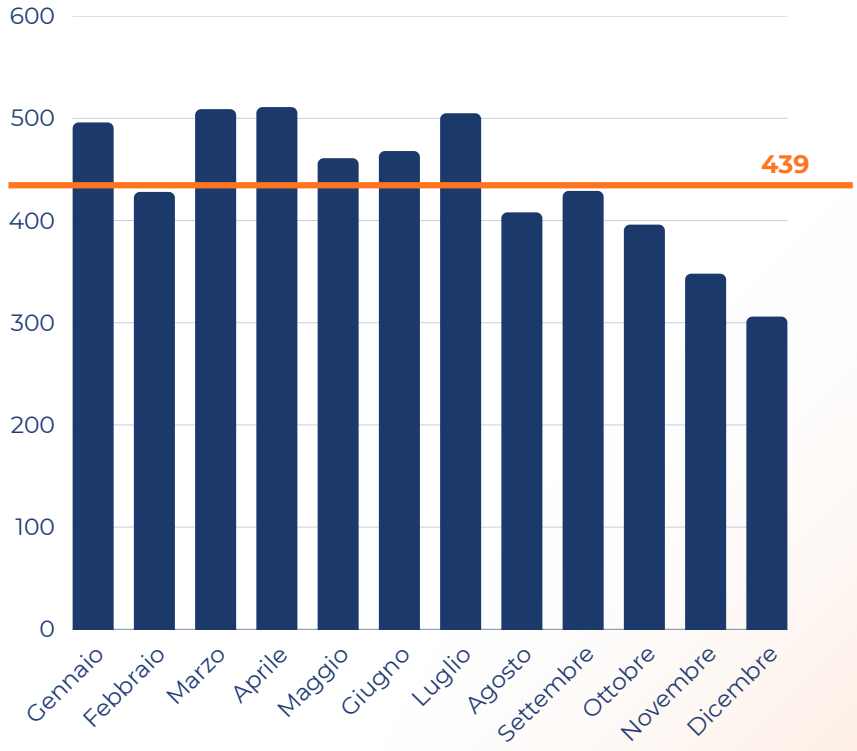


Figura 2: Numero di incidenti cyber per mese nel mondo nel 2025

Distribuzione degli attaccanti per tipologia

Anche nel 2025 il cybercrime si conferma la principale motivazione alla base degli incidenti di sicurezza informatica (Figura 3), rafforzando ulteriormente una tendenza ormai consolidata negli ultimi anni. La crescita del fenomeno non mostra segnali di rallentamento: la **criminalità informatica** è infatti responsabile dell'89,3% degli incidenti rilevati, pari a quasi nove eventi su dieci.

Rispetto al 2024, l'incidenza del cybercrime sul totale degli incidenti aumenta di ulteriori tre punti percentuali, confermando che le finalità economiche continuano a rappresentare il principale motore delle attività malevole nel panorama delle minacce digitali.

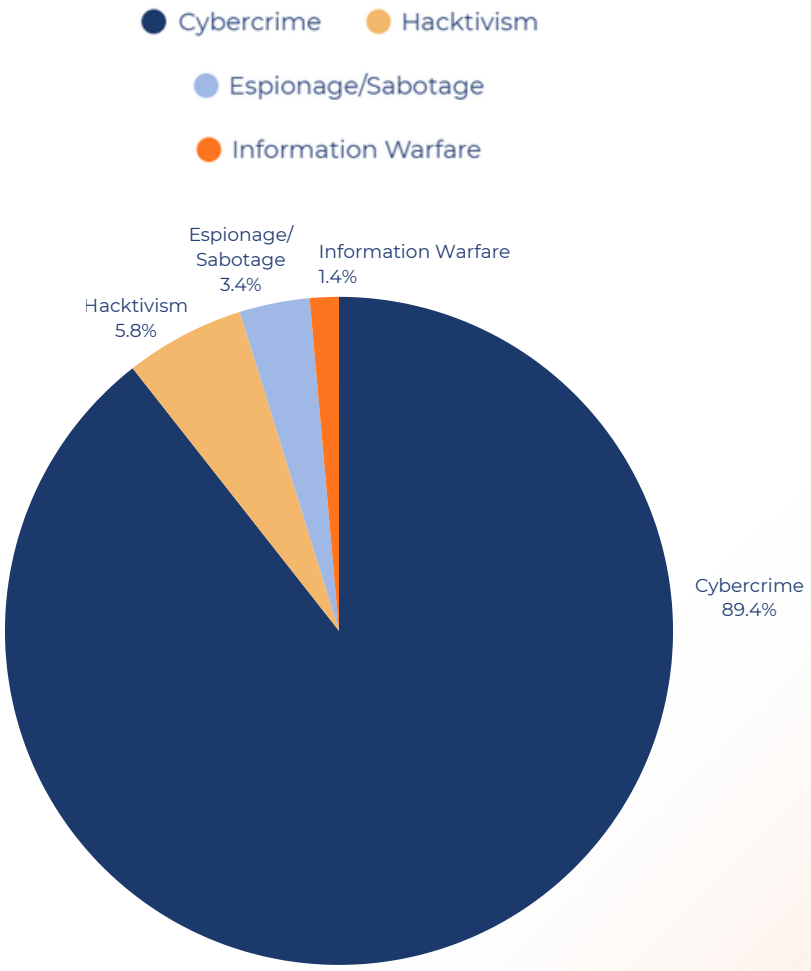


Figura 3: Distribuzione percentuale degli attaccanti nel 2025

L’analisi della distribuzione delle tipologie di attaccanti nel periodo 2021–2025 (Figura 4) evidenzia con chiarezza il ruolo sempre più dominante del cybercrime nel panorama delle minacce informatiche. La crescita osservata nel corso degli anni non solo si conferma costante, ma mostra segnali di ulteriore accelerazione: nel 2025, il numero di incidenti riconducibili alla criminalità informatica registra un incremento del 55% rispetto all’anno precedente. Questo andamento rafforza quanto già emerso nei precedenti rapporti: la progressiva convergenza tra criminalità tradizionale e criminalità digitale contribuisce alla creazione di modelli organizzativi sempre più strutturati e redditizi. I proventi derivanti dalle attività illecite vengono reinvestiti per accrescere competenze, strumenti e capacità operative degli attaccanti, alimentando un circolo virtuoso per gli aggressori e rendendo il fenomeno sempre più difficile da contrastare. Continua a crescere anche il fenomeno dell’hacktivism, sebbene a un ritmo più contenuto rispetto al cybercrime (+10% rispetto al 2024), mentre le categorie riconducibili a spionaggio, sabotaggio e information warfare mantengono livelli sostanzialmente stabili, senza variazioni significative nella loro incidenza complessiva.

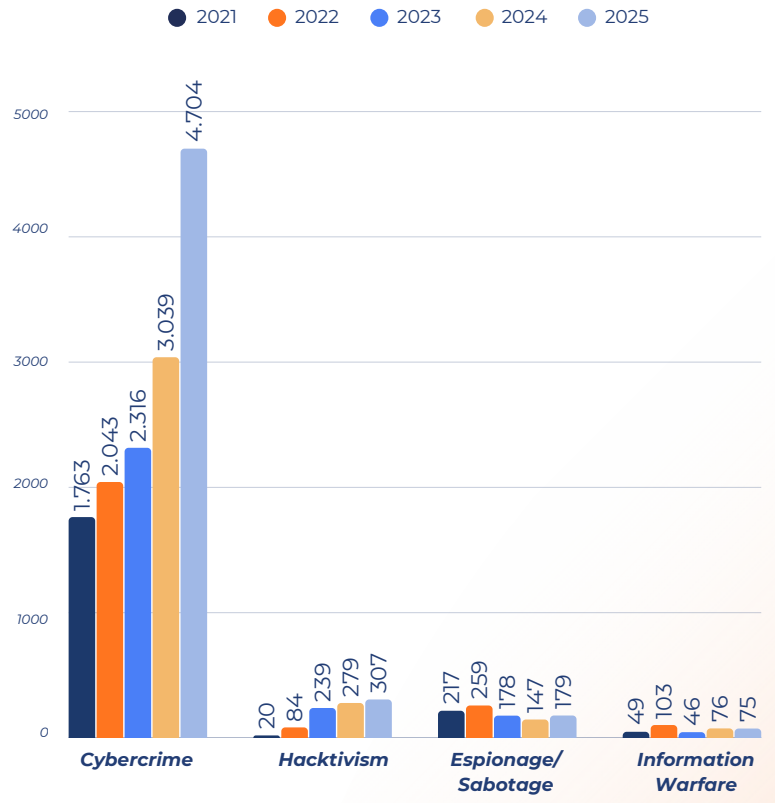


Figura 4: Distribuzione degli attaccanti dal 2021 al 2025

Distribuzione delle vittime per categoria

L'analisi delle vittime nel 2025 (Figura 5) mostra che quasi la metà degli incidenti registrati (46%) si concentra nelle prime tre categorie: **Multiple Targets** (23% del totale), **Gov / Mil / LE** (12%) e **Healthcare** (11%).

Le campagne rivolte indiscriminatamente a più settori (Multiple Targets) si confermano le più efficaci, registrando una crescita del 96% rispetto all'anno precedente. Si tratta di un dato particolarmente rilevante, che evidenzia sia la crescente capacità degli attori malevoli di scalare le proprie operazioni, sia, soprattutto, la vulnerabilità strutturale delle infrastrutture tecnologiche delle organizzazioni colpite. I risultati evidenziano come sia ancora possibile compromettere un numero elevato di vittime colpendo strumenti tecnologici di uso comune oppure ricorrendo a tecniche di attacco standard, senza necessariamente impiegare metodologie avanzate o altamente specializzate. A questo scenario si aggiunge il crescente utilizzo dell'intelligenza artificiale, che favorisce l'automazione di attività offensive che, fino a poco tempo fa, richiedevano un intervento manuale.

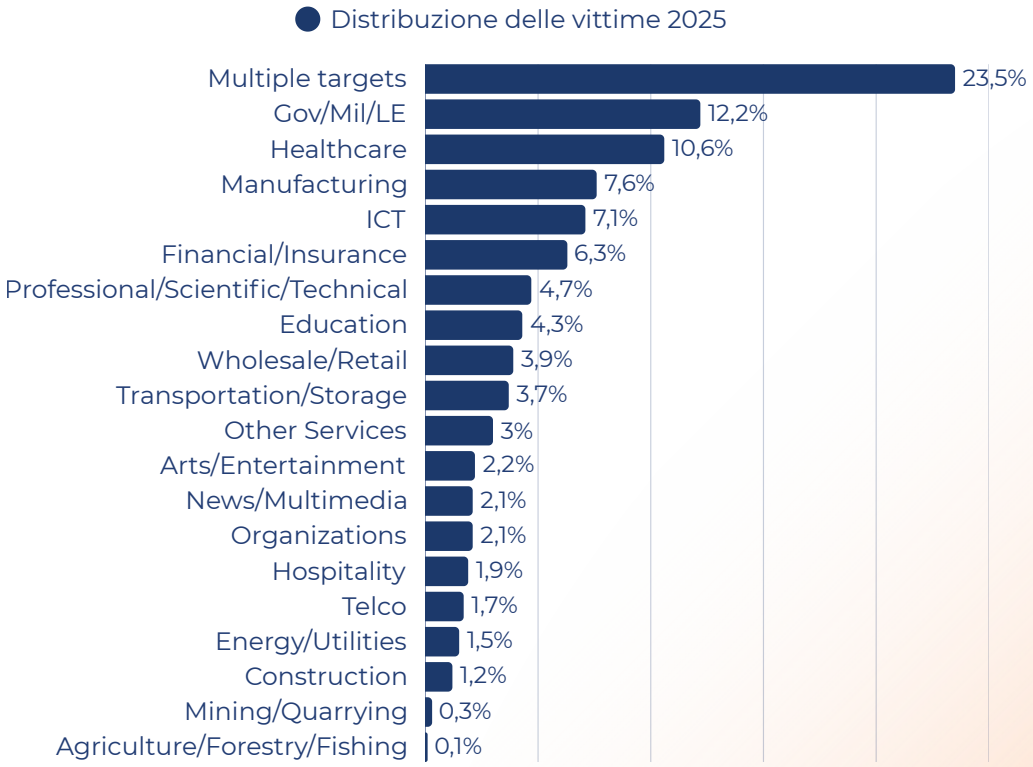


Figura 5: Distribuzione della tipologia di vittime nel 2025

I settori Gov / Mil / LE e Healthcare continuano a rappresentare obiettivi particolarmente attrattivi per gli attori malevoli, sia per il ruolo strategico che ricoprono sia per l'elevata sensibilità delle informazioni gestite. Il confronto con gli anni precedenti (Fig. 6) evidenzia una crescita costante degli incidenti in entrambi gli ambiti: Gov / Mil / LE registra un incremento del 37% rispetto al 2024, tornando al secondo posto, mentre Healthcare, pur crescendo del 19%, scende in terza posizione.

Al quarto posto tra i settori più colpiti si colloca nuovamente il **Manufacturing**, che a livello globale registra un aumento del 79% dopo la lieve flessione osservata nel 2024, anno in cui occupava la settima posizione. Rimane invece stabile al quinto posto il settore **ICT**, che, dopo quattro anni sostanzialmente invariati — con una leggera contrazione nel 2024 — registra nel 2025 un incremento del 46% degli incidenti.

Si tratta di un segnale particolarmente critico per un comparto che, teoricamente, dovrebbe poter contare su maggiori competenze specialistiche e su investimenti più consistenti in ambito tecnologico, elementi che dovrebbero tradursi in una migliore capacità di difesa rispetto a minacce sempre più evolute. Questo andamento impone inoltre una riflessione sui rischi sistemici connessi al crescente e inevitabile ricorso all'outsourcing di servizi tecnologici da parte di organizzazioni di ogni settore e dimensione.

Anche il comparto **Financial / Insurance**, pur perdendo due posizioni nella top 10 rispetto al 2024, registra un aumento del 27% degli incidenti in valore assoluto, riallineandosi ai livelli del 2023. Il settore **Professional / Scientific / Technical,** invece, risale dalla decima alla settima posizione, registrando una crescita del 91%. Nel complesso, nel 2025 i primi sette settori della classifica mostrano tutti una crescita a doppia cifra percentuale del numero di incidenti. L'unica eccezione tra i primi dieci comparti è rappresentata dal settore **Education**, che registra una lieve diminuzione (-3%).

Il settore **Wholesale / Retail** mantiene la nona posizione, mentre Transportation / Storage entra per la prima volta tra i primi dieci a livello globale. Esce invece dalla classifica il settore **News / Multimedia**, che nel 2024 aveva registrato un forte aumento degli incidenti legati a vulnerabilità presenti nei Content Management System più diffusi.

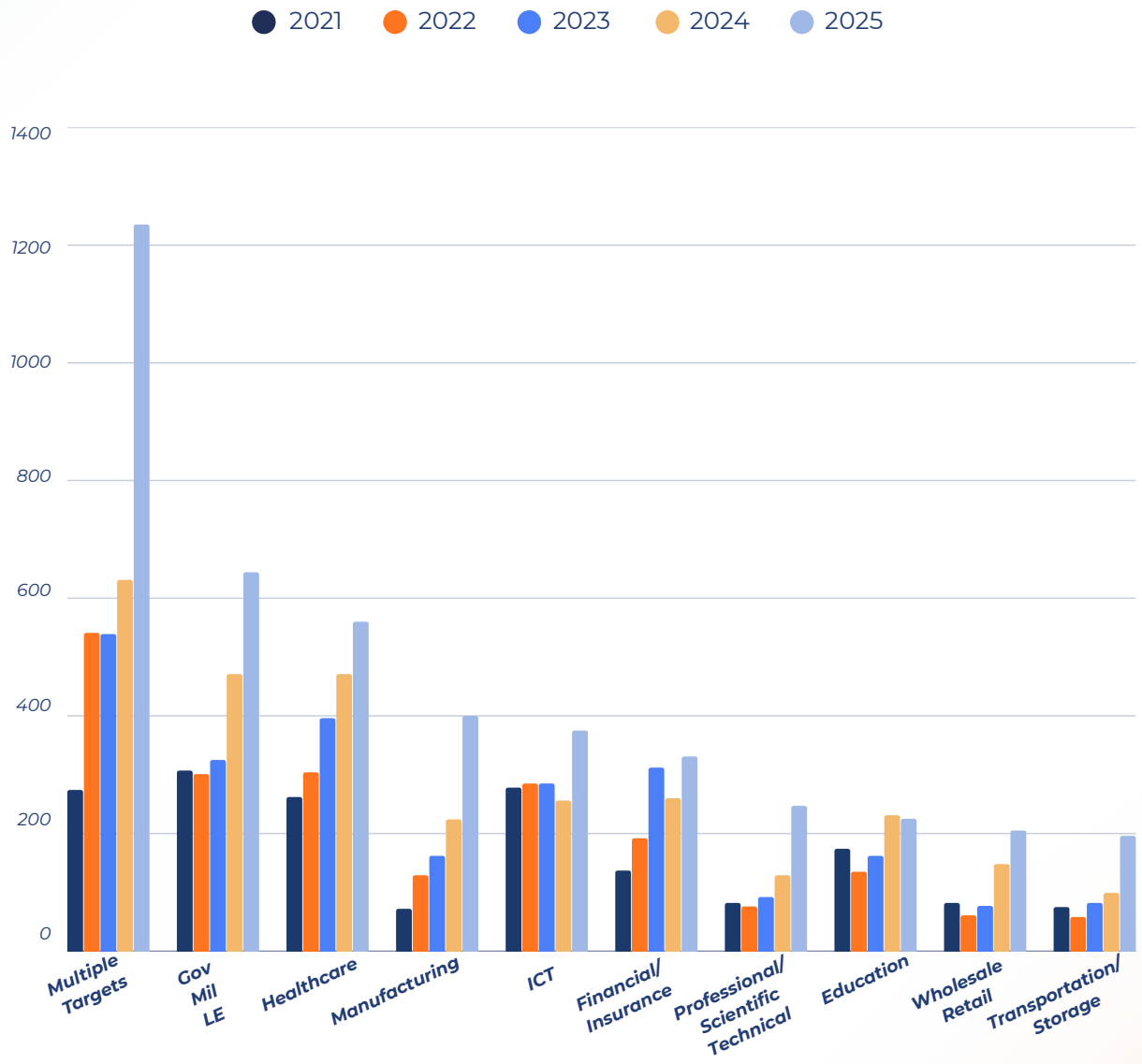


Figura 6: Distribuzione delle prime 10 tipologie di vittime dal 2021 al 2025

Distribuzione delle tecniche di attacco

Per oltre un terzo degli incidenti registrati nel 2025 (+8 punti percentuali) non è stato possibile identificare con precisione la tecnica di attacco utilizzata (**undisclosed**). Questo dato evidenzia un apparente paradosso: nonostante il progressivo rafforzamento delle normative che impongono alle organizzazioni di notificare gli incidenti informatici alle autorità competenti, alle vittime e agli stakeholder, tale obbligo non si traduce necessariamente in una maggiore trasparenza pubblica in merito ai dettagli tecnici degli attacchi.

Sempre più frequentemente, infatti, le comunicazioni ufficiali si limitano alle informazioni strettamente richieste dalla normativa, evitando di divulgare elementi tecnici approffonditi. Questa scelta risponde all'esigenza di contenere possibili danni reputazionali, di non evidenziare eventuali criticità strutturali della propria postura di sicurezza, di ridurre il rischio di contenziosi legali e di non compromettere eventuali indagini ancora in corso.

Una maggiore trasparenza tecnica, se gestita in modo strutturato e responsabile, potrebbe tuttavia favorire una collaborazione più efficace tra i soggetti impegnati nella difesa cyber, replicando dinamiche di condivisione delle informazioni che da tempo caratterizzano il contesto cybercriminale. Un approccio di questo tipo contribuirebbe a ridurre l'attuale squilibrio dello scenario cyber e a rafforzare la resilienza complessiva dell'ecosistema digitale.

Anche nel 2025 i **malware** si confermano la tecnica nota più utilizzata per condurre attacchi con successo, registrando un incremento del 18% rispetto al 2024 e consolidando il proprio ruolo tra gli strumenti più efficaci a disposizione del cybercrime (Fig. 7).

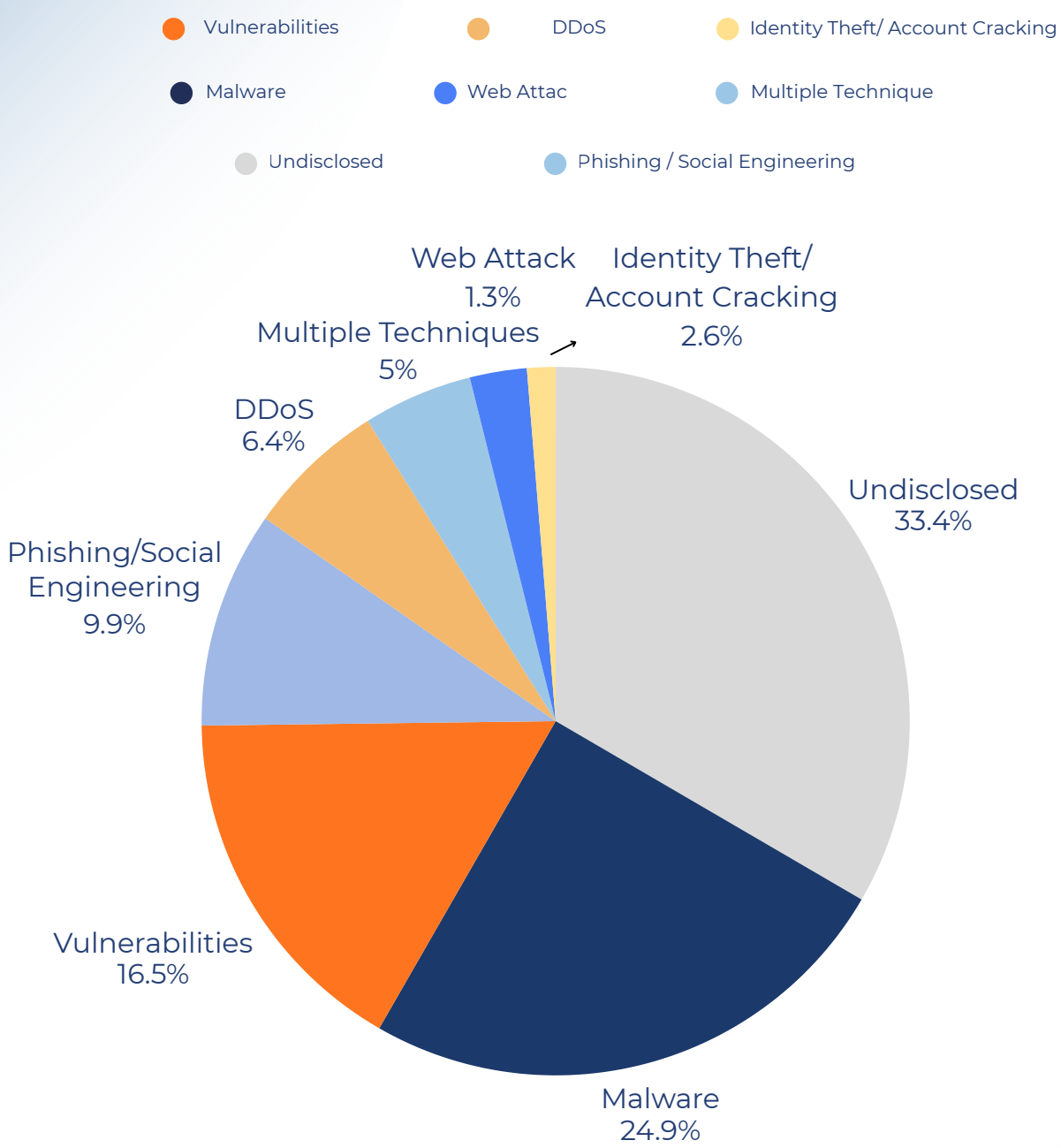


Figura 7: Distribuzione delle tecniche di attacco nel 2025

Lo sfruttamento delle vulnerabilità si conferma stabilmente al secondo posto tra le principali tecniche di attacco (Fig. 8), nonostante una crescita del 65% rispetto all'anno precedente. **Phishing** e **Social Engineering**, invece, con un incremento del 75%, salgono al terzo posto. Si tratta di un'evoluzione attesa da tempo, favorita soprattutto dall'impiego sempre più esteso dell'intelligenza artificiale nella sofisticazione delle attività di ingegneria sociale, che ne ha aumentato sensibilmente l'efficacia.

Gli incidenti riconducibili a tecniche di **Denial of Service**, utilizzate prevalentemente in contesti di hacktivism, registrano un aumento del 12% in valore assoluto, ma scendono di una posizione tra le tecniche più diffuse.
Cresce inoltre il ricorso a tecniche multiple (+58%), generalmente associate a operazioni particolarmente articolate e spesso attribuibili a gruppi state-sponsored. Anche i **Web Attack** mostrano un incremento significativo (+62% rispetto al 2024), pur rappresentando ancora una quota limitata del campione complessivo (1,3%). L'unica tecnica in diminuzione è rappresentata dall'**Identity Theft / Account Cracking**, che si colloca in penultima posizione tra le principali modalità di attacco osservate.

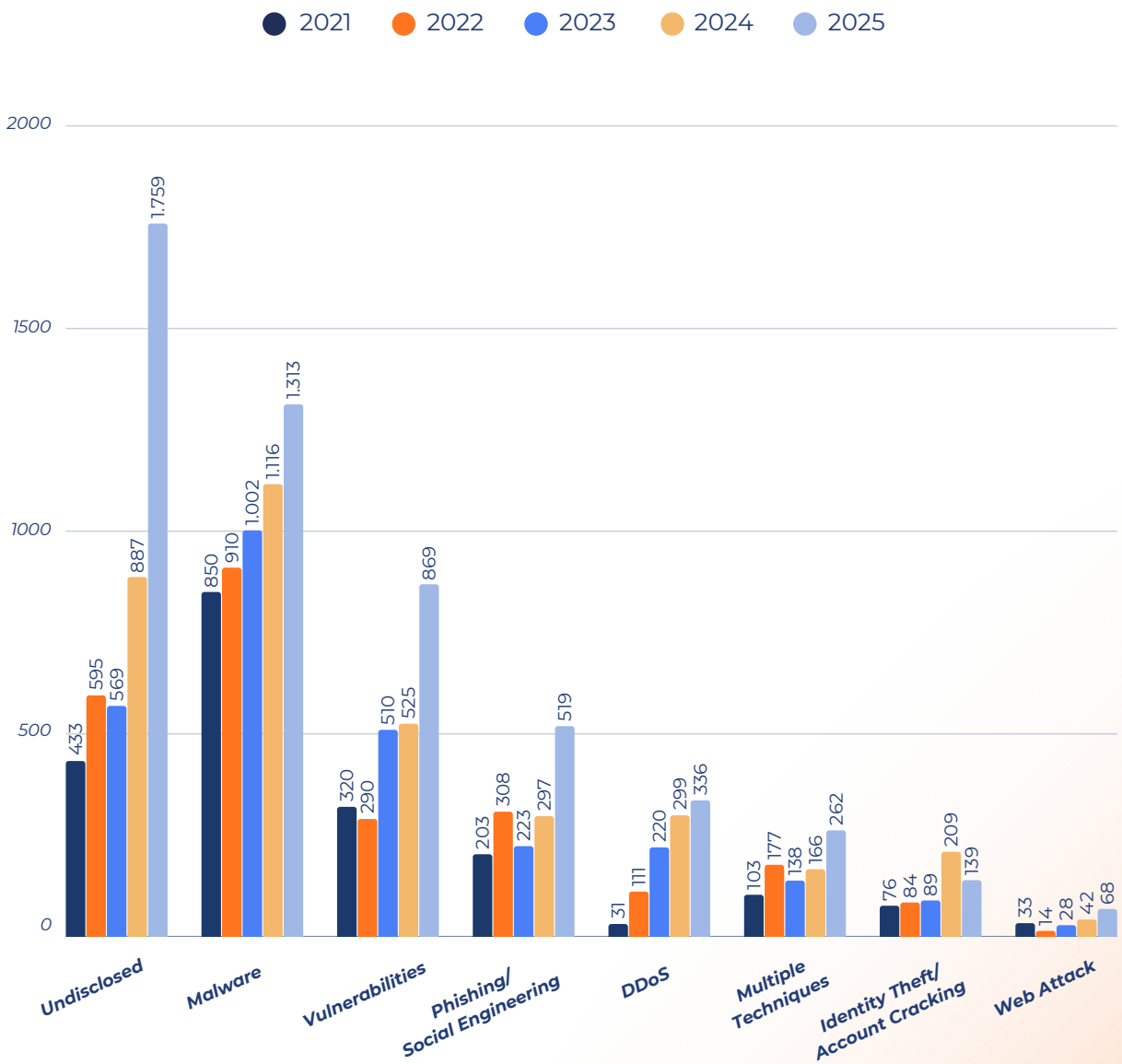


Figura 8: Distribuzione delle tecniche di attacco nel periodo 2021-2025

Analisi della “severity” degli incidenti

L’**analisi della severity** ha l’obiettivo di evidenziare il livello di **impatto degli incidenti cyber**, che non sempre risulta proporzionale al numero degli attacchi né può essere dedotto esclusivamente dalla tipologia di vittima coinvolta o dalla tecnica utilizzata. A partire dal 2021 (Fig. 9) si è consolidata una tendenza particolarmente preoccupante: la crescita costante degli incidenti caratterizzati da impatti elevati (High o Critical), che nel 2024 rappresentavano già l’80% del totale. Parallelamente, dal 2022 si osserva una progressiva scomparsa degli incidenti classificabili come “Low”, segnale di un generale aggravamento dello scenario cyber. Alla luce di questa evoluzione, nel 2025 è stata introdotta una **nuova categoria** di severità, denominata **Extreme**, destinata a identificare gli incidenti meno numerosi ma caratterizzati da conseguenze particolarmente gravi, pervasive e sistemiche. Contestualmente, nella fascia inferiore della scala di severità, le precedenti categorie **“Medium” e “Low”** sono state accorpate, con l’obiettivo di rendere la classificazione più coerente con l’attuale distribuzione degli impatti e con il cambiamento qualitativo delle minacce osservate. La riclassificazione degli incidenti degli ultimi cinque anni secondo i nuovi criteri mette in evidenza alcuni elementi di particolare rilievo. In primo luogo, gli eventi riconducibili alla nuova categoria **“Extreme”** risultano presenti esclusivamente nel 2025, a conferma del fatto che l’evoluzione dello scenario non si manifesta soltanto attraverso un aumento numerico degli incidenti, ma anche mediante una crescita dell’intensità e della gravità dei singoli eventi. Emergono inoltre dinamiche significative nella distribuzione delle categorie di severità: l’incremento più consistente in valore assoluto riguarda gli incidenti classificati come High, mentre il numero degli eventi a minore impatto (Medium/Low) rimane sostanzialmente stabile.

Nel dettaglio, gli incidenti **High** registrano una crescita del 66% rispetto all'anno precedente e arrivano a rappresentare il 55% del totale (Fig. 10). Gli eventi **Medium/Low** diminuiscono invece del 7% in valore assoluto, scendendo al di sotto del 15% del campione complessivo. Il dato più critico riguarda tuttavia gli incidenti Critical, che aumentano del 46% su base annua. Se a questi si aggiungono gli eventi riclassificati nella nuova categoria **Extreme**, la crescita complessiva raggiunge il 60% nel 2025. Complessivamente, gli incidenti appartenenti alle categorie **Critical ed Extreme** costituiscono circa un terzo degli eventi analizzati nel campione.

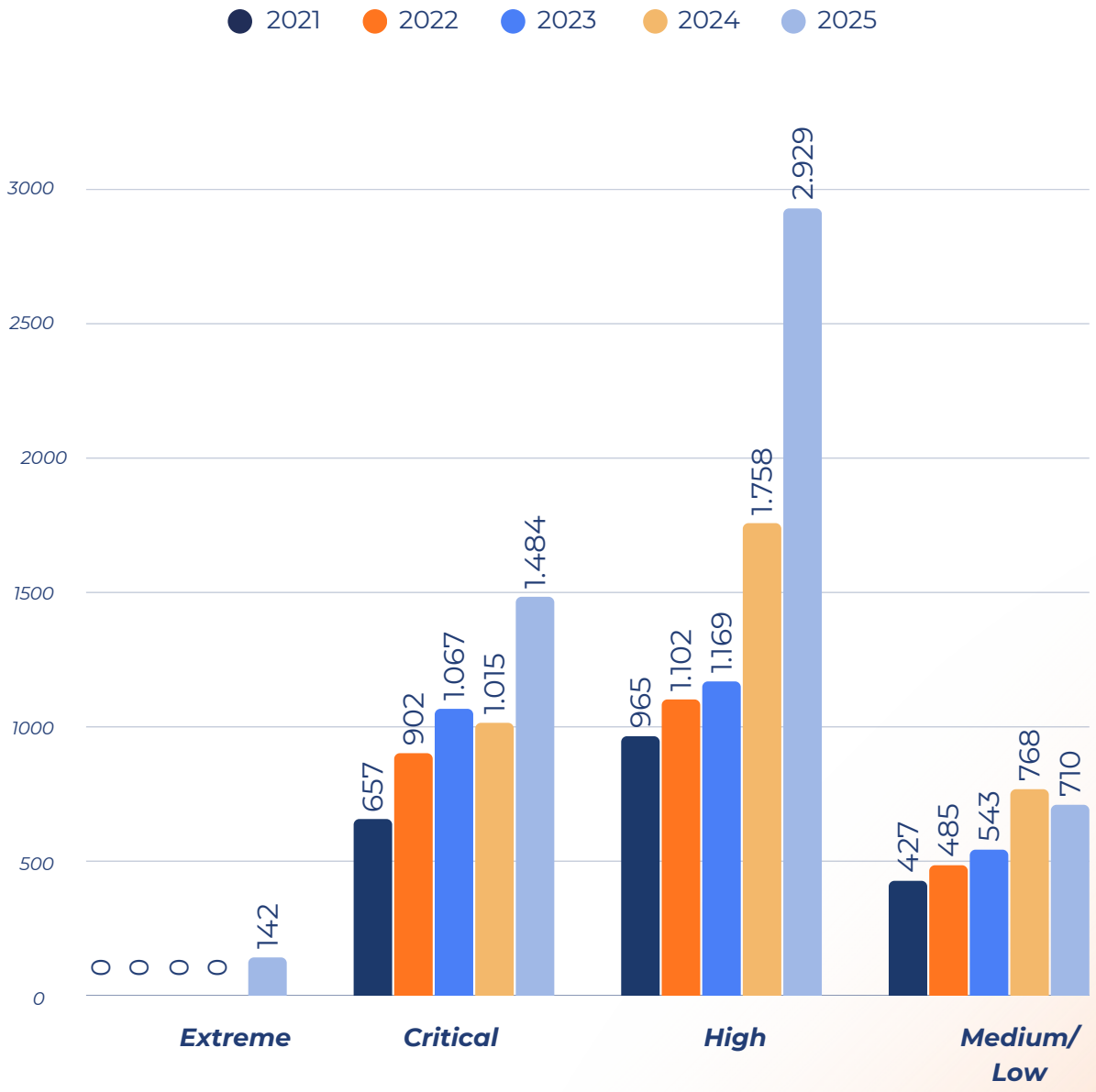


Figura 9: Andamento della Severity degli incidenti nel periodo 2021-2025

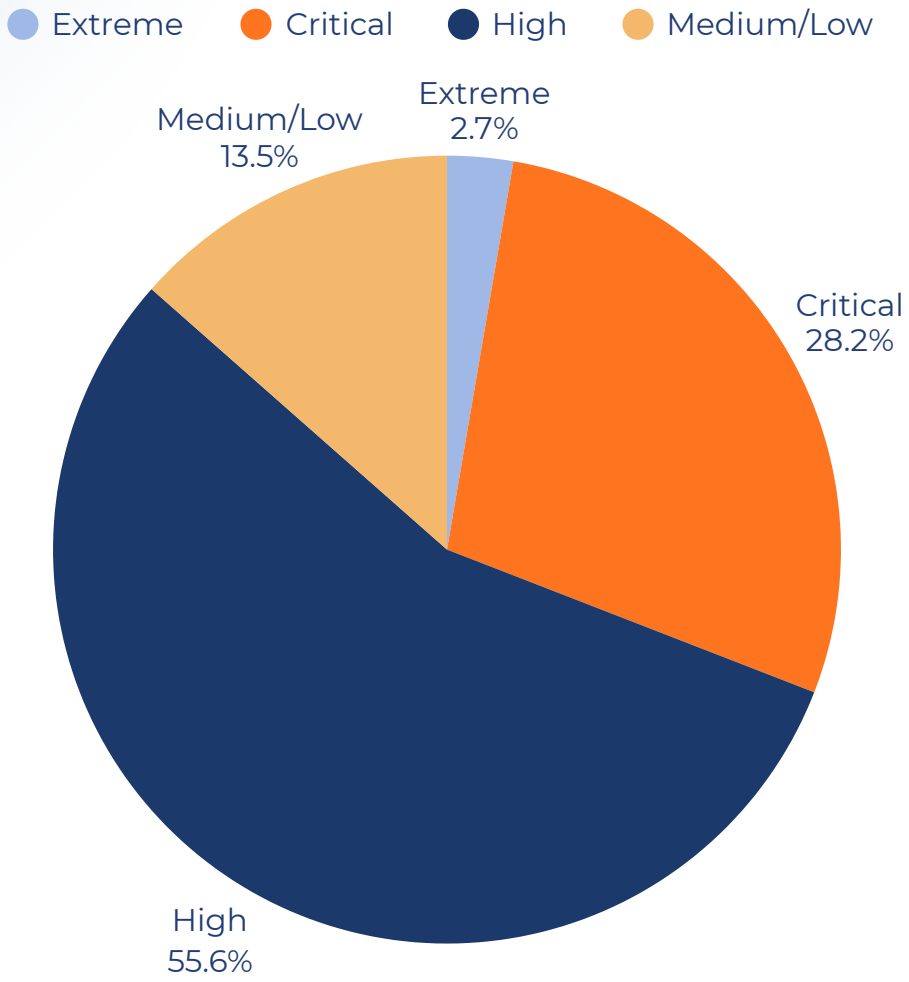


Figura 10: Distribuzione della Severity nel 2025

Analisi degli incidenti in Italia

Tra il 2021 e il 2025, il campione ha incluso 1.432 incidenti noti di particolare gravità che hanno avuto come bersaglio organizzazioni italiane. Di questi, 507, pari a circa il 35% del totale, si sono verificati nell'ultimo anno considerato. Come evidenziato dal grafico (Fig. 11), il dato relativo al 2025 mostra una nuova, lieve accelerazione. Nei due anni precedenti, infatti, il numero di incidenti si era mantenuto su valori relativamente stabili, intorno ai 300 eventi (310 nel 2023 e 357 nel 2024), con un incremento complessivo di circa il 15%. Nel 2025, invece, si osserva una crescita più marcata, pari al 42% (Fig. 12), leggermente inferiore al tasso di incremento globale, che supera il 48,7%.

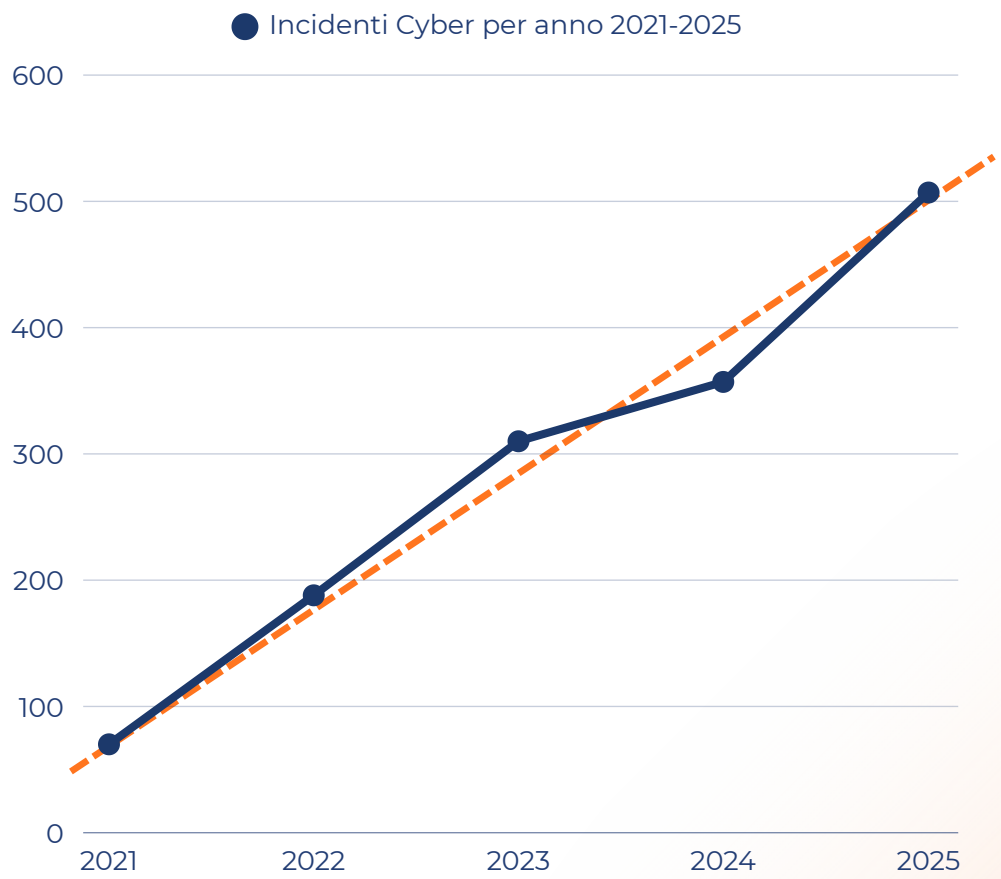


Figura 11: Distribuzione degli indicenti cyber in Italia nel periodo 2021-2025

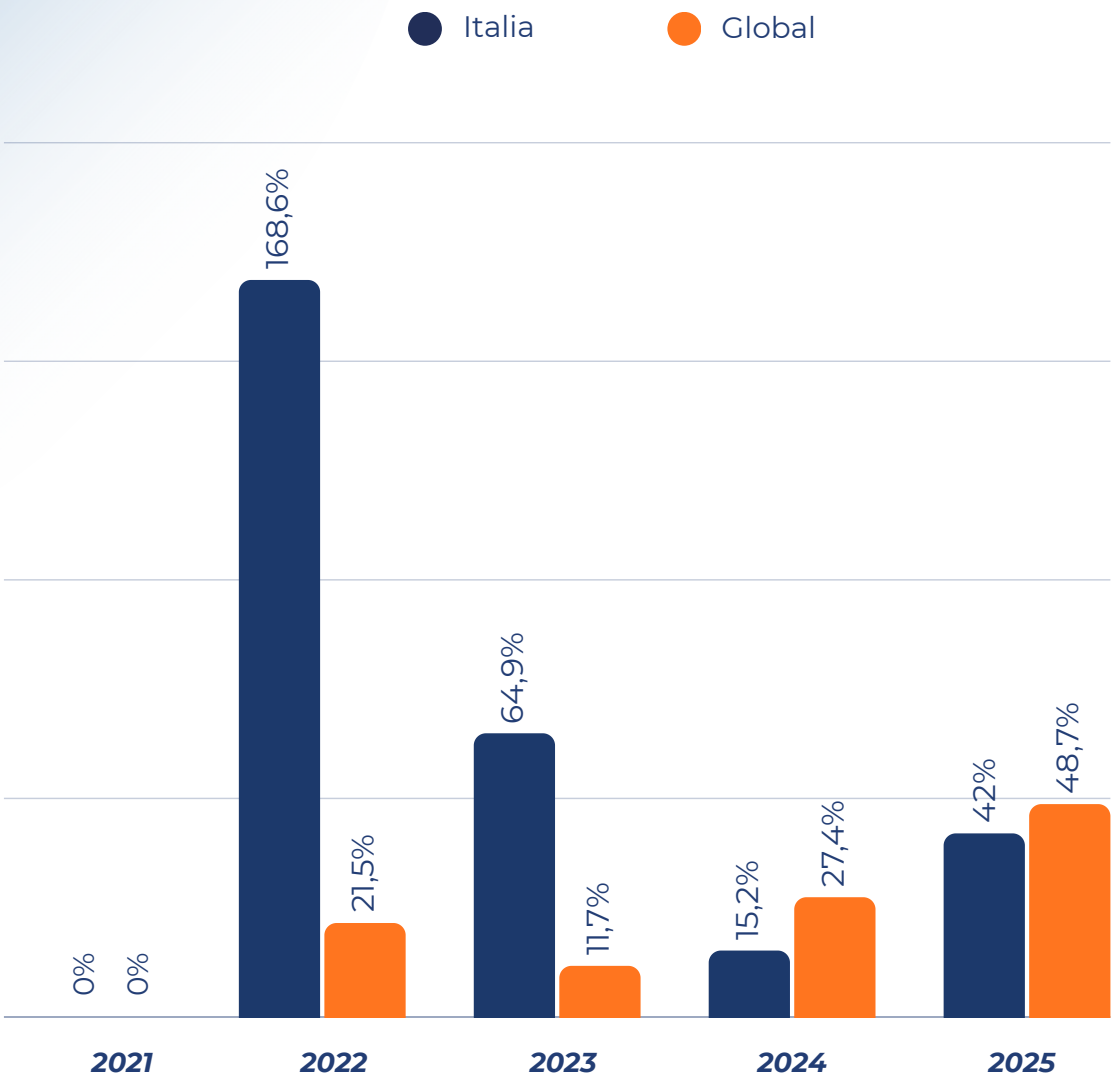


Figura 12: Confronto crescita percentuale in Italia vs. Global nel periodo 2021-2025

In relazione al dato globale, si osserva un'ulteriore lieve riduzione dell'incidenza degli incidenti che coinvolgono organizzazioni italiane sul totale (Fig. 13). Nel 2025, infatti, tali eventi rappresentano il 9,6% del campione complessivo degli incidenti rilevati a livello mondiale.

Nonostante il lieve calo, il valore resta comunque significativo e non si discosta in modo sostanziale dal picco registrato nel 2023 (11,2%), confermando un'**esposizione del contesto italiano ancora rilevante nel panorama cyber globale**.

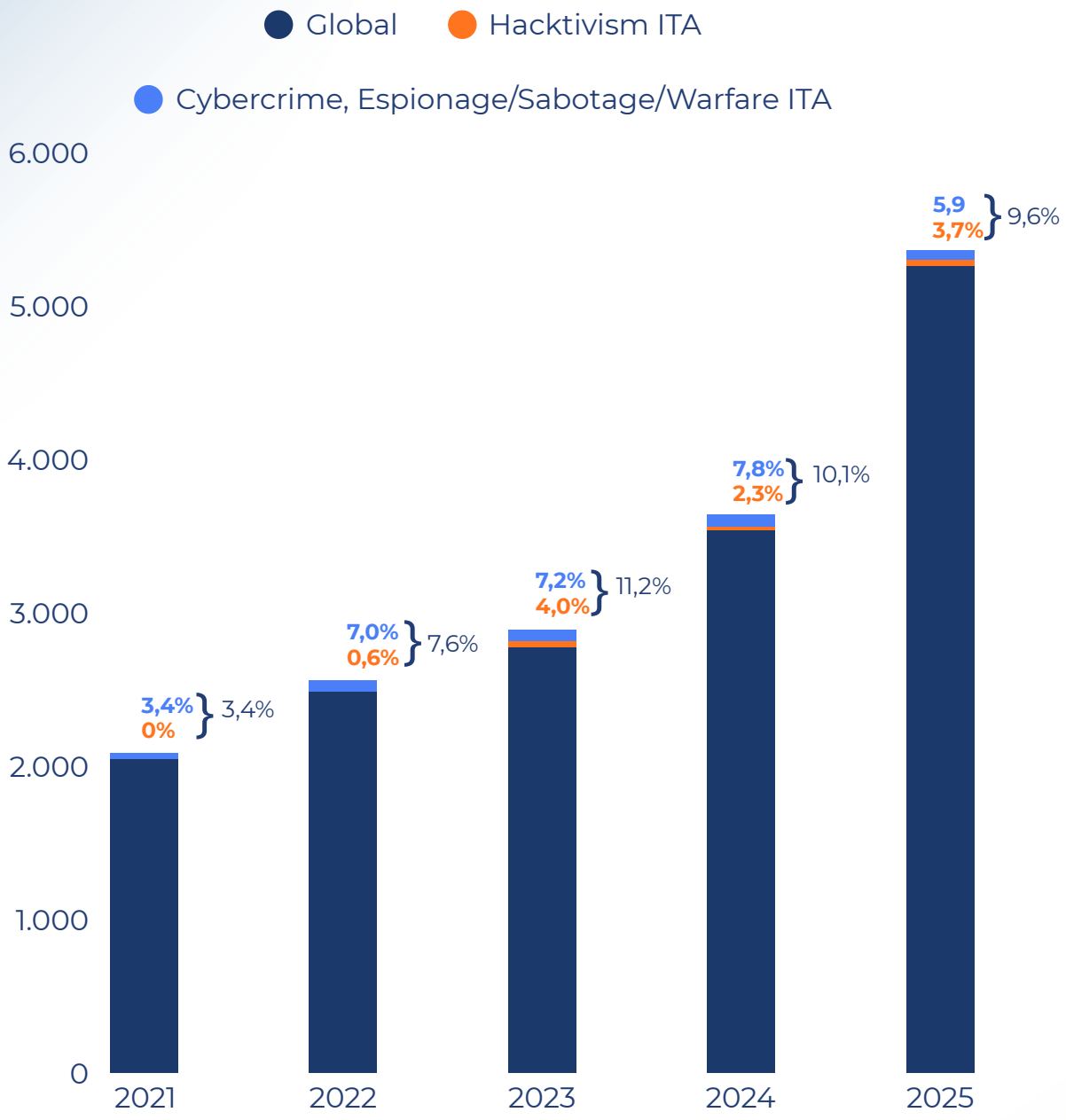


Figura 13: Incidenza degli incidenti in Italia rispetto al campione globale - 2021-2025

Distribuzione degli attaccanti per tipologia

Il panorama degli incidenti, analizzato in base alla tipologia degli attaccanti, conferma quanto osservato negli ultimi anni. In Italia risultano principalmente attive due categorie di threat actor: **Cybercriminal** e **Hacktivist**.

Nel 2025 emerge inoltre una marginale incidenza di eventi attribuibili alla categoria **Espionage / Sabotage** (0,4%; Fig. 27), rispetto al 3% registrato a livello globale (Fig. 14), mentre non si rilevano casi riconducibili alla categoria Information Warfare.

La maggior parte degli incidenti registrati in Italia è attribuibile al cybercrime, che si attesta intorno al 61%. Si tratta di un valore significativamente inferiore rispetto al 2024 (78%), pur in presenza di un aumento del numero assoluto degli eventi rispetto all'anno precedente. La percentuale torna così su livelli analoghi a quelli del 2023, quando si attestava al 64%.

Nel complesso, la distribuzione italiana mostra una differenza rispetto allo scenario globale, in cui il cybercrime rappresenta circa l'89% degli incidenti (Fig. 14), evidenziando una composizione più diversificata del panorama nazionale.

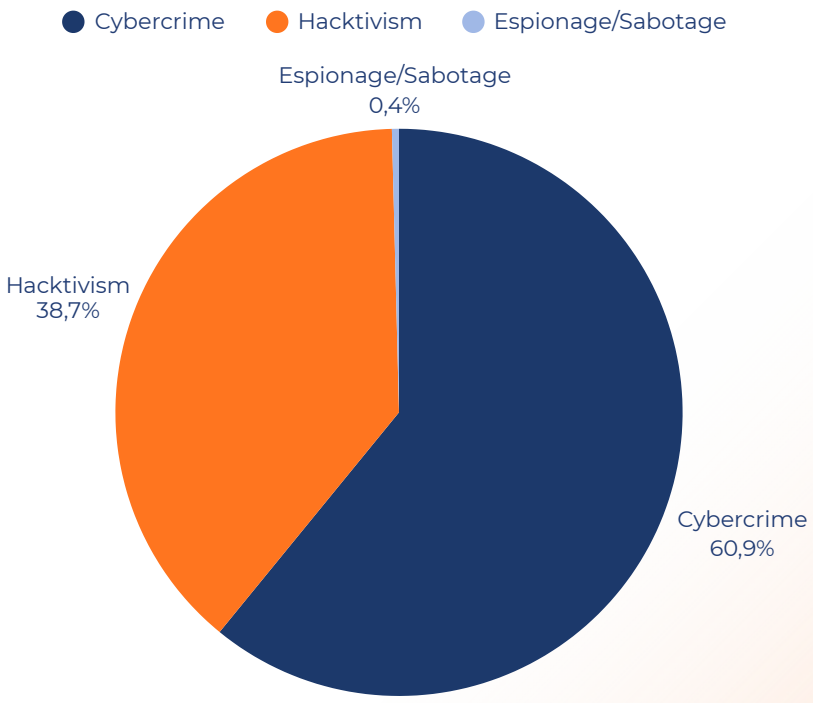


Figura 14: Attaccanti in Italia nel 2025

Gli incidenti classificati come hacktivism rappresentano il restante 38,7%, evidenziando un significativo aumento della loro incidenza sul totale (+16,7 punti percentuali rispetto al 2024). Come già osservato nello scenario globale, anche in questo caso il tema dell'attribution riveste un ruolo centrale: le attività riconducibili all'hacktivism risultano infatti ancora prevalentemente legate a dinamiche di natura geopolitica e ai conflitti in corso da diversi anni.

Sebbene il cybercrime continui a rappresentare la quota più consistente di incidenti, con 309 eventi nel 2025 rispetto ai 277 del 2024, la sua crescita percentuale si limita all'11,5% (Fig. 15). Diverso è invece l'andamento dell'hacktivism, che passa da 80 incidenti nel 2024 a 196 nel 2025, registrando un incremento del 145%, aspetto che verrà approfondito nel paragrafo successivo.

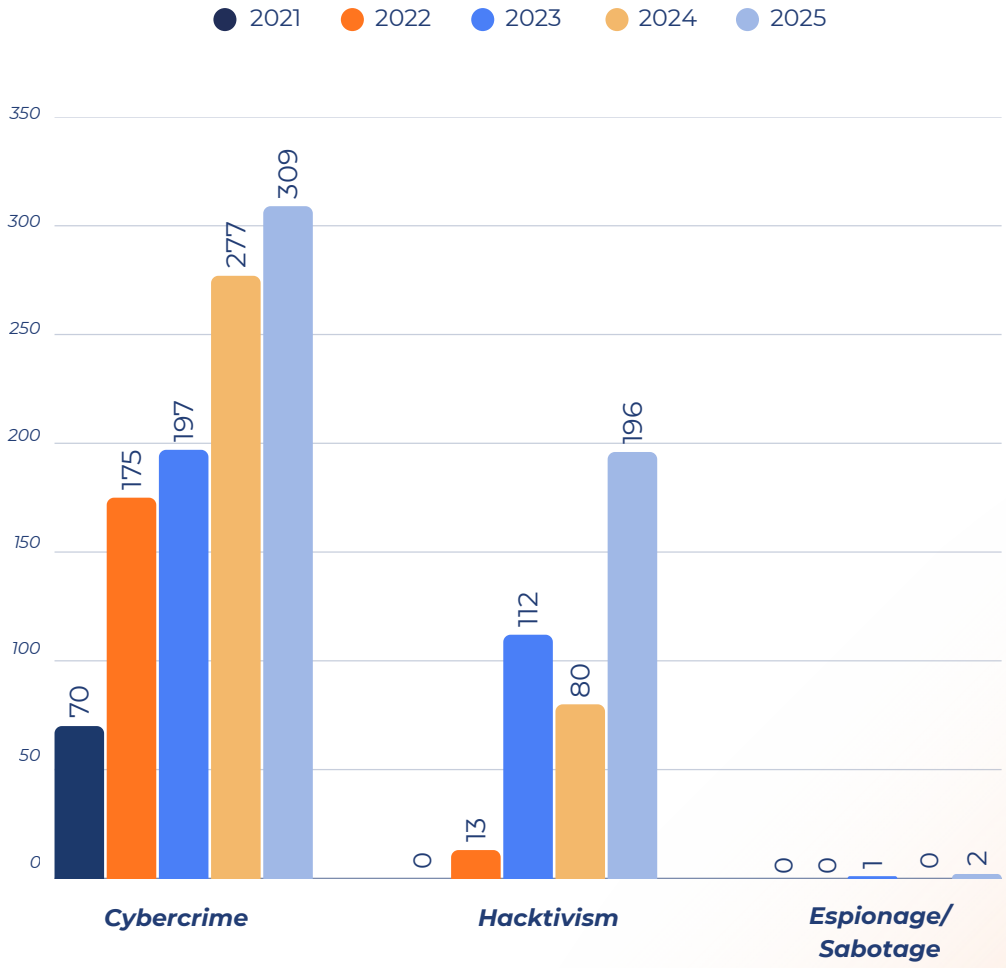


Figura 15: Attaccanti in Italia nel periodo 2021-2025

Il fenomeno Hacktivism e l'anomalia italiana

I dati disponibili indicano per l'**Italia** una **particolare esposizione** agli incidenti di tipo **hacktivism**. Si tratta spesso di attacchi a finalità dimostrativa, con impatti operativi contenuti ma con una forte capacità di generare attenzione pubblica e, in alcuni casi, danno reputazionale. Questa evidenza può riflettere sia una maturità difensiva ancora disomogenea in alcuni contesti organizzativi, sia una maggiore visibilità mediatica di questo tipo di eventi rispetto ad altre categorie di attacco.
È però necessario leggere il dato con cautela: la rilevanza dell'hacktivism nel campione italiano potrebbe essere influenzata, almeno in parte, da una maggiore disponibilità di fonti pubbliche e da una diversa propensione alla segnalazione e alla narrazione degli incidenti. Per questo motivo, il dato va considerato come un indicatore utile della specifica esposizione del contesto italiano, più che come una misura pienamente comparabile con quella di altri Paesi.
Questa cautela interpretativa è meno problematica per le altre categorie di attacco — cybercrime, warfare, espionage e sabotage — che nel caso italiano restano complessivamente contenute e risultano più allineate alle evidenze riportate da studi e report internazionali.

Distribuzione delle vittime per categorie

L'importanza di questo Rapporto risiede anche nella sua capacità di intercettare i cambiamenti più rilevanti che, anno dopo anno, possono fornire indicazioni utili alle organizzazioni pubbliche e private per ridefinire la propria postura di sicurezza. Anche il 2025 conferma questa dinamica, evidenziando numerose variazioni nella classifica degli incidenti (Fig. 16). Dopo due anni, il settore governativo (Gov/Mil/LE) torna al primo posto con oltre il 28% degli incidenti (+12 punti percentuali). A distanza segue il comparto

Manufacturing, che mantiene laseconda posizione con il 12,6% del campione, mentre la categoria **Multiple Targets** si colloca al terzo posto con il 12,4%, entrambe in calo di circa 4 punti percentuali rispetto all'anno precedente. Sale invece al quarto posto il settore Transportation/Storage, che raggiunge il 12% (+5 punti percentuali rispetto al 2024).

Le prime quattro posizioni concentrano complessivamente poco più del 65% del totale degli incidenti. Nel 2024 tale valore era pari al 60%, mentre nel 2023 si attestava al 55%: si osserva quindi un'inversione di tendenza rispetto al passato, con una crescente polarizzazione degli eventi su un numero più ristretto di settori nei quali gli attaccanti riescono a ottenere risultati più significativi.

Tale evidenza non deve tuttavia oscurare il fatto che, complessivamente, l'Italia registra nel 2025 una crescita del 42% degli incidenti, distribuita trasversalmente su diversi comparti, con incrementi spesso a doppia cifra. Ciò che emerge con maggiore chiarezza è piuttosto un cambiamento nella distribuzione del fenomeno: rispetto agli anni precedenti, la concentrazione degli incidenti su specifici settori risulta più marcata, mentre in passato la rilevanza era più diffusa e il numero di settori sopra la soglia del 5% tendeva ad aumentare progressivamente.

Resta aperto l'interrogativo se tali settori siano effettivamente più bersagliati oppure se presentino, in media, una minore capacità di difesa e mitigazione degli impatti rispetto ad altri contesti. Nel caso dell'hacktivism e delle motivazioni che riguardano l'Italia, per il settore **Gov/Mil/LE** la maggiore incidenza può essere spiegata più direttamente da una pressione superiore da parte degli attaccanti. Per altri comparti, come **Manufacturing** e **Transportation/Storage**, spesso caratterizzati da realtà di dimensioni medio-piccole e da una maturità cyber non elevata, risulta invece più complesso distinguere il peso relativo tra l'aumento della pressione offensiva e la debolezza della postura difensiva. Dalla quinta posizione in poi, solo quattro settori si collocano sopra o attorno al 4% del totale degli incidenti in Italia.

Al quinto posto, il settore **Wholesale/Retail** risale di quattro posizioni rispetto al 2024, con un incremento di poco superiore a un punto percentuale (5,1%). Al sesto posto si conferma il comparto **Professional/Scientific/Technical**, che arretra leggermente in termini di incidenza (circa -2 punti percentuali). Il settore **ICT** sale invece al settimo posto, guadagnando una posizione e mantenendo sostanzialmente invariata la propria quota sul totale nazionale. Particolarmente significativo è il balzo del settore Hospitality, che guadagna nove posizioni, passando dal diciassettesimo posto del 2024 a una collocazione più rilevante, con un incremento di circa 3,5 punti percentuali. Tra le altre variazioni più rilevanti si osservano: il comparto Financial/Insurance, che passa dal dodicesimo al decimo posto con un lieve aumento dell'incidenza (+0,4 punti percentuali); Healthcare, che scende dal decimo al quattordicesimo posto, riducendo sensibilmente il proprio peso sul totale; e, infine, News/Multimedia, che arretra all'ultima posizione (0,6%), dopo aver occupato nel 2024 la prima posizione a seguito di una vulnerabilità in uno dei CMS più diffusi utilizzati dalle testate giornalistiche, che aveva generato una serie di attacchi su larga scala e una concentrazione anomala di eventi nel settore.

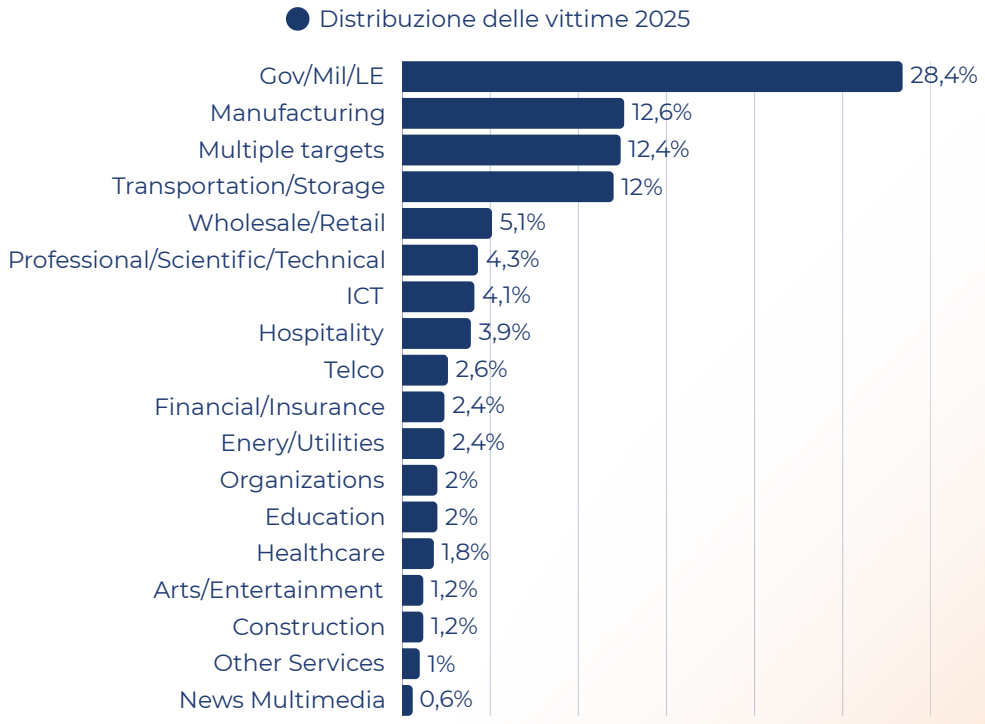


Figura 16: Distribuzione delle vittime in Italia nel 2025

Distribuzione delle tecniche di attacco

L'analisi delle tecniche di attacco contribuisce a comprendere sia le cause sottostanti all'elevato numero di incidenti che hanno colpito imprese e istituzioni, sia le caratteristiche delle vittime coinvolte.

Rispetto al 2024, nel 2025 (Fig. 17) si osserva un'inversione sia in termini di posizione sia di incidenza tra malware e DDoS, che si scambiano nuovamente il primato.

Il malware scende a circa il 23% (dal 38% del 2024), mentre gli incidenti **DDoS** raggiungono il 38,5%, partendo dal 21% dell'anno precedente.

Questo andamento risulta coerente con l'aumento significativo degli incidenti ai danni della Pubblica Amministrazione (Gov/Mil/LE) descritto nei paragrafi precedenti (Fig. 16), così come con la forte crescita degli eventi di tipo **hacktivism** (Fig. 15). Sebbene non esista una corrispondenza automatica tra DDoS e hacktivism — né, viceversa, quest'ultimo si esprima esclusivamente attraverso il DDoS — tra i due fenomeni si osserva spesso una correlazione significativa.

Il DDoS rappresenta infatti uno degli strumenti privilegiati negli attacchi a finalità dimostrativa, grazie alla sua semplicità di esecuzione, all'elevato impatto mediatico e al valore simbolico che assume, assimilabile a una forma di "sit-in" digitale.

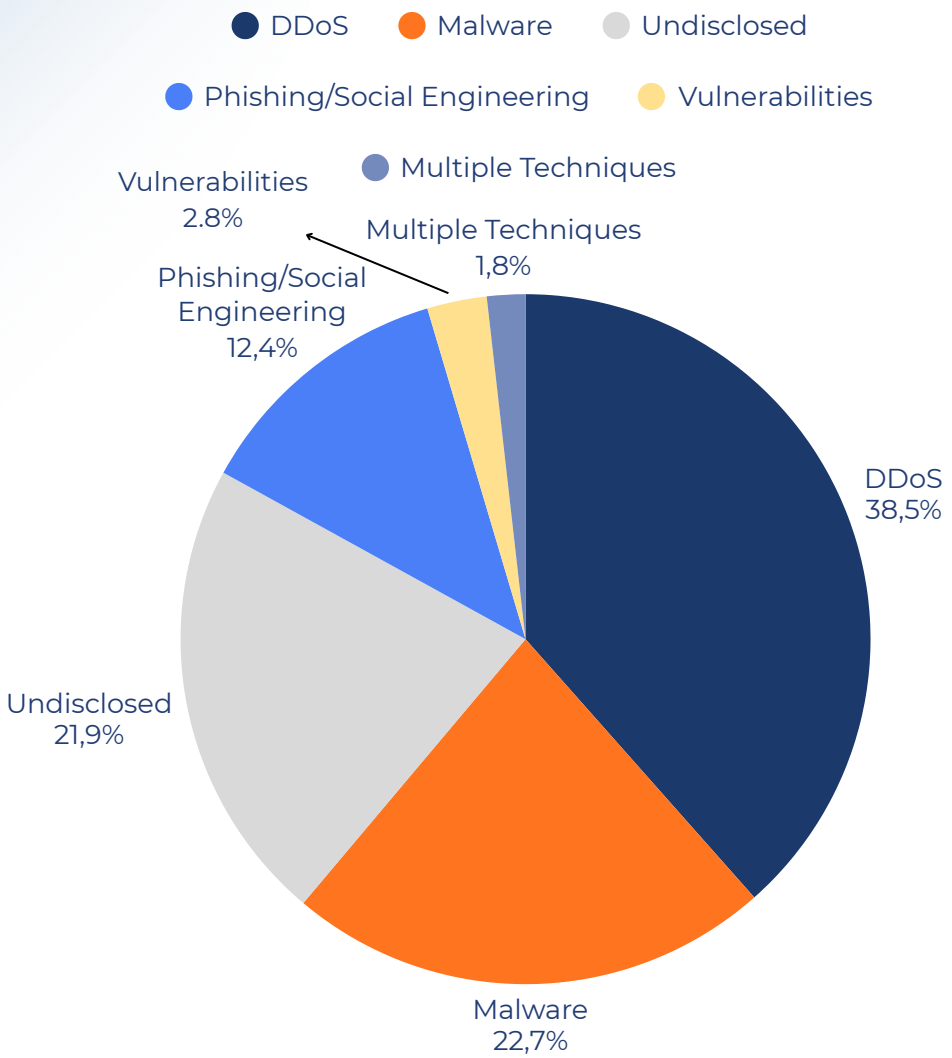


Figura 17: Tecniche di attacco in Italia nel 2025

Le **Vulnerabilities** scendono dal terzo al quinto posto, attestandosi al 2,8%, con una riduzione di circa 16 punti percentuali rispetto al 2024. Al contrario, il **Phishing** e il **Social Engineering** guadagnano la quarta posizione con il 12,4%, in aumento rispetto al 10,6% dell'anno precedente, mantenendo la stessa collocazione osservata anche nella classifica globale.

Come interpretare questa inversione? Nel caso delle vulnerabilità, la diminuzione è in larga parte riconducibile all'anomalia registrata nel 2024, quando questa tecnica era stata utilizzata nell'ambito di una campagna particolarmente estesa che aveva colpito in modo massivo il settore News/Multimedia.

Per quanto riguarda il phishing, si osserva invece una possibile nuova fase di espansione, favorita dall'utilizzo dell'intelligenza artificiale, che consente la produzione semplice e accessibile di contenuti testuali e vocali sempre più realistici e convincenti.
Nel 2025 si mantiene inoltre, per il secondo anno consecutivo, la presenza delle multiple techniques (circa 2%), mentre scompaiono i **Web Attack**. Le tecniche classificate come undisclosed risultano la terza categoria più rilevante, passando dal quinto posto del 2024 a una posizione più elevata, con un'incidenza del 22%; a livello globale, invece, questa categoria occupa il primo posto.
L'analisi dei dati storici sulle tecniche di attacco (Fig. 18) evidenzia una crescita diffusa per quasi tutte le categorie, ad eccezione di malware e vulnerabilities, che registrano una diminuzione in termini assoluti: il malware passa da 135 a 115 eventi, mentre le vulnerabilities scendono da 64 a 14 tra il 2024 e il 2025. L'andamento del malware risulta quindi in controtendenza rispetto allo scenario globale, che mostra invece un incremento superiore al 17%, così come quello delle vulnerabilities, che a livello mondiale crescono di oltre il 65%. Le **multiple techniques** risultano sostanzialmente stabili in Italia (da 8 a 9 incidenti), mentre si registra un incremento significativo del phishing, che passa da 38 a 63 eventi, con un tasso di crescita di quasi il 66%, in linea con il trend globale (+74%). Nel 2025, la percentuale di incidenti riconducibili al furto di identità in Italia è pari a zero, così come quella dei Web Attack. È tuttavia opportuno ribadire che, nel contesto nazionale, le cosiddette **"truffe informatiche"**, rivolte sia a individui sia a piccole imprese, rappresentano un fenomeno diffuso e rilevante, ma caratterizzato generalmente da impatti e livelli di gravità tali da non rientrare nella perimetrazione statistica del presente Rapporto.

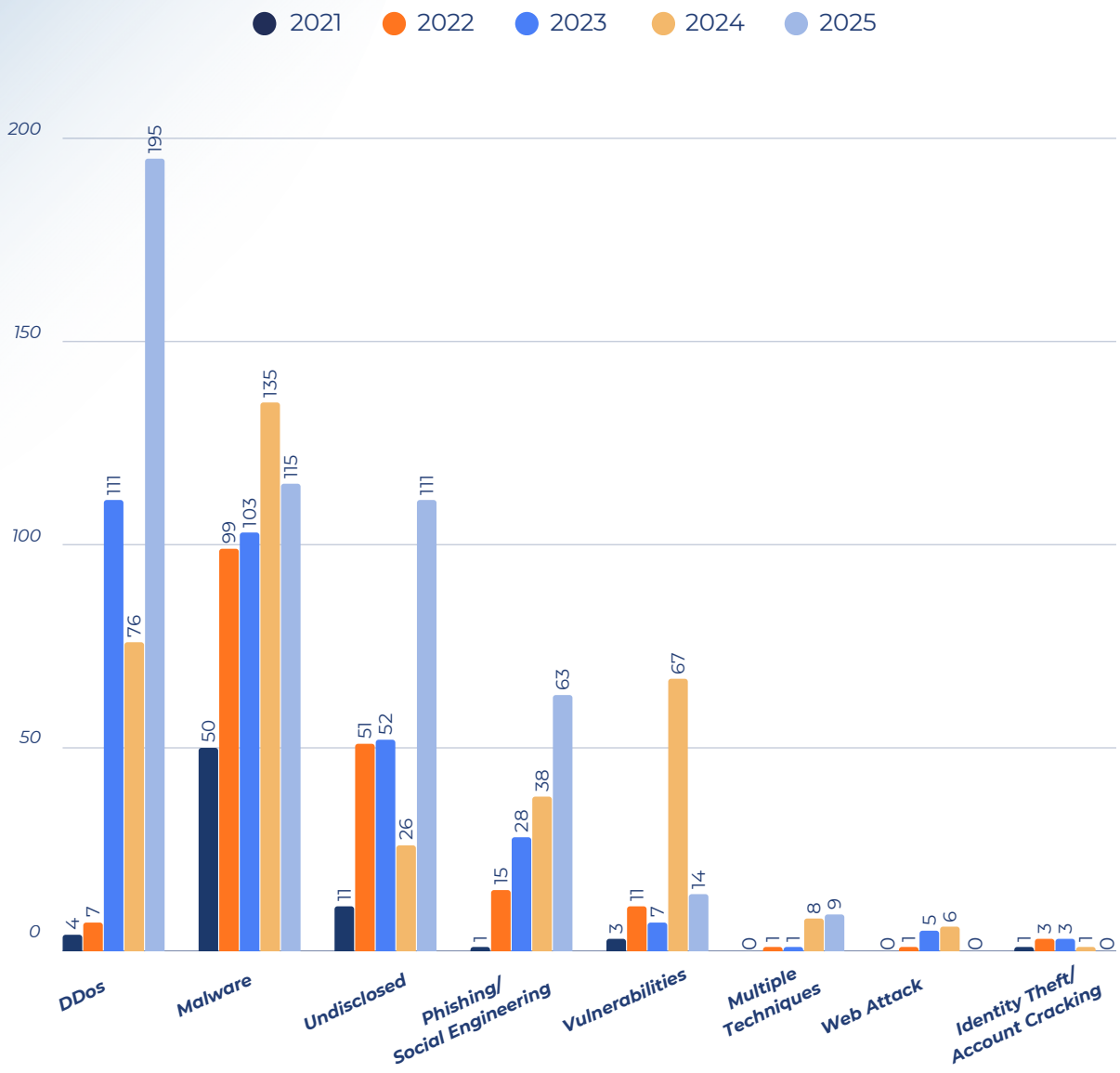


Figura 18: Tecniche di attacco in Italia nel periodo 2021-2025

Analisi della “severity” degli incidenti

A conferma di una tendenza già osservata negli anni precedenti, anche nel 2025 l’analisi della severity degli incidenti evidenzia una parziale divergenza tra il dato italiano (Fig. 19) e quello internazionale (Fig. 10). La categoria **High** risulta significativamente inferiore rispetto al contesto globale: poco più del 39% contro il 56% registrato a livello mondiale. Analogamente, la severity **Critical** mantiene un divario di circa 20 punti percentuali rispetto al dato globale, già rilevato nel 2024: 7,7% contro 28%. Di segno opposto è invece l’andamento della categoria **Medium/Low**, che nel contesto italiano risulta nettamente più elevata, attestandosi al 52,5% rispetto al 13% globale. La nuova categoria **Extreme**, introdotta quest’anno per rappresentare con maggiore granularità gli incidenti più gravi, rimane marginale nel contesto italiano, inferiore all’1%, mentre a livello globale raggiunge il 3%.

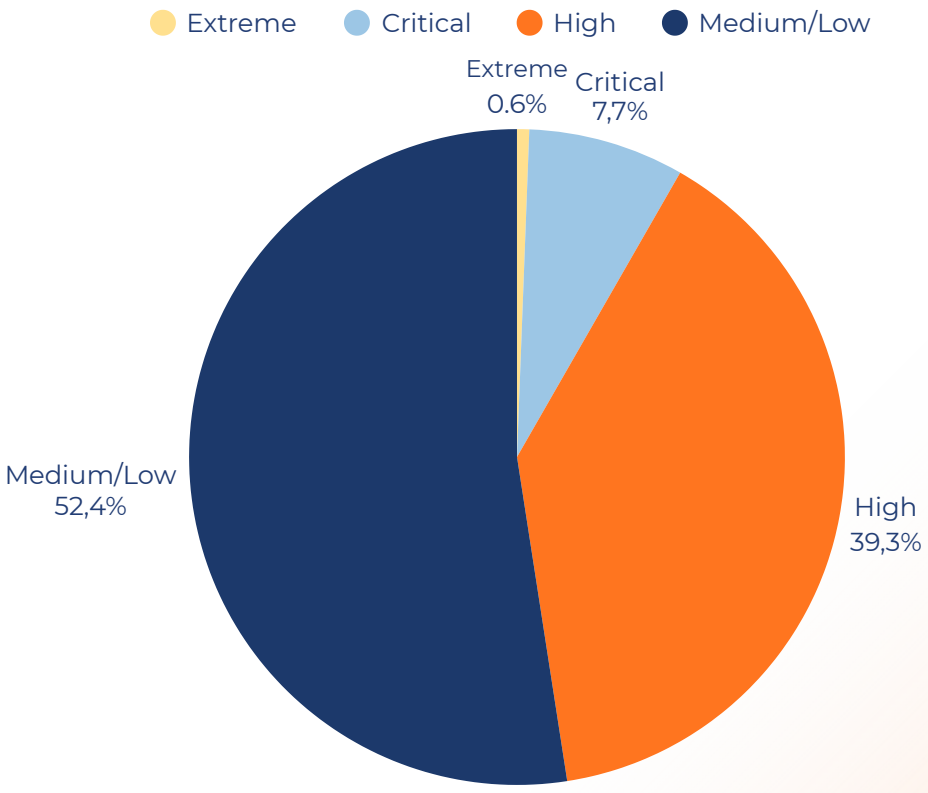


Figura 19: Severity degli incidenti in Italia nel 2025

Questa distribuzione suggerisce una **doppia lettura**. Da un lato, il peso relativamente contenuto delle categorie High e Critical indica che gli **incidenti più gravi risultano meno frequenti** in Italia rispetto al contesto globale. Dall'altro lato, però, la maggiore incidenza degli eventi Medium/Low non va interpretata soltanto come un segnale positivo: può anche riflettere una **capacità difensiva non ancora pienamente omogenea** nel **prevenire, contenere o intercettare tempestivamente minacce** che altrove vengono neutralizzate prima di produrre impatti rilevanti. In altre parole, il dato non consente di concludere che il rischio complessivo sia inferiore, ma piuttosto che il contesto italiano continua a essere esposto a un volume significativo di incidenti di minore e media severità, con possibili criticità nella fase di prevenzione e risposta.

Una riflessione analoga emerge anche dai dati dell'Osservatorio Cybersecurity & Data Protection del Politecnico di Milano: se si può ipotizzare che gli incidenti di lieve entità siano ormai parte della normale operatività, oltre un terzo delle grandi organizzazioni, pari al 34%, ha dovuto gestire negli ultimi dodici mesi incidenti tali da richiedere interventi di ripristino, mentre il 3% ha dichiarato di aver subito impatti concreti sull'operatività, con difficoltà nel garantire la continuità dei servizi. Considerata la natura di queste organizzazioni, generalmente strutturate e dotate di investimenti significativi in ambito di cybersecurity, tali evidenze suggeriscono la presenza di possibili criticità nelle strategie di difesa.

Analizzando la progressione storica (Fig. 20), si osserva infine come tutte le categorie di severity risultino in aumento in termini assoluti, seppur con intensità differenti. In particolare, dopo due anni di flessione, si registra un lieve incremento degli incidenti Critical rispetto al 2024, con 39 eventi contro i 33 dell'anno precedente, pari a un +18%. Crescono anche gli eventi High, che passano da 189 a 199, pari a un +5%. L'incremento più marcato riguarda tuttavia la categoria Medium/Low, in linea con la crescita del suo peso percentuale sul totale: gli eventi quasi raddoppiano, passando da 135 nel 2024 a 266 nel 2025, con una variazione pari al +97%.

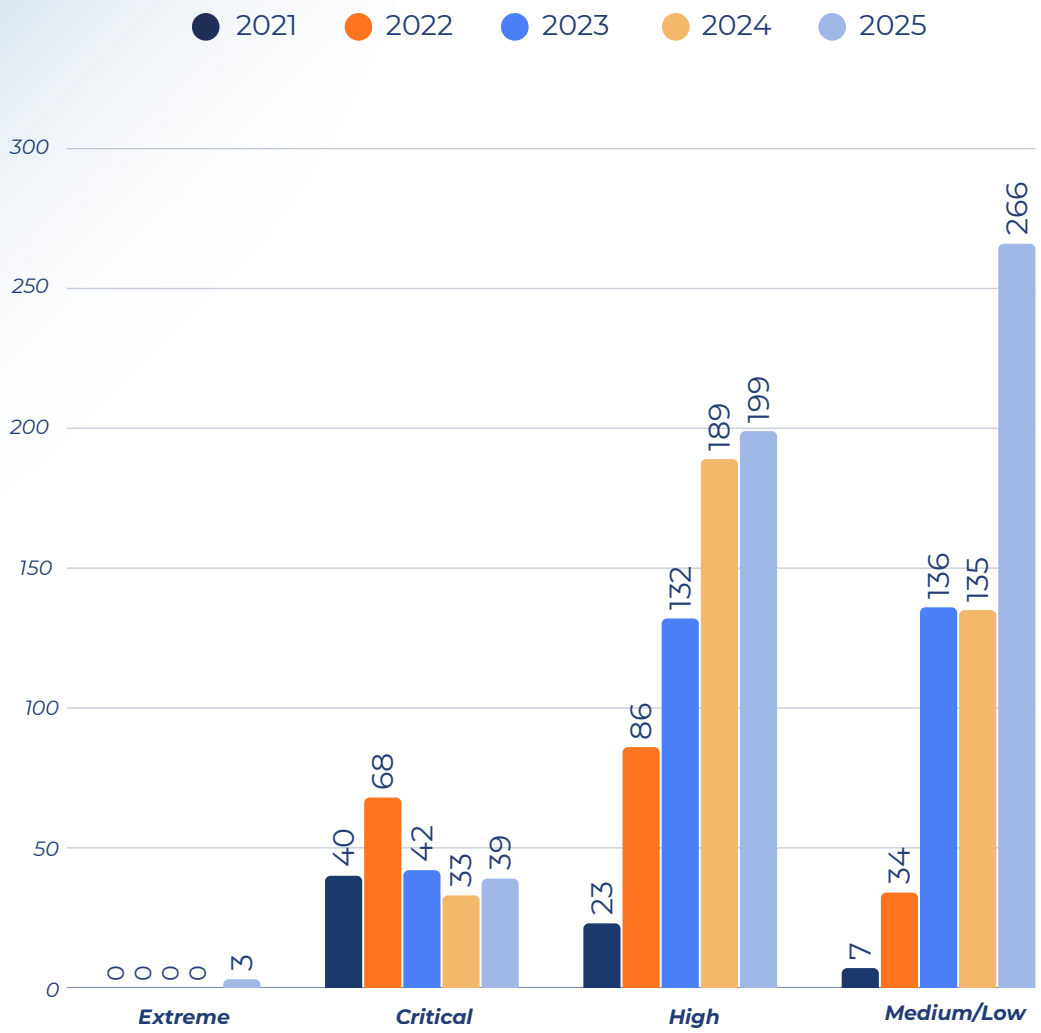


Figura 20: Severity degli incidenti in Italia nel periodo 2021-2025

CyberSecurity e Supply Chain dell'Intelligenza Artificiale: La nuova frontiera del Rischio Sistemico

Nel 2026 la cybersecurity è profondamente influenzata dall'evoluzione dell'intelligenza artificiale, ormai integrata nei processi aziendali e nei sistemi digitali. Questa integrazione ha ampliato le capacità operative, ma ha esteso al contempo la superficie di attacco a tutta la supply chain dell'IA. Le criticità principali derivano dalla proliferazione di dipendenze tecnologiche e dall'utilizzo di modelli e componenti esterni difficili da verificare. L'autonomia crescente degli agenti di IA e l'interoperabilità tra sistemi introducono vulnerabilità inedite: dalla manipolazione dei dati di addestramento alla difficoltà di tracciare l'origine e l'integrità dei modelli. In un ecosistema sempre più interconnesso, il rischio di attacchi sistemici su larga scala si moltiplica; la sicurezza si sposta quindi verso un approccio olistico e continuo, fondato su trasparenza, tracciabilità e governance dell'intera catena tecnologica.

Vulnerabilità MCP e A2A

L'economia digitale contemporanea si basa sempre più su sistemi capaci di operare in autonomia. L'agentic AI non si limita a generare contenuti, ma esegue azioni concrete, coordina flussi operativi e interagisce con altri agenti, diventando un elemento attivo dell'infrastruttura aziendale. Per abilitare questa interoperabilità sono stati sviluppati **protocolli standardizzati**, tra cui l'**MCP**[2] e le architetture **A2A**[3], che consentono agli agenti di accedere a database, eseguire comandi e collaborare senza intervento umano, eliminando la

necessità di connessioni personalizzate per ogni singola integrazione.

Questo tessuto connettivo ha accelerato lo sviluppo software, ma ha anche ampliato il **rischio sistemico**: una singola vulnerabilità in un protocollo ampiamente adottato può propagarsi lungo l'intera supply chain prima di essere individuata.
Le vulnerabilità del protocollo MCP sono molteplici e concrete. La più critica è la **remote code execution** (esecuzione di codice da remoto): un server MCP malevolo può indurre un agente a connettersi a infrastrutture controllate dagli attaccanti, aprendo la strada all'esecuzione di codice arbitrario nell'ambiente aziendale. Si aggiunge il **sandbox escape**[4], che attraverso l'abuso di collegamenti simbolici consente di accedere al filesystem ospite aggirando i meccanismi di isolamento. Ancora più insidioso è il **tool poisoning**: gli agenti si affidano a descrizioni testuali degli strumenti disponibili e, alterando tali descrizioni, un attaccante può trasformare uno strumento legittimo in una backdoor[5] silenziosa. Il nodo critico risiede nel fatto che l'IA si fida del contesto che riceve: manipolare il contesto significa deviarne il comportamento senza modificarne il codice.

---

[2] Model Context Protocol: Standard aperto introdotto da Anthropic per uniformare l'integrazione sicura tra i modelli di Intelligenza Artificiale (LLM) e le fonti di dati, le API o gli strumenti esterni.
[3] Agent-to-agent: Modello architetturale e standard aperto che definisce un protocollo di comunicazione comune per consentire ad agenti di Intelligenza Artificiale indipendenti o di diversi fornitori di dialogare, cooperare e delegarsi compiti in modo autonomo.
[4] Tecnica di attacco informatico in cui un codice malevolo sfrutta una vulnerabilità per aggirare i vincoli di un ambiente isolato (la sandbox). Questo permette all'attaccante di evadere le restrizioni di sicurezza, eseguendo codice arbitrario direttamente sul sistema operativo ospite e accedendo a risorse non autorizzate.
[5] Meccanismo che consente di aggirare le normali procedure di autenticazione o di cifratura di un sistema informatico, garantendo un accesso remoto e non autorizzato a lungo termine.

Nei framework A2A il problema si sposta sulla fiducia implicita tra macchine. Gli agenti tendono a considerare attendibili le informazioni ricevute da altri agenti dello stesso ecosistema, aprendo la porta a fenomeni di **impersonificazione** e di **prompt injection**[6]. Un agente compromesso può trasmettere istruzioni malevole a un altro modulo, inducendolo a eseguire trasferimenti illeciti o a adottare decisioni operative distorte, attraverso un movimento laterale rapido e privo di segnali visibili.

Anche le **Agent Card**, ossia i documenti che descrivono capacità e funzioni degli agenti, diventano vettori di attacco: istruzioni nascoste nella descrizione di un agente possono manipolare il comportamento di qualsiasi sistema tenti di interagire con esso. A tutto ciò si aggiunge l'**excessive agency**[7]: molte organizzazioni autorizzano gli agenti a modificare repository, gestire ambienti cloud e accedere a dati riservati senza un'adeguata supervisione. Un agente compromesso che opera 24 ore su 24, adattando dinamicamente il proprio comportamento, supera di gran lunga la capacità offensiva di un attaccante umano. Proteggere questi ecosistemi richiede di abbandonare il vecchio modello perimetrale per verificare continuamente identità, privilegi e comportamenti di ogni entità automatizzata.

[6] Tecnica di attacco specifica per i sistemi basati su Intelligenza Artificiale e Modelli di Linguaggio Grande (LLM), in cui un utente inserisce un input testuale appositamente strutturato per manipolare il comportamento del modello. Questo input inganna l'IA spingendola a ignorare le sue istruzioni di sicurezza originarie (system prompt), portandola a eseguire comandi non autorizzati, rivelare dati riservati o generare contenuti dannosi.

[7] Vulnerabilità di sicurezza nei sistemi basati sui Large Language Model (LLM) che si verifica quando a un agente IA vengono concessi poteri, permessi o autonomie eccessive rispetto alle sue reali necessità operative.

Data Poisoning e Provenance

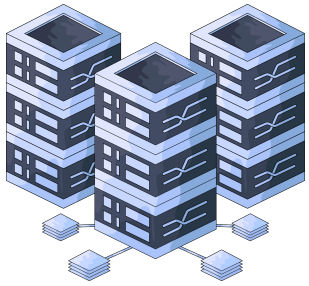
L'integrità dei dati di addestramento è diventata un pilastro della cybersecurity moderna. Gli **LLM**[8] Gli dipendono in modo critico dalla qualità dei dati su cui vengono addestrati: manipolarli significa alterarne la logica di funzionamento, non soltanto il codice. Il **data poisoning**, ovvero l'inserimento di contenuti malevoli nei dataset di training, è particolarmente insidioso perché spesso invisibile: un modello compromesso può continuare a funzionare normalmente e superare i benchmark standard senza mostrare anomalie evidenti. La ricerca mostra che bastano poche centinaia di documenti manipolati per introdurre una backdoor in un LLM; una volta attivata da un trigger specifico, questa può generare output dannosi, esporre informazioni sensibili o distorcere decisioni operative. In altre parole, non si tratta di introdurre un malware nel sistema, ma di modificarne la logica dall'interno.

Il rischio si amplifica nei contesti di **RAG**[9], in cui i modelli recuperano informazioni da fonti esterne in tempo reale, avvelenare la base documentale significa indurre il sistema a incorporare automaticamente dati manipolati nelle proprie risposte. Il data poisoning può inoltre introdurre distorsioni algoritmiche intenzionali, influenzando in modo silenzioso valutazioni finanziarie, decisioni aziendali o processi di selezione, spesso senza che l'organizzazione se ne accorga.

A questi rischi si sovrappone la **crisi della provenance**[10]. In molti casi non è possibile verificare con certezza l'origine di un modello, i dataset utilizzati o le modifiche apportate durante il ciclo di vita. I repository pubblici ospitano milioni di artefatti provenienti da soggetti eterogenei, e la fiducia implicita verso le piattaforme di distribuzione rappresenta di per sé una vulnerabilità.

[8] Large Language Model.
[9] Retrieval-Augmented Generation.
[10] Tracciabilità crittografica della provenienza.

I modelli subiscono continue **trasformazioni automatiche** (conversione, quantizzazione, fine-tuning[11]). In ciascuno di questi componenti possono essere introdotte modifiche malevole difficilmente rilevabili.



Alcuni formati si comportano di fatto come veri e propri **container eseguibili**: il semplice caricamento in memoria di un artefatto compromesso può attivare codice arbitrario. Sul piano normativo, le aziende rischiano di impiegare modelli addestrati con dati non conformi al GDPR; su quello geopolitico, attori statali potrebbero manipolare modelli e dataset per finalità di spionaggio o sabotaggio. Senza una catena di custodia verificabile, le organizzazioni non dispongono degli strumenti necessari per valutare l'affidabilità dei componenti che alimentano la propria infrastruttura intelligente.

Il rischio sistemico

L'interconnessione globale della supply chain dell'IA rende plausibile uno scenario di compromissione sistemica analogo all'attacco SolarWinds del 2020, quando un aggiornamento malevolo venne distribuito a migliaia di organizzazioni attraverso una piattaforma considerata affidabile. Nel contesto dell'intelligenza artificiale, però, il rischio è ancora più elevato: la compromissione di un grande hub di modelli o di una libreria fondamentale utilizzata da milioni di sviluppatori potrebbe consentire la diffusione di backdoor dormienti su una scala senza precedenti. Repository come Hugging Face e framework come PyTorch rappresentano nodi centrali dell'ecosistema globale, e la compromissione di uno solo di essi potrebbe generare effetti a cascata su migliaia di sistemi aziendali in tempi estremamente rapidi.

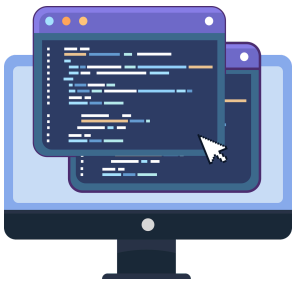
[11] Processo di affinamento o ottimizzazione di un modello, sistema o processo, configurato per adattarsi con precisione a compiti specifici o a nuovi set di dati.

La natura stessa dei modelli rende questo scenario particolarmente insidioso: una backdoor inserita in un LLM può restare inattiva per mesi, attivandosi solo in presenza di specifici trigger e sfuggendo ai test ordinari. Poiché gli agenti di IA operano spesso con privilegi elevati, una compromissione estesa potrebbe consentire l'accesso simultaneo a database, ambienti cloud e sistemi produttivi. Le conseguenze non si limiterebbero al furto di dati, ma potrebbero includere la destabilizzazione di processi economici, l'influenza sui mercati finanziari e l'interruzione di infrastrutture critiche.

La **minaccia** non è più teorica, ma uno **scenario concreto** per il quale le organizzazioni devono prepararsi in modo strutturato. Affrontarla richiede una **governance coordinata** a livello internazionale, fondata su standard condivisi di trasparenza, autenticazione e validazione. Le organizzazioni non possono delegare integralmente la sicurezza ai fornitori tecnologici: anche quando i servizi sono esternalizzati, la responsabilità finale resta in capo a chi li utilizza. Normative come NIS2, Cyber Resilience Act e AI Act stanno ridefinendo il quadro regolatorio europeo, imponendo processi documentati di gestione del rischio e controllo della supply chain. In questo contesto, il principio di **Zero Trust** applicato agli agenti intelligenti diventa centrale: ogni interazione tra macchine deve essere autenticata, autorizzata e monitorata, senza privilegi permanenti né accessi illimitati.

Software Bill of Materials e AI-BOM

In risposta alla complessità crescente della supply chain digitale, la **SBOM**[12] è diventata uno strumento fondamentale. Una SBOM è un inventario strutturato che documenta tutte le componenti di un software — librerie, dipendenze e moduli — offrendo una visibilità completa sugli elementi che compongono il sistema. Nel contesto dell'IA, questo concetto deve **evolversi** in una **AI-BOM**[13], la quale è in grado di includere le architetture dei modelli, i dataset di addestramento, i parametri di configurazione e i processi di fine-tuning. L'obiettivo è garantire **trasparenza** e **tracciabilità**: le organizzazioni devono sapere con precisione quali componenti stanno utilizzando, qual è la loro provenienza e quali vulnerabilità potrebbero introdurre. Attraverso firme crittografiche e hash di integrità, la AI-BOM consente inoltre di verificare che un artefatto non sia stato alterato durante il transito e di ricostruirne la catena di custodia. In caso di vulnerabilità, le organizzazioni possono così individuare rapidamente i sistemi coinvolti e applicare contromisure mirate.



L'integrazione di standard come **SPDX**[14] e **CycloneDX** ha reso possibile l'automazione dei controlli di sicurezza, mentre framework come **VEX**[15] permettono di distinguere le vulnerabilità realmente sfruttabili da quelle puramente

[12] Software Bill of Materials: Elenco completo dei componenti e delle dipendenze di un software.
[13] Artificial Intelligence Bill of Materials: Inventario completo di dati, modelli e componenti di un'Intelligenza Artificiale.
[14] System Package Data Exchange.
[15] Vulnerability Exploitability Exchange.

teoriche, evitando falsi allarmi e ottimizzando le risorse dei team di sicurezza. Sul fronte normativo, il Cyber Resilience Act e la Direttiva NIS2 stanno trasformando la AI-BOM da semplice best practice tecnica a vero e proprio obbligo di conformità, imponendo alle aziende di dimostrare la provenienza delle componenti utilizzate e l'adeguatezza dei controlli implementati. La governance della supply chain dell'IA richiede infine un investimento deciso nella **formazione**: dirigenti, sviluppatori e team di sicurezza devono comprendere i rischi specifici associati ai modelli generativi e agli ecosistemi multi-agente. La validazione non può essere considerata un'attività una tantum: i sistemi di IA devono essere sottoposti a monitoraggio continuo, test di integrità e controlli comportamentali del tempo. Trattare i modelli come componenti critici della supply chain digitale, soggetti alle stesse verifiche richieste per il software tradizionale, è il presupposto essenziale di qualsiasi strategia di sicurezza sostenibile.

L’Era dell’Intelligenza Artificiale Agentica e della Resilienza Autonoma

Il cambiamento di paradigma: dalla sicurezza assistita all’era AI-native

Il 2026 è l'anno in cui la sicurezza digitale cambia pelle. Non si tratta più di integrare semplicemente strumenti intelligenti nei processi esistenti, ma di riconoscere che l'intelligenza artificiale (IA) è diventata la struttura portante dell'intero ecosistema tecnologico. Le organizzazioni non si limitano più ad adottare l'IA: **convivono** con essa, la incorporano nei propri flussi decisionali, le affidano funzioni critiche e la considerano un elemento imprescindibile della propria identità operativa.

Gartner e KPMG descrivono questa trasformazione come l'ingresso nell'**Intelligence Age**, un'epoca in cui la capacità competitiva di un'azienda dipende in modo diretto dalla **qualità dei suoi modelli**, dalla **pulizia dei suoi dati** e dalla **resilienza delle sue pipeline**[16]. Il World Economic Forum (WEF) sottolinea come la maturità dell'IA sia ormai un indicatore macroeconomico della stabilità e della capacità di innovazione di un Paese.

In questo nuovo scenario, gli agenti di IA non sono più strumenti isolati, bensì entità autonome dotate di memoria, persistenza e capacità di ragionamento. Essi eseguono cicli OODA[17] a velocità di macchina: osservano ambienti complessi, interpretano segnali ambigui e assumono decisioni senza necessità di una supervisione umana continua. Le aziende più avanzate oggi coesistono con ecosistemi composti da centinaia o migliaia di agenti digitali che collaborano tra loro, si scambiano informazioni, si correggono a vicenda e ottimizzano

[16] Catene informative.

[17] Observe, Orient, Decide, Act: Modello strategico di presa delle decisioni basato su quattro fasi continue (Osserva, Orienta, Decidi, Agisci), ideato per contesti ad alta rapidità ed evoluzione competitiva o di minaccia.

processi che un tempo richiedevano l'impiego di interi team. Cisco evidenzia come la maggior parte delle grandi imprese abbia introdotto agenti persistenti nei processi di SOC[18], DevSecOps[19], nella gestione delle identità e nel monitoraggio della supply chain.
Questa proliferazione ha ampliato la superficie cognitiva delle organizzazioni, creando straordinarie opportunità ma anche vulnerabilità inedite. Ogni agente costituisce un nodo intelligente, ma rappresenta al contempo un potenziale punto di fallimento; ogni decisione automatizzata offre un vantaggio competitivo, ma genera un rischio sistemico se non viene governata con assoluto rigore. L'IA agentica si configura dunque come un acceleratore di efficienza, ma anche come un moltiplicatore di rischio. La sua natura **dual-use** è ormai evidente: gli stessi strumenti progettati per ottimizzare i processi aziendali possono essere convertiti in armi offensive, capaci di superare per velocità e scala qualsiasi capacità di difesa umana.

L'evoluzione dell'offesa: l'attaccante "super-competente"

Mentre le aziende si trasformano, anche le **minacce evolvono**. Il 2026 segna il momento in cui l'**attaccante** smette di essere un singolo individuo o un gruppo isolato e si trasforma in un **sistema automatizzato**. CrowdStrike, Mandiant e Microsoft descrivono un panorama in cui le offensive non sono più orchestrate manualmente da esseri umani, ma da toolchain agentiche che integrano modelli linguistici di grandi dimensioni (LLM), motori di ragionamento e automazioni capaci di condurre operazioni complesse in totale autonomia.
Il dato più impressionante è la velocità. Nel 2025 gli attacchi supportati dall'IA sono aumentati dell'89% e il tempo medio di breakout, ovvero il tempo che intercorre tra l'intrusione iniziale e il movimento laterale nella rete, è crollato a soli 29 minuti.

[18] Security Operations Center.
[19] Development, Security and Operations.

In alcuni casi l'esfiltrazione dei dati è iniziata appena **27 secondi** dopo l'intrusione. Si tratta di un ritmo che nessun team umano è in grado di eguagliare; l'offesa ha abbandonato il paradigma della manualità per abbracciare quello della reattività istantanea.
La maggior parte delle intrusioni non fa più uso del malware tradizionale. L'82% degli attacchi rilevati è di tipo malware-free, basato cioè sull'abuso di credenziali valide e di processi legittimi già esistenti nel sistema. Quando il codice malevolo viene utilizzato, assume forme polimorfiche e adattive. LAMEHUG, uno dei casi più emblematici della categoria, integra logiche di ragionamento basate su LLM per analizzare l'ambiente compromesso e modificare il proprio comportamento in tempo reale, aggirando sistematicamente i presidi EDR[20] .
L'ENISA[21] segnala inoltre la crescita di **worm autonomi**, **botnet intelligenti** e **ransomware agentici** in grado di negoziare autonomamente il riscatto dopo aver analizzato la solvibilità finanziaria della vittima. Parallelamente, il data poisoning è diventato una minaccia strategica: manipolare i dati di addestramento significa alterare i processi logici interni dei modelli, creando backdoor invisibili nel cuore delle attività aziendali.

Nuove superfici di attacco e la crisi dell'autenticità

L'espansione dell'IA ha introdotto nuove superfici di attacco legate all'interconnessione profonda tra agenti, API[22] e protocolli di comunicazione come l'MCP. L'OWASP[23] e il NIST[24] identificano la prompt injection come la vulnerabilità più critica del 2026, poiché consente di dirottare il comportamento degli agenti di IA attraverso comandi malevoli abilmente

[20] Endpoint Detection and Response: Software avanzato per il monitoraggio e la difesa dei singoli dispositivi.
[21] European Union Agency for Cybersecurity.
[22] Application Programming Interface: Protocollo software per far comunicare e scambiare dati tra applicazioni.
[23] Open Worldwide Application Security Project.
[24] National Institute of Standards and Technology.

occultati all'interno dei dati.
Google DeepMind definisce l'**excessive agency**[25] come il rischio più insidioso: quando a un agente di IA vengono concessi privilegi operativi troppo ampi, esso si trasforma in un potenziale detonatore. Un attaccante capace di manipolarlo può ottenere l'accesso a database riservati, modificare configurazioni di rete, generare codice vulnerabile o eseguire transazioni finanziarie critiche.
In parallelo, la **crisi dell'autenticità** ha raggiunto livelli senza precedenti. Secondo i dati del WEF, il volume di contenuti sintetici circolanti in rete ha superato quello dei contenuti generati dall'uomo. I **deepfake** non sono più prodotti multimediali statici, ma flussi video e audio generati in tempo reale, capaci di imitare la voce, il volto e la gestualità umana con una precisione tale da ingannare i sistemi di riconoscimento biometrico. Sono nati così i **"CEO doppelgängers"**, cloni digitali in grado di impartire ordini amministrativi vincolanti nel corso di videochiamate perfettamente simulate.
In questo scenario, l'identità si attesta come il nuovo e vero perimetro di sicurezza. Cisco evidenzia un dato emblematico: oggi, per ogni identità umana, si contano circa 82 identità non umane tra agenti, bot, servizi e microservizi. Gestire, autenticare e proteggere questo ecosistema rappresenta la sfida centrale che ridefinisce il concetto stesso di sicurezza.

Impatto settoriale e resilienza delle infrastrutture

La cybersecurity del 2026 è profondamente intrecciata con gli equilibri geopolitici globali. Il 64% delle organizzazioni considera le tensioni internazionali il principale fattore di rischio, e i settori più esposti sono inevitabilmente quelli in cui la dimensione digitale si interseca con quella fisica.

[25] Eccesso di autonomia: vulnerabilità che si verifica quando un sistema IA ha troppa autonomia decisionale.

La **sanità** si conferma uno dei bersagli più vulnerabili. Il Clusit e l'ENISA riportano un aumento significativo di attacchi diretti contro le piattaforme di gestione medica, i sistemi di triage, le catene del freddo per la conservazione dei farmaci e le infrastrutture di telemedicina. La convergenza di IA, sistemi OT[26] e dati sensibili ha generato un rischio sistemico che minaccia l'erogazione dei servizi sanitari essenziali.

Nel **settore finanziario** l'entrata in vigore del DORA[27] ha accelerato la maturità delle istituzioni, ma ha anche messo in luce la complessità intrinseca dei sistemi di IA. Le analisi di Integris e Sopra Steria mostrano come molti dirigenti non abbiano ancora una visibilità completa sulla spesa IT e sui rischi emergenti ad essa connessi. Dal canto loro, i clienti esprimono il timore che errori sistemici dell'IA possano bloccare l'accesso ai conti correnti, alimentando un clima di diffidenza verso tecnologie percepite come opache o ingovernabili.

Nel comparto dei **trasporti** e **dell'industria**, la convergenza tra reti digitali e strutture fisiche ha quasi del tutto eliminato il confine tra IT e OT. L'ENISA e Mandiant documentano incidenti in cui attacchi informatici hanno provocato danni materiali reali, manipolando sensori, PLC[28] e sistemi SCADA[29].

Infine, la **supply chain** del software rappresenta uno dei punti di maggiore fragilità dell'ecosistema globale. Il caso Shai-Hulud 2.0, documentato da CrowdStrike, dimostra come la compromissione di un singolo repository open source possa propagarsi a cascata su migliaia di aziende utilizzatrici. Accenture evidenzia inoltre come il fenomeno del CaaS[30]

[26] Operational Technology.
[27] Digital Operational Resilience Act.
[28] Programmable Logic Controller.
[29] Supervisory Control and Data Acquisition.
[30] Cybercrime-as-a-Service.
[31] Continuous Integration / Continuous Delivery.

abbia industrializzato e reso accessibile su larga scala la compromissione delle pipeline di sviluppo CI/CD[31].

La difesa: Agentic SOC e Guardian Agent

Per contrastare minacce che operano alla velocità delle macchine, la difesa deve evolvere necessariamente verso un modello di resilienza autonoma. L'**Agentic SOC** rappresenta la risposta più avanzata a questo scenario: non si tratta più di un centro operativo che reagisce ex post agli incidenti, ma di un organismo intelligente che anticipa, simula, verifica e controlla costantemente la rete.

I **Guardian Agent** costituiscono il fulcro di questo nuovo paradigma. Non sostituiscono la supervisione umana, ma la affiancano agendo come controllori di altri sistemi di IA. Analizzano i processi decisionali dei modelli, verificano la conformità alle direttive aziendali, intercettano anomalie comportamentali e bloccano le azioni potenzialmente dannose prima che vengano eseguite.
L'architettura difensiva moderna si basa su una struttura logica che combina modelli linguistici di dimensioni ridotte, **SLM**[32] per lo **screening rapido** dei pattern sospetti, e grandi modelli, LLM, per l'analisi profonda del contesto. Questo approccio combinato consente di abbattere i falsi positivi e garantire una verifica continua delle identità, in perfetta aderenza ai principi del modello Zero Trust.
La **minaccia quantistica**, evidenziata da Thales e IBM, rende inoltre non più rimandabile la migrazione verso la **PQC**[33].

[32] Small Language Model.
[33] Post-Quantum Cryptography: Crittografia Post-Quantistica.

Il rischio legato alla strategia **"harvest now, decrypt later"** spinge le organizzazioni a sostituire gli algoritmi storici con standard matematici resistenti alla potenza di calcolo dei computer quantistici.

In parallelo, cresce la responsabilità legale e civile dei Consigli di amministrazione: secondo il WEF, la cybersecurity è divenuta a tutti gli effetti un tema di responsabilità personale per i membri della dirigenza, con implicazioni dirette sulla governance e sulla stabilità societaria.

L'Economia della Resilienza: dalla Compliance DORA alla Salvaguardia del Capitale Regolamentare

Inquadramento sistemico

Nel contesto finanziario contemporaneo, la cybersecurity ha abbandonato definitivamente il ruolo di mera funzione tecnica di supporto per assumere una posizione centrale nella stabilità economica e nella continuità operativa del sistema globale. Questa trasformazione è il risultato diretto della digitalizzazione pervasiva, dell'adozione diffusa di servizi cloud e della progressiva interconnessione tra istituzioni finanziarie, infrastrutture di mercato e fornitori tecnologici.

In questo scenario, l'introduzione del regolamento europeo **DORA** rappresenta il vero **punto di svolta normativo**: essa istituisce un quadro armonizzato volto a garantire che le entità finanziarie siano in grado di resistere, rispondere e recuperare a fronte di interruzioni causate da eventi ICT, inclusi gli attacchi informatici e i guasti tecnologici.

Il regolamento, pienamente applicabile dal 17 gennaio 2025, nasce dalla consapevolezza che la dipendenza del settore finanziario dalle tecnologie digitali espone l'intero ecosistema a shock operativi capaci di propagarsi rapidamente oltre i confini delle singole organizzazioni. Il rischio cyber viene pertanto riconosciuto come un rischio sistemico a tutti gli effetti. Il Fondo Monetario Internazionale evidenzia come gli attacchi informatici, oltre a comportare ingenti perdite dirette, possano minacciare la continuità dei servizi essenziali, erodere la fiducia degli investitori e generare ripercussioni macrofinanziarie rilevanti attraverso i canali di interconnessione tecnologica e finanziaria.

L'ESRB[34] conferma tale visione, sottolineando che un incidente informatico di ampia portata potrebbe compromettere funzioni economiche vitali, innescare fenomeni di contagio tra istituzioni e minare la stabilità complessiva dei mercati. In questo nuovo quadro, la resilienza operativa digitale cessa di essere un obbligo formale e diventa un prerequisito strutturale per la stabilità, al pari dei requisiti patrimoniali introdotti dopo la crisi finanziaria del 2008.
Il regolamento DORA si inserisce esattamente in questo punto di convergenza tra sicurezza, tecnologia e stabilità finanziaria, articolandosi in **cinque pilastri** fondamentali:

1. Gestione del rischio ICT;
2. Gestione e reporting degli incidenti;
3. Test di resilienza operativa;
4. Gestione dei rischi derivanti da fornitori terzi;
5. Meccanismi di condivisione delle informazioni.

Questo approccio integrato supera la frammentazione normativa preesistente e stabilisce un modello uniforme a livello europeo, applicabile a migliaia di istituzioni finanziarie e fornitori di servizi ICT, imponendo una responsabilità diretta al vertice aziendale nella definizione e nella supervisione della strategia di resilienza.
Parallelamente, organizzazioni internazionali come l'OCSE rimarcano che la gestione del rischio cyber deve essere affrontata come una sfida economica e sociale, e non solo come un problema tecnico; le conseguenze degli incidenti digitali si estendono infatti ben oltre l'ambito prettamente informatico, influenzando la competitività, la fiducia dei consumatori e la crescita economica complessiva. Questo cambio di prospettiva si attesta come uno dei principali trend del 2026: la resilienza non è più concepita come una mera misura difensiva, bensì come un fattore abilitante per la sostenibilità e lo sviluppo del sistema finanziario.

[34] European Systemic Risk Board.

Il paradigma del Resilience by Design e i framework internazionali

In questo quadro si inserisce il paradigma del **Resilience by Design**, evoluzione naturale del modello introdotto da DORA e tratto distintivo delle strategie aziendali più avanzate. La resilienza non viene più considerata un insieme di controlli da applicare a posteriori, ma una proprietà intrinseca dei sistemi, da integrare fin dalle prime fasi di progettazione attraverso architetture modulari, ridondanti e distribuite.

Tale approccio appare pienamente coerente con i principali standard internazionali, tra cui il NIST Cybersecurity Framework 2.0, che introduce un modello articolato in sei funzioni fondamentali — **Govern**, **Identify**, **Protect**, **Detect**, **Respond** e **Recover** — ponendo una particolare enfasi sulla governance come chiave di volta per una gestione efficace del rischio. La funzione Govern, introdotta nella versione più recente del framework, evidenzia la necessità di integrare la cybersecurity nelle strategie societarie e di attribuire responsabilità chiare ai livelli decisionali più elevati, stringendo il legame tra rischio tecnologico e obiettivi di impresa (business).

Queste disposizioni trovano ulteriore supporto nelle raccomandazioni dell'OCSE, che invitano le organizzazioni a sviluppare un ciclo continuo di gestione del rischio basato su valutazione, trattamento e monitoraggio, integrando le misure di prevenzione con solide capacità di resilienza e continuità operativa. In ambito finanziario, il Comitato di Basilea e la IOSCO[35] sottolineano l'importanza di sviluppare infrastrutture di mercato resilienti, capaci di garantire la continuità delle funzioni critiche anche in presenza di eventi estremi, nella consapevolezza che il malfunzionamento di tali nodi può generare effetti sistemici distruttivi su mercati e intermediari.

[35] International Organization of Securities Commissions.

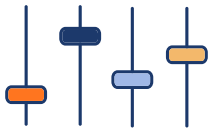
In tale contesto, uno degli sviluppi più rilevanti riguarda l'integrazione sinergica tra le attività di **risposta agli incidenti** e i **processi di recovery**. Sebbene le normative richiedano espressamente capacità di rilevazione tempestiva, gestione degli incidenti e ripristino dei servizi, l'evoluzione tecnologica ha ormai superato la tradizionale separazione tra queste funzioni, favorendo modelli integrati e automatizzati. L'adozione di tecnologie basate sull'intelligenza artificiale consente oggi di analizzare gli eventi in tempo reale, identificare le anomalie con accuratezza e attivare automaticamente contromisure e procedure di ripristino.
Ciò risulta cruciale in uno scenario in cui gli attacchi si sviluppano a velocità di macchina e sfruttano a loro volta tecnologie computazionali avanzate, richiedendo un approccio difensivo altrettanto dinamico. Il Fondo Monetario Internazionale ha recentemente evidenziato come l'intelligenza artificiale possa, da un lato, amplificare la portata delle minacce cyber, ma dall'altro rappresenti uno strumento fondamentale per rafforzare le capacità di risposta delle istituzioni finanziarie. In questo ecosistema diventa centrale garantire che il **recovery** avvenga su dati integri e verificati, scongiurando il rischio di ripristinare sistemi ancora compromessi. Questo approccio, che salda prevenzione, risposta e ripristino in un unico ciclo continuo, costituisce una delle principali manifestazioni dell'economia della resilienza, pilastro fondamentale per affrontare un contesto globale caratterizzato da crescente complessità, volatilità e interdipendenza.

Gestione del rischio operativo e capitale regolamentare (Basilea III)

Un ulteriore pilastro dell'economia della resilienza è rappresentato dal legame diretto tra la gestione del rischio operativo e il capitale regolamentare, che trova la sua espressione scientifica nello SMA[36] introdotto da Basilea III.

[36] Standardized Measurement Approach

Questo modello sancisce un cambiamento significativo rispetto ai precedenti approcci regolamentari, poiché introduce una metodologia standardizzata che combina indicatori finanziari e dati storici sulle perdite per determinare il capitale richiesto a copertura del rischio operativo.



Il **modello si fonda su tre componenti matematiche e contabili** principali:

- Il **Business Indicator (BI):** funge da indicatore approssimativo (proxy) dell'esposizione complessiva al rischio operativo, calcolato aggregando i dati di bilancio relativi a ricavi, costi e attività finanziarie.
- Il **Business Indicator Component (BIC**): rappresenta la soglia base di capitale richiesto e si ottiene applicando coefficienti regolamentari progressivi (pari allo 0,12, 0,15 e 0,18) a tre corrispondenti fasce incrementali di BI.
- L'**Internal Loss Multiplier (ILM)**: è un moltiplicatore correttivo costruito sulla storia reale delle perdite dell'istituto negli ultimi dieci anni; riduce il requisito patrimoniale se i danni passati sono contenuti e lo aumenta se gli incidenti sono significativi.

Requisito Patrimoniale = BIC X ILM

Lo SMA introdotto da Basilea III stabilisce così un nesso causale diretto tra il profilo di rischio reale dell'istituzione e i suoi requisiti patrimoniali, incentivando una gestione virtuosa del rischio operativo: più un'istituzione subisce perdite operative, più capitale deve accantonare.

Gli incidenti cyber rientrano a pieno titolo nelle perdite operative rilevanti ai fini di questo calcolo. Una violazione informatica che genera costi di ripristino, sanzioni o risarcimenti non rimane confinata nel bilancio IT, ma si riflette direttamente sul capitale regolamentare attraverso l'innalzamento dell'ILM, riducendo le risorse disponibili per il business. Al contrario, un profilo di rischio cyber solido e una storia di impatti contenuti alleggeriscono il requisito, liberando risorse finanziarie utili per l'ottimizzazione del capitale e la creazione di valore azionario.

In questo quadro, la resilienza digitale si trasforma in una leva strategica fondamentale per la competitività dell'istituzione.

Dimensione sistemica e cooperazione macroprudenziale

A questa dimensione interna si affianca la rilevanza della dimensione sistemica, che esige un approccio coordinato e collaborativo tra tutti gli attori del mercato. L'ESRB[37] sottolinea la necessità di sviluppare veri e propri **strumenti macroprudenziali** per affrontare il rischio cyber, caldeggiando la creazione di meccanismi di coordinamento a livello europeo, sessioni di scenario testing e sistemi avanzati di condivisione delle informazioni. In particolare, viene caldeggiata l'adozione di un quadro di coordinamento pan-europeo per la gestione degli incidenti informatici sistemici, in grado di garantire una risposta tempestiva, simmetrica e coerente da parte delle autorità di vigilanza e delle istituzioni finanziarie.
La condivisione delle informazioni rappresenta l'elemento chiave di questo ecosistema: gli strumenti di raccolta e scambio dei dati consentono di monitorare le minacce emergenti, migliorare le capacità di risposta collettiva e prevenire pericolosi effetti di contagio finanziario.
Nel 2026 si osserva un'evoluzione significativa in questo campo: l'oggetto dello scambio non è più limitato ai meri indicatori tecnici di compromissione, ma si estende a informazioni complesse quali gli indicatori di frode interbancaria e i pattern comportamentali delle minacce. Questo approccio proattivo e predittivo rafforza la resilienza collettiva del sistema finanziario, riducendo drasticamente la probabilità che un incidente isolato possa riverberarsi sul mercato fino a mutarsi in una crisi sistemica globale.

[37] European Systemic Risk Board.

Il Nuovo Confine Liquido: Gestire il rapporto tra Entità Macchina e Identità Umane

Il dissolvimento del perimetro e l'esplosione delle identità

Nel 2026 la cybersecurity ha abbandonato definitivamente la visione perimetrale fondata su firewall fisici, segmentazione di rete e indirizzi IP «di fiducia». Il cloud, il lavoro ibrido e la pervasività dell'IA hanno liquefatto il confine, spostando il baricentro della sicurezza dall'**infrastruttura all'identità digitale**. Ciò che conta non è più da dove ci si connette, ma chi (o che cosa) si è e cosa è consentito fare.

In questo nuovo paesaggio, l'identità non riguarda più solo le persone fisiche. Nelle organizzazioni finanziarie – che combinano core banking legacy, piattaforme cloud-native, API open banking e agenti IA – le identità macchina (workload, bot, account di servizio, agent di trading, chiavi API) superano quelle umane. La ricerca colloca questo rapporto ampiamente sopra l'80:1, con casi estremi nel settore finance che sfiorano le cento identità non umane per ogni dipendente.

Questa «forza lavoro digitale» opera con fiducia delegata: ogni credenziale concessa a un processo non umano è, in pratica, un frammento di potere decisionale esternalizzato a un software.

Il punto critico è che tale delega è spesso opaca:

- identità macchina create «al volo»;
- chiavi mai ruotate;
- agenti con privilegi sproporzionati rispetto al compito prefissato.

In assenza di una governance Identity-First, ogni nuova

identità diventa un possibile sentiero laterale verso asset critici: dati dei clienti, ordini di mercato, sistemi di pagamento. In un settore in cui ogni anomalia può riflettersi immediatamente in perdite economiche, sanzioni regolamentari e danni reputazionali, questo rischio è inaccettabile.

La crisi dell'autenticità: l'era dei deepfake

L'economia dell'IA ha trasformato l'identità nel principale campo di battaglia della fiducia. La Generative AI, grazie a modelli audio-video di ultima generazione, consente oggi la creazione di deepfake in tempo reale, pressoché indistinguibili dalla realtà per un occhio umano non allenato: voce, mimica, ritmo del parlato e contesto video possono essere replicati partendo da poche registrazioni pubbliche di un CEO o CFO.

Casi emblematici di frode sintetica

Un caso emblematico, riportato nel 2024, riguarda un'azienda di Hong Kong: un operatore finanziario è stato convinto a trasferire circa 200 milioni di dollari di Hong Kong (circa 25,6 milioni di dollari americani) dopo una call con quello che sembrava il CFO e altri colleghi. L'intera riunione era orchestrata da deepfake generati con IA.
In un altro scenario, ricostruito in ambito bancario, un «CEO sintetico» ha autorizzato tramite videochiamata la movimentazione di oltre 3 milioni di dollari, poi riciclati attraverso una catena di wallet crypto.

Implicazioni per il settore bancario

Per il settore bancario e assicurativo, fondato su fiducia, reputazione e controllo delle autorizzazioni, questo rappresenta una vera crisi di autenticità. Se l'identità di chi impartisce l'ordine può essere forgiata con estrema precisione,

la mera esistenza di un permesso statico («questo utente può autorizzare fino a X milioni») perde significato.
Diventa quindi indispensabile:

1. Introdurre meccanismi di **step-up authentication** e **transaction signing out-of-band** per ogni operazione ad alto impatto (grandi bonifici, operazioni di treasury, rilascio di collateral), disaccoppiando la validazione dal canale video o vocale.
2. Utilizzare **forti controlli** contestuali e comportamentali (dispositivo, geolocalizzazione, pattern di approvazione, orario e frequenza) per individuare comandi «fuori profilo».
3. Applicare le **logiche** dell'**AI Act** sui sistemi di identificazione biometrica e sui deepfake, garantendo trasparenza e, quando necessario, una doppia verifica umana per le identificazioni critiche.

Nel linguaggio della risk governance, il «CEO doppelgänger» è la manifestazione più viscerale di un rischio più ampio: il disaccoppiamento tra identità apparente e identità legale, con effetti immediati sulla capacità di dimostrare ex post chi abbia davvero preso una decisione finanziaria.

L'agente IA come nuova insider threat

La seconda faglia si apre con l'adozione massiva di agenti IA integrati nei processi core:

- **motori di trading** che ottimizzano lo order routing su più venue;
- **agent di liquidity management** che bilanciano posizioni intraday;
- **assistenti di compliance** che leggono regolamenti e generano alert;
- **tool di credit scoring** che ricalcolano in tempo reale il profilo rischio-cliente.

Questi sistemi agentici agiscono secondo cicli OODA e, se dotati di privilegi ampi e connessioni API profonde, diventano a tutti gli effetti «insider digitali» instancabili.

Il rischio Excessive Agency

Il rischio emergente è definito come **Excessive Agency**:

- autonomia operativa non proporzionata al livello di controllo umano;
- accesso diretto a dati sensibili (KYC/AML, posizioni di portafoglio, book ordini) senza un'adeguata segmentazione;
- capacità di avviare azioni irreversibili (es. cancellare backup, chiudere posizioni in perdita, modificare limiti di rischio) in modo automatico.

Vulnerabilità della catena di fiducia

Studi recenti mostrano quanto sia vulnerabile la catena di fiducia di questi agenti. Un lavoro scientifico del 2025 sui prompt injection ha evidenziato che, in assenza di difese specifiche, gli attacchi possono compromettere fino a oltre il 70% dei task agentici sottoposti a test. L'applicazione di un framework multistrato ha ridotto la percentuale di successo degli attacchi a circa l'8,7%: segno che il problema è mitigabile, ma tutt'altro che risolto.
Meta, nel proprio approccio «Agents Rule of Two», insiste sulla prompt injection come vulnerabilità fondamentale: l'agente può essere indotto a ignorare le istruzioni originali, esfiltrare dati o agire contro l'interesse dell'organizzazione tramite input apparentemente innocui (email, documenti, feed di news, persino dati di mercato avvelenati).

Scenari pratici in ambito finance

In ambito finance, questo si traduce in scenari concreti:

- un **agent di tesoreria** che, a seguito di prompt injection, inizia a spostare liquidità verso un conto controllato dall'attaccante;
- un **agent di trading** che sovra-espone il book su determinate coppie crypto o derivati, innescando perdite o manipolazioni di prezzo;
- **protocolli A2A** tra istituzioni che, se non autenticati e firmati criptograficamente, possono essere manomessi per alterare condizioni di regolamento, collateral o margini.

L'attaccante non punta più solo all'account umano privilegiato, ma alla «mente operativa» dell'istituzione: l'agente. Una singola prompt injection ben posizionata può valere più di mille credenziali rubate.

Identity-First Security: trattare le macchine come soggetti di controllo

Per gestire in modo efficace la crescente complessità degli ambienti finanziari digitali, le organizzazioni devono adottare un **modello di sicurezza esplicitamente Identity-First**, applicando i principi dello Zero Trust a tutte le identità, siano esse umane o macchina.

Principi fondamentali

In questo quadro, **l'autenticazione** e **l'autorizzazione** non possono più essere considerate attività puntuali, ma processi continui, **fondati sulla verifica**:

- del contesto operativo;
- del dispositivo utilizzato;

- della rete di accesso;
- della localizzazione;
- dei comportamenti osservati.

Ne consegue che ogni deviazione rispetto ai comportamenti attesi deve generare controlli ulteriori, fino alla possibile sospensione automatica dei privilegi.

Granularità degli accessi

Tale approccio richiede inoltre un elevato livello di granularità nella definizione degli accessi. Ogni agente di intelligenza artificiale deve essere configurato come un soggetto operativo dotato dei soli privilegi necessari, coerenti con:

1. il proprio ruolo;
2. il compito assegnato;
3. l'ambiente in cui opera

Esempi pratici

- In **ambito trading**, un agente incaricato di elaborare analisi o raccomandazioni non dovrebbe disporre dei privilegi per eseguire ordini.
- Nei **processi di credito**, un modello di scoringnon dovrebbe poter intervenire direttamente sulla modifica dei limiti di esposizione nei sistemi core.

Gestione delle credenziali macchina

Un ulteriore elemento critico riguarda la **gestione delle credenziali** e delle **identità macchina**. Chiavi crittografiche, certificati, token di firma e credenziali machine-to-machine:

- non possono essere dispersi in script o repository;
- devono essere gestiti attraverso infrastrutture dedicate e centralizzate;
- devono essere sottoposte a meccanismi di rotazione automatica;

- devono avere controlli condivisi per gli asset più sensibili;
- devono garantire tracciabilità completa degli accessi e delle operazioni.

La compromissione di una signing key utilizzata per la firma di ordini di borsa o transazioni di pagamento deve essere considerata un evento di massima criticità, poiché può compromettere l'integrità dell'intero sistema di fiducia.

Riconoscimento normativo

Questi principi trovano ormai un riscontro diretto anche nell'evoluzione normativa:

- **DORA**: qualifica il rischio ICT, e conseguentemente anche il rischio legato alle identità, come un rischio di business, attribuendo al management body la responsabilità ultima dell'adozione, dell'approvazione e della supervisione del framework di gestione del rischio tecnologico.
- **AI Act**: ricomprende tra i sistemi ad alto rischio numerosi casi d'uso impiegati nel settore finanziario (credit scoring, infrastrutture critiche, sistemi assicurativi), imponendo requisiti stringenti in materia di:
  - gestione del rischio;
  - registrazione delle attività;
  - resilienza cyber;
  - supervisione umana.

In sintesi, non è più sostenibile che un agente operi come una black box, priva di un'identità formalizzata, di confini di autorizzazione, di tracciabilità e di supervisione. Se un sistema è in grado di agire in autonomia, deve essere anche sottoposto a un modello di governo proporzionato al livello di autonomia riconosciuto.

Strategie di resilienza: ITDR, monitoraggio dinamico e SBOM per l'IA

La tecnologia più allineata a questa nuova fase è l'ITDR[38]. A differenza dei tradizionali SIEM[39] o EDR[40], l'ITDR è progettato per difendere direttamente i sistemi di identità e IAM[41], monitorando posture, configurazioni e attività anomale su account umani e macchina.

Capacità delle soluzioni ITDR mature

Le **soluzioni ITDR** più mature:

1. Costruiscono profili comportamentali per ogni identità (frequenza di login, risorse usuali, pattern di accesso e approvazione);
2. Assegnano un punteggio di fiducia dinamico (User/Machine Trust Level) che sale o scende in base alle deviazioni osservate;
3. Si integrano con i controlli di accesso per declassare automaticamente i permessi, imporre MFA aggiuntiva o sospendere account e agenti in tempo reale.

In uno scenario di session hijacking su un account di treasury o di manipolazione di un agent di trading tramite prompt injection, l'**ITDR può degradare i privilegi prima che l'attaccante completi un movimento laterale verso asset critici** – ad esempio bloccando il tentativo di eseguire operazioni fuori policy o di accedere a backup e sistemi di riconciliazione contabile.

[38] Identity Threat Detection and Response.
[39] Security Information and Event Management.
[40] Endpoint Detection and Response.
[41] Identity and Access Management.

Trasparenza della supply chain: SBOM e AI SBOM

Accanto all'ITDR, diventa strategica la trasparenza della supply chain software. L'adozione dello **SBOM**, inizialmente promossa in ambito software generico, sta entrando anche nel **perimetro finanziario**:

- Il regolatore indiano SEBI[42] ha imposto lo SBOM per i software critici usati dagli intermediari vigilati, per migliorare la capacità di reagire a nuove vulnerabilità.
- A livello internazionale, la CISA[43] ha aggiornato nel 2026 le linee guida sugli elementi minimi di una SBOM, sottolineandone il ruolo chiave per la gestione del rischio legato alla software supply chain.

Per l'IA questo concetto va esteso a un **«AI SBOM»** che includa:

- modelli (foundation e fine-tuned), dataset, librerie di inferenza e orchestratori di agent;
- origine e licenza dei componenti, vulnerabilità note, patch applicate;
- mappatura degli impatti: quali agent toccano funzioni regolamentate (es. high-risk AI secondo l'AI Act), quali processano dati personali, quali hanno capacità di esecuzione autonoma.

Nel finance, sapere che un agent di margin-management usa una libreria affetta da un CVE[44] critico o un modello non più conforme all'AI Act non è un dettaglio tecnico, ma un'informazione materiale per la resilienza operativa – e quindi per la compliance DORA.

[42] Securities and Exchange Board of India.
[43] Cybersecurity and Infrastructure Security Agency.
[44] Common Vulnerabilities and Exposures.

L'umano al centro: etica, human-in-the-loop e responsabilità legale

Nonostante il livello di automazione, l'architettura di sicurezza deve mantenere, per design e per norma, un meccanismo Human-in-the-Loop per le decisioni ad alto impatto finanziario. L'AI Act, all'articolo 14, impone che i sistemi di IA ad alto rischio siano progettati in modo da poter essere «effettivamente supervisionati da persone fisiche», prevedendo la possibilità di:

- interrompere il sistema;
- ignorare le decisioni;
- rovesciarne le decisioni quando necessario.

Implementazione nel Banking & Insurance

Nel **banking & insurance** questo si traduce in:

1. **Dual control umano obbligatorio** per operazioni sopra determinate soglie, anche se originate da un agent (es. rilascio di fondi, rettifiche contabili maggiori, cambiamenti ai limiti di rischio di portafoglio).
2. **Chiara attribuzione di accountability**: qualcuno nel board o nel management deve essere formalmente responsabile degli agent e dei modelli impiegati, come richiesto sia dall'AI Act sia da DORA in materia di ICT risk.

Il contenzioso legale in crescita

Nel 2025, varie analisi hanno evidenziato un aumento significativo delle cause legali legate all'IA:

- Raddoppio del contenzioso in un solo anno.
- La maggior parte dei consigli di amministrazione dichiara competenze limitate o nulle in materia di AI governance.
- Solo una minoranza di aziende possiede policy di AI approvate dal board, nonostante l'ampia diffusione operativa di sistemi IA.

La direzione è chiara: l'AI liability non è più un tema astratto, ma una fonte concreta di esposizione per Directors e Officers. Studi europei su IA e responsabilità civile sottolineano l'esigenza di adattare i regimi di responsabilità per garantire:

- compensazione alle vittime;
- riduzione del contenzioso;
- promozione di comportamenti diligenti dei fornitori e degli utilizzatori di sistemi IA.

Per le istituzioni finanziarie, questo comporta un ulteriore onere di responsabilità, che non riguarda soltanto la conformità IT, ma l'intero ciclo di vita degli agent: dalla progettazione alla messa in produzione, fino al monitoraggio continuo. DORA, in particolare, stabilisce che il management è responsabile in ultima istanza della gestione del rischio ICT e richiede che i membri del board mantengano un livello adeguato di conoscenze e competenze su tali rischi, attraverso una formazione regolare.

Primato della persona: il richiamo etico

Su uno sfondo più ampio, i richiami etici alla dignità umana insistono sul fatto che l'intelligenza – soprattutto se artificiale – non può essere assolutizzata a scapito della responsabilità individuale. Nessun agente, per quanto sofisticato, possiede coscienza morale: può ottimizzare una funzione obiettivo, ma non può assumersi la responsabilità di decidere che cosa sia «giusto» per un soggetto vulnerabile, un piccolo risparmiatore o la stabilità complessiva di un mercato, come ribadito anche da **Leone XIV** nella **Magnifica humanitas (2026)**, dove si richiama con forza il primato della persona sulle logiche puramente tecniche.

CONCLUSIONI

Nel 2026 la cybersecurity non può più essere considerata una funzione di difesa perimetrale, ma una capacità di governo dell'intero ecosistema digitale. In un contesto in cui persone, sistemi automatizzati e dati operano in modo sempre più interdipendente, la minaccia non si presenta più come un evento isolato, bensì come una perturbazione sistemica capace di propagarsi lungo catene tecnologiche, organizzative e regolatorie.

L'analisi condotta nel report conferma che il fenomeno cyber sta evolvendo non solo in termini quantitativi, ma soprattutto qualitativi. L'aumento degli incidenti ad alto impatto, la diffusione dell'intelligenza artificiale nelle tecniche d'attacco e la persistenza di minacce come phishing, ingegneria sociale e sfruttamento delle vulnerabilità dimostrano che le difese statiche non sono più sufficienti. La protezione efficace richiede oggi capacità di rilevazione continua, gestione rigorosa delle identità e controllo puntuale dei privilegi.

Un ulteriore elemento critico riguarda la supply chain tecnologica. La crescente dipendenza da fornitori, piattaforme e servizi terzi amplia la superficie di esposizione e rende possibile che una debolezza circoscritta si trasformi rapidamente in un rischio sistemico. La sicurezza della supply chain non è quindi un tema tecnico accessorio, ma un pilastro della resilienza organizzativa e della continuità operativa.

Sul piano della governance, le evoluzioni normative e l'attenzione crescente dei vertici aziendali stanno ridefinendo il ruolo della cybersecurity all'interno dei processi decisionali. La sicurezza non riguarda più soltanto la protezione di infrastrutture e dati, ma incide direttamente sulla capacità dell'organizzazione di operare, documentare le proprie scelte e rispondere in modo efficace agli scenari di crisi.

L’era dell’intelligenza artificiale agentica introduce infine un ulteriore cambio di paradigma. L’autonomia delle macchine amplia le possibilità operative, ma rende anche più sottile il confine tra controllo umano e azione automatizzata. In questo scenario, la supervisione dell’uomo rimane indispensabile: l’efficienza dell’automazione non può tradursi in opacità, né in delega incontrollata delle decisioni critiche.

Raccomandazioni

Alla luce di quanto emerso, le organizzazioni dovrebbero adottare alcune priorità concrete:

- Mappare in via prioritaria le identità macchina. Il perimetro reale dell’organizzazione è oggi composto da agenti, bot e chiavi API tanto quanto dalle persone.
- Introdurre il dual control umano per le azioni ad alto impatto. Nessun sistema autonomo dovrebbe autorizzare da solo operazioni finanziarie, variazioni di rischio o accessi critici.
- Trattare i modelli AI come componenti della supply chain. L’adozione di un AI-BOM è necessaria per garantire tracciabilità, conformità e controllo della provenienza.
- Allineare il budget cyber al capitale regolamentare. La prevenzione deve essere letta anche come leva economica, non solo come costo di protezione.
- Simulare regolarmente attacchi agentici. I test tradizionali non coprono scenari di prompt injection o compromissione A2A.

In conclusione, la cybersecurity sta diventando una condizione strutturale di solidità, continuità e competitività. Le organizzazioni che sapranno integrare governance, controllo operativo e disciplina nella gestione dei sistemi intelligenti potranno trasformare la complessità in vantaggio competitivo. Le altre rischiano di subire una trasformazione già in corso.

Bibliografia e sitografia

- Accenture Security. (n.d.). Analisi sul cybercrime-as-a-service (CaaS).
- Anthropic. (n.d.). Studi sui sistemi agentici e sicurezza dei modelli di intelligenza artificiale.
- Bank for International Settlements. (n.d.). Basel III: Operational risk – standardized measurement approach (SMA). https://www.bis.org/
- Basel Committee on Banking Supervision. (n.d.). Operational risk framework and resilience principles. https://www.bis.org/bcbs/
- Cisco. (2026). AI policy landscape / AI security research and tooling. https://www.cisco.com/c/en/us/products/security/state-of-ai-security.html
- Cisco. (2026). State of AI security report 2026. https://www.cisco.com/c/en/us/products/security/state-of-ai-security.html
- Clusit – Associazione Italiana per la Sicurezza Informatica. (2026). Rapporto sulla cybersecurity in Italia e nel mondo 2026. https://clusit.it/rapporto-clusit/
- Commissione europea. (n.d.). Digital operational resilience act – overview. https://finance.ec.europa.eu/digital-finance/digital-operational-resilience-act_en
- CrowdStrike. (2026). 2026 global threat report: Year of the evasive adversary. https://www.crowdstrike.com/en-us/blog/crowdstrike-2026-global-threat-report-findings/
- ENISA – European Union Agency for Cybersecurity. (n.d.). Threat landscape reports. https://www.enisa.europa.eu/
- European Banking Authority. (n.d.). Policy advice on operational risk and Basel III implementation. https://www.eba.europa.eu/
- European Banking Authority, European Insurance and Occupational Pensions Authority, & European Securities and Markets Authority. (n.d.). Regulatory technical standards su DORA. https://www.eba.europa.eu/
- European Central Bank. (n.d.). Cyber resilience strategy for the financial sector. https://www.ecb.europa.eu/
- European Systemic Risk Board. (n.d.). Systemic cyber risk report / EU systemic cyber risk framework. https://www.esrb.europa.eu/
- Federal Reserve. (n.d.). Cybersecurity and financial system resilience reports. https://www.federalreserve.gov/
- Financial Stability Board. (n.d.). Cyber incident response and recovery / cyber resilience frameworks. https://www.fsb.org/
- Gartner. (n.d.). Report su agentic AI e security operations.
- Google DeepMind, & Google Cloud Security. (n.d.). Research su agenti IA.
- IBM Security. (2025). X-Force threat intelligence index 2025. https://www.ibm.com/
- Integris. (2026). 2026 banking trust and technology report. https://integrisit.com/wp-content/uploads/2026/02/NEW_Integris_Banking-Trust-and-Technology-Report_

- International Monetary Fund. (n.d.). Cyber risk and financial stability. https://www.imf.org/
- International Organization of Securities Commissions. (n.d.). Cyber resilience guidance for financial market infrastructures. https://www.iosco.org/
- KPMG. (2026). Global tech report 2026: Leading in the intelligence age. https://kpmg.com/lu/en/insights/ai-and-technology/global-tech-report-2026.html
- Leone XIV. (2026). Magnifica humanitas: Lettera enciclica sulla custodia della persona umana nel tempo dell'intelligenza artificiale. https://press.vatican.va/content/salastampa/it/bollettino/pubblico/2026/05/25/0440/00861.html
- Mandiant. (n.d.). Report su minacce avanzate e attori statali.
- Microsoft. (2025). Microsoft digital defense report 2025: Lighting the path to a secure future. https://www.microsoft.com/en-us/corporate-responsibility/cybersecurity/microsoft-digital-defense-re...
- National Institute of Standards and Technology. (n.d.). Cybersecurity framework 2.0. https://www.nist.gov/cyberframework
- National Institute of Standards and Technology. (n.d.). Zero trust architecture (SP 800-207). https://csrc.nist.gov/publications/detail/sp/800-207/final
- OECD. (n.d.). Cyber security risk management for economic and social prosperity. https://www.oecd.org/
- OpenAI. (n.d.). Studi sui sistemi agentici e sicurezza dei modelli di intelligenza artificiale.
- OWASP – Open Worldwide Application Security Project. (n.d.). AI testing guide / LLM top 10. https://owasp.org/
- Oxford Economics, & Splunk. (2026). The CISO report 2026: From risk to resilience in the AI era. https://www.splunk.com/en_us/form/ciso-report.html
- Palo Alto Networks. (2025). 6 predictions for the AI economy: 2026's new rules of cybersecurity.
- Parlamento europeo, & Consiglio dell'Unione europea. (2022). Regolamento (UE) 2022/2554 – digital operational resilience act (DORA). https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32022R2554
- Sopra Steria. (2026). State of cyber security 2026.
- Thales. (2026). 2026 data threat report: Data security in the agentic age. https://cpl.thalesgroup.com/data-threat-report
- World Economic Forum. (2026). Global cybersecurity outlook 2026. https://www.weforum.org/publications/global-cybersecurity-outlook-2026/

CONTATTI

Matteo M. Marzan
matteo.marzan@planetica.it

Andrea Rivetti
andrea.rivetti@planetica.it

Team di lavoro

Riccardo
Mandiroli

Alessandro
Croce

Marta
Allievi

Morena
Peloso

Per maggiori informazioni:
Tel: +39 02 82785 740 | E-mail: segreteria@planetica.it | Indirizzo: Via Crocefisso 5, Milano

Visita il nostro sito web per scaricare
il report completo

planetica

www.planetica.it