Perché il deep fake preoccupa l’intelligence? Disinformazione e attacchi psicologici con l’uso illecito dell’IA

Author : Francesco Arruzzoli

Date : 3 Giugno 2019

"We're entering an era in which our enemies can make anyone say anything at any point in time", ovvero "stiamo entrando in un’era nella quale i nostri nemici possono far dire qualsiasi cosa, a chiunque, in qualsiasi momento", è il sottotitolo di un video presente su Youtube:

In tale video l’ex presidente degli Stati Uniti Barack Obama mette in guardia su possibili manipolazioni dell’informazione: l’unico problema è che **Obama non hai mai detto quelle**

parole, né ha mai fatto quello specifico video… ma per potervi spiegare il perché i *deep fake* preoccupino tanto le intelligence di mezzo mondo, dobbiamo fare un passo indietro e cominciare dall'inizio.

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, come post o link diretti), soprannominato "Deepfakes", comincia a pubblicare video di alcuni spezzoni di film dove i volti di attori vengono sovrapposti agli attori originali, realizzando così dei video fake, tutto sommato divertenti e molto realistici. Nel giro di poche settimane si passa, però, da divertenti e innocui video a spezzoni di video hard dove il volto dei pornoattori viene sostituito con quello di celebrità del mondo dello spettacolo, come ad es. il caso di Taylor Swift, in cui la faccia della cantante è stata inserita in diversi video fake a contenuto pornografico.

I video di Deepfakes diventano virali; l'accuratezza e la qualità dei fake prodotti è notevole, prima su Reddit, subito dopo su altri social network e, infine, su tutte le più importanti piattaforme pornografiche online. Il successo è tale che Deepfakes crea una "subreddit" - una specifica categoria all'interno di reddit - dove i post sono organizzati in categorie e argomenti, seguita da migliaia di utenti. Il successo dei post di Deepfakes non è solo virale per il "prodotto finale" che pubblica ma soprattutto perché, oltre ai video, alcuni utenti cominciano a postare link ad applicazioni che permettono a tutti di poter creare i propri video senza nessuna conoscenza di computer grafica e di intelligenza artificiale: già, perché la componente tecnologica alla base della produzione di queste tipologie di video si basa proprio sull'**intelligenza artificiale**.

Di per sé la manipolazione video dei volti denominata *"face-swap"* non è una novità, in quanto è una tecnologia utilizzata da tempo nell'ambito del mondo cinematografico dove a volte, a causa dell'impossibilita di poter girare delle scene con specifici attori (ad esempio, perché deceduti) il loro volto viene sovrapposto a quello di altri attori, come accaduto all'attore Peter Cushing, morto nel 1994 e redivivo nei panni del comandante della Morte Nera nell'episodio *Rouge One* di Star Wars del 2016, grazie all'elaborazione grafica del volto dell'attore Guy Henry.

Prima di Deepfakes, quindi, la possibilità di eseguire dei video credibili di *face-swap* rimaneva appannaggio esclusivo di professionisti della computer grafica che avessero a disposizione strumenti e software potenti e costosi. I *deep fakes* rappresentano un primo **salto evolutivo del *"face-swap"*** perché permettono a utenti senza alcuna esperienza in computer grafica e intelligenza artificiale (ma con una buona scheda video e un po' di tempo a disposizione) di utilizzare applicativi in grado di realizzare video fake di ottima qualità: ed è così che la parola *"deep fake"* diventa sinonimo di video falsi prodotti con l'intelligenza artificiale.

La star delle applicazioni di *deep fakes* diviene **Fakeapp**, gratuita e scaricata più di 150.000 volte, in grado di ottenere un *face swapping* accurato, privo di elementi di manipolazione. Uno dei suoi primi "elaborati" pubblicati è stato un video fake che ritraeva Michelle Obama mentre faceva uno *striptease*.

A dicembre del 2017 il caso *deep fakes* esplode: un articolo del magazine online **Motherboard** mette in luce l'uso della tecnologia, i rischi relativi alla manipolazione dell'informazione e i rischi di potenziali attacchi anche di natura politica; l'account dell'utente Deepfakes viene bannato da Reddit.

Il **2018** è stato caratterizzato da una grande attenzione da parte dei giornalisti al fenomeno *deep fakes* e ai suoi potenziali pericoli, in particolare la preoccupazione che potessero essere prodotti video fake costruiti *ad hoc*, veicolo di specifici messaggi, in grado di influenzare persone con ideologie radicalizzate e farle cadere in "trappola" come avviene con i *bias* cognitivi (pregiudizi).

A distanza di due anni dalle apparizioni dei primi video fake su Reddit la minaccia, però, non si è ancora materializzata. Twitter e Facebook nel frattempo hanno smascherato migliaia di account falsi di troll e haters creati per condurre campagne di vario tipo, ma non è mai stato

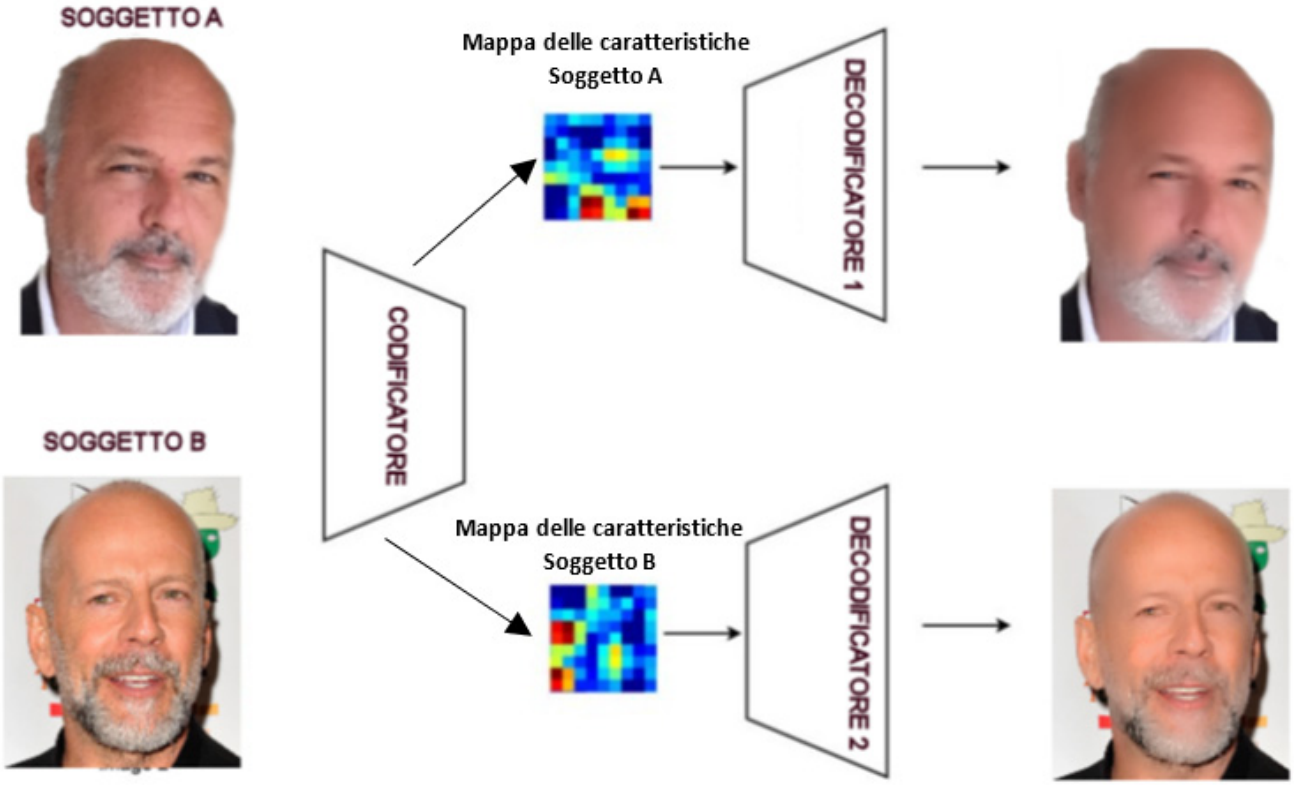
prodotto un solo video *deep fake* specificatamente progettato per manipolare un determinato tema politico; altresì l'eccesiva attenzione dei giornalisti sul tema dei video *deep fakes* ha finito per ingenerare un alone sinistro sui sistemi di intelligenza artificiale.

La tecnologia alla base del *deep fake*

La prima tecnica utilizzata nei video *deep fake*, come dicevamo, sfrutta il *face-swap:* celebri i video falsi con il volto dell'attore Nicolas Cage sovrapposto a quello di altri personaggi.

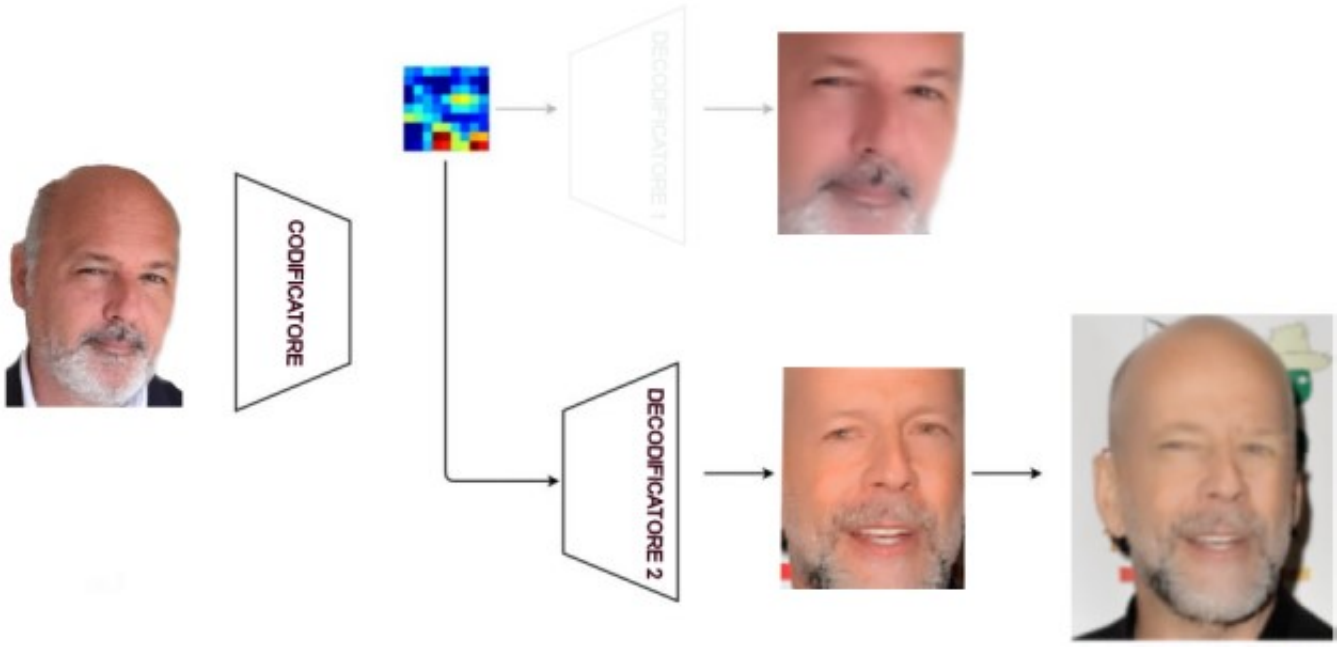

https://youtu.be/dh-QM54RuAs
https://youtu.be/DIZf7eRlD4w
https://youtu.be/v0zFR0ElRd4

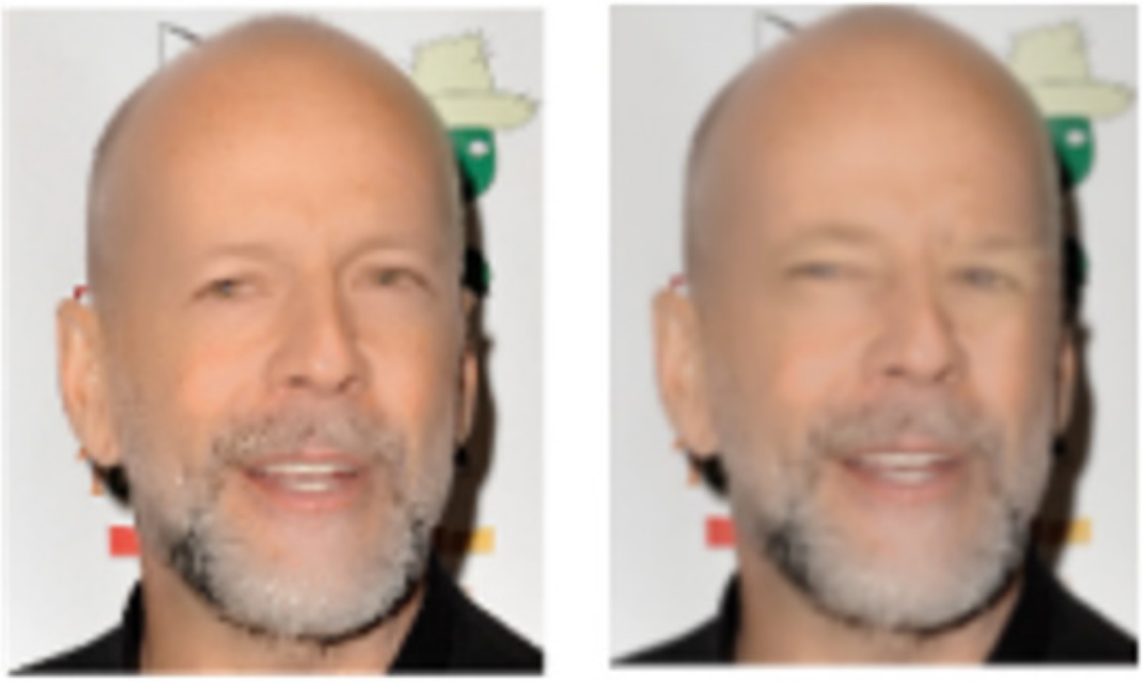
La tecnica *face-swap* richiede una **base dati** di centinaia o migliaia di immagini di entrambi i soggetti che oggi, grazie a internet, non è così difficile da reperire. Una volta popolata la base dati le immagini vengono codificate tramite un "encoder", un codificatore che elabora tutte le immagini utilizzando una rete neurale convoluzionale CNN (*ConvNet - convolutional neural network*). Lo scopo dell'encoder è individuare le peculiarità dei volti - deve cioè estrarre le caratteristiche più significative per ricreare l'input originale, nelle varie espressioni e angolazioni - non memorizzare tutte le immagini acquisite. Per decodificare le caratteristiche viene poi utilizzato un "decoder" o, più specificatamente, un decoder per ogni persona.



Una volta impostati dati, codificatore e decodificatori si deve avviare una procedura di **addestramento della rete neurale** utilizzando l'algoritmo di *backpropagation*, che confronta il valore in uscita dal sistema con il valore desiderato (obiettivo). Ho provato sia l'applicazione Fakeapp sia un'applicazione simile denominata DeepFacelab (che consiglio a chi si volesse cimentare nei test), nel mio caso l'addestramento è durato circa 4 giorni (con un PC dotato di una buona scheda grafica con GPU), durante i quali la rete neurale ha effettuato più di 10 milioni di elaborazioni.

Completato l'allenamento, si procede alla sovrapposizione del volto del soggetto A con quello del soggetto B. Il video viene elaborato fotogramma per fotogramma: il volto del soggetto A viene inserito nell'encoder che però, invece di alimentare il decodificatore 1 del soggetto A, alimenta il decodificatore 2 del soggetto B per ricostruire l'immagine finale.





ORIGINALE FAKE

Per quanto riguarda il mio esperimento, l’effetto finale è stato ottimale solo con il primo piano del volto; la vista prospettica non era perfetta, probabilmente a causa di una base dati iniziale non completa e soddisfacente. È infatti necessario preparare migliaia di immagini di ottima qualità, in diverse pose per entrambe le persone, e numerose riprese video per poter ottenere un video realistico.

L’evoluzione, il salto della specie e il reale pericolo

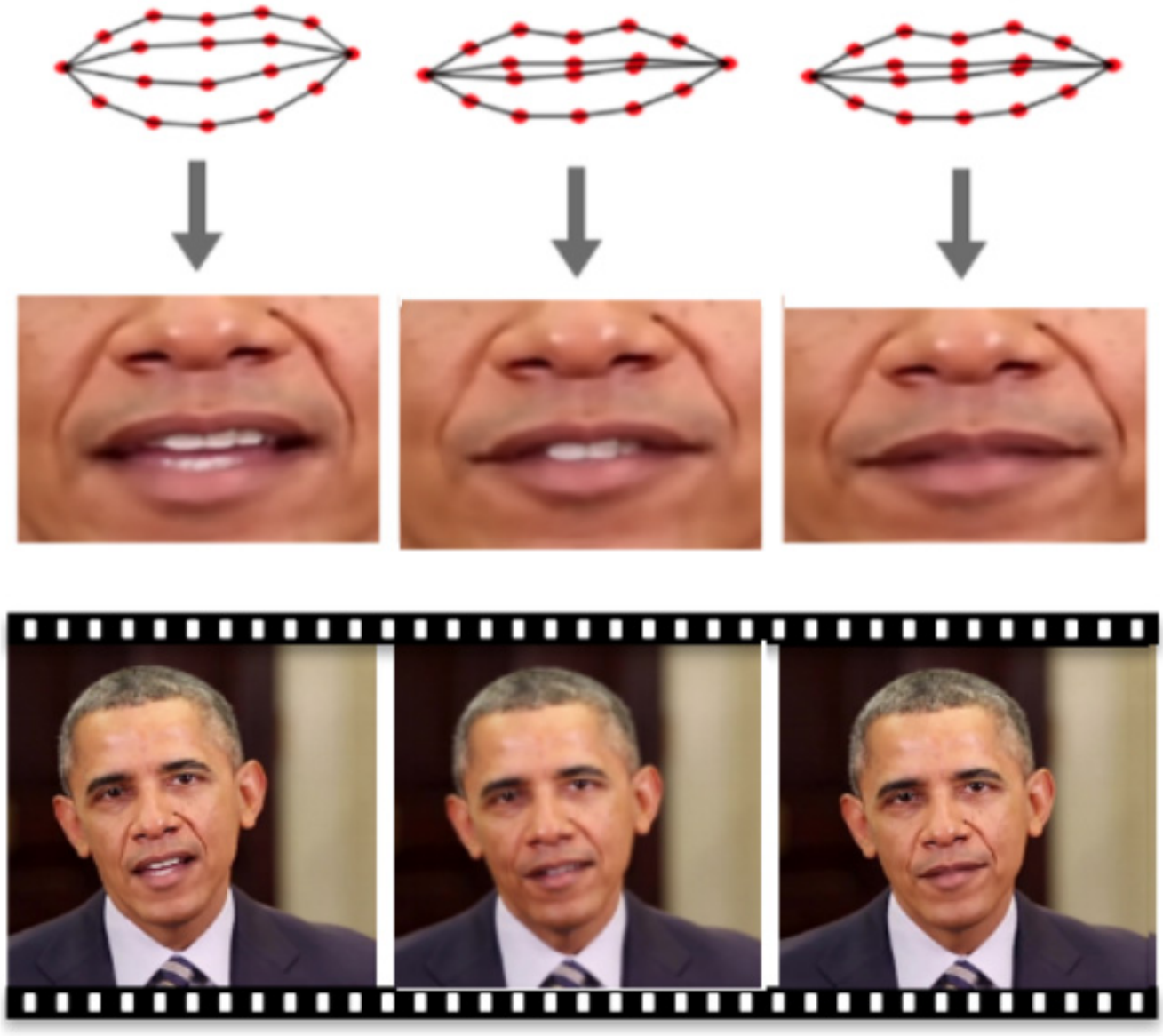
Nell’ultimo anno, però, il *deep fake* ha fatto un notevole balzo in avanti nella sua evoluzione, impiegando sempre meglio l’utilizzo dell’intelligenza artificiale e applicando nuove metodologie di elaborazione delle immagini. Il video di Obama all’inizio di questo articolo è stato realizzato dal regista Jordan Peele, ed è uno dei più difficili da identificare come falso. La **tecnica** utilizzata in questo caso è diversa da quella che ha reso famosi i primi video *deep fake*. Qui le sorgenti dati non appartengono a soggetti diversi ma allo stesso soggetto: se si analizza nel dettaglio il video, il labbro inferiore di Obama è più sfocato rispetto alle altre parti del viso. Invece di scambiare il volto, in questo caso viene sovrapposta sempre la bocca di Obama, rielaborata in sincronia labiale con un **audio falso**.

La creazione di un falso audio che imita la voce del soggetto target oggi è possibile in maniera rapida, veloce ed economica anche grazie ad aziende come Lyrebird, che ha sviluppato l'omonima applicazione in grado di imitare - con un grado di accuratezza stupefacente - la voce di qualunque persona. La tecnologia della Lyrebird, insieme a quella sviluppata all'Università di Washington (UW), permette quindi di sincronizzare il movimento labiale con un audio e di elaborare un **video in cui il soggetto “parla” con un audio ri-elaborato e mai veramente pronunciato.**

Sorgente audio:

Sorgente video:

encoding, ri-elaborazione, sincronizzazione delle labbra e output finale:



Un ulteriore salto di qualità nei deep fake si ha quando nuove classi di algoritmi di intelligenza artificiale usate per l'apprendimento automatico non supervisionato, dette **reti antagoniste generative** (*generative adversarial networks* - GAN), vengono utilizzate per creare immagini sempre più realistiche. Le GAN introducono un *deep discriminator* di rete (un classificatore CNN) per distinguere se le immagini facciali siano originali o create dal computer. Fornendo immagini reali al discriminatore, il sistema impara a riconoscere meglio le immagini. Il discriminatore a sua volta addestra gli encoder a catturare le peculiarità sempre più realistiche delle immagini e i decoder a generare, oltre alle immagini, anche delle maschere. Le maschere che vengono costruite con i dati di allenamento mascherano meglio l'immagine, creando una transizione più fluida del volto di destinazione. Questa continua analisi, ripetuta milioni di volte, genera alla fine immagini così realistiche che non sono più distinguibili da quelle reali.
Questa **evoluzione della tecnica** porta a controllare anche altre parti del volto: non più solo le

labbra ma anche occhi ed espressioni facciali. Uno degli esempi più esplicativi è il video realizzato da un team di studenti di un'università tedesca, che sono riusciti a far sorridere anche Putin...

Ma quello che più stupisce e inquieta è **l’interazione in tempo reale con il video**: le espressioni del loro volto vengono “imitate” dal clone riprodotto artificialmente e questo vuol

dire, ad esempio, che in un'ipotetica intervista in videoconferenza con dei giornalisti, il clone del presidente Trump potrebbe rispondere in tempo reale (con la sua voce) alle domande ma con le parole e l'espressione di un'altra persona, oppure un redivivo Bin Laden potrebbe apparire in video e parlare cordialmente con George Bush, entrambi fake controllati come burattini virtuali, dichiarando che l'11 settembre era tutto un complotto. Il tutto fatto prendendo video e foto da internet, senza particolari tecnologie né strumenti costosi e, soprattutto, senza nessuna particolare capacità tecnica... **un incubo vero e proprio** per chi combatte la disinformazione.

Ed è a questo punto che l'attenzione delle intelligence di tutto il mondo si è focalizzata con particolare attenzione sulla problematica dei *deep fake*. Provate a immaginare il seguente scenario: un gruppo di cyber terroristi trasmette un video dove il leader della Corea del Nord Kim Jong-un, in collegamento diretto con alcune TV nazionali, dichiara che nell'arco di un'ora scatenerà un attacco nucleare contro la Corea del Sud. Il video verrebbe trasmesso contemporaneamente in diretta sui social network e ripreso dai principali *mainstream*. Sarebbe il caos: l'intero network mondiale dell'informazione lo diffonderebbe e lo amplificherebbe esponenzialmente ed eventuali immediate smentite non farebbero altro che aumentare dubbi e paure. Probabilmente la Corea del Sud non si difenderebbe attaccando per prima ma la notizia getterebbe comunque la popolazione di entrambi gli stati nel panico più totale, paralizzandoli per giorni.

La prima struttura di intelligence a muoversi formalmente per contrastare attacchi basati sul *deep fake* è stata il DARPA (l'agenzia nordamericana di ricerca per la difesa) che, per trovare soluzioni in grado di mitigare il gli attacchi a "effetto sorpresa", ha impiegato un primo **investimento di 68 milioni di dollari** per lo sviluppo di tecniche in grado di identificare automaticamente i video manipolati, nonché un programma denominato *DARPA's Media Forensics* (MediFor) che coinvolge diversi centri di ricerca e università nel mondo per lo studio di soluzioni anti-*deep fake* e fake news.

Nel 2018, menzione d'onore all'Italia: l'Università Federico Secondo di Napoli, che studia il fenomeno dei *deep fake,* si è rivelata la più "prolifica" nelle pubblicazioni di studi e approfondimenti in merito alla problematica. L'esecutivo statunitense si è formalmente mosso attraverso il senatore Ben Sasse che ha firmato, a dicembre 2018, il "Malicious Deep Fake Prohibition Act", disciplinando implicazioni giudiziarie e responsabilità connesse alla creazione e distribuzione di contenuti fraudolenti. Anche nei meandri di internet le agenzie per la sicurezza si stanno muovendo informalmente, eseguendo una vera e propria "epurazione" dei codici software pubblicati dagli utenti per generare *deep fake* e disinformazione. Sappiamo per esperienza che nel dominio cibernetico gli attacchi dell'avversario sono comunque destinati, prima o poi, ad avere successo: quindi uno degli obiettivi da perseguire è anche quello di complicare almeno la vita ai potenziali attaccanti.

Uno studio della DEEPTRACE sul fenomeno - effettuato a fine 2018 - mostrava l'andamento della pubblicazione di applicazioni e codici per realizzare *deep fake* inizialmente in crescita esponenziale, poi stabilizzatosi negli ultimi mesi del 2018 per iniziare, successivamente, una decrescita nel primo semestre del 2019. Molte applicazioni e codici pubblicati non sono più disponibili, altri sono disponibili solo nelle prime versioni spesso altamente instabili (ad esempio la versione 2.2 dell'applicazione FakeApp è molto difficile da trovare, mentre la versione 1.1 è

presente su diversi siti di download), altre ancora sono state infettate e quindi, oltre a non funzionare, sono pericolose da utilizzare.

Cumulative number of code commits to the most popular deep learning Github repository for performing face swaps in videos, *deepfakes/faceswap*:

Fonte: DEEPTRACE Report – 2018

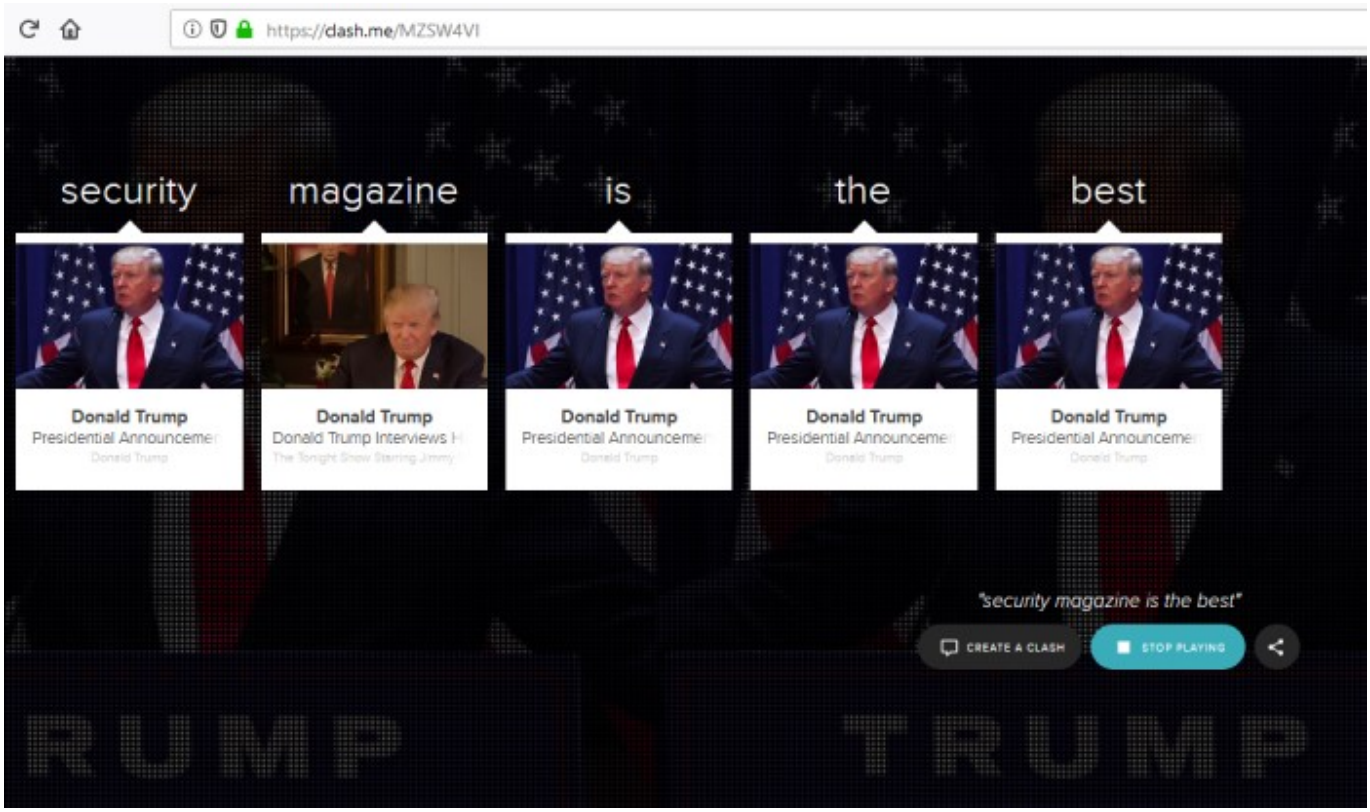
Al momento **non esiste un software in grado di identificare in maniera certa un *deep fake*:** sono state però implementate delle tecniche che sfruttano alcune “mancanze” temporanee di informazione, come il fatto che gli esseri umani sbattono gli occhi ogni 2-10 secondi, ogni battito di ciglia richiede tra uno e quattro decimi di secondo e in rete non si trovano facilmente foto di persone a occhi chiusi. Così le reti neurali dei video *deep fake* hanno poco materiale a cui attingere e, di conseguenza, raramente nei video *deep fake* il protagonista sbatte le palpebre. Facendo analizzare il video ad altre applicazioni, anch'esse basate su I.A. e specializzate nell'individuare il “battito di ciglia”, si riesce attualmente - nel 95% dei casi - a individuare video *deep fake*: ma questa, come si può immaginare, è solo una vittoria temporanea.

Nonostante l’azione di epurazione in atto su codici e applicazioni *deep fake*, esiste una quantità incredibile di pubblicazioni e materiale sull’argomento.
Ho voluto **provare a realizzare un semplice editore di *deep fake*** partendo da zero e utilizzando delle librerie neurali in cloud già pronte e a disposizione sulla piattaforma Azure di Microsoft per manipolare video. In questo caso l’algoritmo utilizzato suddivide il video sorgente in tante piccole sequenze separandole in base alle espressioni o al movimento del viso, che vengono poi associate al movimento di un utente ripreso da una telecamera: il risultato è un video collage di frame che l’I.A. utilizza per imitare i movimenti del volto dell’utente, senza modificare il movimento labiale ma solo cercando i frame più “simili” alla sincronizzazione dell’audio.

Per la fonte audio ho utilizzato il sito https://clash.me, che applica una tecnica molto semplice simile a quella utilizzata per il video: isola le parole in un audio e poi permette di creare nuove frasi con la voce originale del soggetto.

L’audio che ho realizzato lo trovate a questo link:

https://clash.me/MZSW4VI



Il mio video *deep fake* finale, il soggetto del video (il presidente Trump) che fa i complimenti alla nostra mitica rivista:

https://vimeo.com/337776890

Un esperimento rozzo ma divertente che rende l’idea delle potenzialità coinvolte. L’intelligenza artificiale è una componente fondamentale nel processo dei *deep fake* e nei prossimi anni lo sarà sempre di più: in questo e in moltissimi altri ambiti, è difficile a dirsi ma estremamente probabile come ciò possa concretizzare minacce in Internet. Tra i possibili rischi veicolati da un uso illecito dell’IA (*Malicious Use of Artificial Intelligence* - MUAI) vi è appunto la **destabilizzazione psicologica attraverso la disinformazione** manipolata dall’I.A., in grado di operare a livelli di complessità e velocità di esecuzione che la mente umana non è in grado di comprendere: questo potrebbe condurre facilmente anche a conflitti di ordine mondiale.

Per certi versi, i *deep fake* hanno evidenziato ancora una volta che viviamo un’epoca in cui i contenuti multimediali e l’informazione non controllata possono avere un impatto devastante sul comportamento delle persone, accentuando in modo particolare quanto sia sempre più vicino il momento in cui avverrà il **sorpasso della macchina** nei confronti dell’essere umano.

Per ora l'I.A., con i *deep fakes,* è arrivata a confutare anche Eraclito, che affermava: *"Gli occhi sono testimoni più precisi delle orecchie".*

Note

1. https://www.govinfo.gov/content/pkg/BILLS-115s3805is/pdf/BILLS-115s3805is.pdf
2. https://s3.eu-west-2.amazonaws.com/rep2018/2018-the-state-of-deepfakes.pdf
3. https://arxiv.org/pdf/1803.09179.pdf
4. https://github.com/deepfakes/faceswap
5. https://github.com/iperov/DeepFaceLab

Articolo a cura di **Francesco Arruzzoli**