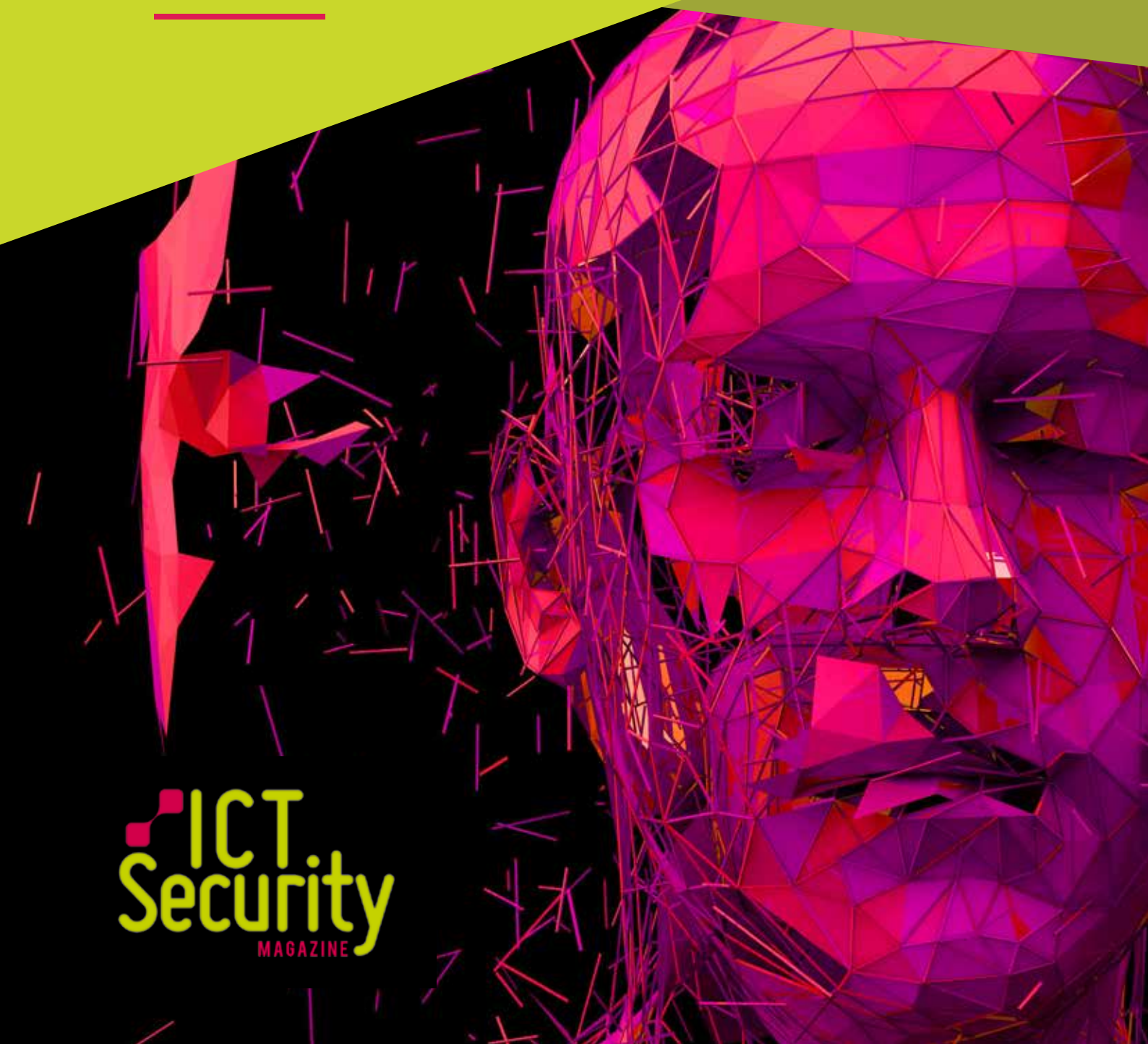
DEEPFAKE & CYBER INTELLIGENCE

Tecniche di creazione, rilevamento e prevenzione

Francesco Arruzzoli, Sr. Cyber Security Threat Intelligence Analyst

WHITEPAPER 06/2022



ICT Security MAGAZINE

SOMMARIO

ICT SECURITY MAGAZINE

1° rivista italiana di sicurezza informatica, attiva da oltre 20 anni, dedicata in forma esclusiva alla cyber security e alla business continuity, si pone l'obiettivo di coinvolgere i più importanti attori del settore, aziende e istituzioni pubbliche, per la diffusione degli elementi conoscitivi legati a tutti gli aspetti della information security.

ABOUT THE AUTHOR

Francesco Arruzzoli

Sr. Cyber Security Threat Intelligence Analyst

Con oltre 30 anni di esperienza nell'ambito della sicurezza delle informazioni Francesco Arruzzoli è Sr. Cyber Security Threat Intelligence Analyst presso la Winitalia di cui è cofondatore. Responsabile del Centro Studi Cyber Defense Cerbeyra presso il polo di cyber security del Gruppo Vianova, coordina le attività di R&D, analisi delle cyber minacce e progettazione di nuove soluzioni per la cyber security di aziende ed enti governativi. Progettista di sistemi esperti, software developer, network e system engineer, è stato tra i primi ethical hacker italiani certificati. Autore di libri ed articoli su riviste di settore, in passato ha lavorato per multinazionali, aziende della sanità italiana, enti governativi e militari. In qualità di esperto di Cyber Intelligence e contromisure digitali ha svolto inoltre attività di docenza presso alcune università italiane.

ICT Security MAGAZINE

Storia ed evoluzione dei deepfake

I **deepfake** sono immagini, audio e video falsificati, creati da algoritmi di intelligenza artificiale. Sebbene la fabbricazione e la manipolazione di immagini e video digitali non siano una novità, i rapidi sviluppi delle reti neurali (Deep Neural Network) negli ultimi anni hanno reso il processo di creazione di immagini/audio/video falsi convincenti sempre più facili e veloci da realizzare.

La parola deepfake finisce sempre più spesso per essere associata allo sconcertante problema della **disinformazione online**.

In realtà **la tecnologia deepfake** è molto di più dell'accezione negativa che gli è stata etichettata, nei prossimi anni **sarà uno dei driver principali nella creazione ed elaborazione delle informazioni**, ma per comprendere meglio il suo potenziale tecnologico e produttivo è necessario capire la sua storia.

Nei primi mesi del 2017 un utente di Reddit (social network e sito di notizie in cui i membri registrati possono pubblicare contenuti di tutti i tipi, sotto forma di post o link diretti), soprannominato "Deepfakes", comincia a pubblicare video di alcuni spezzoni di film dove vengono sostituiti i volti degli attori originali con i volti di altri attori, realizzando così dei video fake, abbastanza realistici, che suscitano fin da subito il divertimento e l'interesse di molti utenti.
Il successo è tale che Deepfakes crea una "subreddit", una specifica categoria all'interno di reddit, dove i post sono organizzati in argomenti e seguiti da migliaia di utenti.
Nel giro di poche settimane si passa, però, da video curiosi e divertenti, a spezzoni di video hard dove il volto dei pornoattori viene sostituito con quello di celebrità del mondo dello spettacolo, cominciando ad ingenerare perplessità e allarme, come ad es. il caso di Taylor Swift, in cui la faccia della cantante è stata inserita in diversi video fake a

contenuto pornografico.
In questa categoria - subito ridenominata "**deep-porn**" - nonostante i video siano realizzati in modo artificiale, la loro diffusione può ledere la reputazione e la dignità personale delle vittime assumendo così rilevanza penale. Reddit si trova rapidamente a dover gestire diverse azioni legali intentate dalle vittime, così alla fine blocca l'account Deepfakes, motivando la decisione con la violazione delle policy della community per pubblicazione di video **NSFW** (**Not Safe For Work**), ossia non sicuro per il lavoro, acronimo usato principalmente nei siti di blog e forum per indicare materiale sessualmente esplicito, volgare o potenzialmente offensivo.
Ormai però il numero di utenti interessati a questa tipologia di video è tale da rendere l'argomento virale, e nonostante le restrizioni imposte anche da altri social network come Twitter, Facebook e siti tematici come Pornhub - che per la prima volta usano la definizione di "**video sintetici**" - si sviluppano canali alternativi ad es. presso 4chan e 8chan. Canali social più permissivi in termini di regole e censura, dove molti utenti criticano aspramente la scelta di Reddit, condividendo oltre ai video anche post con link ad altri siti tematici dedicati ai deepfakes porno che nel frattempo sono sorti (Fig. 1).

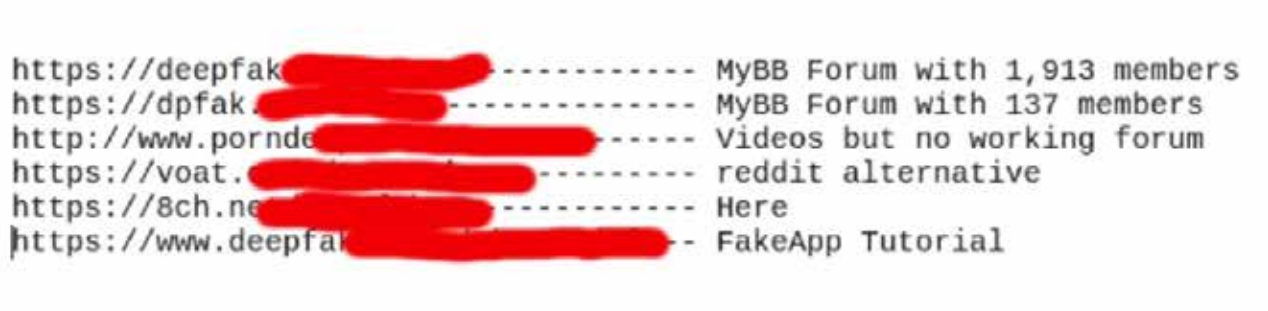


Fig. 1

I post dei video di Deepfake diventano così sempre più virali e il loro successo non è dovuto solo ai "prodotti finali" ma soprattutto perché, oltre ai video, alcuni utenti cominciano a postare link ad applicazioni che permettono a tutti di poter creare i propri deepfake facilmente, senza nessuna conoscenza di computer grafica e di i**ntelligenza**

artificiale, componente tecnologica alla base della produzione di queste tipologie di video. Ed è questa nuova tecnologia la vera novità, poiché la manipolazione video dei volti denominata "**face swap**" non è in realtà qualcosa di nuovo, in quanto è una tecnologia utilizzata da tempo nell'ambito del mondo cinematografico dove, a volte, a causa dell'impossibilità di poter girare delle scene con specifici attori (ad esempio, perché deceduti) il loro volto viene sovrapposto a quello di altri attori; come nel caso di Peter Cushing, morto nel 1994 e redivivo nei panni del comandante della Morte Nera nell'episodio Rogue One di Star Wars del 2016, grazie all'elaborazione grafica del volto dell'attore Guy Henry (Fig. 2).

Fig. 2

Prima di Deepfakes, quindi, la possibilità di eseguire dei video credibili di face swap rimaneva appannaggio esclusivo di professionisti della computer grafica che avevano a disposizione strumenti e software potenti e costosi.

I deepfake rappresentano un primo salto evolutivo nel "face-swap" casalingo, perché permettono a utenti senza alcuna esperienza, ma con una buona scheda video e un po' di tempo a disposizione, di utilizzare applicativi in grado di realizzare video fake di buona qualità, ed è così che la parola **"deepfake" diventa sinonimo di video falsi prodotti con l'intelligenza artificiale**.

La parola deepfake, infatti, deriva dall'inglese "**deep learning**", ed intende una tipologia di algoritmi che sfruttano l'intelligenza artificiale per generare dei falsi, "fake", per creare un'immagine, un audio voce e/o un video umano sintetico.

A dare un primo forte scossone all'opinione pubblica sul tema dei deepfake, con lo scopo di aumentare la consapevolezza dei pericoli di questa tecnologia, è stato, nel 2018, l'attore e regista statunitense Jordan Peele con un video, diventato subito virale, in cui il presidente Barack Obama sembra insultare personaggi pubblici, tra cui Donald Trump e Ben Carson. Il video, pubblicato ad aprile 2018, è visionabile al seguente link: www.youtube.com/watch?v=cQ54GDm1eL0

Il video realizzato attraverso la tecnica deepfake di sincronizzazione delle labbra (Lip-Sync), genera enorme interesse e preoccupazione anche nelle agenzie di intelligence governative che vedono subito il potenziale utilizzo e pericolo di questa tecnologia in scenari di **propaganda politica, disinformazione e cyberwarfare**.

Un altro aspetto importante nell'evoluzione della tecnologia deepfake è che si è passati, da video processati attraverso un grande lavoro fatto di lunghi tempi di "montaggio", a deep fake in tempo reale, in grado di cambiare ad es., l'immagine del volto di un soggetto con un altro, esistente o immaginario, durante l'utilizzo estemporaneo di applicazioni video come ad esempio una video call o in un live streaming.

Questi fattori combinati tra loro, sono in grado di generare dei falsi multifattoriali (audio, video ed interazione), sollevando un'enorme attenzione da parte degli esperti di sicurezza delle informazioni che hanno classificato la **tecnologia deepfake tra le nuove e più pericolose cyber minacce**.

Casi d'uso e nuovi scenari

Da tecnologia di intrattenimento a potente strumento di comunicazione, business e controllo. I vantaggi della tecnologia Deep Fake sono vari, può essere utilizzata in molti e diversi settori: servizi di comunicazione, pubblicitari, moda, campo medico, intrattenimento, dove ad es. le celebrità possono vendere il loro "modello virtuale" senza così avere la necessità di dover interpretare e/o viaggiare per andare sulle riprese di un set cinematografico.

Un altro aspetto interessante è ad es. quello di realizzare applicazioni per **rappresentazioni umane sintetiche** immaginarie, cioè utilizzare la tecnologia deep fake per creare modelli di persone immaginarie, mescolando ad es. le immagini di varie persone reali al fine di generare un modello di persona sintetica non reale (Facial synthesis). Un esempio è il sito: https://generated.photos/face-generator.

Qui è possibile creare un modello di umano (partendo da un database di immagini di persone reali), selezionando razza, età, colori degli occhi, profili, postura, stato d'animo, etc. e disporre così di un set fotografico di un modello virtuale unico da utilizzare, ad esempio, per una pubblicità, tra l'altro l'immagine generata è royalty free (Fig. 3).

Fig. 3

Nel prossimo futuro, sicuramente uno degli obiettivi principali dell'utilizzo del deep fake sarà la creazione di contenuti. Anche per realtà governative il potenziale positivo di questa tecnologia risiede nella capacità di **comunicare in modo più diretto, emotivo e naturale** con le varie parti interessate, ad es. un annuncio di servizio pubblico che può essere trasmesso contemporaneamente in diverse lingue con modelli umani "sintetici" di varie etnie per rendere il messaggio più efficace e tranquillizzante.

Nel 2020 il canale MBN ha trasmesso un notiziario utilizzando un deep fake della giornalista Kim Joo-Ha (Fig. 4), che solitamente presentava le notizie.

Fig. 4

Il pubblico avvertito dell'esperimento reagì principalmente in due modi: sbalordito da una parte e preoccupato sullo stato di salute della giornalista dall'altro.

Per quanto riguarda l'aspetto negativo dei deep fake, a parte "vaporizzare" professioni (si pensi al software VoCo di Adobe, ancora in fase prototipale, in grado di soppiantare i doppiatori cinematografici), la manipolazione delle informazioni eseguita tramite questa tecnologia apre **scenari di minaccia estremamente inquietanti**.

In generale, per le organizzazioni, si prevede che le aziende potranno essere vittime di **diffamazione** e **inganno**, gli **attacchi di phishing** attraverso audio e video saranno sempre più frequenti, complessi e soprattutto efficaci. Ad es. nell'agosto 2019, l'amministratore delegato di una società europea, ingannato da un deepfake audio, ha assecondando la richiesta dei criminali effettuando un trasferimento di 243.000 dollari.

Poiché Internet ha trasformato il mondo in un "villaggio globale", anche le nazioni in conflitto utilizzano la diffusione di notizie false per portare avanti i loro programmi all'estero, minando la reputazione dei paesi avversari nel resto del mondo. Molti paesi ormai gestiscono account di social media, siti Web ed applicazioni sponsorizzate dal governo, contribuendo alla propaganda politica globale. In particolare, si ritiene che i governi di Cina, Israele, Turchia, Russia, Regno Unito, Ucraina, Corea del Nord e specifiche organizzazioni governative come CIA e NSA, siano coinvolte nell'uso di "fantasmi digitali" per diffamare gli oppositori e diffondere disinformazione.
Anche in ambito governativo/militare l'utilizzo della tecnologia **deep fake come arma da utilizzare in conflitti di cyberwarfare** ha un enorme potenziale, la guerra ibrida di quinta generazione utilizza la disinformazione come ulteriore dimensione di conflitto; dalla propaganda politica alla disinformazione per destabilizzare i governi avversari, dall'ingerenza elettorale alla privazione delle persone della cosiddetta "autodeterminazione informativa" ("ciò che voglio far sapere di me lo decido io"), e della loro libertà decisionale ("quello che penso e faccio è una scelta su cui gli altri non possono interferire").
La tecnologia deepfake sarà sempre più facilmente utilizzata ed il suo prodotto, sempre più sofisticato e realistico, sarà sempre più presente in due ambiti: nell'informazione falsa (Misinformation), imprecisa, comunicata indipendentemente dall'intenzione di

ingannare, e nella disinformazione (Disinformation) attraverso strategie ben precise, utilizzate da lobbies economiche e politiche per fabbricare informazioni, per il loro vari obiettivi politici, finanziari, etc. Si prevede che il deepfake diventerà il principale processo di diffusione intenzionale di notizie manipolate per influenzare l'opinione pubblica o oscurare la realtà. Grazie all'uso sempre più intensivo ed estensivo delle piattaforme social media, la divulgazione di notizie false generate con tecnologia deepfake audiovisiva ha effetti molto più devastanti, in quanto i soggetti, vittime a loro insaputa di un deepfake, non solo subiscono una perdita di controllo sulla loro immagine, ma vengono anche private del controllo sul loro pensiero, adducendogli false idee e comportamenti.
Le fonti di Misinformation/Disinformation aumenteranno, grazie al deepfake, le loro capacità "offensive".

Tra le principali fonti troviamo:

Troll: Spesso confusi con gli haters, i troll hanno strategie diverse da questi ultimi, non vogliono solo insultare un soggetto in particolare (persona, organizzazione, Nazione, etc.) ma vogliono generare una rissa mediatica, in modo da attirare più persone sul soggetto da colpire. I troll sono sempre più spesso vere e proprie "aziende" della disinformazione, fabbriche di troll al soldo di soggetti politici, lobbies ed aziende che li ingaggiano regolarmente per falsificare notizie relative ai loro concorrenti e diffonderle nel mercato.

Bot: I bot sono software o algoritmi automatizzati, utilizzati per diffondere contenuti fabbricati o fuorvianti. In genere sono sistemi adottati per una diffusione massiva e trasversale di contenuti, ad es. durante la campagna elettorale statunitense del 2016, i bot sono stati impiegati per generare un quinto di tutti i tweet postati durante l'ultimo mese di attività elettorale. I deepfake potenziati dai sistemi bot possono aumentare in maniera esponenziale il loro impatto negativo. Ne è un esempio il caso del bot sull'app di messaggistica Telegram che, qualche tempo fà, utilizzando il codice di un'altra

app denominata Deepnude (ritirata dai vari app store ma disponibile sul sito Github), riproduceva l'immagine di un corpo femminile nudo basandosi sulle informazioni analizzate nella foto originale (dimensioni, ombre, proporzioni, etc.) inviata dagli utenti. La combinazione Bot+Deepfake può aprire scenari di produzione e diffusione massiva di informazioni false estremamente allarmanti.

Teorici della cospirazione: I teorici della cospirazione, in genere, credono che alcune importanti comunità gestiscono le informazioni nascondono delle verità, come ad es. le teorie del complotto sul controllo delle masse o la cospirazione dell'attuale pandemia di COVID da parte degli Stati Uniti e della Cina. In una situazione del genere, l'uso di contenuti deepfake audiovisivi fabbricati da parte di questi teorici, o da parti avverse che li sfruttano attraverso la Misinformation, possono innescare dannose azioni di disinformazione su controversie politiche a livello globale in pochissimo tempo.

Hyper-partisan Media: con il termine media "iper-partigiani" ci si riferisce a siti Web, blog di notizie, più o meno tematici, che diffondono intenzionalmente informazioni false. Grazie ai social media e ad attori come i troll ed i teorici della cospirazione, i media "iper-partigiani" sono dei veri e propri incubatori per la diffusione di notizie false tra le persone. I convincenti contenuti falsi generati dall'Intelligenza Artificiale permettono la facile diffusione della disinformazione attirando a sé nuovi visitatori aumentandone la visibilità, ed informazioni basate su deepfake ne incrementano inoltre anche l'"autorevolezza". Poiché i media "iper-partigiani" e le piattaforme social sono in gran parte gestite da organizzazioni indipendenti basate sulla pubblicità, la diffusione di informazioni fabbricate può essere puramente una strategia a scopo di lucro.

Politica: Principale fonte storica della disinformazione. A causa del gran numero di follower sulle piattaforme social, i politici sono nodi centrali nelle reti online, usando la loro fama e notorietà sfruttano il sostegno pubblico per diffondere notizie false tra i loro seguaci. Con lo scopo di diffamare gli avversari politici, possono arrivare a pubblicare contenuti controversi sui loro concorrenti anche su media convenzionali.

Esempi attuali di deepfake

I deepfake hanno preso d'assalto Internet, trascinandosi dietro celebrità, politici e gente comune, non solo per divertire ma anche per scopi malevoli, ad es. creando disinformazione attraverso il racconto di cose mai accadute.
Negli ultimi mesi stiamo assistendo all'utilizzo sempre più frequente di tecnologia deep fake, dall'intrattenimento alla cyberwarfare nel **conflitto tra Russia e Ucraina**.

Altro aspetto su cui è bene porre attenzione è la qualità dei prodotti realizzati, sempre più realistici oramai e quasi tecnicamente indistinguibili da un originale, se non per il contesto: contenuti e situazione spesso non coincidono con la persona o la realtà rappresentata.

Un esempio qualitativamente notevole di deep fake, che ultimamente è diventato virale, sono i video su TikTok dell'attore Tom Cruise (Fig. 5). Creati da Miles Fisher, che grazie ad una fisionomia simile all'attore, realizza video deep fake di una elevata qualità.

Fig. 5

La sua somiglianza aiuta significativamente l'algoritmo nel creare un modello sintetico estremamente realistico. Proprio la qualità derivata da questa sua somiglianza lascia immaginare che, in un prossimo futuro, sosia di attori e celebrità potrebbero trovare un nuovo lavoro come interpreti deep fake al posto di attori reali o persone famose in film, trasmissioni o comparsate pubbliche su canali mainstream ed internet.

L'applicazione della tecnologia in *real time* del deep fake sta rapidamente entrando nell'uso quotidiano sui social network e nelle trasmissioni in live streaming. Tra i prodotti attualmente più utilizzati troviamo l'**applicazione DeepFaceLive** (utilizzata da Miles per il deep fake di Tom Cruise), che nasce dalle basi del suo famoso predecessore DeepFaceLab (Fig. 6).

Fig. 6

Grazie all'applicazione DeepFaceLive gli utenti non hanno più bisogno di passare attraverso il laborioso processo di raccolta del materiale facciale della celebrità preferita - mediante, quindi, l'utilizzo di un set di dati che necessita poi dai 3 ai 5 giorni di elaborazione - per ottenere una corrispondenza ottimale, ma addirittura, nella guida all'applicativo gli utenti trovano link ed indicazioni per utilizzare "**modelli di volti pubblici pronti all'uso**", già compilati da altri creatori e disponibili sul Forum nella sezione Trained Models del sito porno MrDeepFakes (Fig. 7).

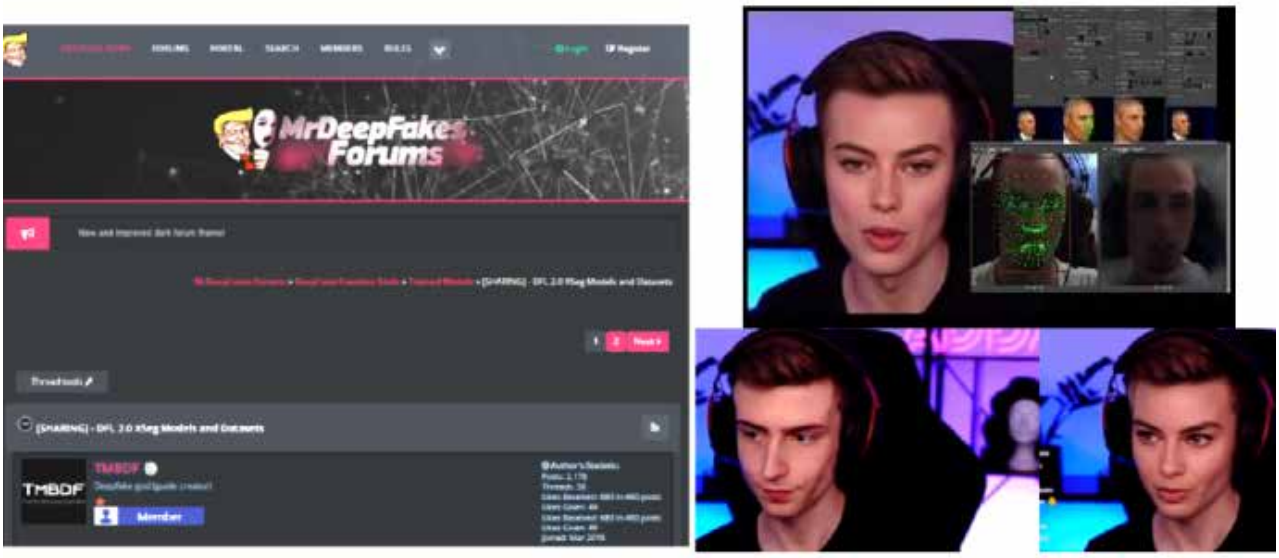


Fig. 7

I progressi della tecnologia deepfake sono entrati anche nelle attenzioni dei cyber criminali che cominciano a sfruttarli in combinazione con altre tipologie di attacchi. L'utilizzo nell'ingegneria sociale, metodologia di attacco rimasta sostanzialmente invariata e comunemente perpetrata attraverso attacchi di impersonificazione e phishing, grazie all'impiego del deep fake aggiunge un aspetto multifattoriale.

Al momento segnalazioni di questa tipologia di attacchi informatici rimangono relativamente minimi, mentre sono in crescita e sono state individuate diverse campagne **BEC** (**Business Email Compromise**) atte a manipolare trasferimenti di soldi in modo fraudolento utilizzando il phishing tramite email associato al **deep fake audio** per impersonare telefonicamente una figura di fiducia/cliente della vittima.

Recentemente un manager di una banca degli Emirati Arabi Uniti è stato indotto con l'inganno, attraverso una mail di phishing, a chiamare il CEO di una loro azienda cliente per confermare un trasferimento bancario e, poiché la voce all'altro capo del telefono suonava come l'effettiva voce del CEO, il manager della banca ha effettuato un trasferimento di 35 milioni di dollari sul conto dei cyber criminali.

L'audio deepfake potenzia notevolmente la percentuale di successo degli attacchi BEC ed è facile da realizzare, bastano infatti meno di quattro secondi di audio sorgente del target per poter catturare i tratti personali della voce come pronuncia, tempo, intonazione, altezza e risonanza, necessari per creare un deepfake convincente, e più audio sorgente e campioni di addestramento saranno disponibili, più convincente sarà l'output.

In ambito governativo/militare, il recente conflitto tra Russia e Ucraina sta mostrando quanto la tecnologia deepfake può essere efficace nelle **guerre ibride di quinta generazione**, dove cioè oltre al combattimento tradizionale cinetico si affianca quello del cyberwarfare, in cui la dimensione di combattimento è il dominio cibernetico e delle informazioni su internet.

Un esempio di questi cyber attacchi, mirati a sviluppare propaganda politica e disinformazione tra le popolazioni coinvolte e l'opinione internazionale, sono i deep fake dei due leader di Russia e Ucraina, realizzati a breve tempo l'uno dall'altro.

Video in cui vediamo prima Putin (Fig. 8) annunciare la fine degli scontri (https://v.

redd.it/ww7nr8706ll81/DASH_720.mp4) e poi Zelensky (Fig. 9) annunciare la resa dell'Ucraina (https://www.youtube.com/watch?v=X17yrEV5sl4)

Fig. 8

Fig. 9

Cognitive Security

La **sicurezza cognitiva** (**COGSEC**) è l'insieme delle tecniche e delle metodologie di difesa dagli attacchi di ingegneria sociale, manipolazioni intenzionali e non di interruzione della cognizione e delle funzioni sensoriali.
La sicurezza cognitiva, tuttavia, in contesti di cybersecurity, si riferisce solitamente all'applicazione di tecnologie di **intelligenza artificiale**, **machine learning** e **cognitive computing** che sono modellate sui processi del pensiero umano per rilevare le minacce e proteggere i sistemi fisici e digitali.

La sicurezza cognitiva è quindi particolarmente utile per prevenire gli attacchi informatici che manipolano la percezione umana. Tali attacchi detti anche "**hacking cognitivi**", sono progettati per influenzare i comportamenti degli utenti al fine di raggiungere gli scopi dell'attaccante.
Gli sforzi per la sicurezza cognitiva in questo settore includono anche approcci non tecnici, ad es. il concetto di "**firewall umano**" che mira a rendere le persone meno vulnerabili alla manipolazione, tecnicamente più preparate, nonché soluzioni tecniche progettate per rilevare dati fuorvianti, disinformazione, così come metodi per prevenirne la diffusione.

Come altre applicazioni di cognitive computing, la sicurezza cognitiva, utilizza algoritmi metodi e processi di apprendimento automatico che consentono ai sistemi cognitivi di estrarre, elaborare e analizzare enormi volumi di dati strutturati e non strutturati, in maniera costante, identificando informazioni significative, connessioni tra punti, dati e tendenze - che sarebbero diversamente impossibili da rilevare per un essere umano - attraverso analisi avanzate, con lo scopo di permettere ai sistemi di imparare come anticipare le minacce, rendendoli in grado di generare soluzioni proattive.

La manipolazione delle immagini ha una sua storia considerevole come propaganda in tempi di conflitto, oggi la facile disponibilità di strumenti digitali, la natura altamente realistica del contenuto falsificato e l'esistenza di nuovi canali mediatici per distribuire la disinformazione hanno trasformato i deepfake in un meccanismo di attacco privilegiato e, nell'ambito della COGSEC, gli attacchi basati su deepfake sono diventati tra le principali e più pericolose minacce da affrontare nell'attuale contesto storico e, ancor di più, nel prossimo futuro.

Il risvolto psicologico del fenomeno dei deepfake è proprio il **mancato riconoscimento tra ciò che è vero e ciò che è falso**, i deepfake si possono definire come "***true lies***", bugie vere, in grado di mettere in crisi la fiducia verso l'Altro e contestualmente aumentare la frammentarietà del Sé.

Gli attacchi che riescono a generare i deep fake mirano, oltre al **furto di identità, credenziali** e **frodi finanziarie**, anche a scenari di più ampia e varia natura: **cambiamenti dell'opinione pubblica, risultati elettorali distorti, confusione di massa, violenza etnica, guerra**. Tutti questi eventi potrebbero essere facilmente innescati da deepfake, alimentati dai progressi dell'intelligenza artificiale e diffusi attraverso i tentacoli dei social media. I deepfake possono rivelarsi tra le forze più destabilizzanti che l'umanità ha dovuto affrontare da generazioni.
Le competenze di cui la disciplina della Cognitive Security deve disporre non sono solo meramente tecniche ed informatiche ma soprattutto umanistiche, psicologiche e geopolitiche. Sviluppare in profondità la tecnologia di rilevamento dei falsi è importante, ma è solo una parte della soluzione, il **fattore umano**, le debolezze nella nostra mente rendono i deepfake efficaci strumenti di offesa.
Molto presto diverrà impossibile riconoscere ad occhio ed orecchio umano se un video o una clip audio sono autentiche o meno, e mentre la propaganda non è una novità, l'immediatezza viscerale della voce e dell'immagine danno ai falsi una profondità d'impatto sulle coscienze e negli animi delle persone senza precedenti. Per questo motivo **la COGSEC negli ultimi anni è sempre più richiesta e finanziata sia dai governi che dall'industria**.

Possibili minacce

Nei prossimi anni, **i deepfake utilizzeranno tecnologie e processi basati sull'Intelligenza Artificiale sempre più sofisticati, accelerando e potenziando gli attacchi di ingegneria sociale**, attacchi semplici da sferrare da parte degli aggressori che, è possibile ipotizzare, impersoneranno con una frequenza ed un realismo allarmante e accuratamente individui - ad esempio dirigenti, persone dal forte potere decisionale - per diffondere notizie false e disinformazioni altamente credibili.
Questi attacchi causeranno gravi danni finanziari, ingannando l'opinione pubblica con video e immagini falsi al fine di manipolare i mercati, promuovere agende politiche e ottenere vantaggi competitivi.
Le persone e le organizzazioni vittime di questi attacchi saranno esposte a gravi danni reputazionali causati dalle loro identità compromesse, trovandosi così ad affrontare una nuova generazione di sfide alla sicurezza ed integrità delle informazioni.
Stati, nazioni, attivisti, gruppi di hacker e cyber criminali utilizzeranno i deepfake per diffondere la disinformazione su larga scala, lasciando le vittime incapaci di distinguere la realtà dalla finzione, la verità dalle bugie.

Considerata l'evoluzione tecnologica e la sempre maggiore facilità d'uso di questa tecnologia, le minacce deepfake saranno sempre più diffuse anche nella sfera personale e privata delle persone comuni.
Tra i giovani già sono stati registrati atti di **cyberbullismo** volti a denigrare e screditare le vittime, spesso utilizzando deepfake con contenuti pornografici, ma anche nella popolazione più adulta, attraverso le piattaforme social, sono stati condivisi deepfake di vittime rappresentate nude, in situazioni compromettenti o pornografiche. Come ad es. i deepfake realizzabili attraverso l'applicazione Deepnude (precedentemente citata) in grado di effettuare l'associazione del volto della vittima a corpi di altri soggetti, nudi o in pose di natura esplicitamente sessuale.

Questa tipologia di deepfake e la loro azione di diffusione e condivisione sulle piattaforme social media denominata "***revenge porn***", generata per motivi di varia natura - denigrazione, ricatto, vendetta da parte di ex fidanzati, amanti, etc. - può causare **danni reputazionali e psicologici** estremamente gravi.
Inoltre la possibilità di generare simili contenuti, a totale insaputa della vittima, spesso alimenta anche la pratica del ***sexting***, cioè lo scambio e diffusione di immagini di nudo e pornografia illegale, arrivando perfino a commettere reati estremamente gravi come la pedopornografia.

I deepfake rischiano di diventare "la tempesta perfetta" sulle fonti di informazione odierne, una dimensione dove i mainstream hanno perso la loro autorevolezza e le notizie vengono sempre più frammentate e micro orientate, dove è facile direzionare il pensiero del lettore proprio su ciò che desidera ascoltare, escludendogli ciò che non gli piace. Questo è il tavolo da gioco preferito da coloro che cercano di disinformare capaci di dirottare gli utenti dei social media amplificando la desiderabilità di un pregiudizio.
Non è possibile, in questo scenario, guardare alla tecnologia come la sola soluzione per affrontare il problema, il deepfake può arrivare nella profondità delle coscienze consentendo ai falsi media di prosperare e solo **capendo come individuare ed affrontare i pregiudizi, in noi e negli altri, possiamo aumentare la possibilità di mitigare gli effetti di un attacco deepfake**.

Gli psicologi hanno scoperto che ripetere lo stesso messaggio aumenta l'effetto di propaganda attraverso un processo chiamato ***priming***, più sei esposto ad un'affermazione, più è probabile che la valuti come vera e se poi quell'informazione proviene da una fonte ritenuta "attendibile" il processo di esposizione subisce un fattore moltiplicatore.

Realizzazione di un deepfake con la tecnica del Generative Adversarial Network (GAN)

La rapida evoluzione della tecnologia deepfake non risiede solamente nella qualità dei suoi elaborati ma soprattutto nella facilità e possibilità di utilizzarla attraverso applicazioni sviluppate per diverse tipologie di *device* e sistemi operativi: sistemi fissi, PC, smartphone, oppure direttamente online tramite un web browser.

Lo smartphone rappresenta la tipologia di dispositivo più utilizzato per la diffusione e l'utilizzo di applicazioni, le sempre più elevate capacità computazionali dei *device* mobili e la loro "naturale" capacità di interazione e creazione di contenuti per le piattaforme social lo rendono un driver tecnologico ideale per le applicazioni deepfake.
Oggi troviamo molte applicazioni deepfake per dispositivi mobili iOS e Android.

La diffusione di "contenuti sintetici" è in continua crescita e spesso la loro tipologia e finalità è influenzata dal canale di distribuzione, ad es. su TikTok i video hanno un limite di alcuni secondi e seguono dei *trend* precisi, le tecnologie deepfake più utilizzate in questo caso sono meno sofisticate in termini realistici ma più immediate nel creare brevi e divertenti meme, questo ha decretato ad es. il successo dell'applicazione **Wombo** disponibile su Android e iOS.

Fig.10

Wombo (Fig.10) è un'applicazione basata su tecnologia deepfake che sfrutta l'AI per la sincronizzazione del labiale (Lip-Sync), per animare l'immagine di un volto in modo che "canti" una canzone selezionata per creare video divertenti.

Esistono poi altre applicazioni utilizzabili direttamente online, tramite web browser, che permettono di effettuare elaborazioni video di swap face in modo veloce e facile, come ad es. il sito **DeepfakesWeb,** dove è sufficiente effettuare l'upload di un breve video del soggetto da cui prendere il volto ed un secondo video del soggetto su cui trasferire il volto (Fig.11).

Fig.11

A prescindere dallo scopo, divertimento, intrattenimento, lavoro, etc. la realizzazione di contenuti sintetici richiede notevoli risorse computazionali, siano esse locali, all'interno dei dispositivi, siano esse distribuite in cloud.
La tecnologia deepfake come abbiamo precedentemente detto, utilizza algoritmi di Intelligenza Artificiale che richiedono grandi capacità computazionali e di risorse, in generale la tecnica più utilizzata è chiamata **Generative Adversarial Network** (**GAN**).

La tecnica del GAN fa parte di un ramo di apprendimento automatico basato sulle **reti neurali**. Queste reti sono progettate per emulare i processi neuronali del cervello umano e possono essere addestrate a riconoscere o manipolare specifiche attività e informazioni.

La caratteristica di un modello GAN risiede nella rivalità tra due o più reti neurali (Fig.12).

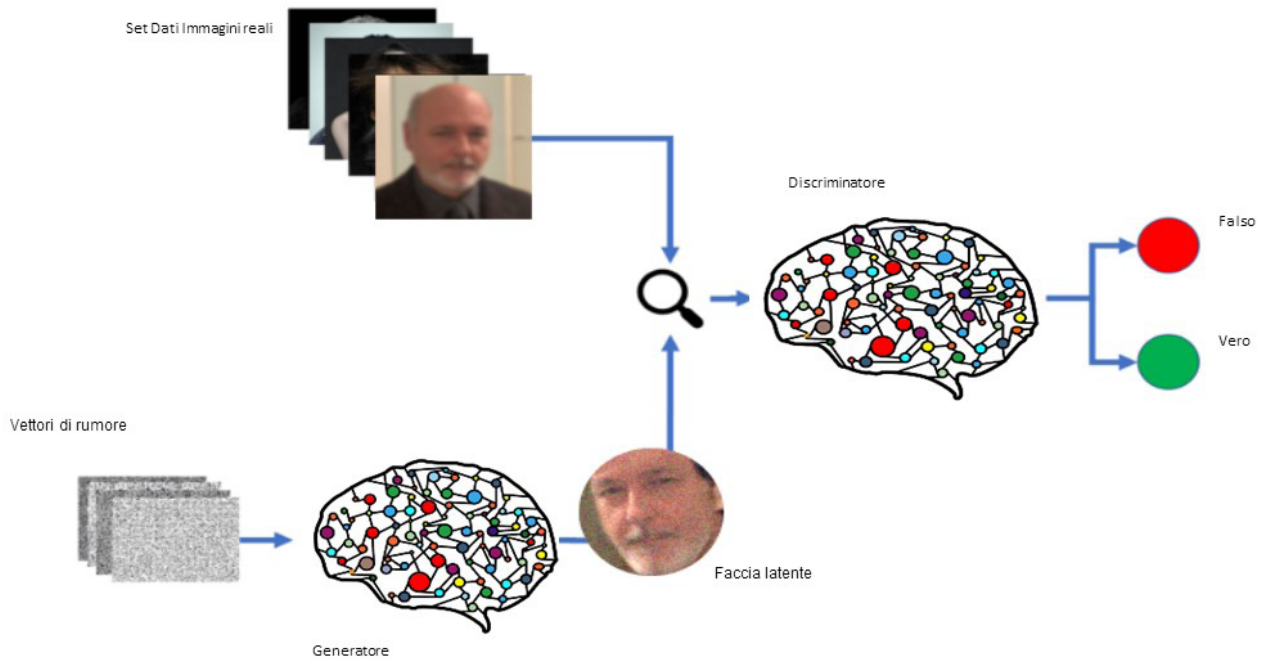


Fig.12

Il modello di GAN utilizzato per la generazione dei deepfake, sfrutta due reti neurali messe l'una contro l'altra con l'obiettivo di generare un output realistico. Lo scopo è garantire che i deepfake creati siano i più reali possibile.
Nel GAN utilizzato per il deepfake, il falsario di immagini e il rilevatore di falsi tentano ripetutamente di superare le capacità dell'altro ed entrambe le reti neurali vengono addestrate utilizzando lo stesso set di dati.

Diverse aziende stanno sviluppando dei modelli personalizzati di GAN come ad es. StyleGAN di Nvidia e BigGAN di Google.
La prima rete è chiamata **generatore**, il cui compito è appunto generare un'immagine contraffatta utilizzando vettori di rumore - un elenco di numeri casuali - che sembrino il più realistici possibile. Il risultato è una rappresentazione dimensionale, di qualità inferiore, di quella stessa faccia che, a volte, viene definita vettore di base o faccia latente.
La seconda rete, denominata **discriminatore**, determina la veridicità delle immagini generate. Confronta l'immagine contraffatta, generata dal generatore, con le immagini autentiche nel set di dati per determinare quali immagini sono reali e quali false.
Sulla base di questi risultati, il generatore, varia il parametro per la generazione delle immagini. Questo ciclo continua fino a quando il discriminatore non riesce ad accertare che un'immagine generata è falsa e viene quindi utilizzata nell'output finale.
Questo è il motivo per cui i deepfake sembrano così reali, con il progredire della tecnologia e delle capacità computazionali non sarà più possibile distinguere un'immagine falsa da una reale.

Per poter generare un deepfake e sfruttare le potenzialità dei GAN - ad es. per un face swap dal soggetto A al soggetto B - è necessario addestrare una rete di codificatori basati su **reti neurali convoluzionali**, anche dette **CNN** (**Convolutional Neural Networks**).

Una CNN utilizza molti strati convoluzionali, ovvero strati dove viene impiegata un'operazione matematica chiamata convoluzione che filtra gli input per trovare le

informazioni più utili. La rete CNN (Fig.13) deve essere addestrata prima di poter generare deepfake e per questo utilizza centinaia di immagini, set dati, del soggetto A e del soggetto B.

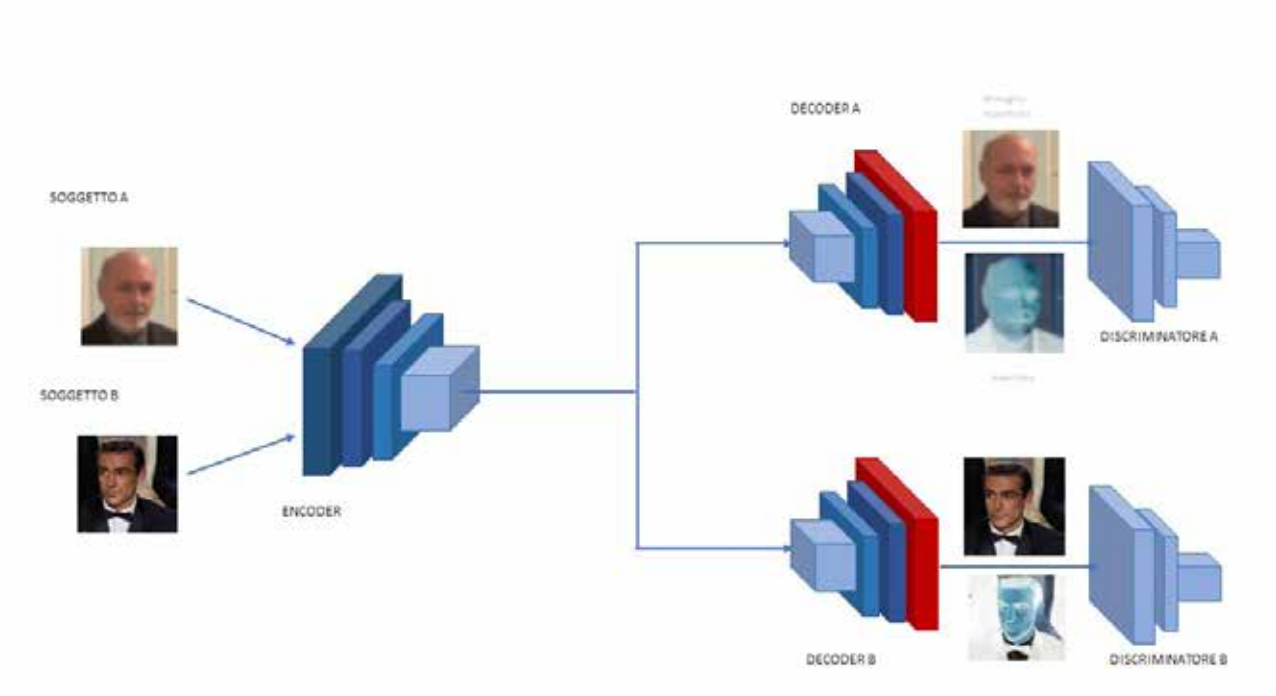


Fig.13

L'encoder impara a codificare le caratteristiche di ciascuna immagine come le espressioni facciali, la forma del viso, etc. producendo una rappresentazione efficiente dell'immagine. L'uscita dell'encoder viene presentata a una rete di decodificatori sulla rete CNN. La rete di decodificatori impara a ricostruire le immagini.
Oltre a generare immagini, i decoder generano anche maschere che aiutano a produrre immagini più realistiche dopo lo scambio di volti.

La rete quindi utilizza due diverse distribuzioni corrispondenti e due decoder separati che imparano a ricostruire rispettivamente le facce del soggetto A e del soggetto B.

L'encoder deve estrarre le caratteristiche più importanti per ricreare l'input originale affinché i decoder funzionino come desiderato.
La combinazione encoder-decodificatore è comunemente chiamata **autoencoder** e costituisce la rete di generazione del nostro GAN.

Pertanto, si hanno due GAN: GAN A (costituito da encoder e decoder A) e GAN B (costituito dall'encoder e decoder B). Due discriminatori separati per A e B, imparano a distinguere meglio tra immagini reali e false. Quando inseriamo le immagini generate nel rispettivo discriminatore, la rete GAN spinge il generatore a realizzare immagini più realistiche, un ciclo continuo che termina quando le immagini generate non sono distinguibili da quelle reali.

Concluso l'addestramento, per generare il deepfake inviamo l'immagine del soggetto A all'encoder e al decoder B per ricostruire l'immagine. Poiché il decodificatore B ha imparato a generare il volto del soggetto B, genererà il volto del soggetto B con le caratteristiche del volto del soggetto A dell'immagine originale (Fig.14).

Fig.14

La “fabbrica” dei fake può essere classificata principalmente nelle seguenti categorie:

- **Falsi economici** (**Cheap Fakes**):
  sono elaborazioni di immagini e video dove il processo è più legato al fotoritocco e all’elaborazione diretta dell’immagine esistente (ad es. software come Photoshop, Adobe Premiere, etc.), questi non rientrano propriamente nella categoria dei deepfake;

- **Sincronizzazione delle labbra** (**Lip-Synching**):
  utilizza algoritmi di apprendimento GAN (ad es. Wav2Lip) in grado di sincronizzare il labiale di un video con qualsiasi audio contenente parlato;

- **Sintesi del viso** (**Facial Synthesis**):
  l’obiettivo è creare volti realistici che prima non esistevano (Fig.3);

- **Cambia faccia** (**Face Swap**):
  inserisce il volto di un soggetto sul corpo di un altro soggetto (Fig.11);

- **Rievocazione facciale** (**Reenactment**):
  i falsari possono trasferire le espressioni facciali da una persona all’altra. Grazie a questa tecnica, si può giocare con l’aspetto di una persona facendola sembrare disgustata, arrabbiata o sorpresa (Fig.15);

- **Voice Cloning:**
  l’obiettivo è creare cloni di tracce audio identiche alle originali.

Fig.15

(link al video: https://www.youtube.com/watch?v=ohmajJTcpNk)

Oltre alle immagini i modelli GAN sono in grado di elaborare anche segnali audio. La differenza nell'analisi da parte dei GAN tra audio e immagini è molto simile: le immagini possono essere viste come matrici, mentre i segnali audio possono essere considerati come semplici vettori. Nel 2018 è stato pubblicato l'algoritmo WaveGAN, un'architettura di Generative Adversarial Network in grado di sintetizzare l'audio utilizzando strati convoluzionali sia nel generatore che nel discriminatore. Sempre nel 2018 il gigante tecnologico cinese Baidu ha sviluppato un nuovo algoritmo di intelligenza artificiale che con soli 3,7 secondi di audio poteva clonare una voce con risultati decisamente credibili.

Tools per generare un deepfake audio video

Tool	Type	Reference/Developer	Technique
CHEAP FAKES			
Adobe Premiere	Commercial Desktop Software	Adobe	Audio Video Editing, AI-powered video reframing
Corel VideoStudio	Commercial Desktop Software	Corel	Proprietary AI
LIP-SYNCHING			
Dynalips	Commercial Web App	dynalips.com	Proprietary
CrazyTalk	Commercial Web App	reallusion.com/crazytalk	Proprietary
Wav2Lip	Open source implementation	github.com/Rudrabha/Wav2Lip	GAN with pre-trained discriminator network and visual quality loss function

Tool	Type	Reference/Developer	Technique
FACIAL ATTRIBUTE MANIPULATION			
FaceApp	MobileApp	FaceApp Inc	Deep generative CNNs
Adobe Premiere	Commercial Desktop Software	Adobe	DNNs + filters
Rosebud	Commercial Web App	rosebud.ai	Proprietary AI
FACE SWAP			
ZAO	Mobile App	Momo Inc	Proprietary
Reface	Mobile App	Neocortext, Inc	Proprietary
Reflect	Mobile App	Neocortext, Inc	Proprietary
Impressions	Mobile App	Synthesized Media, Inc	Proprietary

Tool	Type	Reference/Developer	Technique
FACE SWAP			
FakeApp	Desktop App	malavida.com/en/soft/fakeapp	GAN
FaceSwap	Open source implementation	faceswapweb.com	Employed two pairs of encoder-decoder. Shared encoder parameters
DFaker	Open source implementation	github.com/dfaker/df	For face reconstruction DSSIM loss function [34] is utilized. Keras library-based implementation.
DeepFaceLab	Open source implementation	github.com/iperov/DeepFaceLab	- Provide several face extraction methods, e.g. dlib, MTCNN, S3FD etc. - Extend different Faceswap model i.e. H64, H128, LIAEF128, SAE [33].
FaceSwapGAN	Open source implementation	github.com/shaoanlu/faceswap-GAN	Uses two loss functions namely adversarial loss and perceptual loss to the auto-encoder
DeepFake-tf	Open source implementation	github.com/StromWine/DeepFake-tf	Same as DFaker however, used tensor-flow
Faceswapweb	Commercial Web App	faceswapweb.com	GAN

Tool	Type	Reference/Developer	Technique
FACE REENACTMENT			
Face2Face	Open source implementation	web.stanford.edu/~zollhoef/papers/CVPR2016_Face2Face/page.html	Uses 3DMM and ML technique
Imitator	Mobile App		Proprietary (AI based)
Dynamixyz	Commercial Desktop Software	dynamixyz.com	Machine-learning
FaceIT3	Open source implementation	github.com/alew3/faceit_live3	GAN
FACE GENERATION			
Generated Photos	Commercial Web App	generated.photos	StyleGAN

Tool	Type	Reference/Developer	Technique
VOICE CLONING			
Overdub	Commercial Web App	descript.com/overdub	Proprietary (AI based)
Respeecher	Commercial Web App	respeecher.com	Combined traditional digital signal processing algorithms with proprietary deep generative modeling techniques
SV2TTS	Open source implementation	github.com/CorentinJ/Real-Time-Voice-Cloning	LSTM with Generalized end-to-end loss
ResembleAI	Commercial Web App	resemble.ai	Proprietary (AI based)
Voicery	Commercial Web App	voicery.com	Proprietary AI and Deep Learning
VoiceApp	Mobile App	Zoezi AB	Proprietary (AI based)

In queste tabelle [1] sono state riportate le principali applicazioni e progetti open source per la generazione di deepfake audio e video.

Come realizzare un video deepfake

Una delle applicazioni più interessanti da utilizzare, che dimostrano il livello tecnologico raggiunto e la sorprendente facilità nell'utilizzo degli algoritmi di deepfake, è sicuramente **DeepFaceLive**. Un progetto nato dalle basi di DeepFaceLab, una tra le prime e più utilizzate applicazioni per Windows che sono state sviluppate per soddisfare gli utenti interessati al mondo del deepfake, disponibile gratuitamente come DeepFaceLive su GitHub:

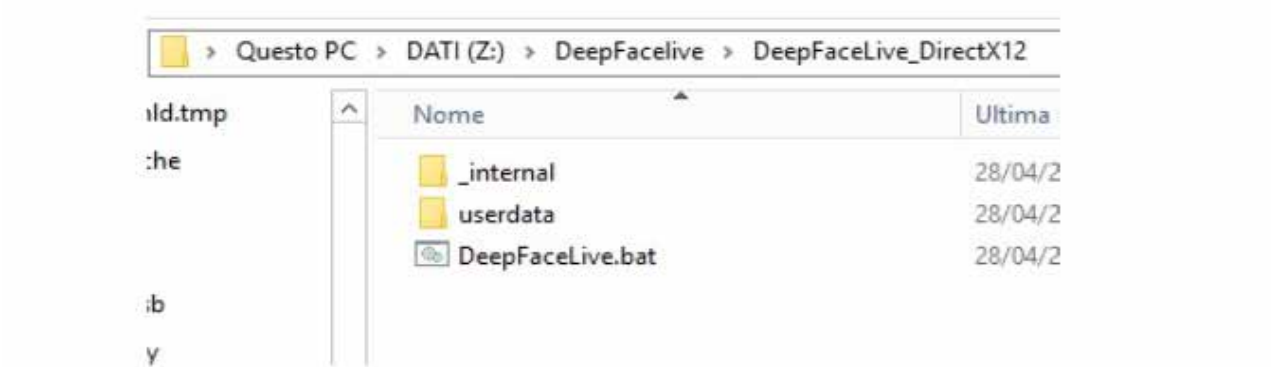
- DeepFaceLive: https://github.com/iperov/DeepFaceLive
- DeepFaceLab : https://github.com/iperov/DeepFaceLab

Attraverso DeepFaceLab è possibile addestrare il modello neurale del proprio volto per avere una maggiore qualità nella generazione del deepfake. DeepFaceLab è in grado di creare modelli addestrati e utilizzarli per creare deepfake estremamente credibili, la maggior parte dei deepfake presenti in rete sono stati creati con questa applicazione.

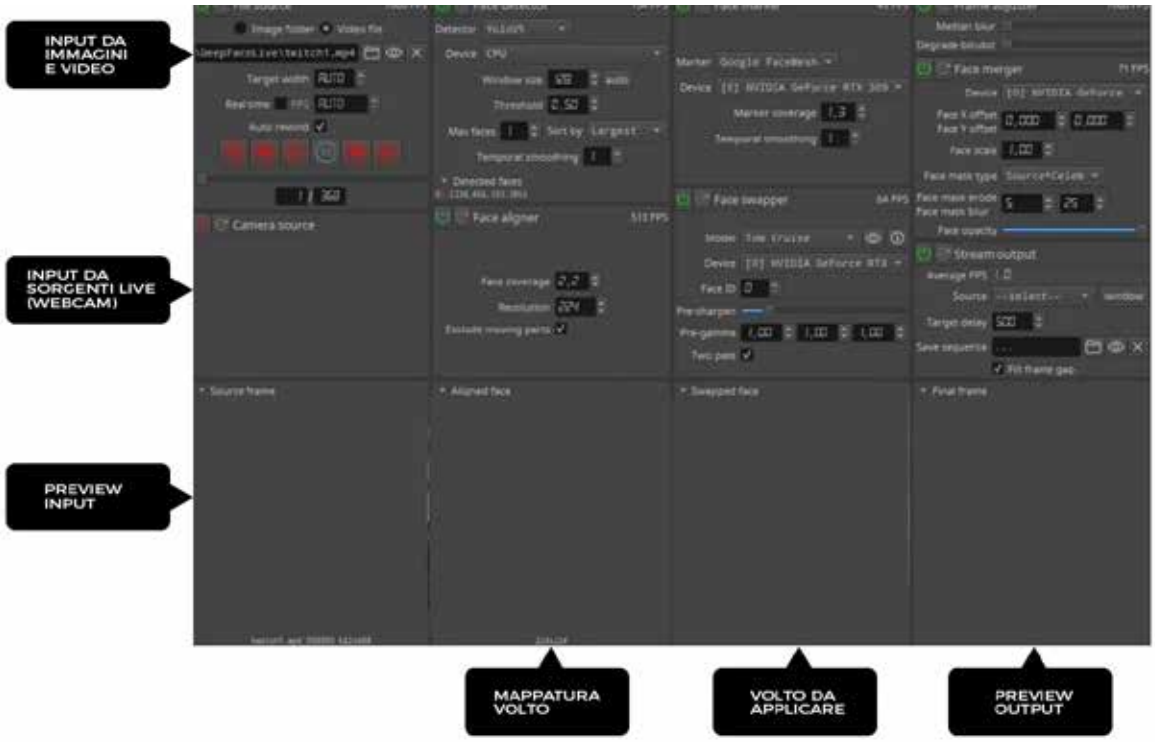
DeepFaceLive permette di importare i modelli generati da DeepFaceLab ed effettuare "face swap" in *real time* da sorgenti come immagini, video e live streaming da videocamera. Ed è proprio da quest'ultima sorgente che è possibile apprezzare la qualità applicativa nel realizzare dei deepfake in tempo reale.

L'installazione di DeepFaceLive è molto semplice, è sufficiente scaricare l'applicazione dal link sul sito Github. È possibile scegliere di effettuare il download della versione per Windows e per Linux, nel nostro caso eseguiremo il download della versione per Windows. Effettuato il download, si esegue il programma e si decomprime il contenuto in una cartella selezionata dall'utente.

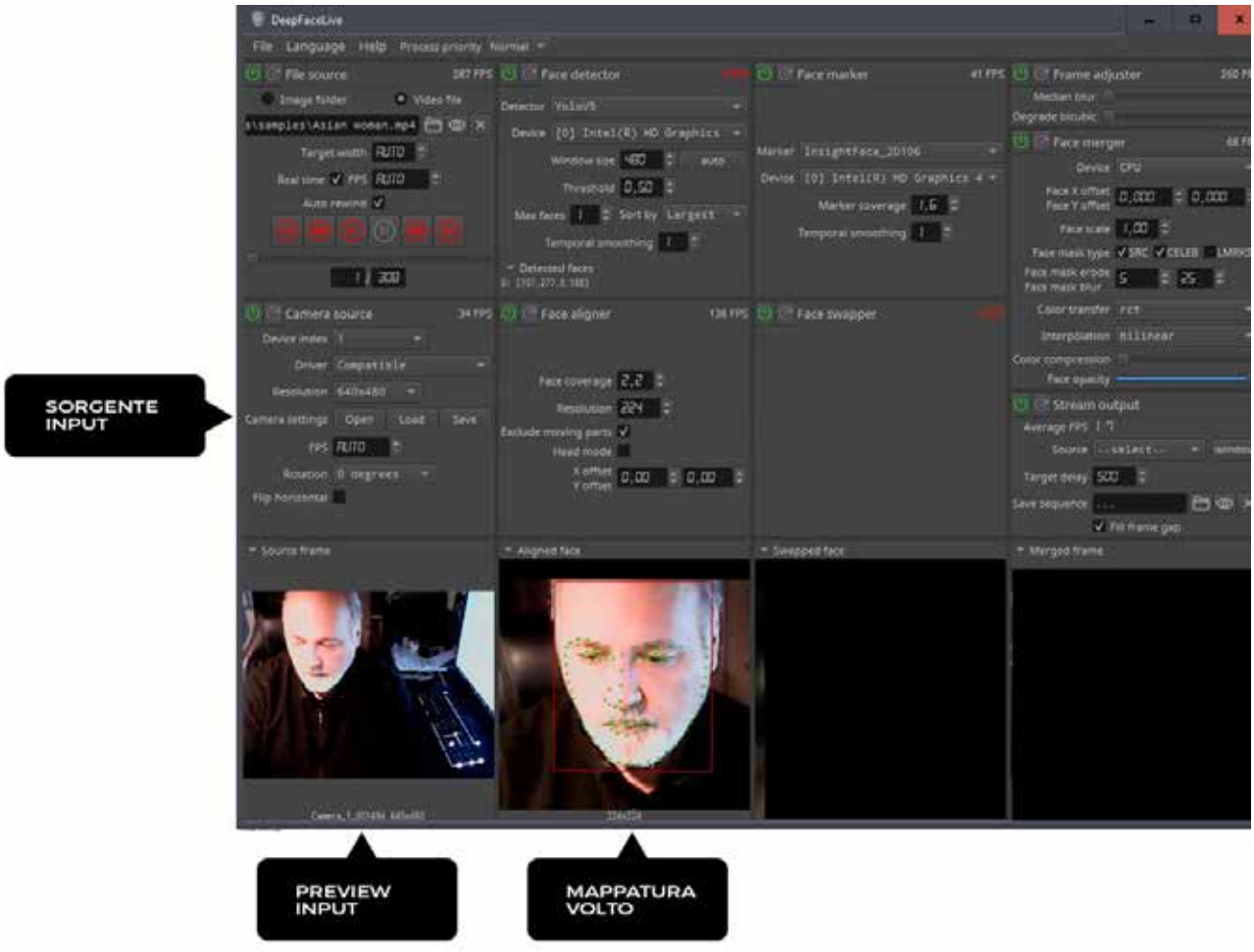
All'interno della cartella è presente il file DeepFacelive.bat:



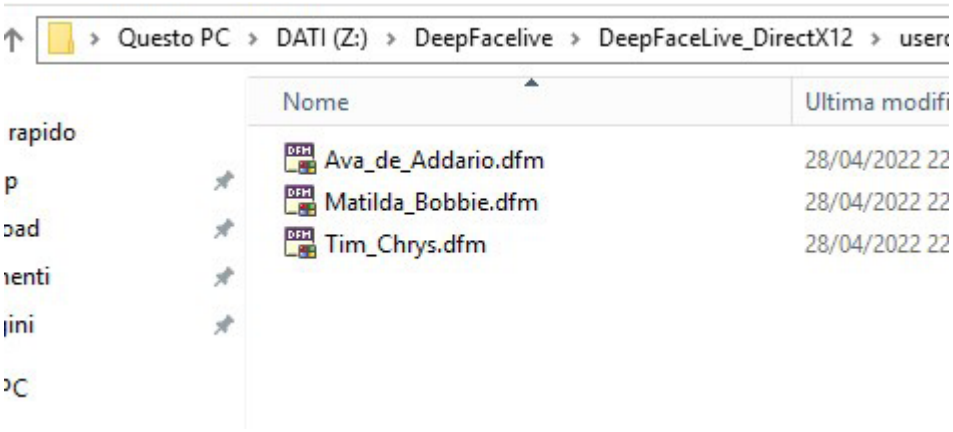
Per avviare il programma è sufficiente eseguire il DeepFacelive.bat, tuttavia prima di lanciare l'applicazione è necessario verificare che il computer abbia i requisiti hardware e software minimi, generalmente un computer con caratteristiche da gaming soddisfa i requisiti.
Una volta eseguita l'applicazione l'interfaccia appare sufficientemente intuitiva nelle funzionalità principali:



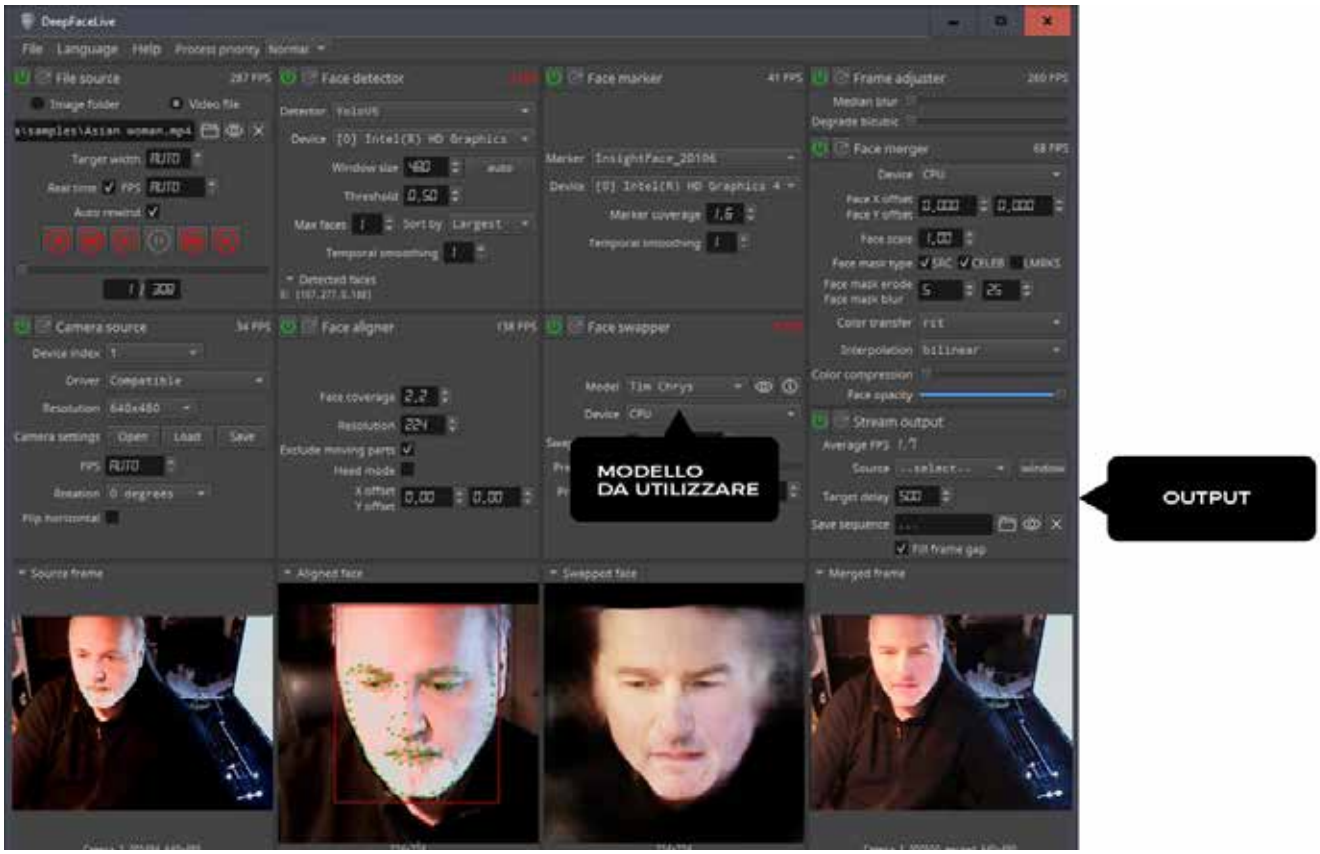
Selezionando la sorgente di input della videocamera si visualizza subito l'immagine nella preview dell'input e si attiva automaticamente la mappatura del volto per individuare i punti di allineamento:



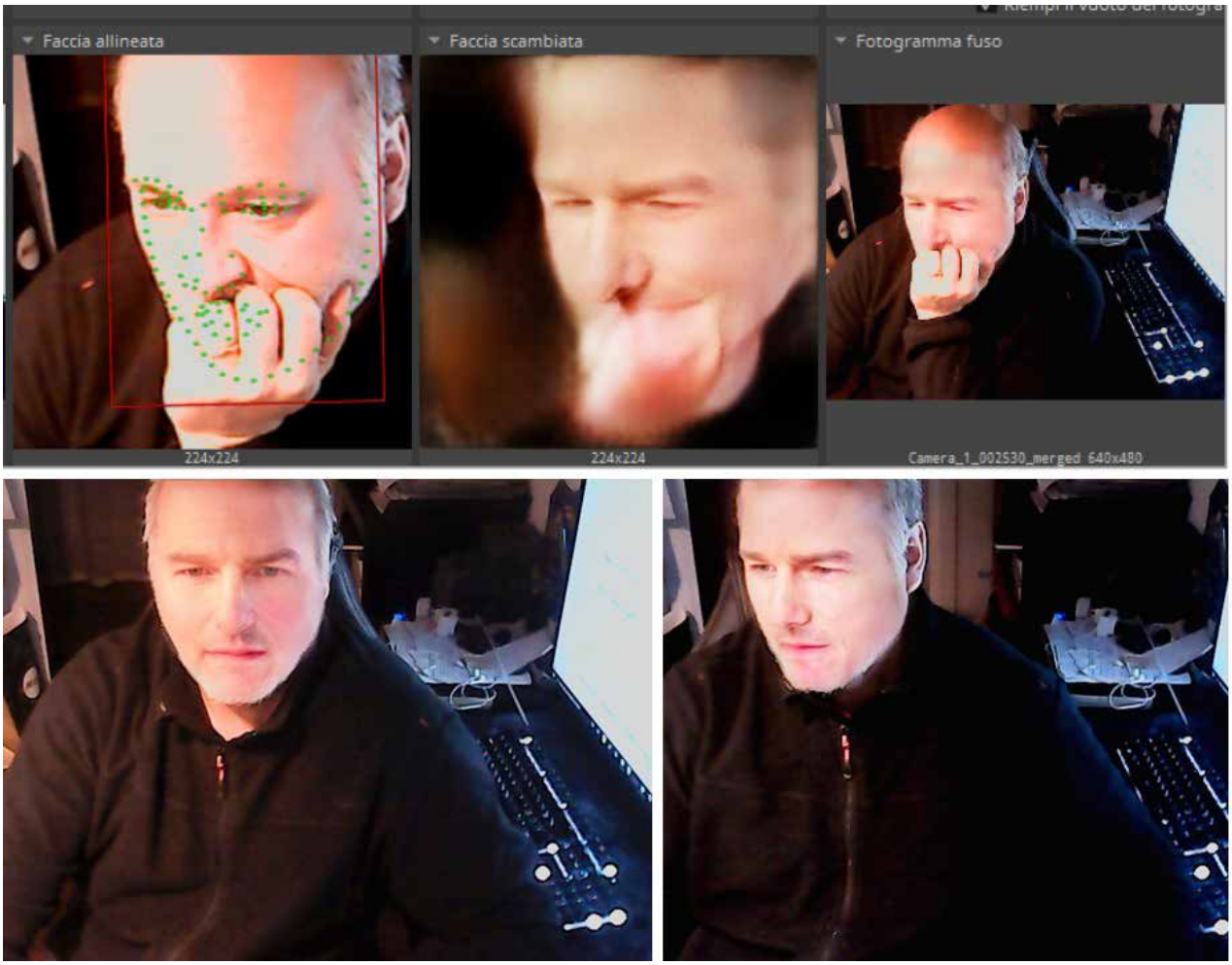
Il passo successivo è selezionare il modello da utilizzare.
L'applicazione dispone di modelli scaricabili, i modelli sono stati generati dall'applicativo DeepFaceLab e si possono aggiungere altri modelli semplicemente inserendoli nella sotto cartella userdata/dfm_models:



Dopo aver selezionato il modello da utilizzare, nel nostro caso l'attore Tom Cruise, DeepFaceLive comincerà ad effettuare immediatamente lo streaming applicando il face swap alla sorgente video. Selezionando l'output è possibile salvare il deepfake su file:



La facilità d'uso ed il livello qualitativo del deepfake prodotto risultano estremamente interessanti.



Soluzioni per il rilevamento e il contrasto del deepfake

L'emergere di deepfake sempre più sofisticati e realistici presenta maggiori potenziali minacce, sia per gli individui che per le organizzazioni in generale.

Se fino a qualche anno fa, il video e/o audio di un individuo che compiva una determinata azione o dichiarava determinate cose era una delle prove più evidenti che l'evento si fosse effettivamente verificato, con i deepfake si rischia di non poter più utilizzarli come prova.

I primi deepfake contenevano alcuni "bug" tecnici che permettevano ad un occhio umano di capire facilmente se il video poteva essere un falso e, siccome la manipolazione di deepfake di fascia alta è quasi sempre relativa a trasformazioni facciali, l'analisi si concentrava sul volto andando a verificare le seguenti principali anomalie:

1. **Battito delle ciglia:** nella realtà, gli esseri umani sbattono gli occhi ogni 2-10 secondi, ed ogni battito di ciglia richiede circa 2 decimi di secondo. Difficilmente quindi si trovano in rete foto di persone con gli occhi chiusi. Le reti neurali che generano i video deep fake non hanno quindi molto materiale da utilizzare per questo stato del volto, il risultato è che raramente nei primi video deep fake la vittima sbatteva le ciglia. Uno dei primi sistemi automatici di rilevamento di deepfake, analizzava i video utilizzando algoritmi in grado di identificare automaticamente e con precisione la posizione degli occhi nei video, tenendo traccia di ogni battito di ciglia. Questo "bug" è stato successivamente superato dagli sviluppatori di applicazioni deepfake di fascia alta, inserendo un battito di ciglia sintetico, anche se a volte questi algoritmi

generavano frequenze innaturali nello sbattere le palpebre provocando l'effetto contrario e indicando la presenza di una deepfakeness.

2. **Parti nascoste del volto**: nei deepfake generati da una GAN può capitare che parti del volto come la gola i denti non siano riprodotti fedelmente perché il materiale a disposizione per il soggetto non contiene dettagli sufficienti. Anche le orecchie possono risultare simmetricamente diverse e la presenza di accessori come ad es. un orecchino visibile in una sola parte, sono indizi di un probabile falso.

3. **Trasformazioni non naturali**: I deepfake possono aggiungere o rimuovere baffi, basette o barba, ma spesso non riescono a rendere completamente naturali le trasformazioni dei peli del viso. Il colore eccessivamente differente delle guance rispetto alla pelle del volto, la pelle troppo rugosa o troppo liscia, le dimensioni asimmetriche delle labbra rispetto al viso, sono tutti indizi di un deepfake.

Anche per queste tipologie di anomalie sono stati sviluppati algoritmi per identificare i falsi video ma l'evoluzione tecnologica, la capacità computazionale sempre maggiore dei sistemi, permette ai software che elaborano i deepfake di riuscire a realizzare immagini e video sempre più realistici in tempi sempre più brevi.

4. **Postura del volto**: è stato osservato che le espressioni facciali ed i movimenti del capo sono fortemente correlati fra loro, in maniera diversa da soggetto a soggetto. La mancanza di questa coerenza è indice di manipolazione. Questo approccio però richiede un database di immagini e video sul soggetto molto esteso e quindi si presta solo alla protezione di un numero ristretto di "*very important people*".

Secondo alcuni ricercatori del MIT, i deepfake di alta qualità non sono facili da rilevare, ma con la dovuta pratica ed esperienza le persone possono costruire uno spirito critico tale da intuire ciò che è falso da ciò che è reale.
Per mettere alla prova gli utenti, i ricercatori hanno creato una pagina Web dove le persone possono provare e addestrare le proprie abilità: detectfakes.media.mit.edu

Negli ultimi mesi però la crescente qualità di realismo dei deepfake e la loro modalità di diffusione attraverso Internet, in particolar modo sui social network, superano largamente le capacità di discriminazione degli esseri umani nel distinguere un vero da un falso.

Per affrontare questa sfida, un approccio consiste nel creare sistemi automatici di rilevamento basati sulla stessa tecnologia di intelligenza artificiale utilizzata per generare i deepfake. Per addestrare un sistema GAN a riconoscere un deepfake è necessario fornirgli un dataset di immagini e video contraffatte in modo che possa estrarre le caratteristiche per riconoscere le anomalie.

Il progetto **FaceForensics++** (github.com/ondyari/FaceForensics) mette a disposizione un set di dati forensi composto da oltre 3000 sequenze video originali che sono state manipolate dai più utilizzati programmi per generare deepfake e pubblicate su Youtube. Tuttavia, se queste reti GAN analizzano video manipolati per i quali non erano state addestrate (ad es. un nuovo software di deepfake che esegue il face swap con un algoritmo modificato) le prestazioni calano drasticamente e lo stesso accade su dati fortemente compressi. Ad esempio ciò accade quando si effettua l'upload di un video

su piattaforme di social network che rielaborano il video comprimendolo per motivi di ottimizzazione.

Un metodo alternativo, per migliorare la robustezza di queste reti e la loro capacità di adattamento a nuovi tipi di manipolazioni, sfrutta la medesima tecnologia ma da un punto di vista diverso. Si utilizzano reti progettate per la visione artificiale, in grado di superare le capacità di esperti umani, utilizzate in vari campi come ad es. la diagnostica medica nell'*imaging* radiologico.

Le reti vengono addestrate su dataset di immagini e video originali al fine di generare un *fingerprint* [2] (impronte digitali) dell'immagine e rilevare qualunque anomalia come possibile manipolazione, in particolare, in vari studi è stato dimostrato che le reti GAN lasciano nelle immagini che generano un marchio digitale caratteristico, specifico per ogni architettura, che può essere utilizzato per classificare le anomalie [3].
Questo metodo permette di rilevare anche deboli anomalie nelle immagini e sequenze video, perfino se caricate sui social network dove vengono ridimensionate e compresse, operazioni queste che ne diminuiscono la qualità.
La **tecnica delle fingerprint,** in pratica, elabora alla fine del processo una mappa colorata (heatmap) di ogni immagine e associa ad ogni pixel un valore (colore) legato alla probabilità che sia stato manipolato.

Esempio heatmap di un'immagine reale

Esempio heatmap di un'immagine modificata

I metodi di deep learning basati su fingerprint sono stati usati con successo per il rilevamento di immagini false, tuttavia, per quanto riguarda gli stessi metodi applicati direttamente per il rilevamento di video falsi la percentuale di successo cala significativamente a causa della disponibilità di una significativa perdita di informazioni sui fotogrammi dopo la compressione video.

I deepfake producono dunque artefatti, alcuni riguardano le incongruenze, le irregolarità sullo sfondo, le impronte generate dalle reti GAN e sono esempi di **artefatti spaziali**, mentre il rilevamento delle fluttuazioni nel comportamento di una persona, dei segnali fisiologici, della coerenza e della sincronizzazione dei frame video sono tutti esempi di **artefatti temporali**.

Un nuovo approccio basato sull'utilizzo di reti neurali per valutare diversi ambiti [4] è attualmente tra gli studi più promettenti per il rilevamento dei deepfake, oltre all'analisi "tecnica" dell'immagine analizzano altri **modelli di rilevamento come quello comportamentale e biologico**.

Nei volti sintetizzati, gli artefatti del segnale biologico forniscono evidenze per il rilevamento del falso. I segnali biologici analizzati sono suddivisi nei seguenti gruppi:

- incoerenza audio-visiva
- incoerenza visiva
- segnale biologico nel video

L'**irregolarità audio-visiva** nei DeepFake è un indizio molto importante, rileva le anomalie nella dinamica della forma della bocca (visemi) ed il fonema pronunciato. Ad es. le parole mamma, baba e papà sono fonemi i quali richiedono che le labbra siano completamente chiuse per essere pronunciati correttamente, se analizzando il video l'algoritmo rileva una incongruenza tra le parole nell'audio e la posizione delle labbra viene evidenziata un'anomalia.

La **mancanza di coerenza visiva** viene utilizzata per analizzare la forma, i tratti del viso, i punti di riferimento dei volti, per verificare se rispettano le regole di simmetria naturale attributi come occhi, denti e caratteristiche facciali. Gli artefatti visivi in genere hanno una mancanza di coerenza globale, dovuta ad una inadeguata illuminazione incidente e/o da una stima imprecisa della geometria effettiva.

I **segnali biologici nei video** hanno dimostrato di essere degli indicatori affidabili per il riconoscimento dei deepfake poiché difficili da replicare. Negli studi è stato dimostrato che la frequenza cardiaca è utile per rilevare i video deepfake. Estrarre la frequenza cardiaca dai video è un compito impegnativo.
Sfruttando l'equazione differenziale ordinaria neurale (Neural-ODE) uno studio [5] ha proposto un modello DeepRhythm in grado di evidenziare i segnali del ritmo cardiaco.

Anche nella **rilevazione dei deepfake audio** gli studi stanno sviluppando nuovi algoritmi simili. Tra i più recenti ed interessanti sicuramente c'è il lavoro fatto da Joel Frank e Lea Schonherr, dell'Horst Gortz Institute for IT Security presso la Ruhr-Universitat Bochum[6].
I due ricercatori hanno accumulato circa 118.000 campioni di registrazioni vocali audio sintetizzate per quasi 196 ore di registrazioni vocali false, sia in inglese che in giapponese.
Per garantire che il set di dati fosse diversificato, il team ha utilizzato sei diversi algoritmi di intelligenza artificiale durante la generazione dei frammenti audio, ogni file audio artificiale è stato poi confrontato con registrazioni di discorsi reali.
Questo confronto ha rivelato sottili ma significative differenze nelle alte frequenze tra file reali e falsi, tali da consentire la determinazione tra un file reale ed uno falso.

Sviluppi futuri e riflessioni

In conclusione, è evidente che il mondo della ricerca, grazie anche allo stimolo dei grandi player dell'Information Technology, sta investendo notevoli risorse economiche ed umane per sviluppare nuovi strumenti in grado di rivelare le manipolazioni dell'informazione.

Ma è altrettanto vero che la tecnologia deepfake è sicuramente diventata uno dei *driver* principali nella generazione di contenuti, si pensi ad es. al mondo dell'intrattenimento ed artistico dove diventa possibile realizzare l'editing di riprese video senza la necessità di ripetere il girato o ricreare artisti che non sono più con noi per eseguire le loro performance dal vivo.

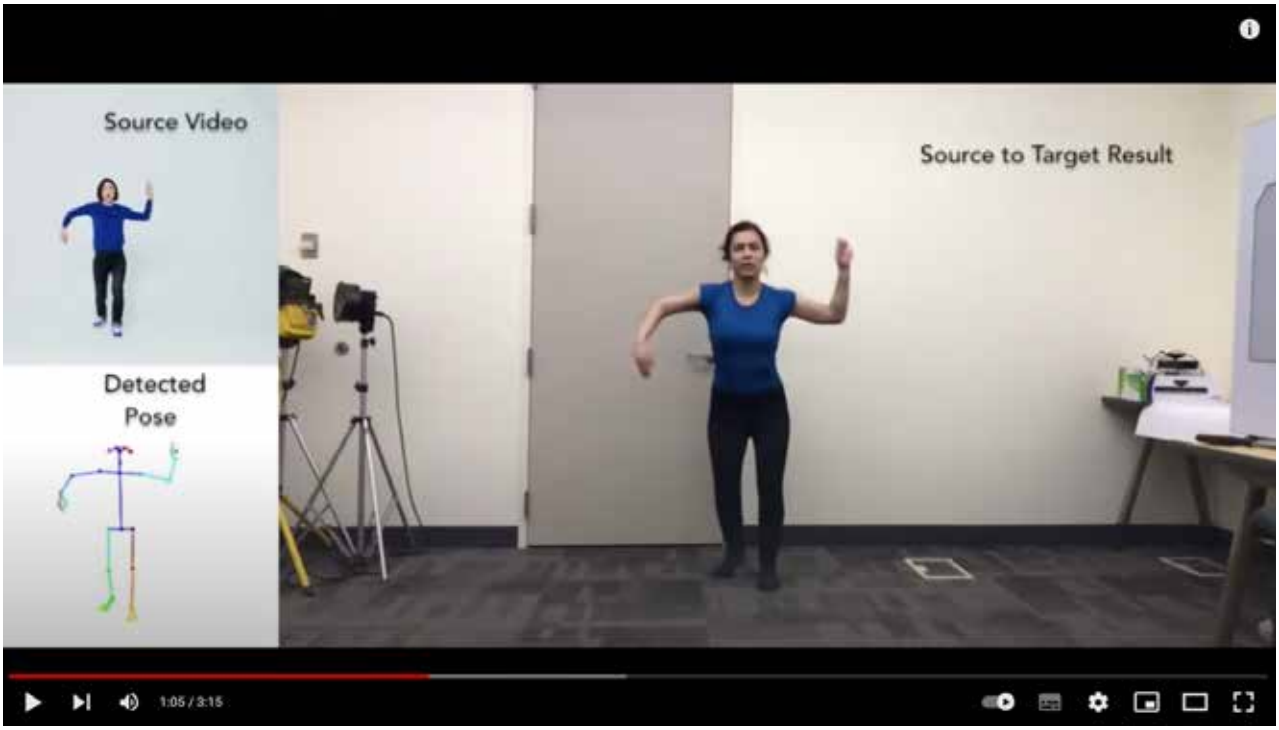
Il software VoCo di Adobe (ancora in fase prototipale), consentirà di produrre il parlato dal testo e modificarlo a proprio piacimento, potremmo ad es. scegliere di ascoltare un qualsiasi libro letto con la voce dei migliori narratori o personaggi famosi, Gigi Proietti, Orson Welles, etc.

VoCo renderà possibile prendere una registrazione audio e modificarla includendo parole e frasi che il narratore originale non ha mai detto.

Un esempio curioso di applicazione della tecnologia deepfake è l'App della Humen AI che crea deepfake per ballare, utilizzando una rete GAN è in grado di leggere i passi di danza di qualcuno e "copiarli" sul corpo di un'altra persona.
Con questa tecnologia è possibile far ballare chiunque.

Il sistema può essere utilizzato per tutti gli stili di danza, vengono registrati i video del

ballerino sorgente e del ballerino target, dopodiché, l'App esegue lo scambio (www.youtube.com/watch?v=PCBTZh41Ris).



La tecnologia Deepfake può avere molti vantaggi anche in altri ambiti, come la diagnosi medica ad esempio, dove può essere utilizzata per sintetizzare dati realistici ed aiutare i medici ed i ricercatori a sviluppare nuovi modi per curare malattie senza dipendere dai dati effettivi dei pazienti.

Un team di ricercatori della Mayo Clinic, del MGH & BWH Center for Clinical Data Science in collaborazione con NVIDIA, hanno utilizzato una rete GAN per creare risonanze magnetiche cerebrali sintetiche.

Il team ha addestrato la rete GAN su due dataset: uno contenente circa duecento risonanza magnetica cerebrali di pazienti affetti da tumore, l'altro contenente migliaia di risonanza magnetica cerebrale di pazienti affetti da Alzheimer.
Secondo i ricercatori, gli algoritmi addestrati con una combinazione di immagini mediche "false" e il 10% di immagini reali sono diventati altrettanto abili nello scoprire i tumori quanto gli algoritmi addestrati solo con immagini reali.

L'altra faccia dei deepfake, quella che non fa dormire sonni tranquilli, è la pericolosità di un suo uso illegale e malevolo.

L'anno scorso, più di 100.000 immagini deepfake raffiguranti nudi sono state generate dall'ecosistema di bot Telegram citato nel capitolo precedente.
Secondo un report realizzato della società di intelligence Sensity sulle minacce visive, la maggior parte delle immagini originali sembrava essere stata presa da pagine di social media o direttamente da comunicazioni private, persone probabilmente ignare di essere state prese di mira, gente comune ma anche celebrità e, cosa peggiore, un certo numero di immagini presentava come target minorenni lasciando ipotizzare che alcuni utenti utilizzassero il bot principalmente per generare e condividere contenuti pedofili.

Anche le agenzie governative sono particolarmente preoccupate sul fatto che i deepfake possano essere utilizzati per propagare disinformazione e condurre crimini.
Gli sviluppatori di deepfake hanno la capacità di far dire o fare alle persone tutto ciò che vogliono e successivamente divulgare i falsi contenuti manipolati online. Ad esempio, nel 2021, la commissione per gli affari esteri del parlamento olandese è stata indotta con l'inganno a condurre una video chat con qualcuno che impersonava Leonid Volkov, il dissidente politico che guida lo staff di Alexei Navalny, leader dell'opposizione russa e principale avversario di Putin.

C'è il potenziale non solo per distribuire notizie false, ma anche per causare disordini politici, un aumento della criminalità informatica, del *revenge porn*, scandali fasulli e un

aumento delle molestie e degli abusi online, inoltre in tribunale video presentati come prova potrebbero essere resi inutili dal utilizzo di tale tecnologia.

Eppure il futuro dei deep fake può essere davvero luminoso. Ma per garantire che le applicazioni false non vengano utilizzate in modo improprio dovremmo creare dei limiti, anche se, considerando la velocità del progresso tecnologico in questo campo, è difficile pensare ad una protezione totale.

Ci saranno sempre nuovi attacchi e nuovi sviluppi da inseguire.

E' quindi importante che cresca anche la consapevolezza delle organizzazioni e degli utenti finali, aumenti la loro attenzione alla gestione dei dati personali, non solo password e credenziali ma la loro stessa immagine, voce e video, la rappresentazione digitale del loro "IO" reale, ricordandosi sempre che come recita la frase tratta dal film "La migliore offerta" di Giuseppe Tornatore:

"In ogni falso si nasconde sempre qualcosa di autentico"

Note e Bibliografia

[1] Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward - 4 marzo 2021 - Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javad , Aun Irtaza,Department of Computer Science, University of Engineering and Technology-Taxila, Pakistan,Department of Computer Science and Engineering, Oakland University, Rochester, MI, USA Electrical and Computer Engineering Department, University of Michigan-Dearborn, MI, USA

[2] Extracting camera-based fingerprints for video forensics - in CVPR Workshops, 2019- D. Cozzolino, G. Poggi, and L. Verdoliva.

[3] Do GANs leave artificial fingerprints? - in 2nd IEEE International Workshop on Fake MultiMedia, 2019 - F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi.

[4] Deep Learning for Deepfakes Creation and Detection: A Survey - 6 febbraio 2022 - Thanh Thi Nguyena, Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-Thec,Saeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyeng

[5] Predicting heart rate variations of deepfake videos using neural ODE - Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), pp. 1721-1729, Oct. 2019 - S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic.

[6] WaveFake: A Data Set to Facilitate Audio Deepfake Detection - 4 Nov 2021 - Joel Frank, Lea Schönherr, Ruhr University Bochum

ICT Security MAGAZINE

ISCRIVITI ALLA NEWSLETTER

per ricevere aggiornamenti sulle prossime iniziative. Seguici sui canali social: Linkedin, Facebook, Twitter