LUISS

Corso di laurea in Economa e Management

Il Deepfake: un innovativo veicolo di diffusione di fake news

Relatore

Prof. Pietro Falletta

Candidato

Danilo Fratangeli

(matr. 238641)

Anno accademico 2022/2023

Sommario

Capitolo 1. Definizione, storia, funzionamento ed utilizzo del deepfake

1.1 definizione data dal Garante per la Protezione dei Dati Personali

Il Garante per la Protezione dei Dati Personali definisce i deepfake come "foto, video e audio (i) creati grazie a software di intelligenza artificiale (ii) che, partendo da contenuti reali (iii), riescono a modificare o ricreare (iv), in modo estremamente realistico (v), le caratteristiche e i movimenti di un volto o di un corpo e a imitare fedelmente una determinata voce".

i. Per stabilire cosa sia un deepfake bisogna partire dunque dal tipo di output proposto. Di fatti la definizione del Garante individua i contenuti deepfake come foto, video o audio. A differenza di altri metodi di disinformazione, quali ad esempio le più generiche fake news, i deepfake consistono in contenuti che ritraggono effettivamente il soggetto in una situazione diversa da quella reale. Non si tratta infatti di un testo ma di un file audio-visivo che quindi crea una falsa sensazione di autenticità e credibilità.

ii. I programmi per la creazione dei deepfake devono dunque sfruttare l'Intelligenza Artificiale (Artificial Intelligence, comunemente abbreviato in AI). L'intelligenza artificiale è quella disciplina, appartenente all'informatica, che studia i fondamenti teorici, le metodologie e le tecniche che permettono di progettare sistemi hardware e sistemi di programmi software capaci di fornire all'elaboratore elettronico delle prestazioni che sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana, come la capacità di apprendimento, di ragionamento e di problem solving. Grazie alle tecniche di deep learning e di apprendimento automatico, i software di intelligenza artificiale possono analizzare grandi quantità di dati e di immagini, imparare a riconoscere pattern e a manipolare l'immagine o il suono in modo da creare un contenuto che sembri autentico o verosimile.

Pertanto, la creazione di un deepfake non può essere ottenuta semplicemente modificando una foto con programmi come Photoshop o simili, oppure utilizzando un video o un audio in maniera tale da invertire il senso di una scena do di un discorso.
Affinché possa effettivamente parlarsi di deepfake sarà necessario che i file impiegati per la produzione dell'output finale vengano processati da un software che sia in grado di imitare l'intelligenza umana. Questo perché l'obiettivo principale dei deepfake è quello di creare contenuti falsi e ingannevoli che possano trarre in inganno le persone e influenzare la loro percezione della realtà. La manipolazione di immagini, video e audio tramite tecniche di intelligenza artificiale consente di ottenere risultati molto realistici e difficili da riconoscere come falsi.

iii. I deepfake partono da contenuti reali, e dunque necessitano di file che fungano da informazioni di partenza, quali foto, video, o audio (a seconda della tipologia di deepfake che si intende elaborare). Da questi input verrà ricavato l'output finale. Tali contenuti sono definiti come "contenuti reali" poiché si riferiscono a file effettivamente esistenti, di persone realmente esistenti. L'obiettivo principale dei deepfake è quello di mostrare persone, principalmente persone famose, in contesti insoliti o mente rilasciano dichiarazioni false o assurde. Non si tratta di deepfake nel caso in cui vengano utilizzati volti fittizi o voci non ben riconoscibili.

iv. Lo scopo del software dopo l'immissione dei dati è quello dell'elaborazione di questi ultimi al fine di produrre nuovi output modificati o anche ben diversi da quelli di partenza. La definizione specifica le tre categorie principali di prodotti deepfake: foto, video e audio vocali. Tra le tre, i video sono quelli che attirano maggiormente l'attenzione, in quanto risultano essere più coinvolgenti e spesso meno sospetti. Approfondiremo meglio sia il funzionamento sia la tipologia di risultati ottenibili più avanti in questo capitolo.

v. Ultimo fattore fondamentale che si ricava dalla definizione data dal Garante per la protezione dei dati personali è che un deepfake deve essere realistico.

Ciò richiede una grande attenzione ai dettagli ed un alto livello di precisione nell'elaborazione dei dati di partenza. Nel caso dei video, ad esempio, l'AI deve essere in grado di analizzare e replicare con precisione i movimenti, le espressioni e le sfumature del viso delle persone coinvolte nel video originale. Questo richiede una notevole capacità di elaborazione e l'impiego di algoritmi sofisticati per ottenere un risultato realistico e credibile. In sintesi, si tratta di uno degli elementi chiave dei deepfake, poiché rappresenta il mezzo per raggiungere l'obiettivo principale di questo tipo di contenuti: risultare realistico, verosimile e coerente, per lo meno a livello visivo. Non è infatti necessario che il video sia di per sé realistico e che venga scambiato per autentico dagli spettatori; ciò, di fatto, escluderebbe dalla categoria la maggior parte dei video ironici. È sufficiente che sia verosimile e fluido nella propria componente visiva. Evitando il più possibile sbavature, imperfezioni e incoerenze nella sovrapposizione dei volti.

1.2 Introduzione ai deepfake: i tre tipi di video

Dalla definizione data si giunge ad una pluralità di informazioni tecniche per la definizione di deepfake ma non ad una vera e propria visione di insieme. È necessario quindi indagare più a fondo.
Come abbiamo già detto, la forma più diffusa di deepfake è quella dei video, probabilmente poiché si tratta di quella maggiormente intrattenente e realistica. Da sempre è possibile imitare una voce, da anni ormai è possibile ritoccare una foto, ma solo da pochi anni è possibile (a patto di avere delle conoscenze informatiche adeguate) modificare i soggetti dei video. Pertanto, quando si parla di deepfake quasi sempre ci si riferisce ai video. Più specificatamente, ci si riferisce a video che presentano il volto di una persona sovrapposto a quello di qualcun altro.

I motivi per cui vengano ritratti personaggi famosi sono intuibili con facilità.
In primo luogo, è più facile trovare foto e video di attori, cantanti, celebrità o anche di politici su internet. Per creare un buon video sarà necessaria una grande mole di fotogrammi del volto da sovrapporre.
Ritrarre celebrità rende inoltre il video più interessante e spesso virale; il deepfake è infatti un fenomeno che, come vedremo avanti, ha trovato la sua fama grazie ad internet.

Esistono vari tipi di video, ma sono tre le categorie principali di video deepfake: i video umoristici (i), le interviste fake (fake news) (ii) ed i video porno o deep porn (iii)

1.2.1 I video ironici o meme

I primi sono video meme: *meme* è un neologismo coniato dal biologo inglese Richard Dawkins nel 1976, nell'undicesimo capitolo del libro *The Selfish Gene* per indicare i «nuovi replicatori», ovvero replicatori di natura culturale, non biologica. Il termine è un'abbreviazione, su modello di *gene*, di *mimeme* (dal greco antico *mivmhma*, 'imitazione, copia'). Il meme è una "unità di trasmissione culturale" o, in altre parole, un 'gene culturale'. I meme di internet (Internet Memes) sono testi appartenenti a diverse sostanze espressive (immagini, statiche o animate, comprendenti porzioni di testo, e video), derivati dalla segmentazione, manipolazione, risemantizzazione di testi preesistenti in base a regole di pertinenza e corretta formazione, dotati di un qualche significato collettivamente riconosciuto, a carattere ludico (umoristico, parodico, satirico).

Tra i meme possono dunque rientrare i video deepfake con finalità umoristica dal momento che può ben trattarsi di video che derivano dalla manipolazione di contenuti preesistenti. Un esempio tra i più celebri su internet (con all'incirca 7,8 milioni di visualizzazioni solo su YouTube) è la parodia del film 'Mamma ho perso l'aereo' (titolo originale: Home Alone) nella quale viene sovrapposto al volto del giovane protagonista, Macaulay Culkin, il volto dell'attore Sylvester Stallone; il titolo del video sfrutta un gioco di parole: "Home Stallone" https://www.youtube.com/watch?v=2svOtXaD3gg .

1.2.2 Le fake news

La seconda tipologia di video deepfake è costituita dai video di interviste o dichiarazioni false. Nella stragrande maggioranza dei casi, i protagonisti di questa categoria di deepfake sono politici e questi video possono a loro volta rientrare nella categoria dei memi. La differenza sostanziale con i video memi sta nel fatto che, a differenza dei video umoristici assurdi, le dichiarazioni false possono essere diffuse come fake news.

È facilmente intuibile che un video in cui l'ex premier Matteo Renzi sostituisce l'attore Toni Servillo nel ruolo di Gep Gambardella nel film 'La grande bellezza' sia solo un

video ironico. Ben diversa è la situazione nel caso in cui lo stesso soggetto venga ritratto nel suo ufficio a fare dichiarazioni assurde ma plausibili, come nel caso del video trasmesso da *Striscia la notizia* il 23/09/2019 intitolato 'Il fuorionda di Matteo Renzi". Nonostante questo possa essere considerato un video ironico, il cui scopo non era quello di screditare il soggetto rappresentato ma di fare satira, il video è stato scambiato per vero da parte di molti telespettatori. È importante specificare che tutt'oggi al 24/03/2023 il video è presente sul sito ufficiale di *Striscia la notizia* (https://www.striscialanotizia.mediaset.it/video/il-fuorionda-di-matteo-renzi_59895/) con lo stesso titolo e senza scritte o banner che ne indichino la natura fake ed ironica, ma solo riportato nella categoria "DEEPFAKE" in un tag sottostante il video. Se cliccato, tale tag rimanda al catalogo con tutti video deepfake prodotti da *Striscia la notizia*, senza dare una definizione di cosa sia un video Deepfake e senza indicare che si tratti di video fasulli, dando per scontato che chiunque sia a conoscenza del fenomeno. Solo a fine video la natura satirica di questo viene rivelata in maniera poco chiara dal conduttore Ezio Greggio, ciò però non è ovviamente bastato ad evitare polemiche nei confronti di Renzi da parte degli spettatori inconsapevoli.

Sebbene anche i video ironici possano promulgare fake news, queste ultime spesso sono divulgate attraverso video seri, pensati appositamente con l'intenzione di diffondere notizie false. Quando nel 2022 è scoppiata la guerra in Ucraina, è diventato molto popolare il video ritraente il presidente ucraino Volodymyr Zelensky intento ad ordinare alle forze armate del suo Paese di deporre le armi di fronte ai militari russi. https://video.repubblica.it/tecnologia/tech/il-video-deep-fake-di-zelensky-che-ordina-agli-ucraini-di-arrendersi/410911/411616
A differenza delle normali fake news, che spesso non hanno fonti attendibili e si basano sulla diffusione di informazioni false, i deepfake sono in grado di generare il loro contenuto e di farlo apparire come autentico. Inoltre, il loro impatto emotivo può essere molto forte, portando gli spettatori a non dubitare della veridicità del messaggio e a prendere posizione sulla questione proposta, anche se falsa.

1.2.3 I deep porn

Infine, la terza tipologia è rappresentata dai video a contenuto pornografico. Si tratta a tutti gli effetti di video a luci rosse, modificati tramiti appositi software, così da far figurare in luogo delle attrici porno altre celebrità, spesso provenienti dall'industria

cinematografica di Hollywood, attraverso una sovrapposizione di volti. Secondo una ricerca citata da CNN Business, al 7/10/2019 su internet esistevano almeno 14678 video deepfake dei quali la maggior parte erano video porno. Inoltre, rispetto a dicembre 2018, in meno di un anno, questi video sono aumentati addirittura del 84%. https://edition.cnn.com/2019/10/07/tech/deepfake-videos-increase/index.html

Questa particolare configurazione del deepfake è denominata "deep porn" e risulta essere particolarmente preoccupante per le ripercussioni che può avere sulla vita privata e professionale delle celebrità coinvolte. Questi video vengono diffusi online senza il consenso delle persone ritratte, che si trovano costrette a fare i conti con conseguenze estremamente negative derivanti dalle altrui azioni, tra cui il danneggiamento dell'immagine pubblica e la difficoltà nel far rimuovere il materiale dal web.

Peraltro, un altro aspetto di sicuro meno rilevante ma comunque degno di nota è la violazione del diritto di copyright nei confronti dei video utilizzati come base, da parte di chi crea e divulga il deep porn. Difatti, il vero video pornografico dopo essere stato scaricato, probabilmente in maniera illegale, e modificato viene diffuso senza il permesso di chi ne detiene i diritti, recando dunque anche un danno economico.

1.3 Storia del deepfake

La forma più diffusa di deepfake, nonché la prima a essere stata sviluppata, è quella video. Un primo progetto di riferimento è certamente il "Video Rewrite" pubblicato nel 1997 da Christoph Bregler, Michele Covell, and Malcolm Stanley. Il progetto in questione aveva la funzione di modificare un video esistente, raffigurante una persona nel mezzo di un discorso, allo scopo di alterarne i movimenti facciali cosi che coincidessero con le parole appartenenti a una diversa traccia audio, sostituita a quella originale.
Tale sistema utilizzava dunque video esistenti per crearne di nuovi, dove il soggetto inquadrato dice cose diverse da quelle realmente pronunciate. "Video Rewrite" utilizzava tecniche di computer vision: il tracking, per tracciare i movimenti effettuati dalla bocca nel video utilizzato, ed il morphing, per trasformare i movimenti e trasportarli nell'output finale.

Il ‘Video Rewrite’ fu il primo programma che riuscì completamente ad automatizzare questo tipo di animazione facciale, usando le tecniche di machine learning per creare connessioni tra i suoni prodotti dal soggetto del video e le sue espressioni facciali.

I primi deepfake, i cosiddetti “face-swaps”, si basavano su un tipo di sistema di deep learning chiamato “auto encoder”.
Una più evoluta tipologia di sistemi di deep learning è quella scaturita dagli studi del ricercatore statunitense Ian J Goodfellow. Nel 2014, Goodfellow pensò e programmò un nuovo sistema di deep learning ora conosciuto come GAN (Generative Adversarial Network). L’invenzione si basa su un’intuizione di Goodfellow: le macchine non erano capaci di realizzare immagini rappresentanti volti umani convincenti ma erano molto performanti nella categorizzazione dei dati (come nei programmi di riconoscimento facciale, ad esempio). Dunque, l’idea di Goodfellow consisteva in una sorta di allenamento trasversale, che comunemente è utilizzato dagli atleti, ma applicato alle intelligenze artificiali. Il sistema GAN si risolve nella contrapposizione di due diversi deep learning network, i quali si sfidano in una sorta di gioco basato sulla generazione di volti umani credibili e coerenti. Un programma “generatore” di volti umani tenta di battere un programma “detector” e migliora di volta in volta che tenta, imparando dai propri fallimenti. Solo poche ore dopo aver avviato il GAN, il programma di Goodfellow fu in grado di produrre immagini artificiali di volti umani migliori di quelle mai prodotte prima da un’intelligenza artificiale.

Un’applicazione del sistema del GAN, che ognuno di noi può sperimentare, è quella utilizzata del sito ‘thispersondoesnotexist.com’ (https://thispersondoesnotexist.com). Collegandosi al link è possibile avviare un generatore di foto iperrealistiche, come quella riportata nella pagina seguente. Questa immagine può essere scaricata, insieme ad altri esempi al seguente link:
https://drive.google.com/drive/folders/1WVuF4c19CNyRGLwTiakdKqr5K2iQRqI0?usp=sharing

Ogni foto generata è unica e richiede pochissimi secondi di elaborazione, a patto di disporre di una buona connessione internet. Risulta ovviamente impressionante come chiunque abbia a disposizione un dispositivo in grado di connettersi ad internet possa essere in grado di generare, tramite l'ausilio di programmi di machine learning, immagini di volti realistici, in pochissimo tempo, con un solo click e soprattutto senza avere delle basi di programmazione e, più in generale, senza avere conoscenze informatiche.

Nonostante il GAN ed il 'Video Rewrite' fossero grandi innovazioni nel mondo della visione artificiale -al punto tale da mettere le basi per la creazione di veri programmi per la creazione di contenuti deepfake-, questa tecnologia ha acquistato maggiore notorietà soltanto intorno al 2017, quando molti video di questa tipo iniziarono ad essere diffusi sui vari social network.

Il 2 novembre 2017, un utente anonimo del social network “Reddit” conosciuto con il nickname “Deepfakes” (al quale si attribuisce appunto la denominazione corrente) iniziò una discussione sul forum da lui creato: “r/deepfakes”. Sul forum venivano postati video porno falsi ritraenti le attrici più famose di Hollywood, o meglio video modificati in maniera tale da sovrapporre i volti di queste ultime su quelli delle attrici originali. L’utente utilizzava, per la creazione dei video, un codice opensource. È allora agevole comprendere che chiunque sia dotato di una buona conoscenza di machine learning e programmazione ha la possibilità di creare facilmente altri video.

1.4 Tecnologia alla base e Funzionamento

1.4.1 Machine Learning

Come già accennato, i programmi deepfake sfruttano il Machine Learning.
Il machine learning (abbreviato ML), conosciuto anche come apprendimento automatico, è una sottospecie dell’intelligenza artificiale che si occupa di creare sistemi di apprendimento o di miglioramento delle performance sulla base dei dati ricevuti. In questo modo, il machine learning emula il modo in cui gli esseri umani imparano e migliora gradualmente le proprie performance.
Attualmente, il machine learning è ampiamente utilizzato ed ha diverse applicazioni, tra cui banche, siti di shopping online e social media. Ad esempio, le banche lo utilizzano per rilevare frodi e prevenire transazioni sospette, mentre i siti di shopping online per fornire raccomandazioni di prodotti personalizzati ai propri clienti. Infine, i social media utilizzano il machine learning per suggerire contenuti rilevanti ai propri utenti in base ai loro interessi e comportamenti passati.

1.4.2 Deep learning

Il deep learning, noto anche come apprendimento profondo, è una tecnica di apprendimento automatico che utilizza reti neurali artificiali composte da tre o più livelli per l'elaborazione dei dati. Questi modelli cercano di imitare il funzionamento del cervello umano, apprendendo da grandi quantità di dati e migliorando le proprie prestazioni ad ogni esecuzione dell'algoritmo.

A differenza delle reti neurali con un singolo livello, che sono in grado di modellare solo funzioni molto semplici, l'aggiunta di diversi livelli permette di risolvere problemi

più complessi e di ottenere risultati più soddisfacenti. Il deep learning è una tecnologia molto potente e versatile, utilizzata in vari settori, tra cui la visione artificiale, il riconoscimento del parlato, la traduzione automatica, il riconoscimento di oggetti e molto altro ancora.

1.5 Come funziona un deepfake?

Dopo aver introdotto la tematica dei video deepfake, averli definiti ed aver analizzato la loro storia, risulta doveroso spiegare come questi funzionino.
Come già sottolineato in precedenza, questi video si ottengono dalla sovrapposizione di un volto su un altro. Il processo di creazione di un video deepfake richiede la presenza di un dataset di immagini o un video del soggetto da imitare, che vengono utilizzati come input per l'algoritmo di machine learning. L'algoritmo cerca di individuare i tratti somatici distintivi del volto del soggetto e di creare una maschera digitale in grado di riprodurli fedelmente.
Una volta che la maschera digitale è stata creata, questa viene sovrapposta al volto della persona di base nel video, in modo da creare un video che appare autentico, ma che in realtà è il risultato di un'elaborazione digitale.
Il volto "di base" è quello che definisce i movimenti; questo volto viene tracciato dal programma punto per punto e funge appunto da base.
Il secondo volto è quello che funge da "maschera"; esso va a coprire il primo volto e ne ricopia i movimenti

1.5.1 Scelta dei dati

Nella realizzazione di un deepfake e, specificatamente, di un video deepfake, la prima scelta da effettuare è quella del video.
A seconda del video scelto la qualità del risultato finale può variare. Per un effetto migliore, è preferibile che nel video sia presente un solo soggetto o che comunque non vengano ripresi più soggetti nella stessa scena. Ciò al fine di risparmiare al software maggiori difficolta, ma anche per evitare che lo spettatore possa effettuare paragoni (anche solo impliciti) tra il soggetto modificato e gli altri e notare particolari innaturali.
Il soggetto utilizzato come base deve avere un volto quanto più simile possibile a quello da sovrapporre, proprio perché il deepfake funziona come una via di mezzo tra una maschera ed un trucco. Dunque, se il soggetto base ha una folta barba mentre la maschera no, si avranno problemi di coerenza dell'immagine nella sovrapposizione. La

scelta di due soggetti con caratteristiche facciali simili renderà possibile il raggiungimento di un migliore risultato senza affaticare il software.

Oltre al video di base è ovviamente necessario individuare il volto da sovrapporre. Come spesso ricordato dal Garante per la Protezione dei Dati Personali, le vittime di deepfake sono solitamente persone famose. Queste sono d'altronde le vittime perfette, non soltanto perché suscitano un grande interesse da parte degli utenti del web, ma anche perché, come abbiamo già detto, per creare la maschera che il programma utilizzerà è necessaria una grande quantità di dati, ed in particolare numerosi fotogrammi del volto che si vuole utilizzare. Detti fotogrammi (anche di elevata qualità, peraltro) possono essere ricavati con grande facilità dalla scena di un film o di un'intervista e, grazie ad internet, sono divenuti molto semplici da reperire.

L'utilizzo di video di bassa qualità o addirittura di un singolo programma produrrebbe, infatti, risultati troppo limitati, più simili ad output ottenibili già nel 1997 tramite programmi come "Video Rewrite". Dunque, attingere a quanti più dati di alta qualità possibile è assolutamente necessario per la buona riuscita del progetto. Il volto tracciato potrebbe non essere sempre nella stessa posizione: variando angolazione è necessario che ci siano a disposizione quanti più fotogrammi possibili e mutevoli del volto-maschera, così che il programma li possa elaborare in maniera coerente e fluida.

Per quanto riguarda la popolarità del video, è evidentemente più semplice che un contenuto diventi virale o comunque d'interesse per un gran numero di persone se al suo interno è rappresentato qualcuno di conosciuto. Nel caso di Reddit, infatti, le vittime dei deepfake erano attrici famose, mentre nelle fake news sono quasi sempre coinvolti politici.

Capitolo 2. Leggi relative al deepfake

Trattandosi di tema assolutamente recente, non esiste una vera e propria regolamentazione in vigore sui deepfake, soprattutto se si considera il contesto italiano. A riguardo si è espresso anche il Garante della Protezione dei Dati Personali, pubblicando nel 2020 una scheda informativa sui rischi dell'uso malevolo di questa tecnologia. Tale scheda, però, fornisce solo una definizione di deepfake, esempi di utilizzi illegali di tale mezzo e accortezze utili per proteggersi da eventuali utilizzi illeciti di tale tecnologia, senza far riferimento a contromisure legali. Non esistono ancora leggi o proposte di leggi a livello nazionale che si impegnino a contrastare il fenomeno degli illeciti che possono scaturire da un utilizzo poco etico di materiale video o audio deepfake.

2.1 Protezione dei dati personali

Anche gli organi legislativi dell'Unione Europea si sono occupati del tema di cui stiamo trattando, sia pur in via indiretta. Infatti, dal momento che la creazione di un falso contenuto digitale comporta, tipicamente, anche il trattamento di dati personali, trovano applicazione le disposizioni del Regolamento UE 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati. Detto regolamento, che ha abrogato la direttiva 95/46/CE, include tra i dati personali anche i "dati biometrici", definiti come «i dati personali ottenuti da un trattamento tecnico specifico relativi alle caratteristiche fisiche, fisiologiche o comportamentali di una persona fisica che ne consentono o confermano l'identificazione univoca, quali l'immagine facciale o i dati dattiloscopici». Considerando quindi i punti iii e iv della definizione di deepfake data dal Garante per la protezione dei dati personali, tutti i video deepfake risultano essere frutto di uno sfruttamento di dati personali contrario al regolamento 2016/679.

2.2 proposta del 21/04/2021, AI act

Certamente rilevante in materia risulta essere la Proposta di Regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza

artificiale (legge sull'intelligenza artificiale) del 21 aprile 2021, comunemente denominata come “AI Act”.
Il Parlamento europeo ha dato il via libera in data 14/06/2023 all'AI Act, che sta ora entrando nella fase conclusiva. L'approvazione definitiva da parte dell'Unione europea dovrebbe avvenire presumibilmente entro la fine dell'anno mentre l’entrata in vigore è prevista nel 2024.

Come si è detto, l’AI Act è una proposta di regolamento europeo sull'intelligenza artificiale (IA, AI in inglese) - la prima da parte di un regolatore cruciale a cui si ispirano i legislatori di tutti i paesi del mondo. La legge assegna le applicazioni dell'AI a tre categorie di rischio:

- in primo luogo, le applicazioni ed i sistemi che creano un rischio inaccettabile, come la valutazione sociale gestita dal governo (come accade in Cina), sono vietati;
- in secondo luogo, le applicazioni ad alto rischio, come uno strumento di scansione del curriculum che classifica i candidati ad un determinato lavoro, sono soggetti a requisiti legali specifici;
- infine, le applicazioni non esplicitamente vietate o elencate come ad alto rischio sono in gran parte non regolamentate.

Dunque, dal momento che il deepfake costituisce a tutti gli effetti una tecnologia basata sull’intelligenza artificiale, le regole per l’uso delle applicazioni di IA contenute nella proposta di Regolamento potrebbero acquisire grande importanza pratica in futuro.
La proposta di Regolamento europeo sull’Intelligenza Artificiale sopra richiamata, come noto, ha l’obiettivo di consentire un uso affidabile e sicuro dell’Intelligenza Artificiale, nel rispetto dei valori e dei diritti fondamentali degli individui.
A tal fine, stabilisce regole armonizzate per lo sviluppo, l’immissione sul mercato e l’utilizzo dei sistemi IA.
Il quadro normativo proposto adotta un approccio basato sul rischio, distinguendo tra: un rischio inaccettabile, un rischio alto, un rischio basso o minimo. La proposta è volta a vietare l’uso di sistemi che presentino un rischio inaccettabile (quindi i sistemi contrari ai valori dell'Unione, ad esempio in caso di violazione dei diritti fondamentali), mentre per i sistemi che rientrano nella categoria ad alto rischio prevede l’obbligo di effettuare dettagliate analisi e valutazioni d’impatto, nonché di garantire una fase di controllo da parte di supervisori umani.

L'AI act rappresenta un regolamento dell'Unione Europea che ha l'obiettivo di garantire la protezione dei diritti e delle libertà fondamentali dei cittadini dell'UE in relazione all'utilizzo di tecnologie di intelligenza artificiale.
La scelta del regolamento come strumento giuridico per disciplinare l'AI Act è frutto di una volontà concreta, a livello comunitario, di regolare la questione dei deepfake e delle intelligenze artificiali. Il regolamento, infatti, ha una portata più ampia delle direttive e si applica direttamente agli Stati membri: ciò significa che l'AI act ha forza di legge in tutti gli Stati membri dell'UE e non richiede una specifica adozione da parte degli stessi mediante un processo di recepimento.
Il regolamento ha come obiettivo quello di garantire che le tecnologie di intelligenza artificiale siano utilizzate in modo etico e responsabile, salvaguardando i diritti fondamentali delle persone. In particolare, l'AI Act prevede una serie di requisiti che le tecnologie di intelligenza artificiale devono rispettare, tra cui la trasparenza, l'accountability, la privacy e la sicurezza. Il regolamento prevede anche l'istituzione di un'autorità europea per la sicurezza delle intelligenze artificiali, che avrà il compito di monitorare l'attuazione del regolamento e di garantire che le tecnologie di intelligenza artificiale rispettino i requisiti previsti.
Uno degli obiettivi principali dell'AI act è quello di regolare la questione dei deepfake, divenuti ormai una minaccia sempre più diffusa per la sicurezza e la privacy delle persone. Pertanto, l'AI Act prevede che i deepfake siano considerati illegali e punibili a livello europeo.
In conclusione, l'entrata in vigore dell'AI Act rappresenta un importante passo in avanti nella regolamentazione delle tecnologie di intelligenza artificiale in Europa. Inoltre, la creazione dell'autorità europea per la sicurezza delle intelligenze artificiali è un cruciale per garantire la corretta attuazione del regolamento e la protezione dei cittadini dell'UE dai rischi derivanti dall'utilizzo delle tecnologie di intelligenza artificiale.

2.2.1 la proposta sul deepfake, art.52

Con particolare riguardo alla tecnica del deepfake, la proposta di Regolamento consente tale tecnologia, ma articola alcuni requisiti minimi e prevede un obbligo di trasparenza in capo a chi ne fa uso. Nel testo del regolamento, all'art. 52, comma 3, sono specificati tali requisiti. In particolare, gli autori di deepfake sono obbligati ad etichettare il contenuto generato in modo che sia evidente che si tratti di un contenuto artificiale e manipolato digitalmente. In questo modo, chiunque visualizzi il contenuto è

consapevole della sua natura manipolata e può prendere decisioni informate sulla sua veridicità.

Successivamente, lo stesso art. 52 prevede che tale obbligo non si applichi "quando l'uso è autorizzato dalla legge per accertare, prevenire, indagare e perseguire reati o se è necessario per l'esercizio del diritto alla libertà di espressione e del diritto alla libertà delle arti e delle scienze garantito dalla Carta dei diritti fondamentali dell'UE, e fatte salve le tutele adeguate per i diritti e le libertà dei terzi."
Dunque, se vi è una giustificazione basata sull'interesse pubblico o sulla libertà di espressione, gli autori potrebbero essere dispensati dall'obbligo di etichettatura.
Tuttavia, è importante sottolineare che ciò non significa che gli autori di deepfake siano esenti da responsabilità legale per la diffusione di contenuti manipolati o ingannevoli.
È in ogni caso necessario assicurare adeguate protezioni per i diritti e le libertà delle persone coinvolte, che devono essere sempre rispettate e tutelate.

L'introduzione di un obbligo di etichettatura dei deepfake potrebbe rappresentare il primo passo per mitigare gli eventuali impatti negativi del fenomeno. Tuttavia, come evidenziato dalla ricerca dell'EPRS (Servizio Ricerca del Parlamento Europeo), la natura e l'entità della disposizione sono attualmente poco chiare e si nota una certa inerzia regolatoria da parte del legislatore, il quale non ha previsto alcuna sanzione per l'eventuale mancato rispetto dell'obbligo previsto dall'articolo 52.

Inoltre, la possibilità di deroga, prevista dalla norma laddove l'uso del deepfake sia "necessario per l'esercizio del diritto alla libertà di espressione e del diritto alla libertà delle arti e delle scienze", include un raggio di ipotesi così ampio da svuotare di significato l'obbligo di trasparenza sancito.
Pertanto, la risposta normativa tratteggiata, in tale sede, dal legislatore europeo non è attualmente ritenuta sufficiente a contrastare il problema.

Con riguardo, poi, all'aspetto relativo al trattamento dei dati personali, è da rilevare che, nel contesto dei deepfake, i dati personali vengono trattati non soltanto nella fase di creazione dei contenuti, ma anche per addestrare il software che viene utilizzato. Per tale motivo, è bene innanzitutto precisare che il GDPR (Regolamento generale sulla

protezione dei dati), in relazione al suo ambito di applicazione, rileva in entrambe le fattispecie.

Con particolare riferimento al requisito di liceità del trattamento, di un certo interesse è la ricerca della base giuridica.
Secondo quanto affermato dall'EPRS, nel caso specifico della creazione di deepfake, i creatori potrebbero appellarsi a due basi giuridiche: l'interesse legittimo e il consenso esplicito degli interessati. Nel caso in cui il creatore sostenga di avere un interesse legittimo, quest'ultimo deve prevalere sui diritti e le libertà dell'interessato per costituire una base legale valida. In caso contrario, l'uso dei dati personali per la creazione e diffusione di deepfake richiede il consenso informato delle persone raffigurate nei contenuti originali e in quelli "fabbricati". È importante notare che il consenso deve essere ottenuto sia dalle persone raffigurate nei contenuti originali, sia da quelle raffigurate nei contenuti "fabbricati".
Il GDPR offre una serie di diritti alle vittime dei deepfake, tra cui il diritto alla correzione dei dati inesatti e il diritto alla cancellazione dei propri dati.

2.3 Situazione in altri paesi europei

2.3.1 Spagna

Nel 2022, il governo spagnolo ha deciso di istituire un'agenzia per la supervisione sull'intelligenza artificiale responsabile dello sviluppo, del controllo e della vigilanza dei progetti relativi alla strategia nazionale spagnola AI e di quelli dell'Unione Europea relativi allo sviluppo regolamentare dell'IA e ad i suoi possibili utilizzi. L'agenzia, chiamata AESIA (Agencia Española de Supervisión de la Inteligencia Artificial), avrà sede a La Coruña, in Galizia.
Tra le sue funzioni, l'AESIA avrà il potere di vietare l'uso di sistemi AI che potrebbero rappresentare un rischio. In caso di violazioni, l'agenzia potrà infliggere sanzioni ed i relativi proventi saranno destinati al suo budget. Il budget iniziale sarà di 5 milioni di Euro, ma fonti del governo assicurano che l'intenzione è di aumentarlo prima del lancio, anche se ciò dipenderà dalla decisione del Ministero delle Finanze.
Gran parte del budget sarà destinata a pagare i salari di circa 40 persone, tra tecnici, amministrativi e legali. Il governo è stato tuttavia criticato per non aver incluso

professionisti umanistici, che sarebbero stati utili per la definizione di questioni etiche legate all'IA.

L'agenzia gestirà la sandbox nazionale sull'IA, che consentirà alle aziende (che aderiranno su base volontaria da marzo ad aprile 2023) e ai regolatori di collaborare per definire le migliori pratiche per l'implementazione dei sistemi AI. La sandbox aiuterà anche a stabilire il ruolo dell'agenzia presso le aziende private.

L'AESIA istituirà un "bollino AI nazionale", che certificherà che i sistemi AI soddisfano i requisiti richiesti dall'Europa. Almeno fino all'entrata in vigore dell'AI Act, la richiesta del bollino sarà facoltativa. Il certificato potrebbe in futuro diventare collegato all'accesso ai finanziamenti europei per le aziende.

Il governo prevede di avere tutti i documenti pronti entro giugno di quest'anno, il che consentirà di emanare le relative disposizioni entro l'estate e di avere l'agenzia operativa già in autunno.

2.3.2 Regno Unito

Nel Regno Unito è stata presentata la proposta di legge "Online Safety Bill" con la finalità dichiarata di rendere la navigazione in internet più sicura nel paese.

Tuttavia, alcuni rilevanti aspetti sembrano non essere stati affatto presi in considerazione dal legislatore, il che fa sorgere numerosi dubbi circa l'adeguatezza della proposta al raggiungimento degli obiettivi indicati. La proposta di legge è stata avanzata dal governo di Boris Johnson ai legislatori ed una prima bozza del testo stabilisce i limiti e i confini all'interno dei quali il sistema legislativo si inserirà. Sarà il parlamento a cristallizzarne definitivamente il contenuto.

2.3.2.1 Online Safety Bill e i punti critici in Gran Bretagna

Un punto critico emerso riguarda proprio la definizione di contenuti illegali. Tali contenuti non troveranno più spazio sui principali canali di comunicazione online, tra cui piattaforme, social network e motori di ricerca. Bisogna però considerare che la definizione di contenuto illegale non è facile da stabilire in modo univoco, soprattutto se si considera il confronto con ciò che è ritenuto illegale offline. In rete, infatti, esistono situazioni che non hanno riscontro nella realtà, come nel caso dei deepfake. Come ci si

può assicurare che questi contenuti rientrino in quelli che non possono essere pubblicati online?

Secondo alcuni parlamentari, la definizione lacunosa di contenuto illegale proposta dal governo metterebbe a rischio la sicurezza di minori, minoranze e donne sui social network e sulle altre piattaforme web. In particolare, la mancanza di una definizione precisa di contenuto illegale esporrebbe gli utenti a contenuti pedopornografici che, secondo la proposta di legge, non sarebbero disciplinati se condivisi come suggerimenti di visione o collegamenti a chat o canali privati. In questo senso, l'Online Safety Bill potrebbe divenire un'"occasione mancata" se dovesse essere promulgato senza modifiche

2.3.3 USA

Negli Stati Uniti, al momento, si sta esaminando una proposta di legge a livello federale, che sarebbe valida quindi per tutti e cinquanta gli Stati. Solo la Virginia, il Texas e la California hanno attualmente adottato delle leggi relative a questo argomento.

La legislazione dello stato della Virginia considera reato la diffusione di materiale deepfake pornografico non consensuale, mentre in Texas è vietata la creazione e la distribuzione di video deepfake che danneggino i candidati ai pubblici uffici o che possano influenzare le elezioni. Da questo esempio si può notare come i problemi legati ai video deepfake siano percepiti in maniera differente nei vari Stati. Nello stato della Virginia, la legge mira a proteggere gli individui dalla divulgazione non consensuale di materiale deepfake pornografico, mentre in Texas si cerca di vietare la creazione e distribuzione di video deepfake che potrebbero danneggiare la reputazione dei candidati. In entrambi i casi, è evidente che la legislazione sulle intelligenze artificiali non copre tutte le possibili forme di reato che possono essere commesse attraverso questa tecnologia.

2.3.4 Cina

Il governo cinese ha approvato nuove disposizioni per regolamentare la diffusione dei video falsi generati tramite intelligenza artificiale. Tali video dovranno essere esplicitamente contrassegnati con un disclaimer o censurati se danneggiano gli "interessi nazionali" del Partito comunista cinese.

La normativa, che è la prima nel suo genere ad affrontare il fenomeno dei video falsi generati tramite intelligenza artificiale, è entrata in vigore il 10 gennaio. La privacy degli utenti è al centro della regolamentazione, con la possibilità per le vittime di deepfake di contestare l'uso della propria immagine.
Ai fornitori di servizi è richiesto di registrare i propri servizi e di sottoporre regolarmente i propri codici e dati a revisione da parte dello Stato. Secondo la normativa, i gestori delle app di video falsi devono indicare chiaramente la presenza di un video modificato, classificandolo come tale e non come originale. Sono vietati completamente i contenuti "non approvati" dalle autorità governative e quelli considerati illegittimi per la "diffusione di false informazioni".
La Cyberspace Administration of China, l'ente regolatore della sfera online in Cina, è responsabile della supervisione del corretto comportamento dei fornitori di servizi.
Il tema della diffusione di video falsi generati tramite intelligenza artificiale è controverso anche al di fuori della Repubblica Popolare Cinese. Tuttavia, per Pechino, il tema della manipolazione dell'informazione è particolarmente sensibile e il governo controlla da tempo le tecnologie che regolano il flusso di informazioni online.
Non tutte le forme di video falsi generati tramite intelligenza artificiale sono vietate, ma solo quelle che contengono "informazioni illegali o dannose" o che "si oppongono agli interessi nazionali". La formulazione tipica della legislazione cinese non specifica cosa sia considerato "dannoso", ma ribadisce l'inderogabile supervisione del governo sulle nuove tecnologie, soprattutto se queste fanno uso di intelligenza artificiale e raccolgono dati.
La nuova normativa sui video manipolati si aggiunge all'arsenale normativo di Pechino per la protezione del digitale, che include già la legge sulla sicurezza dei dati (Dsl), la legge sulla privacy (Pipl) e la normativa sugli algoritmi. L'obiettivo è ancora una volta duplice: da una parte si proteggono gli interessi degli utenti, salvaguardandone la privacy, dall'altra si riafferma l'intenzione del Partito comunista cinese di supervisionare i contenuti diffusi online.

2.4 Le contromisure dei privati

Società come Meta e Twitch avrebbero vietato questo tipo di contenuti. Meta, in particolare, ha lanciato nel 2019 la Deepfake Detection Challenge, un’iniziativa intrapresa al fine di accelerare lo sviluppo di nuovi metodi per rilevare i video deepfake.

Per tre mesi, infatti, tra il 2019 e il 2020, Facebook (ora Meta) ha co-ospitato questa Challenge, chiedendo ai partecipanti di automatizzare il processo di determinazione circa la manipolazione con l'intelligenza artificiale delle foto. Il concorso ha attirato 2.114 partecipanti e ha assegnato premi per un totale di 1 milione di dollari (USA) alle voci con gli algoritmi più riusciti. Ma, nonostante al concorso avessero partecipato alcune delle menti più brillanti dell'intelligenza artificiale, il miglior programma è stato in grado di rilevare deepfake solo il 65 percento delle volte. Attualmente, la maggior parte dei programmi di rilevamento basati sull'intelligenza artificiale cerca "artefatti visivi" - imperfezioni come incoerenze di illuminazione, posizionamento di ombre strane e disaccordi geometrici - per identificare dove un'immagine potrebbe essere stata manipolata. Tuttavia, a causa della natura in continua evoluzione dell'intelligenza artificiale, questi programmi possono imparare rapidamente come coprire le proprie tracce. Caso noto è lo studio dell'Università di Albany del 2018 che ha scoperto che le persone ritratte nei deepfake tendono a sbattere le palpebre più frequentemente o meno frequentemente rispetto alle persone reali. Un anno dopo, alcuni ricercatori in Corea del Sud hanno notato che i deepfake stavano sviluppando modelli di battito di ciglia sempre più realistici. Sono stati apportati cambiamenti simili con gli occhiali e i denti, in precedenza non dotati di un aspetto verosimile nelle foto generate dall'intelligenza artificiale. Sebbene tale lavoro degli esperti sia fondamentale, questo fornisce, involontariamente, ai creatori di deepfake indicazioni per creare immagini più ingannevoli.

I deepfake non sono ancora in grado di produrre immagini di esseri umani perfetti; quindi, gli strumenti di rilevamento rimangono efficaci. Tuttavia, Andy Parsons, il direttore senior dell'iniziativa per l'autenticità dei contenuti presso Adobe (The Content Authenticity Initiative), che lavora per sviluppare strumenti che aiutano a contrastare la diffusione di fake news, è di diverso avviso. Egli ritiene infatti che non sarà una soluzione valida per sempre, definendola "una battaglia persa".

2.4.1 I cheap fake

Mentre i deepfake rappresentano una minaccia in crescita, Jane Lytvynenko, che lavora sul Media Manipulation Casebook (ricerca rivolta a giornalisti e ricercatori che documenta la disinformazione e le informazioni errate) afferma che la preoccupazione

più grande riguarda i "cheap fake", che sono foto e video modificati senza l'utilizzo dell'intelligenza artificiale.
Prima di unirsi al progetto Technology and Social Change presso il Shorenstein Center dell'Harvard Kennedy School, Lytvynenko si è fatta un nome occupandosi di disinformazione e informazioni errate presso BuzzFeed News. Secondo Lytvynenko, i cheap fake, che utilizzano taglia e incolla, audio rallentato spezzoni di video, forniscono ai manipolatori un modo economico ed efficace per creare media manipolati. "Le persone vengono disinformate da tattiche più semplici dei deepfake, quindi al momento non c'è molta incentivazione a utilizzare approcci complessi", ha detto Lytvynenko.
Un esempio potrà chiarire quanto appena riportato. In un video intitolato "IS SHE DRUNK?!?! Nancy Pelosi Fumbles Words, Struggles Through Press Conference", pubblicato da un canale YouTube noto per le sue teorie del complotto di destra, Pelosi viene mostrata apparentemente balbettante. Il video utilizza metodi di inganno vecchia scuola, rallentando la velocità per far apparire il soggetto in stato confusionale. Nonostante numerosi utenti si siano accorti della falsità del video, è stato comunque condiviso ampiamente e rimane sulla piattaforma. Ma una nuova soluzione, chiamata *provenienza dei contenuti*, potrebbe offrire un modo migliore per adattarsi all'evoluzione del mondo della disinformazione. Con un nome preso in prestito dal mondo dell'arte, questa iniziativa cerca di stabilire una catena di provenienza che documenti ciò che è successo ad un'immagine per l'intera durata della sua vita digitale, incluso chi l'ha scattata, quando è stata scattata e quali modifiche sono state apportate ad essa. Invece di lavorare all'indietro per vedere se un'immagine è stata manomessa, il software cerca di garantire l'autenticità di un'immagine fin dalla sua creazione. Questi dati vengono quindi confezionati e, una volta che l'immagine è stata pubblicata online, condivisi in una casella delle informazioni accanto alla foto.
Adobe ha iniziato a promuovere questo tipo di verifica attraverso la sua Content Authenticity Initiative, annunciata per la prima volta nel 2019. Il programma, che è già stato lanciato in Adobe Photoshop, offre ai creatori un modo per tracciare le modifiche apportate ad una foto, e alle organizzazioni, come Twitter e il New York Times, un modo per essere più trasparenti con il proprio pubblico. Dal momento che su tratta unicamente di uno strumento facoltativo, l'autenticità dei contenuti non permetterà di risolvere le questioni sul deepfake, tuttavia conferirà maggiore credibilità ai media. Lo sviluppo di partnership con piattaforme digitali e organizzazioni mediatiche e

l'implementazione di autenticità dei contenuti sulle immagini di stock hanno messo in gran risalto il programma di Adobe.
Secondo Andy Parsons, la rilevazione dei deepfake e la provenienza dei contenuti sono autenticatori complementari: il primo è reattivo e il secondo costituisce una misura proattiva. L'obiettivo è non solo quello di offrire maggiore trasparenza online, ma anche di incoraggiare il pubblico a pensare in modo più critico ai media che utilizza.

Capitolo 3. La realizzazione di un deepfake

3.1 Premessa

Sebbene questa tecnologia si stia diffondendo velocemente, la creazione di un video deepfake non è ancora alla portata di tutti. Per la realizzazione di tale prodotto, infatti, bisogna essere esperti programmatori o, in alternativa, avere accesso a programmi già pronti all'uso.

In questo capitolo esporremo delle procedure per la produzione di due semplici contenuti deepfake che, seppur molto rudimentali, dimostrano il funzionamento di questa tecnologia. L'obiettivo sarà quello di spiegare passo per passo tutti i procedimenti e le accortezze per la creazione di un audio quanto più realistico possibile e di un video che sicuramente, se diffuso, non verrebbe scambiato per vero ma che rientri comunque nella definizione di deepfake. Per comprendere a pieno i passaggi per la realizzazione del secondo contenuto sarà necessario avere delle conoscenze base di programmazione, mentre per il primo si farà affidamento ad un sito web che svolgerà gran parte del lavoro. In entrambi i casi sarà possibile ottenere i risultati sperati seguendo le istruzioni passo per passo.

3.2 la scelta del soggetto

Nonostante, come già anticipato nel primo capitolo, sia possibile creare contenuti deepfake che ritraggano chiunque, a patto di possedere una mole di dati sufficienti a far girare i vari programmi, le vittime principali dei deepfake sono solitamente personaggi famosi. Ciò avviene sia perché è estremamente più facile trovare foto, audio e video di politici o celebrità che di persone qualunque, sia perché essendo i deepfake un fenomeno virale di internet è necessario che questi vengano diffusi ed è più facile che la diffusione avvenga se si tratta di video ritraente soggetti noti. Dichiarazioni assurde, poco condivisibili o incoerenti di personaggi famosi sono sempre state uno dei fenomeni più condivisi in rete e ciò si denota anche, e soprattutto, nel caso dei deepfake.

Dunque, per la realizzazione del contenuto che vogliamo produrre, sarà necessario scegliere un personaggio famoso ed un testo riguardante un argomento di attualità, che magari riguardi nello specifico notizie recenti o dichiarazioni fatte dal soggetto stesso. Come fatto per l'esempio di fake news generata tramite Intelligenza Artificiale riportato nel paragrafo 1.2.2, ricorreremo all'utilizzo della figura di un politico, in questo caso dell'ex presidenti degli Stati Uniti Donald Trump.

Il caso di attualità a cui faremo riferimento è quello riguardante l'incarcerazione di Trump avvenuta ad Atlanta il 25/08/2023. In tale data, infatti, l'ex presidente degli USA si è consegnato alle forze dell'ordine essendo accusato, insieme ad altri 18 complici, di aver cospirato per ribaltare l'esito della votazione per l'elezione del Presidente degli Stati Uniti in Georgia nel 2020. Si tratta della prima volta in cui un ex presidente viene arrestato.

Per quanto riguarda l'audio deepfake produrremo un'improbabile dichiarazione di Donald Trump sul caso. Il discorso che faremo generare all'IA sarà questo:

"I am guilty and I deserve to be thrown in jail, but I will never go there because I am Donald Trump. I am rich. And I can do whatever I want in this country. Do you think being arrested for 20 minutes will harm me or my image? It's not the case. I've shown everyone that I'm so rich and powerful that I can decide when to turn myself in, do so in my private plane, and leave in less than an hour.".

Tradotto:
"Sono colpevole e meriterei di essere sbattuto in galera ma non ci andrò mai perché sono Donald Trump. Sono ricco. Posso fare quel che voglio in questo paese. Pensate che essere stato arrestato per 20 minuti danneggi me o la mia immagine? Non è così. Ho mostrato a tutti che sono talmente ricco e potente che posso decidere io quando andare a costituirmi, farlo arrivando col mio aereo privato e andarmene in meno di un'ora.".

3.3 Audio deepfake

Nonostante alcune fake news sfruttino la capacità di un imitatore e di montaggi o video deepfake in modo tale da far sembrare il discorso pronunciato dalla vittima, in questo paragrafo vedremo un modo che non richieda capacità umane per la sua riuscita.

Per rimanere nella definizione di deepfake, la quale prevede che i passaggi di elaborazione dati (e quindi l'imitazione) siano frutto di un'Intelligenza artificiale, procederemo utilizzando un software di IA, il sito fakeyou.com (https://fakeyou.com/tts/result/TR:jdk1pv6ra551yc5v1fx75dhp5dger).

Attingendo ai dati presenti in un database privato, in cui sono contenuti terabyte di file audio, il programma renderà possibile trasformare un testo in un audio imputabile ad un personaggio noto.
Il sito in questione offre un'ampia scelta di personaggi famosi di diverse nazionalità e la possibilità di produrre audio in più lingue. Nel catalogo sono presenti oltre 3500 voci tra cui scegliere e ben 8 lingue con cui generarle. La procedura è inoltre molto semplice.

Dopo aver effettuato una veloce registrazione al sito, basterà entrare nella sezione "Testo Alla Voce" (TTS), selezionare la lingua desiderata, cercare e selezionare la voce da utilizzare ed in fine scrivere il testo da generare successiva mente sottoforma di audio deepfake.

Nell'esempio portato dovremo selezionare la lingua inglese e cercare nell'elenco, la voce di Donald Trump. Essendo l'ex presidente degli USA un soggetto estremamente conosciuto ed una ricorrente vittima sia di fake news, ma soprattutto di meme deepfake, avremo una vasta scelta di voci, potendo optare dunque per più varianti, tra cui "arrabbiato", "sarcastico", "casual" etc.

Nel nostro caso potremmo sia utilizzare quella che riteniamo più opportuna sia un mix di voci, in modo da sottolineare parole o concetti in maniera diversa.

Dopo aver scritto il nostro testo sarà possibile generarlo cliccando sull'apposito tasto ed infine scaricare il tutto. Essendo il sito, nella sua versione gratuita limitato, si potranno generare solo parti di un lungo discorso. Non sarà però un problema unire tali parti su un qualsiasi programma di registrazione audio.

Il risultato dell'esperimento è consultabile dal link:
https://drive.google.com/drive/u/0/folders/1yAdq-tK-UpAwmo5fGmJCndsbvct5KERT

3.4 Il video deepfake

Il programma di cui ci avvarremo è, come anticipato, assolutamente user friendly. Pertanto, il nostro secondo output non sarà un normale video deepfake ma un video che mostri la capacità del programma di analizzare i dati forniti e rielaborarli in modo tale da dirigere un volto in base ai movimenti registrati da un altro soggetto. Nonostante il risultato finale sarà molto più semplice ed imperfetto rispetto ad i video analizzati negli scorsi capitoli, verrà generato tramite una tecnologia definibile deepfake per i criteri già elencati nella definizione presente nel paragrafo 1.1.

3.4.1 La scelta dei dati

Come già anticipato nel paragrafo 1.5.1 alla base della realizzazione di un deepfake c'è la scelta dei dati da utilizzare. Normalmente, per realizzare un video in alta qualità, servirebbero molti frame raffiguranti il soggetto che vogliamo ritrarre nel video. Non essendo, però, il programma utilizzato in grado di elaborare una mole così elevata di file, nel nostro caso ci limiteremo a scegliere una singola foto da utilizzare come "maschera".

Per la scelta della foto le accortezze principali da tenere a mente sono poche:
La foto deve essere coerente nel contesto in cui verrà utilizzato il video, deve essere scattata frontalmente e non devono essere inquadrate mani o oggetti che normalmente si muoverebbero nello sfondo, in quanto il software spostando il volto della foto, genererebbe un effetto onda irrealistico. L'espressione deve essere più neutrale possibile ed avere una buona risoluzione.

La foto utilizzata nel nostro esempio è la seguente:

A dettare i movimenti della foto sarà un video "driver". La denominazione "driver" deriva proprio dal fatto che il software scannerizzerà il volto e traccerà i suoi movimenti, in maniera da riprodurli sulla "maschera". Per una migliore riuscita dell'output finale è importante che l'inquadratura del video sia quanto più simile alla foto scelta, che quindi la distanza dalla fotocamera e la direzione del volto non si discostino troppo da quelle della maschera. È importante anche che la persona ritratta nel video "driver" sia quantomeno simile alla maschera. Non è assolutamente necessario che si tratti di un sosia, ma la presenza di differenze fisiche troppo accentuate potrebbe affaticare il programma e non lasciarlo lavorare come dovrebbe. Un esempio è quello di soggetto driver con una folta barba ed una maschera senza peli in volto oppure di una machera che porta gli occhiali: il programma potrebbe facilmente generare errori dovuti alla difficoltà di tracciare il volto.

Una volta prese queste accortezze, basterà girare un video in cui si svolgono movimenti che vorremmo riportare sulla "maschera".

3.4.2 Il programma First Order Model

Disponendo ora dei file necessari possiamo passare alla fase successiva, l'avvio del programma.
Il programma utilizzato nella seguente guida per la parte video è disponibile sul sito di Google Colab. Questo significa che non sarà necessario scaricarlo ma che tutti i passaggi verranno effettuati online utilizzando il proprio account Google, rendendo superfluo l'utilizzo di un computer dalle alte prestazioni. La creazione di deepfake molto elaborati, ha di fatti bisogno di un supporto tecnologico non indifferente. Non è possibile aspettarsi grandi risultati utilizzando un computer dalle prestazioni limitate, dovendo elaborare, come già detto, migliaia di fotogrammi da sovrapporre nel punto giusto al momento giusto con la luce giusta.

Il programma in questione si chiama "First Order Model" ed è presente al seguente link: https://colab.research.google.com/drive/1aJpE2WeuKcCUcA_m2RLyguuQCh3yRZCm?usp=sharing

Il programma è scritto in Python, una delle lingue di programmazione più diffuse. Sebbene questo dato possa sembrare secondario per chi non si intenda di

programmazione, non lo è affatto. Il fatto che si tratti di un programma scritto tramite un linguaggio informatico comprensibile a molti, ne favorisce la diffusione ed il miglioramento da parte degli utenti che ne vengano a contatto. Come accaduto per i primi programmi deepfake, anche questo è stato, e tuttora viene, continuamente modificato per essere migliorato e semplificato

"First Order Model" è pensato e strutturato affinché in soli otto passaggi si possa ottenere dai propri input (la "maschera" ed il "driver") un video deepfake senza audio. I dettagli sul funzionamento del programma saranno elencati ed esplicati nei paragrafi successivi. I passaggi sono tra loro separati, divisi in quanto ognuno è scritto come un programma a sé in una cella. Quasi tutte le celle riportano un tasto di inizio a sinistra della prima riga, che appunto serve per avviare le fasi del programma singolarmente e che avverte se ci sono stati errori. Ciò risulta estremamente utile in quanto sarà più semplice accorgersi in che punto si presenteranno eventuali errori.

Andiamo ad analizzare ora i passaggi del programma "First Order Model".

3.4.3 Clone repository and navigate to the directory

Il primo passaggio del programma di divide in due parti, la prima è "Clone repository". Grazie al comando **git clone** seguito dall'URL del repository è infatti possibile duplicare il depository Git dalla sua posizione remota sul proprio dispositivo locale. Un repository Git è una struttura di archiviazione di dati utilizzata per tenere traccia delle versioni di un progetto software o di un insieme di file. Un repository Git contiene l'intera cronologia delle modifiche apportate ai file nel progetto, consentendo a più persone di collaborare sullo stesso progetto in modo efficiente.

La copia del depository Git è il primo passaggio da effettuare poiché fondamentale per l'esecuzione del programma. Si tratta poi di un elemento comune tra i programmi atti alla realizzazione dei deepfake, in quanto presenta caratteristiche che favoriscono la diffusione, la modifica ed il miglioramento del programma stesso. Un depository Git, infatti, rende possibile accedere a tutte le versioni precedenti dei file, consentendo di recuperare facilmente versioni precedenti o confrontare le differenze tra di esse. Inoltre, più persone possono collaborare su un progetto, caricando e scaricando modifiche dal

repository remoto. È inoltre possibile creare rami separati all'interno del repository, tali rami possono contenere modifiche in corso o esperimenti senza influire sulla versione principale del progetto. Si possono unire le modifiche da un ramo in un altro per combinare il lavoro di diverse persone o per incorporare nuove funzionalità. Può essere ospitato su un server remoto, come nel nostro caso, per consentire la condivisione e la collaborazione su larga scala. Come riportato nel capitolo 1, la nascita dei programmi deepfake si basa sulla condivisione dei file dei primi programmi. File che gli utenti hanno condiviso e modificato a loro volta in modo tale da renderli sempre più efficienti. Dopo aver copiato il repository Git, la seconda parte del primo passaggio, "navigate to the directory" ci consente di spostarci nella cartella appena creata contenete i file ed il codice sorgente del repository.

3.4.4 Download sample file and checkpoints or use your own

In questo passaggio il programma non farà nulla se non indirizzarvi ad un link di Google Drive dove poter scaricare i file utilizzati per questo progetto. È possibile utilizzare in parte file diversi, come a breve illustreremo.
Cliccando sul link si viene reindirizzati ad una pagina di Google Drive contenente quattro file diversi: una foto, un video e due file tar necessari per l'esecuzione dei passaggi successivi. Come anticipato, si possono utilizzare un'altra foto ed un altro video diversi da quelli presenti in questo link ma bisognerà comunque utilizzare i file tar. Ipotizzando di voler seguire la guida passo per passo procediamo utilizzando tutti i file presenti nella cartella. Come prima cosa selezioniamo tutti i file e creiamone una copia che sarò visibile sulla nostra home di Drive, spostiamoci quindi sulla home e creiamo una nuova cartella che intitoleremo "First Order Model" (attenzione, è importante che il nome sia questo, non cambiatelo e fate attenzione anche a maiuscole e spazi, altrimenti il programma genererà errori). Spostiamo quindi le copie della maschera, del driver e dei due tar in questa nuova cartella e rinominiamole tutte eliminando il prefisso "copia_di".

3.4.5 Mount your Google Drive folder on Colab

Tramite il comando from **google.colab import drive drive.mount('/content/gdrive')**, avviando questa terza fase del programma, verrà consentito l'accesso ai file e alle

directory presenti nel proprio account Google Drive direttamente dall'ambiente di sviluppo Colab.
Questo comando richiederà l'autorizzazione e genererà un link che permette di accedere al proprio account Google e concedere a Colab il permesso di accedere a Google Drive. Dopo aver concesso l'autorizzazione, la propria cartella di Google Drive verrà montata in Colab nella posizione specificata, nella directory **/content/drive**. Ora sarà possibile accedere ai propri file e cartelle su Google Drive tramite Colab, quindi anche alla nostra cartella “First Order Program”.

3.4.6 Making a display function

Questo passaggio pone le basi per l’animazione che verrà creata successivamente. La funzione **display** è utilizzata per creare e generalmente visualizzare (non in questo caso) un’animazione. Questa necessita di due elementi: **source,** che funge da immagine d’origine e si comporta come modello da modificare, e **driving**, una sequenza di immagini considerate come immagini guida ovvero il nostro “driver”. Il programma servirà per analizzare frame per frame il video “driver”, tracciarne i movimenti e riportarli sulla “maschera”.

3.4.7Create a model and load checkpoints

Dopo l'esecuzione di queste righe di codice, si otterranno due variabili: **generator e kp_detector.**

La prima rappresenta il generatore del modello. Il generatore è responsabile per creare i frame di output basati sul file che indicheremo al programma. La seconda variabile rappresenta invece il rilevatore di punti chiave (keypoint detector). Questo viene utilizzato per individuare i movimenti del volto nei frame del video driver, in modo da utilizzarli per replicare il movimento sulla sorgente durante il motion transfer.

In sintesi, questo passaggio serve a preparare il modello e i checkpoint necessari per eseguire il motion transfer (il trasferimento dei movimenti del driver sulla maschera).

3.4.8 Choose image and driver path

In questa fase dovremo comunicare al programma i due path, ovvero la maschera ed il driver, che abbiamo scelto. Avremo a disposizione due celle in cui inserire le nostre scelte, una per path. In entrambi i casi dovremo indicare posizione e nome del file da utilizzare, quindi "**First Order Model/NOMEDELFILE**". Nel nostro caso, per quanto riguarda la maschera, andremo a scrivere "First Order Model/Trump.png" nell'input denominato **face_path**, mentre inseriremo "First Order Model/driver.mp4" nel input denominato **driver_path**. Avendoli così indicati ed avviando anche questa fase, la nostra intelligenza artificiale saprà quali file andare ad utilizzare nella successiva.

3.4.9 Start predictions

La settima fase del programma è sicuramente la più importante, essendo quella che presenta l'effettivo uso dell'intelligenza artificiale e che ci permette di ottenere il nostro video svolgendo una funzione di face swapping (scambio di volti) basandosi su un sistema di deeplearning.

Prima di tutto la settima fase svolge l'attività di Importazione delle librerie. Tale parte del programma inizia importando le librerie necessarie alla creazione dell'output finale. Tra le librerie importate ci sono **imageio** per la lettura e riscrittura delle immagini e **skimage** per la manipolazione delle stesse. Inoltre, vengono importate altre librerie, tra cui **HTML,** il cui scopo è illustrato più avanti.

Successivamente avviene il caricamento dell'immagine di origine e del video di guida. Viene caricata l'immagine "maschera" indicata nel sesto punto dal percorso, **source_image,** utilizzando la funzione **imageio.imread**. Viene poi creato un lettore video utilizzando la funzione **imageio.get_reader** per leggere il video driver dal percorso (sempre specificato nel punto 6) **driver_path**.

Da qui è possibile per il programma estrarre informazioni sul video di guida. Il First Order Model estraendo il frame rate (fps) dal video di guida utilizzando **reader.get_meta_data()['fps']** legge il video driver frame per frame, aggiungendo questi ultimi a una lista chiamata **driving_video**. Questa lista conterrà quindi tutti i frame del video di guida.

Dopo un veloce ridimensionamento dell'immagine e del video, l'immagine di origine o maschera e tutti i frame del video driver vengono ridimensionati a 256x256 pixel utilizzando la funzione **resize** di **skimage**; avviene la creazione dell'animazione di face-swapping. Tale funzione nel programma è denominato **make_animation** e contiene tra i suoi argomenti sia **source_image**, in modo da ritrovare la maschera, sia **driving_video,** che è la lista dei frame estratti dal video driver, sia le funzioni **generator** e **kp_detector**. Questi ultimi sono modelli preaddestrati, utilizzati per la generazione di frame del video il face-swapped.
Altri parametri come **relative** e **adapt_movement_scale** influenzano il movimento dell'animazione.

I frame risultanti dall'animazione vengono poi riuniti, convertiti in un video e salvati sottoforma di file MP4 utilizzando **imageio.mimsave**.

Infine, l'animazione risultante viene visualizzata direttamente sullo schermo, nello spazio sottostante la cella contenente il codice del settimo passaggio, utilizzando le funzioni **HTML** e **display**.
In sintesi, la penultima fase del programma carica l'immagine maschera ed il video driver, li ridimensiona, utilizza un modello preaddestrato per eseguire il face-swapping tra i due file indicati nel sesto punto; quindi, genera l'animazione risultante come un video MP4 e la riproduce sullo schermo.

3.4.10 Download the resulted video

Si tratta dell'ultimissima fase del programma First Order Model. La funzione **files.download("/content/output.mp4")** viene utilizzata per scaricare un file chiamato MP4 dalla directory "/content" dell'ambiente di esecuzione direttamente sul computer locale.
Ecco come agiscono nello specifico queste ultime due righe di programma:

la funzione **from google.colab import files** importa il modulo **files** da **google.colab**, che fornisce funzionalità per gestire il caricamento e il download di file all'interno dell'ambiente di sviluppo Google Colab.

Mentre la riga di codice **files.download("/content/output.mp4)** esegue il download del file "output.mp4" (il video generato nel settimo passaggio) dalla directory "/content" all'interno dell'ambiente di esecuzione di Google Colab. Il percorso "/content/output.mp4" specifica il percorso completo del file che si desidera scaricare.

Dopo l'esecuzione di questa riga di codice, Google Colab avvierà il processo di download del file "output.mp4" sul computer locale.
Una finestra pop-up apparirà per consentire di scegliere la posizione di destinazione per il download.
In sostanza, questa riga di codice permette di scaricare il file generato in precedenza su Google Colab direttamente sul computer, rendendolo accessibile per l'uso o la visualizzazione al di fuori dell'ambiente di Colab.

Il file ottenuto corrisponderà a questo:
https://drive.google.com/drive/folders/1A4PynSC_4_46UeFe-FhuwSn3f0AEvTV2?usp=sharing

Capitolo 4. Conclusioni

Quello del deepfake è un fenomeno complesso, sintomo del progresso tecnologico e della componente sociale umana. Tale tecnologia si basa su una materia che sta prendendo sempre più piede nel nostro quotidiano: l'Intelligenza Artificiale. L'IA trova nel deepfake la possibilità di mostrare il proprio potenziale, elaborando dati complessi ed emulando il funzionamento di una rete neurale, ma soprattutto, migliorandosi di volta in volta acquisendo da sola dati in base all'output ottenuto.

Grazie a questi fattori, la tecnologia deepfake si è dimostrata sia una grande risorsa che un grande pericolo. La componente sociale, unita alla relativa semplicità della sua esecuzione, la rendono un'arma pericolosa se usata nel modo sbagliato. Tale fattore è degno di nota in quanto, il fatto che si tratti di una tecnologia sviluppatasi e diffusasi su internet, la rende pericolosamente facile da utilizzare da chiunque, con qualsiasi finalità.

Sin da subito, infatti, l'oggetto dei vide diffusi dagli utenti riguardavano contenuti in voga su internet. Come abbiamo già detto, le tipologie principali di deepfake, fin da subito, potevano essere suddivise in tre gruppi: i meme, le fake news, i deep porn.
Non a caso la nascita della terminologia "deepfake" risale al 2017 anno in cui sul social network "Reddit" fece la propria comparsa una delle prime (se non la prima) pagina che pubblicava dei video deepfake: "r/deepfake". I contenuti pubblicati dalla pagina erano video pornografici in cui i volti delle attrici erano stati sostituita da quelli di celebrità. Il deepfake nasce quindi sfruttando e danneggiando l'immagine di personaggi famosi, in particolar modo delle attrici dei film Hollywoodiani.

Col passare del tempo gli utenti di internet hanno iniziato ad usufruire di questa tecnologia soprattutto per la diffusione di fake news.
In pochi anni di vita i deepfake sono diventati lo strumento più pericoloso per il danneggiamento dell'immagine, specialmente per i politici.

Data la contrapposizione, nata sin da subito, tra deepfake e diritto d'immagine, si rende sicuramente necessaria una chiara regolamentazione. Trattandosi in parte di video che potrebbero avvalersi del diritto alla satira, bisogna infatti ricordare che non tutti i deepfake hanno finalità diffamatoria; non è quindi possibile negare agli utenti l'utilizzo di tale tecnologia. Un primo passo per la soluzione del problema è stato compiuto singolarmente da vari stati, ma con l'approvazione dell'AI act si procederà probabilmente verso una regolamentazione sempre più coerente e sicura per la popolazione dell'Unione Europea.

I deepfake sono solo un esempio di nuovi strumenti tecnologico poco, o per nulla, regolamentati. Col passare del tempo, sempre più tecnologie alla portata di tutti potrebbero generare fenomeni sociali pericolosi come quello delle fake news realizzate con l'ausilio dell'IA. La regolamentazione dei deepfake potrebbe comportare la nascita di un ottimo precedente che renda più facile in futuro la gestione delle nuove tecnologie da parte della legislatura.

Bibliografia e sitografia

Marino G., Corpi Mediali, https://iris.unito.it/retrieve/e27ce430-aa8b-2581-e053-d805fe0acbaa/Marino_2014_Corpi_mediali_Keep_Calm_%26amp%3b_Do_the_Harlem_Shake_Meme.pdf

Somalvico M., Intelligenza Artificiale, https://schiaffonati.faculty.polimi.it/pubblicazioni/H1.pdf

Bregler C., Covell M., Slaney M., Video Rewrite: Visual Speech Synthesis from Video, https://www.isca-speech.org/archive_open/archive_papers/avsp97/av97_153.pdf

IBM cloud education, *What is machine learning?*, https://www.ibm.com/topics/machine-learning

Garante per la Protezione dei Dati Personali, Vademecum dicembre 2020, https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9512226

Commissione Europea, *Proposta di REGOLAMENTO DEL PARLAMENTO EUROPEO E DEL CONSIGLIO CHE STABILISCE REGOLE ARMONIZZATE SULL'INTELLIGENZA ARTIFICIALE (LEGGE SULL'INTELLIGENZA ARTIFICIALE) E MODIFICA ALCUNI ATTI LEGISLATIVI DELL'UNIONE*, *https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:52021PC0206&from=EN*

Garante per la Protezione dei Dati Personali, REGOLAMENTO GENERALE SULLA PROTEZIONE DEI DATI, https://www.garanteprivacy.it/documents/10160/0/Regolamento+UE+2016+679.+Arricchito+con+riferimenti+ai+Considerando+Aggiornato+alle+rettifiche+pubblicate+sulla+Gazzetta+Ufficiale++dell%27Unione+europea+127+del+23+maggio+2018

Longo A., Il Parlamento europeo approva l'AI Act, cosa cambierà per le nostre aziende?, ilsole240re.com, https://www.ilsole24ore.com/art/il-parlamento-europeo-approva-l-ai-act-cosa-cambiera-le-nostre-aziende-AE2aw3gD?refresh_ce=1

Redazione, Arriverà quest'anno l'agenzia spagnola per l'intelligenza artificiale, Notizie.AI, https://www.notizie.ai/arrivera-questanno-lagenzia-spagnola-per-lintelligenza-artificiale/

Sibilla F., Guerreschi E., UK – I nuovi scenari dell'Online Safety Bill e del Digital Competition Bill, dirittodiinternet.it, https://dirittodiinternet.it/uk-i-nuovi-scenari-dellonline-safety-bill-e-del-digital-competition-bill/

Briscoe S., U.S. Laws Address Deepfakes, asisonline.org, https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/

Goldin L., La Cina regolamenta l'uso deepfake, Istitutoconfucio.unimi.it, https://www.istitutoconfucio.unimi.it/2023/01/la-cina-regolamenta-luso-deep-fake/

Canton Ferrer C., Dolhansky B., Pflaum B., Bitton J., Pan J., Lu J., Deepfake Detection Challenge Results: An open initiative to advance AI, ai.meta.com, https://ai.meta.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

Il Sole 24 Ore, Trump arrestato a Atlanta: le immagini della storica foto segnaletica, ilsole24ore.com, https://www.ilsole24ore.com/art/trump-arrestato-storica-foto-segnaletica-e-schedatura-poi-rilascio-cauzione-AFbEnRe

Schick N., Deepfakes: The Coming Infocalypse, Twelve, 2020

Warwick K., Intelligenza Artificiale - Le basi, Dario Flaccovio Editore, 2015

Downey A. B., Think Python: How to Think Like a Computer Scientist, seconda edizione, O'Reilly Media, 2015

Moore A. D., Python GUI Programming - A Complete Reference Guide: Develop responsive and powerful GUI applications with PyQt and Tkinter, Packt Publishing, 2019