Deepfake: le nuove frontiere dei Cyber Attack che ci rubano l'identità

Nicolò Giuli
Dottore in Ingegneria gestionale, Manager Consulting Technology Cybersecurity & Privacy di PwC Italy

1. Il ruolo dell'Intelligenza Artificiale nei nuovi Cyber Attack

L'Intelligenza Artificiale (di seguito, per brevità, anche "IA") consente di programmare e realizzare *software* e macchine dotate di caratteristiche tipicamente umane, quali capacità decisionale, abilità di calcolo e apprendimento continuo dall'analisi ed elaborazione dei dati.

La possibilità di ricondurre l'intelligenza umana a particolari comportamenti, riproducibili da macchine, rappresenta uno strumento estremamente potente utilizzabile in molteplici settori del mondo reale.

Non sempre, però, i fini ultimi delle applicazioni dell'intelligenza artificiale si rivelano meritevoli.

Lo sanno bene gli studiosi e gli esperti dell'Europol che, in collaborazione con l'Istituto internazionale della Nazioni Unite per la ricerca sul crimine e la giustizia, hanno realizzato il report "*Malicious Uses and Abuses of Artificial Intelligence*"[1], in cui si analizzano le potenziali applicazioni future

[1] "*Malicious Uses and Abuses of Artificial Intelligence*" a cura di Trend Micro Research, United Nations Interregional Crime and Justice Research Institute (UNICRI) e Europol's European Cybercrime Centre (EC3), disponibile al seguente link: *http://bit.ly/35iTIXz*.

dell'intelligenza artificiale in ambito *cybercrime*. I criminali, infatti, possono utilizzare e sfruttare l'IA per migliorare i propri attacchi, riducendo le tempistiche di realizzazione, colpendo nuove vittime e introducendo elementi innovativi nelle tecniche di attacco.

Pertanto, senza dimenticare i notevoli benefici che l'IA può apportare in termini di automazione, efficienza e autonomia, non si possono nascondere all'opinione pubblica i rischi e le minacce connessi all'utilizzo di queste tecnologie, sfruttate dai *cyber* criminali sia come vettore che come superficie di attacco.

Vediamo le principali applicazioni dell'intelligenza artificiale in relazioni ai crimini informatici:

- sviluppare *malware* difficilmente rilevabili; per esempio, l'IA può essere utilizzata per generare messaggi di posta elettronica con un elevato livello di qualità semantica in grado di confondere e aggirare i filtri *antispam;*
- effettuare attacchi di ingegneria sociale su larga scala; l'IA consentirebbe all'attaccante di automatizzare e accelerare la fase di studio delle vittime, anticipando le risposte più comuni dei potenziali target e concentrandosi solo sui soggetti più facili da ingannare;
- aggirare sistemi di sicurezza basati sul riconoscimento biometrico o sulla rilevazione di comportamenti anomali da parte degli utenti; mediante l'IA, il criminale è in grado di riprodurre comportamenti in linea con gli standard in termini, a titolo esemplificativo, di orario di utilizzo di un servizio e di indirizzi IP in intervalli credibili;
- affinare gli algoritmi volti a compromettere le credenziali di accesso a servizi online. In particolare, ricorrere all'IA consente di analizzare un ampio dataset di password e generare variazioni che rispettano la distribuzione statistica; questo consente di generare password più mirate ed efficaci;
- violare i sistemi di sicurezza CAPTCHA[2]; l'applicazione dell'IA consente di velocizzare vertiginosamente il processo di risoluzione dei CAPTCHA, facendo leva sulla potenza computazionale nel riconoscere le varietà di forme e di colore che caratterizzano tali sistemi di sicurezza.

[2] I CAPTCHA – acronimo di "*Completely Automated Public Turing test to tell Computers and Humans Apart*" – sono dei sistemi che permettono di accertare la natura umana dell'utente che tenta di compiere un'operazione su una pagina web, quali la richiesta di fruizione di un servizio, la trasmissione di un messaggio tramite form, etc., al fine di evitare l'utilizzo di detto servizio da parte di un bot, ovvero di software in grado di eseguire attività in forma automatizzate.

In tale contesto, tra le principali minacce informatiche che incombono sulle organizzazioni pubbliche e private svettano da qualche tempo i *deepfake,* parola composta che richiama il "*deep learning*", ovvero tecniche avanzate di intelligenza artificiale basate su algoritmi di *machine learning* per estrarre informazioni da dati non strutturati, quali movimenti e caratteristiche del corpo umano. Tali dati sono poi manipolati per creare contenuti multimediali fasulli (da qui il termine inglese "*fake*"), quali video e audio, con l'intento di rappresentare falsamente la realtà e i suoi protagonisti.

2. Deepfake: principali tecniche e perché devono preoccuparci

Il potenziale connesso a questa tecnologia di minaccia informatica è rappresentato dalla possibilità di creare contenuti multimediali falsi, difficilmente distinguibili dalla realtà, in grado di sfruttare velocità e portata della rete internet, dei *social network* e delle app di messaggistica, per raggiungere rapidamente milioni di utenti in tutto il mondo.
Tali caratteristiche fanno sì che la tecnologia *deepfake* si presti efficacemente a molteplici scopi malevoli e criminali, quali:

- rovinare reputazione e immagine di un individuo[3];
- diffondere informazioni false al fine di plasmare l'opinione pubblica;
- sostenere la narrazione di gruppi terroristici e/o associazioni criminali;
- incitare all'odio sociale e fomentare disordini a fini sovversivi;
- creare panico e confusione di massa;
- falsificare o manipolare elementi di prova in ambito giudiziario;
- perpetrare estorsioni e illeciti.

Come vedremo, alla base dei *deepfake* c'è quasi sempre un'applicazione innovativa delle tecnologie di *machine learning*.
Una potenziale applicazione è rappresentata dalla rete generativa avversa (in inglese *Generativa Adversarial Network-GAN)*[4] in cui due reti neurali

[3] Rovinare la reputazione di un personaggio pubblico diventa un obiettivo facilmente perseguibile quando si può disporre di contenuti audiovisivi estremamente realistici; in tale contesto, non è difficile prevedere che nel prossimo futuro gli *hacker* ricorreranno sempre più a questa tecnologia per attaccare le organizzazioni pubbliche e private e ricavarne vantaggi non solo in termini economici.

[4] Per rete generativa avversa si intende una classe di metodi in cui due reti neurali, ovvero dei modelli computazionali composti da "neuroni" artificiali, sono addestrate in maniera competitiva in una serie di giochi di tipo minimax – metodo che, nella teoria dei giochi, è finalizzato a minimizzare la massima perdita possibile – per apprendere come generare un nuovo set di dati a partire da un dataset con la medesima distribuzione statistica.

svolgono i seguenti compiti: la prima rete, con funzione generativa, crea nuovi dati con le stesse caratteristiche di quelli utilizzati in fase di addestramento; la seconda rete, con funzione discriminante, giudica questi nuovi dati in termini di correttezza e valuta se il singolo dato è stato prodotto dal generatore o se è un dato reale.
La logica alla base della GAN trova la sua naturale applicazione con riferimento ai *deepfake*. Prendiamo a titolo esemplificativo una GAN addestrata a generare e analizzare nuovi set di dati in diversi formati multimediali (e.g. foto, video o file audio). Il generatore produce nuovi file multimediali, mentre il discriminatore li riceve e cerca di identificarli. L'addestramento della GAN mira a rendere il generatore sempre più abile nell'ingannare la rete discriminante (generando risultati difficili da distinguere dai dati reali) e la rete discriminante sempre più capace di rilevare le immagini generate artificiosamente.
In linea con i principi base delle GAN, i *deepfake* sono creati partendo da un dato reale che, adeguatamente manipolato, serve a generare un contenuto audiovisivo difficilmente rilevabile come falso. Le tecnologie utilizzate per creare contenuti audiovisivi *deepfake* dipendono dai tipi di falsificazione digitale, classificati come di seguito:

- *face-swap*, in cui si sostituisce il viso di un individuo sovrapponendolo su quello di un'altra persona;
- *face re-enactment*, in cui si manomettono i lineamenti e i connotati dell'individuo target per fare in modo che il volto dica qualcosa che si intende far pronunciare;
- *face generation*, in cui si crea un viso del tutto fittizio;
- *speech synthesis*, in cui si generano contenuti audio addestrando algoritmi per creare voci fittizie;
- *shallowfakes,* in cui si creano contenuti audiovisivi falsi tramite tecniche meno evolute di montaggio.

3. Evoluzione del fenomeno e scenari futuri

Marzo 2019. L'amministratore delegato di un'azienda energetica britannica riceve la telefonata del CEO del Gruppo di cui fa parte che gli chiede di provvedere quanto prima a disporre un trasferimento di denaro a favore di un fornitore ungherese. Come si scoprirà poco dopo, si tratta di una truffa basata proprio su un *deepake* vocale[5]. La vittima della truffa si sarebbe

[5] Per approfondire il caso di cronaca si veda "Thieves" are now using AI deepfakes to trick companies into sending them money" disponibile al link *http://bit.ly/3g09XLa*.

giustificata garantendo di aver riconosciuto il tono, il timbro e la pronuncia tipica tedesca del proprio capo. Falso talmente convincente da indurre la vittima ad agire con urgenza, senza nemmeno controllare l'attendibilità della richiesta pervenuta.

In realtà, video e audio *fake* iniziarono ad arrivare alla ribalta già a partire dal 2017 quando un utente del social network Reddit cominciò a pubblicare spezzoni di film rivisitati in chiave divertente e realistica. Ben presto, però, da innocui spezzoni di film si passò alla produzione di contenuti a sfondo pornografico con protagonisti celebrità del mondo dello spettacolo.

Il fenomeno, in precedenza appannaggio esclusivo di professionisti dotati di strumenti grafici all'avanguardia, è arrivato alla portata di tutti con il lancio nel 2018 di *Fakeapp,* software in grado di creare video e immagini *fake* a partire dagli originali contenuti nella galleria multimediale del dispositivo in cui è installata l'app. Quest'ultima è progettata per comparare diversi elementi come lineamenti, colore degli occhi e della pelle al fine di adattare al meglio il viso dell'ignara vittima su ogni singolo *frame* del video finale contraffatto.

Un altro caso interessante di *deepfake* ha coinvolto nel marzo 2019 la società Tesla: improvvisamente un account intestato a "Maisy Kinsley", fantomatica giornalista di Bloomberg, inviò richieste di collegamento a circa 200 *shortseller* di Tesla. La foto di Maisy Kinsley era stata generata proprio da una GAN secondo le procedure in precedenza descritte. Bloomberg successivamente confermò di non annoverare fra le proprie fila nessuna giornalista di nome Maisy Kinsley e i social coinvolti procedettero con la chiusura degli account. L'obiettivo era quello di contattare le ignare vittime al fine di esfiltrare dati e informazioni sensibili.

Da ultimo, i *deepfake* rappresentano un pericoloso strumento in mano a terroristi o lupi solitari che mirano a mettere a repentaglio equilibri geo-politici e la stabilità dei mercati finanziari. Ne è un caso emblematico la *spy story* di matrice russa che vede come protagonisti due funzionari del governo britannico e lettone: costoro, nel tentativo di mettersi in contatto con Aleksej Navalny, politico russo e noto oppositore del governo Putin, riescono a fissare un incontro con Leonid Volkov, leader del principale partito di opposizione russo, in cui confrontarsi su tematiche di politica interna alla Russia e internazionale, quali il conflitto con l'Ucraina per il controllo della Crimea.

In realtà, il presunto Volkov è soltanto un attore in grado di impersonificare il politico russo mediante l'utilizzo delle tecnologie *deepfake*. In questo caso, la scoperta da parte dei funzionari britannico e lettone è avvenuta soltanto

dopo aver compiuto un'importante *disclosure* di informazioni riservate[6]. Dagli esempi sopra citati emerge la facilità del processo di creazione di falsi audio e video. Come detto in precedenza, le reti avversarie generative (GAN) competono l'una contro l'altra per migliorarsi sistematicamente di un fattore esponenziale. Maggiore è il numero di informazioni e dati in mano ai criminali (in termini di video, immagini, audio), maggiore è la qualità e la verosimiglianza del prodotto finale.

Nel caso della tecnica di *face-swap,* in cui il prodotto finale è ottenuto dalla sostituzione del volto dell'individuo target (soggetto A) con quello di un'altra persona (soggetto B), ai criminali è necessaria una base dati di centinaia o migliaia di immagini di entrambi i soggetti, non così difficilmente reperibile online, mediante il semplice accesso ai più comuni social network.

Successivamente le immagini di entrambi i soggetti sono codificate al fine di individuare le peculiarità dei volti, ovvero lineamenti, singole espressioni e angolazioni. Da ultimo, le procedure di addestramento della rete neurale mediante specifici algoritmi consentono di passare alla sovrapposizione del volto del soggetto A con quello del soggetto B e ottenere il video contraffatto.

In tale contesto, è facile comprendere perché, mediante le nuove classi di algoritmi di intelligenza artificiale in grado di garantire risultati accurati, estremamente realistici e perfino di interagire in tempo reale, i *deepfake* abbiano finito per rientrare a pieno titolo nelle classifiche delle principali minacce informatiche più pericolose e per le aziende e per i singoli cittadini[7].

Il Voice Intelligence Report[8], pubblicato nel 2019 dalla società Pindrop[9], fornisce uno spaccato di estremo interesse circa i numeri del fenomeno. Con riferimento alla minaccia rappresentata dai *deepfake* vocali, i settori economici più a rischio sono risultati essere quello assicurativo (1 su 7.500 chiamate fraudolente), del commercio al dettaglio (1 su 325 chiamate fraudolente), il settore bancario (1 su 755 chiamate fraudolente), il mercato delle società emittenti carte di pagamento (1 su 740 chiamate fraudolente), il settore del brokeraggio (1 su 1.742 chiamate fraudolente) e quello delle cooperative di credito (1 su 1.339 chiamate fraudolente).

[6] Per approfondire il caso di cronaca si veda "European MPs targeted by deepfake video calls imitating Russian opposition" disponibile al link *http://bit.ly/3fWe62H*.

[7] Si veda, a tal proposito, il report 2020 del Clusit sulla sicurezza ICT in Italia che rileva la preponderanza degli impatti dei deepfake sulla sicurezza delle organizzazioni economiche. Disponibile al seguente link *http://bit.ly/35j1AIx*.

[8] Report annuale pubblicato dalla Società Pindrop volto ad analizzare i rischi di sicurezza e gli impatti economici legati all'utilizzo di tecnologie vocali sull'economia conversazionale globale.

[9] Pindrop è una società americana operante in ambito information security. Essa è prevalentemente specializzata nello sviluppo di soluzioni di risk scoring e fraud detection da applicare alle chiamate telefoniche.

Pindrop evidenzia "come gli attacchi vocali sintetici diventeranno presto la prossima forma di violazione dei dati"; la Società è convinta che nel prossimo futuro i truffatori chiameranno i call center servendosi di voci sintetiche per testare se le aziende hanno soluzioni e competenze per rilevare i *deepfake* vocali.
Ulteriori e significativi numeri derivano dal rapporto "*The State of Deepfakes – Landscapes, Threats and Impact*" realizzato e pubblicato dalla società Deeptrace[10]: nel 2019 il numero di contenuti *deepfake* online è passato da 14.678 a 145.227, una crescita del 900% rispetto all'anno precedente.
Come rileva il rapporto CLUSIT 2020, i *deepfake* costituiscono un "pericolo grave, concreto e attuale", che fa leva proprio sull'irrazionalità dell'essere umano che, come sostenuto da Walter Quattrociocchi, docente a contratto presso il "*Center of Data Science and Complexity for Society*" dell'Università La Sapienza di Roma, "ha una visione del mondo che è emotiva e percettiva"[11].
I *deepfake* sfruttano a pieno questa vulnerabilità dell'essere umano, potendo contare su strumenti messi a disposizione dalla rete internet, quali viralità e audience potenzialmente illimitata. Inoltre, dato che il buon esito di un attacco mediante tecnica *deepfake* dipende prevalentemente dall'errore e dalla fiducia umana, rappresenta una delle versioni più mature ed evolute dell'ingegneria sociale.

4. Rischi dei deepfake e come difendersi

Versatilità e caratteristiche dei *deepfake* ci portano a concludere che essi rappresentano una minaccia sia per gli individui singoli, sia per le organizzazioni di vario tipo.
Nel caso, infatti, di persone fisiche, un contenuto *deepfake* potrebbe, ad esempio, costituire uno strumento estremamente pericoloso ed efficace per perpetrare un furto d'identità: in tal caso, l'obiettivo dei cyber criminali è quello di sottrarre l'identità della vittima e ricrearla in luoghi, situazioni o contesti manipolati così da comportarne la perdita non solo dell'immagine, ma anche dei pensieri e delle idee.
Nel caso delle organizzazioni (pubbliche e private), i cyber criminali hanno ormai compreso le potenzialità che queste tecnologie mettono a disposizione

[10] Deeptrace è una società olandese che fornisce soluzioni e tecnologie di cybersecurity per la rilevazione e il monitoraggio dei contenuti multimediali sintetici al fine di proteggere individui e aziende dagli utilizzi malevoli dell'intelligenza artificiale.
[11] Per approfondire la posizione del Prof. W. Quattrociocchi sul tema deepfake e, in particolare, le fake news, si veda l'articolo di G.Bona, W. Quattrociocchi, "*Come ti prevedo le fake news*, in *Scienza in Rete*, *http://bit.ly/3H5pJQT*.

al fine di perfezionare azioni di ingegneria sociale e ricavarne un beneficio economico diretto, estorcere informazioni per manipolare i mercati finanziari o, in caso di gruppi terroristici, manipolare l'opinione pubblica e/o danneggiare la reputazione di un esponente dell'organizzazione stessa o di interi organismi, enti, governi e istituzioni.

In ogni caso, il danno reputazionale conseguente a un attacco mediante tecnologie *deepfake* può essere enorme: a titolo esemplificativo, basti pensare alle potenziali perdite borsistiche o di quote di mercato causata dalla diffusione online di false campagne pubblicitarie di brand commerciali contenenti messaggi diffamatori circa prodotti e servizi.

Alla luce del contesto finora descritto, urge definire e implementare efficaci strategie di prevenzione e contrasto alle tecnologie *deepfake* in termini sia di raccomandazioni e prescrizioni a beneficio degli utenti privati, sia di programmi complessi e maturi da adottare nel contesto delle aziende private e delle amministrazioni pubbliche.

Con riferimento alle modalità di protezione per le persone fisiche, è estremamente utile la scheda informativa predisposta dal Garante per la protezione dei dati personali a fine 2020[12]; con tale iniziativa, l'Autorità ha voluto fornire un set di raccomandazioni semplici e di facile implementazione per gli utenti meno esperti della rete internet.

In particolare, in ottica preventiva, prima di procedere con la pubblicazione di un qualsivoglia contenuto multimediale raffigurante sé e i propri cari, risulta necessario interrogarsi circa i rischi associati alla diffusione indiscriminata, limitando, se necessario, l'audience che potrà visualizzare detto contenuto.

Ragionando, invece, in ottica di controlli *ex-post*, nel caso in cui si nutrano dei dubbi circa l'autenticità di un contenuto, diviene fondamentale astenersi da ulteriori condivisioni e divulgazioni, nonché segnalarlo prontamente al gestore del sito web/social network o, in funzione del contenuto, alle Autorità competenti (*e.g.* Polizia Postale, Garante per la protezione dei dati personali, Garante per l'infanzia, etc.).

Agli utenti maggiormente esperti, il Garante fornisce alcuni consigli utili su come provare a riconoscere un *deepfake*: riscontrare anomalie quali malformazioni al volto, immagini particolarmente sgranate o sfocate, movimenti innaturali degli occhi, luci e ombre anormali sul viso dei protagonisti rappresentano tutti campanelli di allarme che devono spingere a verificare su fonti attendibili la veridicità del contenuto visualizzato.

[12] Il Garante per la protezione dei dati personali ha realizzato e pubblicato una scheda informativa per sensibilizzare gli utenti sui rischi connessi agli usi malevoli delle tecnologie deepfake. La scheda è disponibile al link *http://bit.ly/34erv3S*..

Senza sottostimare l'importanza del fattore umano, è evidente come affidarsi all'attenzione e alle capacità del singolo individuo di riconoscere un contenuto *fake* garantisca un tasso di successo non sempre sufficiente. Pertanto, le grandi aziende ICT operanti nel mondo del *digital* hanno deciso di combattere la minaccia rappresentata dai *deepfake* con lo stesso mezzo che consente a tale tecnologia di proliferare e divenire sempre più sofisticata: l'intelligenza artificiale.

Le tecniche più promettenti per combattere i *deepfake* si basano infatti proprio sull'IA, tramite cui distinguere video, immagini e audio veri da contenuti falsi. Addestrare reti neurali al riconoscimento di contenuti manipolati si sta finora rivelando uno strumento particolarmente efficace, consentendo di ottenere risultati ottimali, di gran lunga superiori rispetto alle capacità di discriminazione degli esseri umani.

In tale contesto, si inserisce a pieno titolo la definizione e l'implementazione di processi automatizzati che fanno leva su software di *Presentation Attack Detection* (PAD): si tratta di programmi che sfruttano l'intelligenza artificiale e l'apprendimento automatico per verificare se le immagini si riferiscono a un essere umano reale o se si tratta di un contenuto contraffatto.

I metodi PAD più semplici sono quelli che prevedono controlli attivi, in quanto richiedono l'azione dell'utente. Si tratta generalmente di richieste di movimenti casuali, come posizionare o spostare un oggetto o eseguire a comando movimenti delle parti del viso. L'obiettivo è verificare che il video non sia stato pre-registrato o non sia opera di un software interattivo per realizzare *real-time deepfake*. L'efficacia di questi metodi basici è accettabile solo in caso di attaccanti dotati di strumentazioni ancora rudimentali: essere in grado di simulare azioni e movimenti in tempo reale è possibile solo se si può disporre di una potenza computazionale significativa.

I metodi più avanzati, invece, sono basati tipicamente su controlli passivi, in quanto non presuppongono alcuna azione da parte dell'utente. Essi sono finalizzati a valutare caratteristiche ed espressioni involontarie come micro-movimenti del volto, degli occhi (*e.g.* dilatazione della pupilla), micro-variazioni dell'intensità del colore della pelle e dell'iride. Tali controlli sono in grado di rilevare un battito di ciglia innaturale o mancante, occhi o denti mal generati, dilatazione anomala della pupilla e altre variabili. I colori casuali possono essere visualizzati sullo schermo per analizzare i riflessi della luce sul soggetto e l'ambiente circostante.

Il mito dell'innovazione a tutti costi, dunque, ci impone di introdurre nuove tecnologie digitali, senza sosta, nella nostra vita quotidiana. Ciò, però, comporta inevitabilmente il configurarsi di nuove minacce sempre più sofisticate e complesse.

Il riconoscimento facciale, a titolo esemplificativo, può sicuramente comportare una facilitazione nella fruizione di servizi online.
Ma ci interroghiamo a sufficienza sui rischi connessi alla possibilità, per un terzo non autorizzato, di ricreare perfettamente il nostro volto e/o la nostra voce?
Il futuro delle minacce informatiche si fa, dunque, sempre più complesso, imprevedibile, misterioso.
Interroghiamoci oggi sull'utilizzo che vogliamo fare delle nuove tecnologie, definendo fin d'ora complessi programmi di mitigazione dei rischi informatici; in caso contrario, perderemo il controllo delle nuove tecnologie. E con loro, delle nostre identità.