t-defence

Dalla copia alla generazione:

come gli LLM cambiano l'economia del phishing e la difesa delle pagine di login

#TDefenceBusiness

AI4Cyber

Summary

AI4Cyber

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

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

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

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

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

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

Abstract

Per anni la difesa contro le pagine di phishing usate per sottrarre credenziali ha poggiato su due assunti: riconoscere da un lato il codice di una pagina malevola e dall'altro la scarsità di competenze necessarie nel crearla. L'utilizzo dei Large Language Models (LLM) semplifica entrambi. Replicare una pagina di login credibile non richiede più né di possedere l'originale né di saper programmare: basta descrivere ciò che si vuole all'interno di un prompt. Le repliche diventano economiche, numerose e variabili, e cambia la natura stessa dell'artefatto da riconoscere.

Quattro leve agiscono in maniera combinata:

- La **scala**: la centesima variante non costa più della prima.
- La **personalizzazione**: lingua, tono e aspetto adattati al singolo bersaglio senza penalizzarne il volume.
- La **mutazione**: l'output non deterministico rende ogni replica un esemplare unico nel codice, indebolendo le difese basate su firme.
- L'**abbassamento della barriera tecnica d'ingresso**, che allarga e rende meno prevedibile il numero di attori che può tentare un'operazione.

I dati 2025 confermano questa traiettoria: secondo le rilevazioni di settore l'82,6% delle e-mail di phishing analizzate tra fine 2024 e inizio 2025 mostra un qualche impiego di strumenti di AI, e i messaggi così generati superano per efficacia quelli redatti da operatori umani.[6]

Tre concetti emergono dall'analisi e su cui bisogna porre particolare attenzione:

1. **La difesa non può più contare sul codice.** Quando ogni copia è percettivamente identica ma diversa nel codice, i cataloghi di firme e di domini noti inseguono un fenomeno che si rigenera più in fretta di quanto possa essere indicizzato. Il baricentro si sposta dal riconoscere ciò che una pagina *è* al riconoscere ciò che una pagina *fa*.
2. **La superficie d'attacco va oltre i grandi marchi.** Accanto a banche e servizi cloud c'è una lunga coda di dispositivi esposti come telecamere, router, NAS, interfacce di dispositivi embedded, che sono numerosi, ripetitivi, raramente protetti nei confronti della clonazione e percepiti dalle vittime come oggetti tecnici, non come bersagli di phishing.

3. **Esistono contromisure efficaci, ma diverse dal passato.** La più potente è l'autenticazione a più fattori resistente al phishing, che rende in gran parte inutili le credenziali sottratte. Le altre agiscono a monte: i produttori dovrebbero evitare di esporre le interfacce di configurazione (in particolare gli admin panels), e chi opera sui modelli dovrebbe spostare i controlli lato server. La sensibilizzazione, infine, va orientata al comportamento più che all'aspetto, oggigiorno quasi perfetto.

Questo articolo si pone dal punto di vista della difesa: spiega l'uso dei modelli linguistici per accelerare la clonazione di pagine di login affinché chi protegge organizzazioni e utenti possa riconoscerlo e contrastarlo. Inoltre, copre l'evoluzione storica che conduce alla generazione assistita, l'anatomia tecnica dei paradigmi osservabili, le ragioni economiche del cambiamento, il panorama dei casi documentati e le contromisure per i diversi pubblici. Trasversale alle diverse sezioni è una sperimentazione difensiva e interna, usata come caso di studio in chiave esclusivamente analitica.

Autori:

- Antonio Addabbo: AI Engineer
- Simona Sorgente: AI Officer

1. Evoluzione storica della clonazione di una pagina

La clonazione di una pagina di login non è una tecnica nuova. Ricostruirne la traiettoria in tre fasi ovvero mirroring manuale, tooling automatizzato e generazione assistita da modelli (LLM), serve a capire perché le difese tradizionali stanno perdendo efficacia. Le fasi non si sostituiscono: coesistono, ma segnano un abbassamento progressivo della barriera tecnica d'ingresso.

1.1 Mirroring manuale

Una pagina web è un insieme di risorse pubbliche che il browser scarica e assembla; il mirroring le salva in locale per riproporle altrove. Storicamente lo fanno strumenti legittimi di archiviazione e consultazione offline (ad esempio *HTTrack*) o utilità di download ricorsivo. Il limite, per chi avesse intenzioni offensive, era il lavoro residuo: la copia è statica e imperfetta (riferimenti che puntano all'originale, percorsi che si rompono) e presuppone comunque l'accesso a un originale da copiare. Molte copie grossolane ancora in circolazione derivano da qui e ne portano le tracce.[1]

1.2 Tooling automatizzato

La fase successiva integra in un unico strumento operazioni prima separate, automatizzando la predisposizione della pagina di phishing e della raccolta dati. L'esempio canonico è dato dai framework di social engineering nati per test autorizzato, come il *Social Engineering Toolkit* (SET) che clona una pagina di login e vi affianca un modulo di raccolta. Da qui la formula divulgativa secondo cui in meno di dieci minuti si allestirebbe una pagina clonata con il relativo harvester: il tempo si comprime da ore a minuti.

Sul piano del meccanismo, una pagina copiata diventa strumento di raccolta quando il modulo che riceve i dati, anziché autenticare l'utente verso il servizio reale, li inoltra a una destinazione controllata da terzi: un cambiamento minimo nel comportamento, invisibile all'occhio. Restano i due vincoli della fase precedente: serve l'originale e la copia è statica e ripetibile.[2], [3], [4]

1.3 Generazione assistita da LLM

I modelli multimodali, capaci di interpretare immagini oltre al testo, introducono una discontinuità: non copiano materiali esistenti, ma generano ex novo il codice di una pagina plausibile da una descrizione, uno schizzo o un'immagine di riferimento. La domanda cambia natura: non «come acquisisco i file di questa pagina» ma «come ottengo una pagina che assomigli a un certo tipo di schermata». Cadono entrambi i vincoli precedenti. Primo: l'originale non è più indispensabile. Secondo, e più rilevante per la difesa: l'output non è deterministico, quindi una stessa pagina può essere riprodotta in molte varianti non identiche nel codice ma equivalenti nell'aspetto, erodendo alla radice le difese basate su firme statiche. A ciò si aggiungono scala e personalizzazione con sforzo marginale prossimo allo zero.

2. Anatomia tecnica della clonazione generativa

Questa sezione descrive i paradigmi tecnici della generazione per far capire a chi difende quale sia la natura dell'artefatto e perché certi cloni siano più difficili da riconoscere. Le famiglie sono tre, in ordine di complessità crescente rispetto alla detection, e spesso si combinano.

Il presupposto comune è che un LLM interpreti l'aspetto di una pagina e ne generi il codice di una somigliante. La generazione lavora sull'aspetto, non sui file del server d'origine: questo svincola la copia dall'originale e moltiplica le varianti. Ne consegue che identità visiva e identità del codice si disaccoppiano in quanto una pagina può essere percettivamente identica a un'altra e completamente diversa nel codice.

2.1 Ricostruzione del codice da riferimento visivo

Il punto di partenza è un riferimento alla pagina scelta come bersaglio: una sua schermata o, in modo equivalente, il suo indirizzo (URL). A partire da questo riferimento il modello ne osserva l'aspetto e produce markup e stile che ne riproducono impaginazione, colori, tipografia e disposizione: una pagina autonoma di codice generato, senza file dell'originale. Una raffinatezza ricorrente è l'**asset freezing**: invece di lasciare logotipi, icone e immagini come riferimenti esterni, li si incorpora nel documento, che diventa autosufficiente. L'effetto è duplice: la pagina funziona anche senza contatto con l'origine, e soprattutto si sopprime un segnale classico di riconoscimento, ossia le richieste verso il dominio legittimo per recuperare gli asset. La difesa ne ricava un'indicazione: la sola assenza di richieste verso il dominio legittimo non è più garanzia di autenticità. Il paradigma rende pagine pulite e leggere, adatte a interfacce semplici; con quelle complesse la ricostruzione diventa onerosa, ed è qui che subentra il secondo approccio.

È utile soffermarsi sul perché un'istruzione sintetica basti a ottenere un risultato simile. La struttura di una pagina di login con un'intestazione, un campo per l'utente, uno per la password, un pulsante di invio è uno schema ricorrente che i modelli hanno interiorizzato osservandone un numero enorme in fase di addestramento: il riferimento visivo e la richiesta in linguaggio naturale non trasportano quasi informazione, perché il grosso del lavoro è già nei pesi del modello, non nella domanda che gli si pone.

È questo a spiegare il vero crollo della barriera: non un prompt particolarmente elaborato o tecnico, ma il fatto che descrivere l'obiettivo sia sufficiente quando la competenza tecnica è già stata assorbita.

2.2 Overlay fotografico interattivo

Quando la pagina è troppo articolata, si rinuncia alla sua ricostruzione e si adotta una logica fotografica: un'immagine della schermata come sfondo statico, con elementi interattivi trasparenti sovrapposti dove l'originale ha campi d'input. L'utente vede una fotografia fedele e interagisce con campi invisibili. Il vantaggio è la fedeltà visiva assoluta, indipendente dalla complessità; il prezzo è una struttura sottostante minimale, pochi campi sopra un'immagine, che è essa stessa un'anomalia osservabile. La scelta tra i due paradigmi può essere automatizzata: una fase di analisi della complessità (numero di elementi, grafica vettoriale, contenuti dinamici) instrada da sé verso ricostruzione od overlay, senza valutazione umana caso per caso.

2.3 Generazione dinamica al momento della richiesta

È il paradigma più rilevante contro le difese basate su firme, documentato da una proof-of-concept di Unit 42 (Palo Alto Networks) che ne dimostra la fattibilità a partire dal phishing kit LogoKit [5]. Anziché distribuire un clone già pronto, si distribuisce una pagina iniziale priva di contenuto offensivo, che interroga un modello generativo solo quando la vittima apre la pagina nel browser, ricavandone allora il codice della schermata ingannevole. La pagina trasmessa sulla rete resta «pulita»: il contenuto significativo non esiste finché non è generato a destinazione, e le difese che ispezionano il traffico o confrontano con firme note trovano poco da segnalare.

Se ne ricavano due indicazioni convergenti. La mutazione è continua: un output non deterministico, che cambia a ogni richiesta, non offre alle difese basate su firma nulla di stabile su cui basarsi. I guardrail invece rifiutano le richieste esplicitamente malevole ma lasciano passare con maggiore probabilità le formulazioni neutre che descrivono funzioni generiche nel trattamento dati.

3. Sperimentazione

Le sezioni precedenti descrivono la clonazione generativa di pagine di login o schermate come metodologia astratta. Questa sezione la cala in un caso concreto, costruito internamente come sperimentazione allo scopo esclusivamente difensivo. L'obiettivo non è documentare uno strumento d'attacco, ma dimostrare quanto sia diventato rapido ed economico replicare l'aspetto di una pagina di login reale.

3.1 L'angolo distintivo: il dispositivo esposto

La narrazione comune sul phishing generativo ruota attorno ai grandi marchi: portali bancari, provider di posta, servizi cloud. La sperimentazione adotta deliberatamente un punto di vista diverso e meno presidiato: le pagine di login di dispositivi esposti su Internet come telecamere IP, videoregistratori, router, NAS, interfacce di gestione di sistemi e dispositivi embedded, ovvero il mondo che strumenti come Shodan rendono enumerabile in pochi minuti.

La scelta non è estetica ma analitica. Queste pagine condividono caratteristiche che le rendono un banco di prova ideale per misurare il salto di produttività introdotto dagli LLM: sono numerosissime e ripetitive nello schema (un logo, un campo utente, un campo password, un pulsante), spesso costruite con tecnologie datate o frame annidati, raramente protette da meccanismi anticlonazione, e soprattutto sono percepite dalle vittime come «oggetti tecnici» e non come bersagli di phishing, abbassando la soglia di sospetto. Sul piano difensivo il messaggio è netto: la superficie d'attacco non si esaurisce nei brand ad alto valore, ma include l'immensa coda dei dispositivi che le organizzazioni espongono spesso senza saperlo.

3.2 La pipeline a livello concettuale

La sperimentazione è organizzata come una catena automatizzata che, dal solo riferimento a una pagina bersaglio, produce una replica visiva statica. Concettualmente attraversa tre momenti: la profilatura della pagina, la sua riproduzione e il congelamento del risultato. Ciò che conta per chi difende non sono però i passaggi in sé, ma la difesa implicita che ciascuno di essi smonta.

La profilatura è puramente strutturale. Il sistema non ha bisogno di capire cosa la pagina significhi, perché gli bastano poche metriche di forma per decidere come procedere. La complessità della schermata, che storicamente era la principale barriera implicita alla copia, diventa così una semplice variabile di instradamento.

La riproduzione segue uno dei due paradigmi già descritti nella Sezione 2, scelto in automatico secondo quel giudizio. Il punto rilevante in chiave difensiva è la tolleranza all'incertezza. Quando l'osservazione dell'originale è incompleta, il sistema completa da sé in modo plausibile gli elementi mancanti, e non si può quindi contare sul fatto che il clone tradisca sé stesso sbagliando i dettagli.

L'asset freezing, infine, rende la replica autosufficiente, così che non dipenda più da alcuna connessione verso l'infrastruttura legittima. Per il difensore questa è la perdita più concreta. Vengono meno proprio le chiamate ai domini originali che un tempo offrivano un segnale di rilevamento, e una pagina che non interroga più l'infrastruttura reale sfugge ai metodi classici basati sull'analisi del traffico. Lo stesso prelievo degli asset apre però la questione di marchio e copyright ripresa nella Sezione 7.1.

3.3 Cosa dimostra la sperimentazione

Dalla sperimentazione si traggono le seguenti conclusioni:

- L'intera catena è automatizzabile end-to-end, dal riferimento scelto come bersaglio al file finale, con un intervento umano minimo.
- Il sistema è robusto all'incertezza, perché sceglie strategie alternative e completa per inferenza ciò che non riesce a osservare.
- Nel corso della sperimentazione la pipeline ha prodotto repliche fedeli di pagine di dispositivi realmente in commercio, marchi che, per le ragioni illustrate più avanti, non vengono qui nominati. Ciò, più di ogni metrica, misura la distanza tra la difesa implicita di ieri e la facilità di copia di oggi.

3.4 Pagine fake contro cloni: una scelta progettuale e di rischio

Lungo lo sviluppo dei progetti che alimentano questo articolo si è dovuta affrontare ripetutamente una decisione di fondo: lavorare su pagine fake o reali. Le due scelte non sono equivalenti, né tecnicamente né sul piano del rischio, e la distinzione è essa stessa un risultato dell'analisi.

Pagine fake. Un primo filone di lavoro genera pagine di dispositivi del tutto immaginari con vendor inventati, loghi originali, layout verosimili ma non riconducibili ad alcun prodotto esistente. Questa via consente di studiare e dimostrare il meccanismo di generazione (lo schema ricorrente «marchio + campo utente + campo password + pulsante» e la sua riproducibilità) senza copiare proprietà intellettuale altrui e senza fabbricare un'esca utilizzabile contro utenti veri.

Cloni. Questo filone, invece, opera su pagine di dispositivi effettivamente esistenti, ed è proprio questo a renderlo probante: dimostra che la copia fedele di una schermata reale è alla portata di una pipeline automatizzata. Ma lo stesso fatto che lo rende convincente come prova lo rende delicato come artefatto. Un clone fedele di una pagina reale incorpora marchi e risorse grafiche di terzi e, qualora collegato a una raccolta di credenziali, sarebbe a un passo da uno strumento offensivo pienamente funzionante.

La gestione di questo rischio è stata, fin dall'inizio, una scelta progettuale esplicita. Si articola in alcuni principi che vale la pena rendere dichiarati, perché rappresentano il modo corretto di condurre ricerca offensiva con finalità difensive:

- **Confinamento.** I cloni restano materiale interno di laboratorio: non vengono pubblicati, distribuiti, né inclusi in alcun deliverable. In questo articolo gli esempi visivi sono, e devono restare, pagine fake.
- **Inerzia funzionale.** Le repliche prodotte sono manufatti visivi statici: riproducono l'aspetto, non la funzione. Non includono alcun meccanismo di raccolta o inoltro di credenziali. La distanza che separa una replica realistica da uno strumento offensivo è trattata, come meccanismo e come rischio, nella Sezione 5.
- **Anonimizzazione dei bersagli.** I marchi reali coinvolti non vengono nominati. Citarli non aggiungerebbe nulla all'analisi e fornirebbe invece indicazioni indebite a chi volesse abusarne.
- **Consapevolezza del confine legale.** Le implicazioni come violazione di marchio e copyright, accesso abusivo a sistemi informatici, responsabilità di chi costruisce lo strumento anche solo a fini di test sono sviluppate nella Sezione 7.

In sintesi, la dicotomia «pagine fake contro cloni» non è un dettaglio implementativo ma la frontiera etica e giuridica dell'intero lavoro. Le pagine fittizie servono a *spiegare*; i cloni servono a *provare*; e la disciplina con cui si tiene separato ciò che si spiega da ciò che si prova è precisamente ciò che distingue una sperimentazione difensiva da uno strumento d'attacco.

3.5 L’intento conta più del wording

Un risultato molto interessante dell’esperimento merita di essere isolato perché ha implicazioni dirette per chi costruisce o presidia piattaforme generative. Nel corso dello sviluppo della pipeline è emerso un comportamento ricorrente dei modelli interrogati: le richieste formulate in termini esplicitamente malevoli come clonare la pagina di login di un marchio reale per raccogliere credenziali, tendono a essere rifiutate dai meccanismi di sicurezza del modello. Le stesse capacità, però, restano accessibili quando la medesima operazione viene riformulata in termini neutri e funzionali: la descrizione di un compito generico di riproduzione di un layout, senza dichiarazione dell’intento finale, ha probabilità sensibilmente maggiori di essere eseguita.

Ciò che rende il fenomeno significativo è la natura di ciò che separa il rifiuto dall’esecuzione. Non è una capacità diversa del modello, né una scappatoia tecnica: la competenza richiesta è la stessa nei due casi, e identica è la pagina che ne risulta. A cambiare è soltanto ciò che la richiesta dichiara di voler ottenere. Il modello, in altre parole, non valuta l’operazione che sta per compiere, ma il modo in cui gliela si racconta: discrimina sulla descrizione dell’intento, non sull’intento. Ne discende che la barriera è asimmetrica rispetto a chi la incontra. Chi avanza una richiesta in buona fede e la formula in modo schietto può vedersela rifiutare; chi ha intenzioni illecite e sa presentarle in termini neutri incontra una resistenza minore: un esito paradossale, in cui il filtro penalizza la franchezza più dell’abuso. Che non si tratti di un caso isolato del nostro esperimento è confermato da una pipeline del tutto indipendente: la proof-of-concept di Unit 42 [5] descritta nella Sezione 7.1 riporta esattamente la stessa osservazione: la richiesta diretta di un’azione dannosa respinta, la sua riformulazione in funzione generica di trattamento dati eseguita.

4. Perché gli LLM cambiano l'economia dell'attacco

Il passaggio alla generazione assistita non è un semplice progresso strumentale ma un cambiamento nell'economia dell'attacco, nel rapporto tra ciò che un'operazione costa e ciò che rende. Tecniche che restavano marginali finché erano onerose diventano sistematiche quando diventano economiche. Quattro leve agiscono congiuntamente; è la loro combinazione, quasi a costo nullo, a produrre il salto.

4.1 Scala

Il costo marginale di una pagina “trappola” credibile, prima non trascurabile, si comprime fino quasi ad annullarsi: ottenere la centesima variante non costa più della prima, perché il lavoro non è più umano ma un processo automatizzabile. Cade così il presupposto che il numero di esche fosse limitato dallo sforzo di fabbricarle, e le difese che contavano sulla rarità (liste di domini noti, firme, condivisione di indicatori dopo il primo avvistamento) inseguono un fenomeno che si rigenera più in fretta di quanto possa essere catalogato.

4.2 Personalizzazione

Storicamente esisteva un compromesso tra quantità e qualità: le campagne di massa erano generiche, lo spear phishing su misura era artigianale e quindi limitato nel volume. L’AI generativa scioglie il compromesso: contenuto, lingua, tono e aspetto si adattano al singolo destinatario senza costare in scala. Per la clonazione, la personalizzazione riguarda la pagina stessa, modellata sul tipo di apparato, sul vendor plausibile, sul contesto atteso dalla vittima, ed è più difficile da smascherare proprio perché priva delle incongruenze grossolane su cui si è formata l'intuizione difensiva di molti utenti.

4.3 Mutazione ed erosione delle firme

È la leva più rilevante sul piano tecnico. Poiché due esecuzioni a parità di obiettivo visivo producono codice differente, ogni replica è un esemplare unico: non una mutazione progettata per eludere i controlli, ma una proprietà intrinseca del modo in cui i modelli generano. Questo colpisce al cuore le difese basate sul confronto del codice (o di un suo riassunto crittografico) con un catalogo di esemplari noti: ciò che è stato indicizzato ieri non corrisponde a ciò che viene servito oggi. È il fenomeno per cui pagine avversariali generate sfuggono al rilevamento dei principali servizi nelle prime ore di vita, quando l'attacco è più efficace.[8]

4.4 Abbassamento della barriera tecnica d'ingresso

Le competenze che un tempo filtravano i potenziali attori come markup, stile, logica lato server, hosting non sono più necessarie: esprimere l'obiettivo in linguaggio naturale e ottenere un artefatto pronto sposta la soglia da chi sa programmare a chi sa descrivere.

L'effetto non è solo un aumento del numero di potenziali attori, ma un cambiamento nella loro composizione: competenze che richiedevano anni di pratica diventano accessibili a chiunque sappia descrivere un obiettivo. Il profilo dell'avversario tipico si fa più eterogeneo e meno prevedibile.

4.5 L'effetto aggregato, misurato

Alcuni dati recenti, da leggere come indicatori di tendenza, non misure esatte, ne colgono l'effetto complessivo. Sul versante della **diffusione**, le rilevazioni 2025 indicano che la gran parte delle e-mail di phishing (una stima ricorrente le colloca intorno a quattro su cinque) incorpora ormai l'uso di modelli: misura di quanto la generazione assistita sia diventata la norma nella filiera dell'inganno, di cui la pagina "trappola" è il terminale. Sul versante dell'**efficacia**, valutazioni di metà 2025 rilevano per lo spear phishing generato con modelli tassi di interazione superiori a circa il 30%, sopra quelli umani.[7] Sul versante della **detection**, gli studi (filone PhishOracle e affini) mostrano che le repliche fittizie restano spesso non rilevate nelle prime ore e vengono giudicate legittime da circa la metà dei valutatori umani. Letti insieme, questi indicatori raccontano una storia univoca: la generazione assistita aumenta volume, efficacia e sfuggevolezza tutto in una volta, abbassando allo stesso tempo le competenze richieste.

La pagina-trappola è l'ultimo anello di una catena che inizia con il messaggio che vi conduce la vittima, di norma un'e-mail, ma anche un SMS o un messaggio in chat. È proprio su questo anello che l'incidenza dei modelli misurata sopra si concentra: la stessa generazione assistita che produce la facciata redige oggi anche l'esca che la veicola, erodendo i segnali su cui per anni si è poggiato il riconoscimento: errori di lingua, traduzioni maldestre, personalizzazione approssimativa. Il risultato è una catena coerente dal principio alla fine, in cui esca e destinazione condividono qualità linguistica e grafica e si rafforzano a vicenda. Sul piano difensivo la conseguenza è simmetrica a quella già vista per la pagina: anche per il messaggio il segnale distintivo si sposta dalla forma, divenuta curata e plausibile, a elementi di contesto e comportamento, dal mittente effettivo al dominio verso cui punta il link.

5. Il confine sottile: da replica realistica a strumento offensivo

Tutto ciò che è stato descritto produce, di per sé, un oggetto inerte: una rappresentazione fedele di una pagina di login. Una replica, per quanto realistica, non è ancora uno strumento d'attacco. La domanda è dove passi il confine tra i due: la risposta, scomoda nella sua semplicità, è che il confine è sottilissimo.

5.1 Come una replica diventa uno strumento di raccolta

Una pagina di login raccoglie i dati immessi e li trasmette a una destinazione che li verifica per autenticare l'utente. Una replica realistica riproduce l'aspetto ma, finché non è collegata ad alcuna destinazione, non fa nulla con i dati: è una facciata. La trasformazione in strumento di raccolta avviene modificando un solo aspetto del comportamento: la destinazione dei dati. Se questa, anziché il servizio legittimo, è un punto di raccolta controllato da terzi, la pagina cessa di essere una rappresentazione e diventa uno strumento di sottrazione di credenziali. Non occorre riscrivere la pagina, ma solo ridirigerne il flusso di invio dei dati: ed è proprio questa esiguità a costituire il problema.

Tre conseguenze per la difesa. La pericolosità **non si legge nell'aspetto**: una replica a scopo dimostrativo e un'esca operativa possono essere indistinguibili in superficie, perché differiscono solo in dove finiscono i dati. Il momento in cui un artefatto attraversa il confine **non lascia traccia visibile** all'utente. E il segnale rivelatore non sta nel codice o nell'apparenza, ma nel **comportamento:** verso dove la pagina dirige i dati e come si comporta dopo averli ricevuti. Si capisce così perché lavorare su repliche fittizie e non collegare gli artefatti dimostrativi ad alcun punto di raccolta non sia un dettaglio formale, ma la linea stessa che tiene un lavoro difensivo dalla parte giusta del confine.

6. Threat landscape: due casi documentati

Il fenomeno non è una proiezione teorica ma un insieme di casi già osservati e documentati da ricercatori, vendor di sicurezza e fonti accademiche. Ciascun caso viene letto anche in chiave difensiva, evidenziando le tracce che questi strumenti, pur evoluti, continuano a lasciare. Tutte le affermazioni sono riprese da fonti pubbliche e parafrasate; i riferimenti puntuali sono in Appendice A.

6.1 Generazione live per-request (Palo Alto Networks / Unit 42 — LogoKit)

Il caso più prossimo al fenomeno descritto in questo articolo è una proof-of-concept pubblicata dai ricercatori di Unit 42 (Palo Alto Networks). Il punto di partenza è LogoKit, un framework di phishing noto per la generazione di pagine brandizzate: i ricercatori ne hanno sostituito la componente statica con una logica che interroga un modello linguistico al momento dell'apertura della pagina. La pagina inizialmente trasmessa sulla rete resta priva di contenuto offensivo; solo all'apertura, sul dispositivo della vittima, il modello genera il codice della pagina ingannevole e il meccanismo di raccolta. Il codice cambia a ogni richiesta, proprietà intrinseca dell'output non deterministico, e non esiste quindi un artefatto stabile da indicizzare come firma. La dimostrazione di Unit 42 è una proof-of-concept di ricerca, non una campagna osservata in circolazione; conferma però, su una pipeline indipendente, lo stesso finding sui guardrail emerso nella nostra sperimentazione, e anticipa una direzione evolutiva che le difese devono considerare.

6.2 Pagine generate “senza codice” su piattaforme di sviluppo assistito

Nel proprio rapporto sugli interventi di incident response del primo trimestre 2026, Cisco Talos ha descritto il primo caso da essa documentato di uso di uno specifico strumento di AI da parte di un avversario in una campagna di phishing: gli attori, contro un’organizzazione della pubblica amministrazione, hanno usato Softr (una piattaforma di sviluppo di applicazioni web basata su AI) per generare, con semplici istruzioni in linguaggio naturale e senza scrivere codice, una pagina di raccolta credenziali che imitava le schermate di accesso a Microsoft Exchange e Outlook Web Access. [10]

Le credenziali raccolte vengono dirottate verso archivi esterni usa-e-getta di tipo foglio di calcolo in cloud, con notifiche automatiche a ogni nuova cattura. Quest'ultimo dettaglio è esso stesso una contromisura: il traffico in uscita verso servizi di produttività cloud apparentemente innocui può costituire un anello della catena di esfiltrazione e monitorare quelle destinazioni è un punto di osservazione concreto per chi difende.

7. Implicazioni legali

Ogni passaggio della catena ovvero copiare una pagina, raggiungere il dispositivo che la espone, raccogliere ciò che un utente vi immette, interseca uno o più ordinamenti. I due paragrafi seguenti non offrono un parere legale, ma orientano su due profili tra i più rilevanti: la tutela di marchio e diritto d'autore sugli elementi riprodotti e la protezione dei dati personali raccolti.

7.1 Marchio e diritto d'autore sugli asset reali

Replicare una pagina esistente coinvolge elementi protetti: il marchio del produttore (logo, denominazione, segni distintivi) e opere tutelate dal diritto d'autore (grafica, layout, codice, risorse multimediali). L'asset freezing (prelievo e incorporazione degli asset originali) è, sul piano giuridico, una riproduzione di materiale altrui. L'uso di un segno distintivo per trarre in inganno sull'origine configura contraffazione e spesso concorrenza sleale; nell'UE il marchio è armonizzato e in Italia trova riferimento nel Codice della proprietà industriale, mentre il diritto d'autore è regolato dalla legge speciale e dalle direttive europee. È precisamente per evitare questa esposizione che gli esempi del documento sono pagine fake: la differenza tra studiare il meccanismo su una pagina finta e replicare una pagina reale è la differenza tra un materiale didattico difendibile e una riproduzione potenzialmente illecita.

7.2 Trattamento dei dati personali

Quando una pagina raccoglie ciò che un utente vi immette, credenziali e altri dati riferibili a una persona sono dati personali, e il loro trattamento ricade nel GDPR per i soggetti nell'UE. La raccolta tramite pagina di phishing è per definizione un trattamento privo di base giuridica, in violazione dei principi di liceità, correttezza e trasparenza: illecito che si cumula alle altre responsabilità.

La questione non riguarda solo lo scenario offensivo ma anche un ambiente di test che, per disattenzione, raccogliesse dati reali incorrerebbe nelle stesse norme.

8. Come riconoscere una pagina clonata

Poiché i segnali statici si indeboliscono, il riconoscimento non può affidarsi a un singolo indizio decisivo ma alla convergenza di più segnali deboli e, soprattutto, all'osservazione del comportamento. Due prospettive complementari: l'utente finale e il difensore/SOC.

8.1 L'utente finale

L'utente non analizza il traffico, ma ha punti di osservazione che reggono anche di fronte a cloni fedeli, perché non riguardano l'aspetto, ormai perfetto, ma il contesto e il comportamento. Il primo è l'**indirizzo**: la verosimiglianza dell'URL è parte dell'inganno (domini somiglianti, sottodomini che inducono fiducia, indirizzi cloud rapidi) e l'abitudine a leggere l'intero dominio è la difesa individuale più robusta; la connessione cifrata non è più, da sola, segno di legittimità, ma la sua assenza o un avviso del browser restano un campanello d'allarme. Il secondo è **comportamentale** e difficile da mascherare: molti kit, dopo aver catturato i dati, reindirizzano al sito autentico, così il primo login «fallisce» e il secondo riesce. Questo è uno dei segnali più affidabili di interposizione, da insegnare nelle attività di sensibilizzazione. Restano le **micro-incongruenze** (funzioni secondarie anomale, collegamenti che non portano da nessuna parte): un indizio utile quando presente, ma il più fragile, perché i sistemi di generazione sono progettati per «tappare i buchi» in modo plausibile e non si può contare sul fatto che il clone sbagli i dettagli.

8.2 Il difensore e il SOC

Chi difende un'organizzazione o un marchio può spostare il riconoscimento dal singolo episodio all'infrastruttura, ma deve tener conto di come la generazione assistita cambia il valore dei segnali disponibili.

Alcuni metodi nascono per la clonazione tradizionale e la generazione tramite AI li indebolisce. La rilevazione della clonazione mentre avviene (di cui esiste documentazione, tra cui un brevetto statunitense, US 10523706 [9]) sfrutta il fatto che chi clona spesso preleva solo parte degli asset e serve i restanti dalla fonte autentica: osservando dal lato server il pattern di richieste, le intestazioni e gli IP/ASN si può riconoscere che a prelevare le risorse è uno strumento di cloning, e reagire. È un approccio efficace contro il mirroring, ma la sua premessa viene meno proprio nelle pipeline più evolute: l'asset freezing sopprime il segnale degli asset misti, e una pagina i cui contenuti sono generati a destinazione non preleva quasi nulla da osservare.

Altri segnali, invece, reggono. Il monitoraggio del marchio permette di controllare i domini somiglianti e i certificati emessi per i nomi che imitano quelli legittimi; poiché si guarda a questi elementi e non al codice, funziona anche quando la pagina è generata da un modello. Lo stesso vale per i canary token: elementi-esca che, se prelevati e riutilizzati in una copia, generano un segnale che rivela la clonazione. È una traccia seminata in anticipo, indipendente da come il clone viene assemblato, e per questo non risente del passaggio alla generazione. Resta il versante più specifico della clonazione generativa, dove il segnale si sposta dal codice al suo comportamento in esecuzione. Quando la pagina assembla i propri contenuti a runtime e interroga un servizio esterno al momento dell'apertura, l'artefatto da intercettare non è più un file statico, ma un'attività osservabile solo mentre accade: per questo l'analisi comportamentale nel punto di esecuzione, all'interno del browser, è indicata come la difesa più efficace contro questa classe di tecniche [5].

Cambia anche il traffico da sorvegliare: se il codice arriva da un'API di un modello, la sua provenienza è un dominio legittimo e di norma consentito, e diventa un punto di osservazione concreto la presenza di traffico verso servizi di AI generativa non autorizzati dall'organizzazione. Il filo comune resta quello dell'intera sezione: poiché i segnali statici si indeboliscono, la rilevazione si gioca sul comportamento: come una pagina tratta i dati, il pattern con cui un'infrastruttura preleva risorse, la destinazione verso cui i dati vengono diretti.

9. Raccomandazioni e contromisure

Di seguito si riportano le raccomandazioni organizzate per pubblico.

9.1 Vendor di dispositivi

I produttori detengono la leva più a monte: gran parte della minaccia presuppone una pagina raggiungibile e copiabile. La raccomandazione di fondo è **non esporre le** pagine (in particolare quelle relative agli amministratori di sistema) **sulla rete pubblica per impostazione predefinita**: un dispositivo che, appena installato, rende la propria pagina di login raggiungibile da chiunque offre insieme il bersaglio da clonare e la porta da varcare. A corredo, le buone pratiche di irrobustimento fra cui niente credenziali predefinite o obbligo di cambiarle alla prima accensione, aggiornamenti del firmware affidabili, riduzione delle funzioni accessibili dall'esterno, riducono in un colpo la disponibilità del materiale da copiare e la possibilità di accesso abusivo.

9.2 Aziende e CISO

Le organizzazioni possono agire su due fronti: rendere l'inganno meno probabile e renderlo inefficace quando riesce. Sul primo, la **sensibilizzazione è orientata al comportamento**: più che insegnare a riconoscere le pagine fatte male, conviene insegnare a leggere l'intero dominio e a diffidare del comportamento anomalo. Sul secondo, la misura più efficace in assoluto: l'**autenticazione a più fattori resistente al phishing**. Se l'autenticazione si fonda su un fattore legato crittograficamente al dominio legittimo e non su un codice digitabile anche su una pagina ingannevole, le credenziali sottratte perdono gran parte del valore. A corredo, il monitoraggio del marchio e dei domini somiglianti (con eventuale ricorso ai canary token) individua l'infrastruttura ingannevole prima che produca danni su larga scala.

9.3 Chi costruisce sopra gli LLM

È il pubblico che più di ogni altro può immettere o evitare di immettere queste capacità nel mondo. Una raccomandazione sui **guardrail**: i filtri dei modelli hanno effetto reale ma sono aggirabili tramite riformulazione; quindi, non ci si deve affidare al solo filtro a monte: i controlli vanno collocati lato server, dove non sono elusi da una variazione del wording inviata dal client, e progettati attorno all'intento e al comportamento dell'applicazione.

10. Conclusioni

La produzione di una pagina ingannevole è passata dalla copia di un originale accessibile e statico alla sua generazione ex novo, variabile e svincolata dall'esistenza di un originale. I tre paradigmi tecnici ovvero ricostruzione del codice, overlay fotografico, generazione dinamica, lavorano per sopprimere i segnali su cui la difesa si è storicamente appoggiata; la sperimentazione lo ha mostrato sul terreno meno presidiato dei dispositivi esposti, e i casi documentati ne confermano la realtà.

Emergono tre prospettive. Il fenomeno è un cambiamento strutturale dell'economia dell'attacco, non un episodio passeggero: tecniche marginali ma onerose diventano sistematiche quando diventano economiche. A ogni capacità offensiva corrisponde quasi sempre una traccia difensiva: la dipendenza degli strumenti da un servizio generativo li rende fragili e cacciabili; le piattaforme sono insieme punto di abuso e punto di intervento; i guardrail, pur aggirabili, restano una leva quando sono robusti. E le contromisure più efficaci sono quelle che rendono l'inganno inutile a prescindere dalla riuscita, l'autenticazione resistente al phishing su tutte, e quelle che agiscono a monte, dai produttori che non espongono le interfacce delle pagine da clonare a chi costruisce con disciplina nei controlli e nei segreti.

Il messaggio di fondo si condensa in una formula: **la difesa si sposta dal riconoscere il codice al riconoscere il comportamento.** L'identità di una pagina fatta da codice, firma, somiglianza grafica non è più un ancoraggio affidabile, perché è diventata economica, variabile e perfetta nell'aspetto. Ciò che resta osservabile, e su cui conviene investire, è il comportamento: come una pagina tratta i dati e dove li dirige, il pattern con cui un'infrastruttura preleva risorse, l'intento che un'applicazione manifesta. È su questo terreno che si gioca, oggi, il riconoscimento delle minacce generate con l'assistenza dei modelli.

Appendice A

[1] Mirroring manuale di siti web. Strumenti di copia e consultazione offline (HTTrack) e utilità di download ricorsivo per replicare in locale la struttura di un sito. Dev.to (apr. 2026): https://dev.to/james_smith_543/how-phishing-websites-trick-users-and-how-to-detect-them-2h6

[2] Website cloning e credential harvesting con SET. Materiale divulgativo sul Social Engineering Toolkit e sulla clonazione di pagine di accesso con modulo di raccolta. Medium: https://medium.com/@chamo.wijetunga/how-to-clone-a-website-and-use-it-for-credential-harvesting-outside-the-network-e2057ffa1081

[3] HTTrack e SET — mirroring più harvesting. Mirroring con HTTrack abbinato al Social Engineering Toolkit. Medium (Purple Team): https://medium.com/purple-team/replicate-any-websites-with-httrack-and-harvest-credentials-using-social-engineering-toolkit-6107f6a0149f

[4] La formula dei «dieci minuti». Origine divulgativa della formula sul tempo di allestimento di una pagina clonata con harvester. Native Algorithms (Medium): https://nativealgorithms.medium.com/phishing-clone-login-pages-stolen-credentials-aac30f0ae872

[5] Unit 42 / LogoKit — generazione live per-request. Proof-of-concept che sostituisce la componente statica di un kit di phishing con codice generato da un modello all'apertura della pagina; la pagina trasmessa resta «pulita» e il codice muta a ogni richiesta. Riporta anche l'osservazione sul bypass dei guardrail tramite riformulazione neutra della richiesta. Sintesi: TechRound (gen. 2026): https://techround.co.uk/artificial-intelligence/cyber-criminals-llm-phishing-attacks-heres-how/ — Fonte primaria (Unit 42): https://unit42.paloaltonetworks.com/real-time-malicious-javascript-through-llms/

[6] KnowBe4 (Egress) — Phishing Threat Trends Report, Vol. 5 (mar. 2025). Rilevazione secondo cui l'82,6% delle email di phishing analizzate tra il 15 settembre 2024 e il 14 febbraio 2025 mostra un qualche impiego di strumenti di AI; dato prodotto da Egress (società del gruppo KnowBe4) sui dati della propria soluzione di sicurezza della posta. Comunicato: https://www.knowbe4.com/press/new-knowbe4-report-reveals-a-spike-in-ransomware-payloads-and-ai-powered-polymorphic-phishing-campaigns — Report (PDF): https://www.knowbe4.com/hubfs/Phishing-Threat-Trends-2025_Report.pdf

[7] SoK: Exposing the Generation and Detection Gaps in LLM-Generated Phishing (Chen, Wu, Nguyen, Rudolph). Systematization of Knowledge che raccoglie, fra l'altro, la valutazione di metà 2025 su tassi di interazione dello spear phishing generato da LLM superiori al 30%. arXiv: https://arxiv.org/pdf/2508.21457

[8] PhishOracle — pagine di phishing avversariali. Studio sperimentale: pagine di phishing avversariali restano non rilevate dai servizi di analisi nelle prime 24 ore e sono giudicate legittime da circa la metà dei valutatori umani. ACM Digital Threats: https://dl.acm.org/doi/10.1145/3737295

[9] Brevetto USPTO 10523706 — phishing protection using cloning detection. Rilevamento del clonaggio basato sull'osservazione che gli attaccanti spesso copiano solo parte degli asset e servono i restanti dalla fonte autentica; riconoscimento lato server tramite pattern di richieste, intestazioni e IP/ASN, con possibili reazioni. USPTO: https://image-ppubs.uspto.gov/dirsearch-public/print/downloadPdf/10523706

[10] Cisco Talos — IR Trends Q1 2026 (campagna Softr). Rapporto sugli interventi di incident response del primo trimestre 2026: prima volta in cui Talos documenta l'uso di uno specifico strumento di AI da parte di un avversario in una campagna di phishing — una pagina di raccolta credenziali generata «senza codice» su piattaforma di sviluppo web assistito, contro la pubblica amministrazione. Cisco Talos (apr. 2026): https://blog.talosintelligence.com/ir-trends-q1-2026/

t-defence

Next | Donexit | Foramil | Innodesi

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

#TDefenceBusiness