Deepfake: quali conseguenze e rimedi se a "metterci la faccia" è l'intelligenza artificiale?

di **CARMINE ANDREA TROVATO E ELISA SIMIONATO**

Abstract

Artificial Intelligence systems, and deepfake technology in particular, play an increasingly central role in everyday life. This often implies important consequences for citizens' rights, both in a public and private way. A key to understanding the phenomenon, which is necessary to navigate the ecosystem of increasingly realistic applications, is not to demonize deepfake as a technology, rather to consider it as a mere tool. The approach of the legislator, who is proposing a systemic regulation aiming to prevent any abuse of the tool and not to ban its use, also seems to move in this direction. The human factor is given a predominant role, that is, the user's use of the technology to pursue his own purposes.

1. La tecnologia di deepfake: inquadramento della fattispecie e analisi delle principali caratteristiche. Duecentoventimila euro. L'equivalente di duecentoquarantatremila dollari. Per perdere (o guadagnare) una cifra del genere a causa di una truffa bisognerebbe essere delle vittime piuttosto sprovvedute, o dei truffatori abbastanza scaltri. Non sarebbe esattamente il pensiero del CEO di una società inglese del settore energetico che, una mattina di marzo del 2019, si è visto passare una telefonata dalla sua segretaria. All'altro capo della cornetta, il *Chief Executive* della compagnia sorella che batte bandiera tedesca gli parla di un'operazione sulla quale i due collaborano, chiedendogli di trasferire dei fondi a un fornitore con sede in Ungheria. Duecentoventimila euro. La *deadline* è stringente, il trasferimento deve essere completato entro un'ora per confermare l'accordo. La controparte scalpita, l'azienda non può perdere

questa occasione. È chiaro che l'operazione è importante, tanto più che l'ordine arriva da una carica così alta in persona. Il CEO inglese invia il denaro. I soldi non sono mai stati recuperati[1].

La compagnia assicurativa Euler Hermes, che ha coperto la cifra pagata dalla società ai truffatori, non intende rivelare il nome dalla vittima del raggiro, sua cliente. Sostiene, però, che si è trattato del primo caso di truffa realizzata sfruttando un "artifizio" così particolare. Si è infatti andati ben oltre il classico "*man in the middle*"[2], che già di per sé rappresenta l'esito di un certo sforzo bellico sul piano tecnologico da parte dei truffatori: è stata utilizzata una tecnologia che ha permesso a questi ultimi di sfruttare un alter ego del CEO tedesco, esistente solo nella sua versione digitale, che appare pienamente padrone delle sue azioni al pari dell'essere umano rappresentato, mentre nella realtà dei fatti è completamente sotto il controllo dei suoi creatori. Ciò è stato possibile sfruttando l'intelligenza artificiale e, in particolare, una sua applicazione: la tecnologia deepfake.

Il tema dell'IA suscita sempre maggiore attenzione in ragione della rapidità con cui, negli ultimi dieci anni, le sue applicazioni pratiche si stanno moltiplicando ed evolvendo. Una di esse è, appunto, il deepfake, un tempo fantasia di qualche regista e, ad oggi, realtà che permette di toccare con mano i frutti (e le conseguenze) di questo particolare impiego dell'IA. Il termine deepfake si riferisce a una serie di contenuti multimediali, quali audio, foto, video aventi ad oggetto esseri umani o animali, alterati digitalmente e fra loro ricombinati allo scopo di creare un nuovo e autonomo contenuto-personaggio che interagisce con i terzi grazie a tecnologie di *machine learning* e *deep learning*[3]. Questi ultimi

[1] C. Stupp, *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*, in *The Wall Street Journal*, 2019. La notizia è stata inoltre riportata e commentata anche all'interno di CLUSIT, *Rapporto Clusit 2020 sulla sicurezza ICT in Italia*, 2020, 156.

[2] Con la locuzione "*man in the middle*" (anche abbreviato in *MITM, MIM, MIM attack* o *MITMA*, in italiano "uomo nel mezzo") si intende una tipologia di attacco informatico che si realizza quando un soggetto, il cosiddetto "uomo nel mezzo", segretamente si intromette all'interno di una comunicazione tra parti che credono di comunicare direttamente tra di loro. L'"uomo nel mezzo", una volta inseritosi nello scambio, può leggere i messaggi, inviarne di nuovi, oppure intercettarli e ritrasmetterli dopo averli modificati, mantenendo le parti del tutto ignare della sua presenza.

[3] Ai fini di una maggiore precisione, è necessario sottolineare che i termini citati, "*machine learning"* e "*deep learning"*, sono spesso utilizzati in modo intercambiabile in un contesto atecnico.

sono sistemi capaci di migliorare le proprie performance o produrne di nuove grazie a un continuo lavoro di analisi dei dati utilizzati.

Si potrebbe affermare senza timore di smentita che il personaggio "viva di vita propria" dal momento che i contenuti prodotti parlano e si muovono con una naturalezza e adesione al modello archetipico da farli percepire non quali meri prodotti digitali, ma come esseri viventi, anzi proprio come quell'essere vivente che stanno imitando[4].

Molti sono gli esempi di deepfake ormai divenuti celebri in rete per la loro verosimiglianza rispetto ai personaggi reali che rappresentano: da Nicholas Cage inserito in film in cui non ha mai recitato come «*Fight Club*»[5] e «*The Matrix*»[6], al video in cui Jim Carrey sostituisce Jack Nicholson in «*Shining*»[7]. Ancora, la finta intervista a Mark Zuckerberg che ha convinto molti ingenui spettatori che il CEO

I due termini, tuttavia, rappresentano in realtà due diverse tipologie di sistemi tecnologici. Semplificando i concetti, si può percepire la differenza tenendo a mente che un sistema di "*machine learning"* identifica una "macchina" che, in modo automatico, "apprende" e migliora le proprie *performance* utilizzando una serie di algoritmi di analisi dei dati; quando si parla di "*deep learning"* si fa, invece, riferimento a un sottoinsieme dei sistemi di "*machine learning"*, caratterizzati dalla peculiarità della presenza, alla loro base, una complessa struttura di algoritmi che richiamano il modello del cervello umano. Al fine di un maggiore approfondimento sul tema, si veda, *ex multis*, L. Zhang, J. Tan, D. Han, H. Zhu, *From machine learning to deep learning: progress in machine intelligence for rational drug discovery*, in *Drug Discovery Today*, 2017, Vol. 22, n. 11, 1680-1685.

[4] Si consenta una digressione di natura letteraria. La situazione descritta riporta alla memoria, per certi versi, l'opera di Andrew Hodges*, Alan Turing - Storia di un enigma: The Imitation Game*. Hodges, nel 1983, descriveva il lavoro e la vita del matematico Alan Turing, che durante la Seconda Guerra Mondiale prestò in suo ingegno a favore dell'Inghilterra presso Bletchley Park, il sito militare in cui l'esercito del Regno Unito istituì la principale unità di crittoanalisi del Paese. Turing contribuì in modo decisivo alla creazione di un macchinario capace di decifrare "Enigma", il complesso sistema ideato dall'esercito tedesco per criptare le proprie comunicazioni. L'apparecchio ideato da Turing e dalla sua squadra è considerato l'archetipo degli odierni computer; se è vero, dunque, che la storia della nascita di quelli che oggi sono i nostri terminali è partita da un "*imitation game*", c'è un che di poetico nel constatare che l'evoluzione della tecnologia sta muovendo proprio sui binari dell'imitazione.

[5] *Fight Club,* diretto da D. Fincher (1999, USA, 20th Century Fox). Il video deepfake che riproduce le scene del film inserendo artificiosamente il volto di Nicholas Cage nei fotogrammi originariamente presenti della pellicola è reperibile al seguente indirizzo: https://www.youtube.com/watch?v=bVy2xwW3MHc.

[6] *The Matrix,* diretto da A. e L. Wachowski (1999, USA e Australia, Warner Bros.). Il video deepfake che riproduce le scene del film inserendo artificiosamente il volto di Nicholas Cage nei fotogrammi originariamente presenti della pellicola è reperibile al seguente indirizzo: https://www.youtube.com/watch?v=bVy2xwW3MHc.

[7] *The Shining,* diretto da S. Kubrick (1980, USA e Regno Unito, Warner Bros.). Il video deepfake che riproduce le scene del film sostituendo il volto di Jim Carrey a quello dell'attore presente nella versione originale della pellicola, Jack Nicholson, è reperibile al seguente indirizzo: https://www.youtube.com/watch?v=Dx59bskG8dc.

di *Facebook* stesse pubblicamente dichiarando che il social network ruba i dati degli utenti[8], oppure, per restare entro i confini nazionali, il video andato in onda nel settembre 2019 durante la trasmissione televisiva «Striscia La Notizia» in cui l'ex primo ministro Matteo Renzi esprime alcune opinioni poco "politicamente corrette" a proposito di alcuni dei suoi colleghi parlamentari[9]. Questi esempi[10] si aggiungono a quelli che avevano suscitato scalpore a partire dal 2017, anno della pubblicazione su *Reddit* dei primi video pornografici artefatti che ritraevano diverse celebrità (fra gli altri, Aubrey Plaza, Daisy Ridley, Gal Gadot, Natalie Portman, Scarlett Johansson, Meghan Markle, Taylor Swift e persino l'ex *first lady* americana Michelle Obama)[11].

Questo iperrealismo dipende dal fatto che il meccanismo di *deep learning* genera i *fake* imparando a replicare fedelmente la mimica, le espressioni del viso, il tono di voce, la cadenza e l'inflessione della parlata dei soggetti: da qui infatti il nome deepfake, combinazione dei termini "*deep learning*" e "*fake*"[12]. Semplificando il concetto, si può dire che un deepfake opera confrontando e mappando le caratteristiche di una persona e sostituendole a quelle di un'altra, normalmente utilizzando come base diverse foto o altri contenuti originali, benché sia possibile realizzare un deepfake anche partendo da un singolo contenuto originale (come un banale *selfie*)[13].

[8] La falsa video intervista a Mark Zuckerberg è stata pubblicata sul social network *Instagram* al link di seguito riportato, diventando poi virale e venendo riportata da diverse testate giornalistiche e canali *YouTube*. Si veda per l'originale: https://www.instagram.com/p/ByaVigGFP2U/?utm_source=ig_embed&utm_campaign=embed_video_watch_again.

[9] L'estratto della trasmissione televisiva *"Striscia La Notizia"* contenente il video deepfake dell'On. Matteo Renzi andato in onda il 23/09/2019 è riportato dal sito internet del programma stesso al seguente indirizzo: https://www.striscialanotizia.mediaset.it/video/il-fuorionda-di-matteo-renzi_59895.shtml.

[10] Si veda L. Guarnera, O. Giudice, C. Nastasi, S. Battiato, *Preliminary Forensics Analysis of DeepFake Images*, in *2020 AEIT International Annual Conference (AEIT)*, 2020, che riporta altri esempi oltre a quelli citati nelle note precedenti.

[11] M. Westerlund, *The Emergence of Deepfake Technology: A Review*, in *Technology Innovation Management Review*, 2019, 9(11): 39-52. Questi esempi sono inoltre citati da M.H. Maras, A. Alexandrou, *Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos*, in *The International Journal of Evidence & Proof* 1–8*,* 2018.

[12] Westerlund, *Ibid*.

[13] Westerlund, *Ibid.;* Maras, Alexandrou, *ibid*.

Da un punto di vista tecnico, nella complessa creazione di un contenuto deepfake, vi sono due algoritmi AI che interagiscono fra loro, uno detto "*generator*" e l'altro "*discriminator*"; si tratta di reti neurali artificiali, dunque di modelli computazionali composti da "neuroni" artificiali che ricordano il cervello umano[14]. L'interazione fra il "*generator*" e il "*discriminator*" rappresenta un vero e proprio metodo di "allenamento" delle due reti, che determina il crearsi di una "rete generativa avversaria" (*Generative Adversarial Network*, "GAN") in cui i due modelli implementano reciprocamente le proprie "abilità". Il primo di questi algoritmi, il *generator,* crea un contenuto multimediale *fake* e "chiede" al secondo di determinare se tale contenuto sia reale o artificiale. Ogni volta che il *discriminator* individua dei connotati che gli permettono di determinare la falsità di un contenuto, trasmette queste informazioni al *generator* che, così, impara letteralmente dai propri errori, progredendo nella creazione di modelli sempre più verosimili. Entrambe le reti neurali, attraverso questo metodo di "allenamento", migliorano le proprie abilità: così come il *generator* rende contenuti sempre più sofisticati, allo stesso tempo il *discriminator* aumenta con ogni test la propria capacità di individuare le "spie di falsità"[15].

La creazione di un contenuto deepfake, dunque, non è particolarmente complessa: è la tecnologia AI, nelle sue accezioni di *deep* e *machine learning*, a svolgere il lavoro necessario, migliorandosi sempre più, senza richiedere alcun tipo di supervisione da parte dell'essere umano dopo l'avvio del processo[16].

Questa caratteristica ha contribuito a diffondere la tecnologia deepfake rendendola facilmente accessibile e consentendo il diffondersi di prodotti che, con il passare del tempo, raggiungono livelli qualitativi sempre maggiori[17].

Questo genere di contenuti iperrealistici, una volta condivisi, può generare non poche ricadute pratiche: a prima vista il personaggio raffigurato appare esattamente come il "se stesso" reale e per lo spettatore è difficile, se non

[14] M. Papadatou-Pastou, *Are connectionist models neurally plausible? A critical appraisal*, in *Encephalos* 2011, 48(1):5-12, 2011.
[15] Westerlund, *ibid*.
[16] Maras, Alexandrou, *ivi,* 2-3.
[17] Maras, Alexandrou, *ivi,* 2.

impossibile, accorgersi che sta osservando un falso. Ciò, a maggior ragione, se il soggetto ritratto è noto al pubblico -o, comunque, a chi visualizza il contenuto fasullo- per essere spesso "sopra le righe"[18]: si pensi, ad esempio, al poliedrico CEO di *Tesla*, Elon Musk, che da sempre è noto per le sue "eccentricità", tanto da aver fumato della marijuana mentre veniva intervistato in diretta televisiva, come testimoniato da un video che venne a prima vista considerato un deepfake e solo successivamente dichiarato reale[19]. Inoltre, alla difficoltà di distinguere il vero dal falso, deve sommarsi la notevole facilità e velocità di diffusione dei contenuti attraverso internet, e l'impossibilità di controllare la loro circolazione.

Oggetto di falsificazione sono, generalmente, i contenuti fotografici, video e audio. Questi ultimi necessitano di una costruzione artificiosa che comprende anche la composizione, la lettura e l'interpretazione del testo, che vengono realizzate secondo le stesse modalità imitative sfruttate per la creazione dei contenuti per immagine[20]. Un esempio di come ciò sia possibile è il celebre discorso di 21 minuti che pochi anni fa il defunto Presidente americano John Fitzgerald Kennedy ha tenuto «*direttamente dalla tomba*»[21].

Ad ogni modo, deepfake è uno strumento e come tale va considerato.

È certamente indubbio che i casi di utilizzo che suscitano maggiore scalpore siano quelli in cui viene posta a rischio la reputazione dei soggetti: casi, dunque, di pornografia non consensuale, bullismo, video compromettenti utilizzati a scopi denigratori o ricattatori, sabotaggi politici[22]. Altrettanto allarme è generato dagli utilizzi che, come è accaduto nel caso del deepfake audio descritto all'inizio di questo paragrafo, determinano elevati livelli di rischio in termini di sicurezza.

Non a caso, le grandi multinazionali hanno reagito alzando gli scudi: *Google*, ad esempio, ha creato un database di video fasulli fruibile dai ricercatori

[18] Guarnera, Giudice, Nastasi, Battiato, *ivi*.
[19] J. Mullen, D. Shane, *Weed, whiskey, Tesla and a flamethrower: Elon Musk meets Joe Rogan*, in *CNN Business*, 2018.
[20] Un approfondimento sulle modalità di creazione dei contenuti audio da un punto di vista tecnologico è disponibile qui: G. Lawto, *Generative adversarial networks could be most powerful algorithm in AI*, in *TechTarget*, 2018.
[21] Y. Steinbuch, *Listen to JFKspeak from beyond the grave*, in *New York Post*, 2018.
[22] Maras, Alexandrou, *ivi.*

per allenare i propri sistemi di riconoscimento dei falsi[23], mentre *Facebook* e *Microsoft* hanno lanciato una *Challenge* che premia i più innovativi sistemi di individuazione delle opere deepfake[24].

Tuttavia, lo strumento tecnologico in sé non deve essere demonizzato in quanto tale: quando non viene sfruttata a fini malevoli, questo tipo di tecnologia permette, ad esempio, di semplificare il lavoro di artisti, storici, operatori dello spettacolo, che possono creare opere viventi di grande impatto comunicativo mediante ricostruzioni per immagini o video. Allo stesso modo, è utilizzata da medici e terapisti che possono beneficiare a vario titolo di creazioni computerizzate: dalle applicazioni che "riportano in vita" i defunti per aiutare chi è rimasto in vita a superare il trauma della perdita, ai software capaci di restituire la voce a chi ha perso la possibilità di parlare, fino ai programmi che, sfruttando l'intelligenza artificiale, sono capaci di predire l'insorgenza di patologie come il diabete[25].

2. Le possibili implicazioni sui diritti fondamentali derivanti dall'utilizzo del deepfake: quali conseguenze in concreto? Dal 2017, anno delle prime apparizioni dei deepfake al pubblico degli internauti, l'utilizzo di questa tecnologia ha avuto un rapido sviluppo sia in termini di miglioramento delle prestazioni, sia relativamente alle conseguenze che i deepfake determinano.

I video in circolazione, nell'ordine delle decine di migliaia, hanno generalmente contenuto di carattere pornografico: secondo alcune stime, ciò avviene in oltre il 90% dei casi[26], tanto da configurare quello che è stato definito un vero e proprio «abuso sessuale perpetrato per immagini» («*image-based*

[23] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, *Faceforensics++: Learning to detect manipulated facial images*, in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, 1–11; Guarnera, Giudice, Nastasi, Battiato, *ivi*.
[24] Guarnera, Giudice, Nastasi, Battiato, *Ibid*.
[25] Maras, Alexandrou, *ivi*.
[26] Questi dati sono riportati da H. Ajder, G. Patrini, F. Cavalli, L. Cullen, *The State of Deepfakes: Landscape, Threats, and Impact*, in *Deeptrace*, 2019, 7; si veda anche Westerlund, *ivi*.

sexual abuse»)[27]. Questi contenuti vengono comunemente realizzati anche tramite lo sfruttamento dell'applicativo DeepNude[28] che permette agli utenti di "rimuovere" digitalmente i vestiti da contenuti che, nella versione originale, rappresentano persone coperte[29].

Ad oggi, l'accessibilità dei software capaci di produrre deepfake espone potenzialmente chiunque abbia mai caricato una propria immagine sul web a scoprirsi, un giorno, protagonista inconsapevole di un contenuto di questo tipo[30]. Infatti, fra i casi entrati nella cronaca più di recente, risale al maggio 2021 la notizia di novantaquattro arresti in Korea, tutti in danno di giovanissimi sospettati fra i diciotto e i vent'anni, accusati di aver compiuto reati di diversa natura mediante l'uso di tale tecnologia, utilizzata in particolare per realizzare numerosi video dai contenuti pornografici che ritraevano le vittime ignare intente in atti osceni. Le centoquattordici vittime che sono state identificate sono ancora più giovani: hanno fra i dieci e i vent'anni. Quelle ragazze non erano ovviamente consenzienti né tanto meno consapevoli di come il loro volto fosse utilizzato da completi sconosciuti che ne avevano recuperato l'immagine sui social network[31].

[27] Un approfondimento sul tema, corredato da una riflessione di notevole interesse, è reperibile qui: C. McGlynn, E. Rackley, *Image-based sexual abuse*, in *Oxford Journal of Legal Studies*, 2017, vol. 37, n. 3, 534–561.

[28] DeepNude rappresenta una particolare tipologia di deepfake che utilizza la tecnologia AI per realizzare delle immagini in cui le persone appaiono ritratte senza vestiti. Definita dai suoi creatori come una *"nudifier app"*, essa permette, infatti, di ottenere ritratti senza veli altamente verisimili, grazie ad una serie di piccoli accorgimenti descritti all'utente interessato ad utilizzare il servizio: in particolare, ai fini di un miglior risultato, gli sviluppatori richiedono il caricamento di immagini in cui il soggetto sia ritratto con una buona risoluzione e indossi vestiti quanto più attillati possibile, in modo da svelare la fisionomia del corpo. I prezzi per fruire del servizio sono assai irrisori - circa 0,30 dollari per immagine -, con la possibilità di testare il servizio anche in versione gratuita. La facilità di utilizzo e l'economicità di DeepNude hanno determinato il successo del software, che, parallelamente, ha potuto migliorare notevolmente le proprie funzionalità, sfruttando set di *training* sempre più numerosi, tanto da poter vantare, ad oggi, oltre duemila ore di "allenamento" dalla sua creazione. Maggiori informazioni sono reperibili sul sito ufficiale di DeepNude al seguente indirizzo: https://app.deepnude.cc/upload.

[29] Ajder, Patrini, Cavalli, Cullen, *ivi*, 8.

[30] Si segnala che, per maggiore approfondimento, una densa riflessione sul tema della *deepfake pornography* e sulle implicazioni etiche e morali delle "perversioni" che trovano sfogo nel mondo virtuale è reperibile qui: C. Öhman, *Introducing the Pervert's Dilemma: A Contribution to the Critique of Deepfake Pornography*, in *Ethics and Information Technology*, 2019.

[31] Redazione, *Police arrest 94 suspects over deepfake crimes in 5 months*, in *The Korean Times*, 2022.

Tuttavia, non sono queste le uniche implicazioni per i diritti fondamentali che possono essere determinate da un utilizzo malevolo e doloso della tecnologia deepfake.

In particolare, quando i video ritraggono artificiosamente personaggi politici, l'impatto che questi possono avere rischia di minare gli stessi pilastri alla base delle libertà e dei diritti garantiti nelle società democratiche. Screditare un avversario politico per azioni che in realtà non ha mai compiuto o fargli rilasciare dichiarazioni che diversamente non renderebbe diviene semplice quando si può istruire una macchina a tenere questi comportamenti. Sono numerosi gli esempi: da Nanci Pelosi derisa da Donald Trump perché "ubriaca" in un video[32], ai membri dell'opposizione russa, nella propria "versione *fake*", inconsapevolmente a colloquio con alcuni parlamentari europei[33]. Questo genere di prodotti malevoli della tecnologia deepfake, la cui creazione e diffusione è diretta a interferire con il settore pubblicistico, ha ricadute su temi di grande risonanza, tanto da avere ripercussioni in materia di democrazia a causa della diffusione di vere e proprie *fake news* che appaiono reali, influendo fino al punto di disturbare le elezioni[34], o a sfociare in colpi di Stato come avvenne del 2018 in Gabon[35], o, ancora, a tentare di influenzare le sorti di una guerra, come nel caso del video deepfake in cui il Presidente Ucraino Volodymyr Zelensky esorta le proprie truppe, impegnate nel conflitto con la Russia scoppiato nel marzo 2022, ad arrendersi[36].

[32] S. Mervosh, *Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump*, in *The New York Times*, 2019.
[33] A. Roth, *European MPs targeted by deepfake video calls imitating Russian opposition*, in *The Guardian*, 2021.
[34] P. Barrett, *Disinformation and the 2020 Election: How the Social Media Industry Should Prepare*, in *NYU Stern Center for Business and Human Rights*, 2019.
[35] A. Breland, *The Bizarre and Terrifying Case of the 'Deepfake' Video that Helped Bring an African Nation to the Brink*, in *Mother Jones*, 2019.
[36] Il video del Presidente Zelensky, caricato il 17 marzo 2022 sul sito di "*Ukraine 24*" (canale televisivo *all news* ucraino), è stato velocemente individuato come un falso e ne è stata quanto più possibile limitata la diffusione: in particolare, *Meta* e *YouTube* sono prontamente intervenuti per contenere la condivisione del video falso nelle bacheche dei profili social degli utenti ucraini. Riportano la notizia, *ex multis,* L. Nicolao, *Zelensky chiede agli ucraini di arrendersi: è il primo (e malriuscito) deepfake sulla guerra*, in *Corriere della Sera*, 2022; J. Rhett Miller, *Deepfake video of Zelensky telling Ukrainians to surrender removed from social platforms*, in *New York Post,* 2022.

Peraltro, i diritti fondamentali garantiti da una società democratica e le pubbliche istituzioni non sono che alcuni dei destinatari delle minacce derivanti dall'incontrollata preminenza della tecnologia sulla realtà. Infatti, i prodotti della tecnologia deepfake minacciano molto da vicino anche le organizzazioni economiche, sebbene il fenomeno sia, ad oggi, numericamente meno rilevante[37].

Una volta individuato il bersaglio da colpire, la possibilità di sfruttare software che permettono di creare contenuti personalizzati di questo tipo è sempre più accessibile, sia da un punto di vista economico che di immediata reperibilità, tanto che è possibile parlare di «*deepfake-as-a-Service*», costruiti su richiesta dei clienti dei mercati online[38]. Il caso citato in apertura del deepfake audio che ha ingannato il CEO inglese non è il solo. Anche *Tesla*, ad esempio, è stata presa di mira nel marzo 2019, quando due account *LinkedIn* e *Twitter* di una fantomatica giornalista di *Bloomberg*, "Maisy Kinsley", hanno cercato di connettersi a 195 *shortseller*[39] di *Tesla*. È risultato poi che non solo Maisy non era mai stata nel libro paga di *Bloomberg*, ma la ragazza raffigurata nella foto non era neppure mai esistita: il suo volto altro non era che un'immagine generata da una GAN, il suo profilo era stato creato artificialmente al fine di portare a termine un tentativo di furto di dati e delle azioni di spionaggio industriale[40].

Il furto o la falsificazione dell'identità rappresenta un pericolo anzitutto per le singole persone. Un volto o una voce artificiale artefatta possono, ad esempio, costituire un vero e proprio lasciapassare per superare i diversi sistemi di sicurezza basati su dati biometrici. Minacce concrete sono inoltre rappresentate dallo *spoofing*, cioè il furto di informazioni realizzato falsificando l'identità di una

[37] Ajder, Patrini, Cavalli, Cullen, *ivi*.
[38] CLUSIT, *ivi*, 156.
[39] Lo *shortselling,* o vendita allo scoperto, è una pratica finanziaria che consiste nella vendita di titoli presi in prestito (quindi non posseduti direttamente), che vengono restituiti al prestatore in un secondo momento, dopo che saranno stati riacquistati nell'ambito di un movimento ribassista del mercato. Il *trader* individua il momento più redditizio per la compravendita fondando le proprie previsioni sulla base dell'interpretazione dei dati di mercato in suo possesso. Nel caso in cui le previsioni siano corrette, il *trader* incasserà dunque un importo, proporzionale al numero di contatti scambiati nell'operazione, derivante dalla differenza tra il prezzo di vendita e quello di riacquisto. Si precisa, ad ogni modo, che la definizione qui resa semplifica nettamente il concetto, che imporrebbe una più ampia e dettagliata analisi degli istituti economici, che in questa sede non sarebbe tuttavia pertinente.
[40] Ajder, Patrini, Cavalli, Cullen, *ivi*, 13; CLUSIT, *ivi*, 157-8.

persona o un dispositivo, in modo da ingannare altre persone o dispositivi e ottenere la trasmissione di dati; dalla clonazione di dati biometrici per interagire con sistemi digitali che utilizzano questi dati come password (si pensi, ad esempio, al banale sblocco di uno smartphone attraverso il riconoscimento facciale, o alle autorizzazioni fornite con un comando dato a voce agli assistenti vocali); ancora, dall'uso dell'immagine o della voce di un soggetto sfruttate per carpire la fiducia dei suoi contatti stretti e convincerli con l'inganno a rivelare informazioni sensibili, o a cliccare su link, o a scaricare allegati ad un messaggio che espongono il computer del destinatario al rischio di pericolose intrusioni[41].

Al tema della sicurezza informatica si associa il fattore umano: come detto, le creazioni dell'AI sono estremamente verosimili, sia che ricalchino i lineamenti di persone realmente esistenti, sia che diano vita a volti che presenti soltanto nel web[42].

A volte i "veri" umani hanno la sensibilità necessaria per interpretare questi fenomeni, individuarne la natura e, ove necessario, prenderne le distanze: un esempio è quello dell'influencer virtuale Rozy[43] che, pur appassionando gli utenti di *Instagram* anche fuori dai confini coreani in cui è "nata", è comunque oggetto di riflessioni da parte dei suoi spettatori, che si interrogano circa il suo ruolo nel processo di de-umanizzazione del marketing[44]. Altre volte, invece, gli utenti non riconoscono o non sono in grado di riconoscere la falsità di queste creazioni, il loro elemento "*fake*", con conseguenze immaginabili, quali la perdita di 243.000 dollari inviati con bonifico, in perfetta buona fede, ad una voce creata al computer. Se, come l'"ex Presidente Americano Barack Obama" ricorda ironicamente nel

[41] GPDP, *Deepfake: dal Garante una scheda informativa sui rischi dell'uso malevolo di questa nuova tecnologia*, Doc. Web 9512278, 28/12/20.
[42] Alcuni esempi grafici della verosimiglianza delle immagini create con tecnologia deepfake sono reperibili, ad esempio, in S. Lyu, *Deepfake Detection: Current Challenges and Next Steps*, in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020, 1-6.
[43] Il profilo di Rozy, influencer virtuale che, nonostante la sua declamata "non-realità", ha guadagnato un importante seguito sui social network e in particolare su *Instagram*, è reperibile al seguente indirizzo: https://www.instagram.com/rozy.gram/.
[44] E. Stringhi, *Deepfake e manipolazione dell'identità digitale: rischi e prospettive etico-giuridiche*, in *Agenda Digitale*, 2021.

celebre video deepfake che lo ritrae durante un'intervista mai avvenuta[45], è necessario mantenere sempre un occhio vigile sui contenuti che si ricevono, si visualizzano e si condividono, la realtà dei fatti è che questo non sempre è possibile e, anzi, è spesso l'elemento umano a mettere a rischio i diritti dei cittadini.

3. L'approccio e le misure adottati dalle Istituzioni: una panoramica degli interventi. Di fronte a una problematica che, scorrendo di ripetitore in ripetitore, oltrepassa i confini nazionali e minaccia potenzialmente ogni cittadino del mondo, i legislatori non potevano restare inermi. Se è vero che la tecnologia, come è noto, si evolve più velocemente delle norme, è altrettanto vero che le istituzioni in molti Paesi si sono mosse per far sentire la loro voce sul tema.

Nel panorama nazionale spicca la presa di posizione del Garante per la Protezione dei Dati Personali. Oltre a stigmatizzare fermamente l'utilizzo malevolo della tecnologia deepfake, in particolare la sua estrinsecazione dal contenuto pornografico, il DeepNude, e a sensibilizzare gli utenti del web su questo aspetto[46], il Garante ha diramato un *vademecum*[47] in cui affronta il tema perseguendo gli obiettivi dell'informazione e della prevenzione relativamente ai furti d'identità, al cyberbullismo, alla diffusione di *fake news*, ai crimini informatici e alla *cybersecurity*.

Oltre a spingere sulla prevenzione, le Istituzioni si sono mosse anche per definire un quadro normativo quanto più possibile condiviso. A livello Europeo, la

[45] Il celebre video deepfake che ritrae l'ex Presidente Americano Barack Obama durante un'intervista è stato pubblicato, *ex multis*, dal canale BuzzFeedVideo con il titolo "*You Won't Believe What Obama Says In This Video!*", disponibile al seguente indirizzo: https://www.youtube.com/watch?v=cQ54GDm1eL0.

[46] Il video-intervento della Vice Presidente del Garante per la Protezione dei Dati Personali, Prof.ssa Ginevra Cerrina Feroni, intitolato *Le parole dell'AI - Deepfake e Deepnude nelle parole di Ginevra Cerrina Feroni* è disponibile al seguente indirizzo: https://www.youtube.com/watch?v=OldG6BtLgc0&ab_channel=Garanteperlaprotezionedeidatipersonali.

[47] GPDP, *Deepfake: dal Garante una scheda informativa sui rischi dell'uso malevolo di questa nuova tecnologia*, Doc. Web 9512278, 28/12/20; GPDP, *Deepfake - Il falso che ti «ruba» la faccia (e la privacy)*, 28/12/20, Doc. Web 9512226.

proposta di regolamento in materia di intelligenza artificiale ("Regolamento IA")[48] dell'aprile 2021 ha definito alcune primarie aree di intervento. Sono state identificate infatti le applicazioni proibite in quanto causa di rischi insopportabili per i diritti e le libertà fondamentali; le applicazioni ad alto rischio, classificazione che comporta la necessità di soddisfare particolari condizioni di gestione dei rischi; quelle definite «a rischio limitato» e, infine, altre applicazioni che presentano un livello di rischio trascurabile.

Fra gli utilizzi dell'IA vietati dal Regolamento figurano «l'immissione sul mercato, la messa in servizio o l'uso di un sistema di IA che utilizza tecniche subliminali che agiscono senza che una persona ne sia consapevole al fine di distorcerne materialmente il comportamento in un modo che provochi o possa provocare a tale persona o a un'altra persona un danno fisico o psicologico»; «l'immissione sul mercato, la messa in servizio o l'uso di un sistema di IA che sfrutta le vulnerabilità di uno specifico gruppo di persone, dovute all'età o alla disabilità fisica o mentale, al fine di distorcere materialmente il comportamento di una persona che appartiene a tale gruppo in un modo che provochi o possa provocare a tale persona o a un'altra persona un danno fisico o psicologico»; le pratiche di social scoring ai fini «della valutazione o della classificazione dell'affidabilità delle persone fisiche per un determinato periodo di tempo sulla base del loro comportamento sociale o di caratteristiche personali o della personalità note o previste»[49]; i sistemi di identificazione biometrica da remoto ove il loro uso non sia strettamente necessario o determini un rischio troppo alto di effetti dannosi per i cittadini.

Sono invece previsti una serie di obblighi di informazione e trasparenza, di conformità, di valutazione, e di rispetto delle specifiche di produzione per i sistemi che, pur essendo classificati ad alto rischio, non rientrano nelle applicazioni dell'AI vietate *tout court*[50].

---

[48] *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, Brussels, 21/04/2021 COM(2021) 206 final 2021/0106 (COD)*.

[49] Regolamento AI, art. 5, co. 1, lett. a, b, c.

[50] Questi obblighi di conformità ai requisiti sono previsti dagli artt. 8-15 del Regolamento AI. In particolare, il Regolamento AI prevede che «in relazione ai sistemi di IA ad alto rischio è istituito,

Queste prese di posizione da parte del legislatore europeo devono essere considerate nell'ambito di un più ampio programma di tutele offerte ai cittadini a salvaguardia dei diritti e delle libertà fondamentali minacciate dagli eventuali utilizzi malevoli degli strumenti tecnologici. Tali presidi poggiano, infatti, su una serie di atti emanati nel tempo e tesi a disciplinare, in modo sempre più completo, l'esperienza digitale del singolo: fin dall'entrata in vigore del noto Regolamento Generale sulla Protezione dei Dati (GDPR)[51], infatti, l'impianto normativo comunitario si è fatto carico del compito di guidare il processo di aggiornamento digitale degli Stati Membri dettando principi comuni per i cittadini europei. Ora che l'utilizzo dell'IA rappresenta una realtà sempre più concreta, continua e si attualizza questo approccio europeo di sistema, che va a coinvolgere dunque anche applicazioni dell'Intelligenza Artificiale come deepfake.

Questo tipo di approccio è, peraltro, sistemico: non solo il Consiglio ma anche diversi organismi europei hanno contribuito e contribuiscono al delinearsi di un programma disciplinare comune. Lo European Data Protection Supervisor (EDPS), già nel 2018, si era espresso sul tema del deepfake all'interno della propria «*Opinion on online manipulation and personal data*»[52], nella quale sottolineava le criticità derivanti da un utilizzo malevolo di tali applicativi e dagli strumenti tecnologici in generale, aggravate peraltro dalla rapida diffusione che i contenuti possono avere sul web: l'impatto veniva definito tale da avere effetti

attuato, documentato e mantenuto un sistema di gestione dei rischi » (art. 9, co. 1) «costituito da un processo iterativo continuo eseguito nel corso dell'intero ciclo di vita di un sistema di IA ad alto rischio, che richiede un aggiornamento costante e sistematico. Esso comprende le fasi seguenti: a) identificazione e analisi dei rischi noti e prevedibili associati a ciascun sistema di IA ad alto rischio; b) stima e valutazione dei rischi che possono emergere quando il sistema di IA ad alto rischio è usato conformemente alla sua finalità prevista e in condizioni di uso improprio ragionevolmente prevedibile; c) valutazione di altri eventuali rischi derivanti dall'analisi dei dati raccolti dal sistema di monitoraggio successivo all'immissione sul mercato di cui all'articolo 61; d) adozione di adeguate misure di gestione dei rischi conformemente alle disposizioni dei paragrafi seguenti» (art. 9, co. 2). L'art. 10 dello stesso Regolamento, inoltre, disciplina le caratteristiche dei set di dati di addestramento per i sistemi di AI ad alto rischio.

[51] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, Brussels, 04/05/2016.

[52] European Data Protection Supervisor (EDPS), *Opinion on online manipulation and personal data, Opinion 3/2018*, 19/03/2018.

sostanziali su un importante numero di diritti e libertà garanti dalla Carta Europea dei Diritti Fondamentali[53].

In particolare, l'EDPS sottolineava come ad essere minacciati fossero il diritto al rispetto della vita privata e della vita familiare (art. 7 Carta Europea dei Diritti Fondamentali) e il diritto alla protezione dei dati personali (art. 8 Carta Europea dei Diritti Fondamentali), ma anche la libertà di pensiero, di coscienza e di religione, di espressione e d'informazione, di riunione e di associazione (artt. 10, 11, 12 Carta Europea dei Diritti Fondamentali), oltre alle libertà connesse ai risvolti già analizzati in tema di politica e vita democratica[54].

Conseguenza diretta è il dilagare di flussi (dis)informativi che, diramandosi lungo la rete internet e raggiungendo via via un numero sempre più elevato di utenti, creano delle vere e proprie correnti di disinformazione. Rilevante dottrina ha sottolineato come l'informazione sia uno strumento fondamentale affinché gli individui possano assumere posizioni e formulare decisioni in grado di rispondere alle sfide che la società propone, basando il proprio ragionamento sulle fondamenta della conoscenza che gli stessi hanno dei fatti[55]. Le decisioni informate rappresentano dunque un elemento chiave dell'autonomia degli individui. Tuttavia, quando gli individui vengono in contatto o sono bersagliati da informazioni false ma dall'aspetto realistico, si ritrovano intrappolati in una c.d. "*echo-chamber*"[56], ossia quella situazione in cui un'informazione, per quanto falsa, viene corroborata dal suo ripetersi all'interno di un ecosistema in cui l'utente, volontariamente o inconsciamente, si è inserito. Una volta compromessa l'attitudine dei cittadini di prendere decisioni informate e dunque fondate su un proprio personale ragionamento - in una parola, l'abilità di prendere delle decisioni libere -, anche la loro capacità di intervenire nel discorso democratico, si tratti di un dibattito o di una votazione, viene irrimediabilmente meno e, di riflesso, si incrina il sistema democratico stesso[57].

[53] Ivi*, par. 4,* 12.
[54] Ivi*, par. 4.i,* 12.
[55] N. Bontridder, Y. Poullet, *The role of artificial intelligence in disinformation,* in *Data & Policy*, e32, 6, Cambridge University Press, 2021, 6.
[56] Ivi, 6.
[57] Ivi, 6.

Il tema delle implicazioni e dell'impatto dell'utilizzo di algoritmi di IA rispetto alla tutela dei diritti fondamentali ha evidenti ripercussioni, peraltro, quanto all'effetto prorompente che essi determinano in punto di libertà di espressione, *surveillance* e *censorship*. Questo fenomeno, invero, può operare in maniera bidirezionale, cioè muovendo, come si è visto, dai cittadini verso le istituzioni (ove i primi perdano la capacità di interpretare criticamente i messaggi provenienti dalle autorità) o, viceversa, in taluni casi, partire dalle istituzioni stesse ed essere indirizzato alla popolazione, rivelandosi dunque un processo connaturato da un'intrinseca natura simbiotica. È il caso, questo, per citare un esempio, del sistema di analisi realizzato dal governo cinese capace di analizzare ed interpretare il contenuto dei post pubblicati su oltre millequattrocento servizi di *social media*, al fine di valutare con un processo automatizzato l'opportunità di censurarne la pubblicazione[58]. Ciò, almeno in linea di principio, non dovrebbe essere possibile nel mondo occidentale, dove il diritto di espressione è regolato dalle costituzioni nazionali e riconosciuto come uno dei principali diritti alla base della democrazia: si pensi all'art. 21 della nostra Costituzione o, spostando lo sguardo oltreoceano, al Primo Emendamento della Costituzione degli Stati Uniti d'America. Rispetto a quest'ultimo, è noto, infatti, il tentativo di censura realizzato dall'ex Presidente americano Donald J. Trump al fine di limitare la proliferazione di critiche in risposta ai contenuti pubblicati sul proprio profilo *Twitter*, che nel 2019 ha richiesto l'intervento dell'autorità giudiziaria che si è espressa in difesa della libertà di espressione (e di critica) dei cittadini[59] (si fa riferimento al caso *Knight First Amendment Institute v. Donald J. Trump*[60]).

Ancora, anche dal punto di vista della Cybersicurezza, le applicazioni malevole dei sistemi di IA, con particolare riferimento ai deepfake, hanno attirato l'attenzione delle Istituzioni Europee e dei Garanti nazionali.

[58] G. King, J. Pan, M. E. Roberts, *How Censorship in China Allows Government Criticism but Silences Collective Expression*, American Political Science Review, 107, 2 (May), 1-18, 2013. Si veda inoltre E. Hine, L. Floridi, *New deepfake regulations in China are a tool for social stability, but at what cost?,* Nature Machine Intelligence, Vol. 4, 2022, 608–610.
[59] K. Henry, *Redefining Censorship in the Digital Age*, Departmental Honors Research Paper, Hood College – English and Communication Arts, 2020, 22-26.
[60] *Knight First Amendment Institute v. Trump*, 928 F.3d 226 (2019).

Uno dei più recenti report dell'Agenzia dell'Unione Europea per la Cybersicurezza (ENISA), il «*Remote Identity Proofing - Attacks & Countermeasures*»[61], inserisce esplicitamente il deepfake fra gli «*attack*» che mettono a rischio i sistemi di sicurezza fondati sull'identificazione remota (*remote identity proofing*), elemento che pure è essenziale nella creazione e nel funzionamento di servizi digitali affidabili. Gli altri elementi di minaccia individuati dall'ENISA accanto al deepfake sono «*photo attack, video of user replay attack, 3D mask attack*»[62]: si tratta di attacchi alla sicurezza realizzati mediante la presentazione di prove facciali dell'immagine di un volto stampata o visualizzata tramite lo schermo di un dispositivo, attacco che può essere compiuto anche attraverso un video; oppure, impiegando delle maschere 3D, capaci di riprodurre i tratti reali di un volto umano includendo persino dei fori per gli occhi, in modo da ingannare il rilevamento della vivacità basato sullo sguardo, l'ammiccamento e il movimento anche più impercettibile. Nella stessa sede, peraltro, ENISA propone una serie di misure di prevenzione come «*extra technical controls, process controls or organisational controls*», pur specificando che la *countermeasure* maggiormente efficiente consiste nell'adozione di più misure di protezione e nella differenziazione delle stesse, in un'ottica di rafforzamento generale della sicurezza del sistema, parametrata in considerazione del tipo di attività, del tipo e del numero di utenti e del grado di sicurezza desiderato. Nella progettazione e implementazione di contromisure, continua ENISA, si dovrebbe seguire un approccio di *security-by-design* supportato da un'analisi dei rischi in costante aggiornamento. Ciò, comunque, non significa paralizzare le attività all'atto pratico: nella scelta delle contromisure è necessario trovare il giusto equilibrio tra *effectiveness* e *usability*[63].

[61] European Union Agency for Cybersecurity (ENISA), *Remote Identity Proofing - Attacks & Countermeasures,* 20/01/2022.
[62] Ivi, 3.
[63] Ivi, 29.

Già nel dicembre 2020, peraltro, ENISA aveva messo in guardia i cittadini e le istituzioni politiche dai rischi derivanti da un utilizzo illecito dei sistemi di IA, affrontando la tematica nel proprio report sulle «*AI Cybersecurity Challenges*»[64].

Sono numerosi i richiami ai rischi, ai pericoli, alle conseguenze di un uso scorretto dei nuovi strumenti che la società si trova a maneggiare, più o meno consapevolmente, ogni giorno, attraverso il mero accesso ai più diffusi e comuni dispositivi elettronici.

All'alba della diffusione dei *social media*, ormai un decennio fa, la dottrina salutava con favore la circolazione pressoché capillare di strumenti di interconnessione fra pari che non necessitavano di filtri, fossero essi imposti dalla stampa, dalle istituzioni o da qualunque altro canale di divulgazione. Gli scambi si facevano più rapidi, le comunicazioni superavano i confini territoriali e i messaggi raggiungevano agilmente destinatari sparsi in ogni dove: attraverso tali piattaforme, "*la democrazia poteva essere vissuta più direttamente e in modo più partecipativo*"[65]. Le potenzialità di questi mezzi di comunicazione sono diventate evidenti con lo scoppio della Primavera Araba che dal tardo 2010 ha attraversato il Medio Oriente e il Nord Africa; tuttavia, non molti anni dopo, con lo scandalo *Cambridge Analytica* è risultato sempre più manifesto il potenziale uso malevolo degli stessi[66], ed è bastato qualche anno perché si palesassero anche agli occhi dei più ingenui le minacce cui l'IA può esporre la democrazia mediante la diffusione di *fake news* – si vedano gli esempi citati al paragrafo precedente[67].

Eppure, di fronte ad un mondo che evolve velocemente, l'atteggiamento che traspare è sempre quello di un legislatore attento sì a normare il fenomeno,

---

[64] ENISA (a cura di A. Malatras, G. Dede), *AI Cybersecurity Challenges - Threat Landscape for Artificial Intelligence*, 2020.

[65] A. Kaplan, *Artificial intelligence, social media, and fake news: is this the end of democracy?*, 150, in A. Akkor Gül, Y. D. Ertürk, P. Elmer (a cura di), *Digital Transformation in Media & Society,* Istanbul University Publication, n. 5270, 2020, 149.

[66] Ivi, 150.

[67] A. Kaplan, ivi, 153-155, sottolinea inoltre come siano almeno tre le aree in cui l'IA potrebbe rappresentare una minaccia per i sistemi democratici, e in particolare "*supervision, manipulation, frustration*": *«Firstly, states now have very advanced means of controlling and supervising their citizens' daily behaviour, which could be abused by governments to limit freedoms. Secondly, citizens can increasingly be manipulated in their voting behaviour by ample use of artificial intelligence and social media. Thirdly and finally, such AI-driven supervision and manipulation can, in short, lead to citizens' frustration and their deciding to no longer take part in democracy.».*

ma nell'ottica di prevenire gli abusi di uno strumento e non di vietarne l'utilizzo perché, ancora una volta, è l'elemento umano a poter fare la differenza avvalendosi di un mero strumento per perseguire le proprie finalità.Inoltre, anche spostando la lente dell'osservatore dal lato delle "istruzioni per l'uso" a quello delle tutele per le vittime di utilizzi malevoli della tecnologia in esame, esse possono essere ricercate fra gli istituti previsti dall'ordinamento, sebbene l'interprete possa incontrare alcune criticità nel compiere tale operazione di natura inevitabilmente interpretativa[68]. A titolo di mero esempio, nell'ordinamento italiano potranno essere invocate le norme sulla tutela dei dati personali, oppure si potranno ottenere risarcimenti di natura civilistica determinati da danni all'immagine o, ancora, potranno verificarsi i presupposti per la realizzazione del reato di diffamazione con l'applicazione della relativa disciplina. Pertanto, anche nella situazione di "interregno" in cui il legislatore sta ancora compiendo i passi necessari per disciplinare puntualmente un tema tanto vasto quanto variegato e in continua evoluzione, sarà comunque possibile ricercare fattispecie potenzialmente assimilabili a livello interpretativo.

4. Portare il deepfake in tribunale: quali prospettive future per la tutela da un uso malevolo? Il 23 ottobre 2020 il Garante per la Protezione dei Dati Personali ha annunciato l'apertura di una istruttoria[69] nei confronti di *Telegram*, la nota piattaforma di messaggistica istantanea: come avvenuto in altri esempi citati nei paragrafi precedenti, diverse donne hanno scoperto e denunciato che alcune foto che le rappresentano nude e talvolta in posizioni particolarmente esplicite stanno rimbalzando sugli schermi di perfetti sconosciuti, fra le chat, i gruppi e i canali presenti sull'applicazione. Quelle foto, però, non sono mai state scattate, né da loro, né da nessun altro.

[68] A questa conclusione giungono, a fronte di un esame del quadro normativo statunitense, M. B. Kugler, C. Pace, *Deepfake Privacy: Attitudes and Regulation*, 116 Nw. U. L. Rev. 611 (2021), Northwestern Public Law Research Paper No. 21-04, 2021, 660.
[69] GPDP, *Deep fake: il Garante privacy apre un'istruttoria nei confronti di Telegram per il software che "spoglia" le donne*, Doc. Web 9470722, 23/10/20.

Il caso della diffusione di contenuti DeepNude su *Telegram* è solo uno degli esempi di casi in cui il Garante per la Protezione dei Dati Personali è intervenuto, negli anni, per tutelare il diritto degli interessati a veder correttamente rappresentata la propria identità personale[70] e rispettate la dimensione sociale e affettiva[71] e la reputazione degli utenti[72]. In tutti i casi precedentemente affrontati dall'Autorità, tuttavia, la tecnologia aveva sempre rappresentato un mezzo di diffusione dei contenuti incriminati, non lo strumento stesso di creazione dell'offesa. Peraltro, il caso non solo solleva profili rilevanti dal punto di vista della tutela dei dati personali, ma pone anche gravi risvolti penalistici, ben potendo configurarsi una serie di reati a sfondo sessuale imputabili agli utenti che hanno creato e messo in circolazione tali contenuti. La pericolosità di DeepNude in questo senso è stata poi riconosciuta dal suo creatore, che l'ha ritirata dal mercato; ciò, tuttavia, non ha impedito ad altri *software developer* di replicare le peculiarità dell'applicazione originaria in altri prodotti tuttora facilmente reperibili online.

Ad ogni modo, una "rivoluzione" del genere ribalta completamente l'ottica con cui deve essere affrontato il caso. Ciò, anche da un punto di vista processualistico.

Ha fatto scalpore oltreoceano la storia, raccontata dai quotidiani americani, di una madre della Pennsylvania che, per agevolare la carriera da *cheerleader* della sua giovanissima figlia, ha realizzato dei video falsi delle compagne di squadra più promettenti in modo da rovinare loro la reputazione, con l'intento di farle espellere dalla squadra[73]. Gli articoli che riportavano la notizia, raccontando con dovizia di particolari della creazione di fotografie e video artefatti in cui le studentesse facevano uso di diverse sostanze e, in alcuni casi, erano state anche sottoposte a una modifica delle loro immagini con DeepNude,

[70] Si vedano, ad esempio: GPDP, Provv. 13/09/1999, in Boll., n. 09, giugno 1999, 94, Doc. Web 1090502; GPDP, Provv. 10/10/2002, in Boll., n. 32, ottobre 2002, 3, Doc. Web 1066415.
[71] Si veda, ad esempio: GPDP, Provv. del 07/07/2005, in Boll. n. 63, luglio 2005, Doc. Web 1148642.
[72] Si vedano, ad esempio: GPDP, Provv. del 12/06/2019, Doc. Web 9126859; GPDP, Provv. del 27/11/2019 Doc. Web 9236677.
[73] Redazione, *Mother charged with deepfake plot against daughter's cheerleading rivals*, in *The Guardian*, 2021.

hanno raccolto commenti sdegnati. Ancora di più, però, ne ha raccolti la successiva notizia, riportata sempre dalla stampa, che, nel processo a carico della signora, l'accusa abbia incontrato notevoli difficoltà a livello probatorio, non riuscendo a provare "oltre ogni ragionevole dubbio"[74] l'artificiosità dei video e delle immagini realizzati con deepfake[75].

Non è la prima volta che i Tribunali statunitensi si pronunciano sull'ammissione di prove digitali considerandole «ammissibili nella misura in cui poteva essere verificata la loro affidabilità»[76] (*Nooner v. State of Arkansas*[77]). Eppure, in questo caso il punto di vista è opposto: è infatti la non affidabilità a dover essere non solo verificata, ma anche dimostrata.

Portare il deepfake in tribunale impone e presuppone che le immagini siano analizzate da personale con una notevole preparazione tecnica nel campo dell'*image processing*, ma che abbia anche l'esperienza e le competenze necessarie a interpretare le informazioni estratte alla luce del contesto dell'ambito forense e giudiziario.

Un contenuto deepfake può infatti introdurre nel processo un certo numero di elementi: prendendo ad esempio un'immagine artefatta, con la giusta tecnica sarà possibile ricavare da essa non solo il dato relativo all'immagine falsata in sé, con l'individuazione delle spie di falsità, ma anche le informazioni lasciate dal dispositivo che l'ha generata, le modalità con cui è stata creata, la possibilità di restauro delle immagini deteriorate e/o di recupero dei contenuti originali[78].

Come una sorta di contrappasso, al fine di dimostrare la falsità di un prodotto deepfake, sono stati sviluppati algoritmi e applicazioni AI studiati per verificare l'integrità di un'immagine digitale. Questi sistemi, in costante fase di sperimentazione, parrebbero fornire risultati affidabili; tuttavia, ciò può non

[74] La Corte Suprema Americana utilizzò l'espressione per la prima volta nel 1880, nel caso *Miles v. United States*: «The evidence upon which a jury is justified in returning a verdict of guilty must be sufficient to produce a conviction of guilt, to the exclusion of all reasonable doubt».
[75] D. Harwell, *Remember the 'deepfake cheerleader mom'? Prosecutors now admit they can't prove fake-video claims.*, in *The Washington Post*, 2021.
[76] *Nooner v State of Arkansas*, riportato da Maras, Alexandrou, *ibid*.
[77] *Nooner v State of Arkansas*, 907 S.W.2d 677 (1995).
[78] S. Aterno, *Deepfake in tribunale: ecco come la digital forensics smaschera il falso*, in Agenda Digitale, 2019.

essere sufficiente. Infatti, in conseguenza dell'uso dell'intelligenza artificiale, potrebbero sorgere nuovi problemi legati alla difficoltà di spiegare come questi risultati siano stati ottenuti, dal momento che ad ottenerli è stato un software che compie deduzioni sulla base di analisi "proprie" e non guidate dall'essere umano. Si arriverebbe poi al paradosso in cui l'AI verrebbe utilizzata per individuare i prodotti dell'AI stessa, con una sorta di coincidenza fra controllore e controllato[79] ma, soprattutto, con un rischio di perdita totale di supervisione da parte dell'essere umano. Tuttavia, questa soluzione sembra essere la più promettente nel contrasto agli utilizzi malevoli del deepfake ed è necessario continuare a svilupparla per rintracciare e disinnescare minacce (altrettanto) tecnologiche.

Per tali ragioni la letteratura ha rilevato come si stia sviluppando un concetto di governance multilaterale, fondato sia su principi giuridici che mettono al centro la persona e la costruzione di un ambiente di fiducia nella tecnologia, sia sull'incentivo a un impiego della tecnologia come antidoto alla tecnologia stessa ogniqualvolta il contesto sia di tale complessità da rendere insufficiente e inefficace la sola prescrizione giuridica[80].

Non deve infatti essere demonizzato uno strumento in ragione dei rischi derivanti dal suo utilizzo. Certamente si verificheranno altri casi di impiegati e dirigenti d'azienda presi di mira da qualche truffatore, come capitò al CEO inglese salvato (si fa per dire) dalla Euler Hermes. Inevitabilmente i fautori di determinati interessi politici più privi di scrupoli continueranno a sfruttare ogni mezzo a loro disposizione per trarne vantaggio per la propria carriera, manipolando l'opinione pubblica e diffondendo notizie false. Passerà del tempo prima di eradicare l'abuso compiuto con tanta facilità tramite DeepNude. Ciò non significa, però, che la diretta conseguenza di questi episodi debba essere un divieto assoluto di utilizzo delle applicazioni dell'AI. Al contrario, la fruibilità dei servizi offerti dalla tecnologia è troppo significativa perché l'uomo se ne privi. Esprimere informazioni false, imprecise o fuorvianti non è di per sé condannabile in modo

[79] Maras, Alexandrou, *ibid*.
[80] G. D'Acquisto; C. A. Trovato; L. De Benedetti, *Alcune riflessioni sul concetto di autonomia decisionale della macchina e sulle sue implicazioni regolamentari*, in *La rivoluzione dell'AI: profili giuridici* (a cura di O. Pollicino e M. Bassini), Il Mulino, Bologna, 2022.

generalizzato ed astratto. Ad essere oggetto di condanna e, di riflesso, di regolamentazione, è e deve essere non tanto la qualità del *quid* oggetto della produzione IA (audio, video o grafica), quanto piuttosto lo sfruttamento di questo *tool* per finalità malevole o, nel peggiore dei casi, dolose[81]. Tuttavia, ove sia la tecnologia IA stessa ad offrire un sistema per arginare i propri *malus*, ciò rappresenta, in una valutazione comparativa, di fatto un incentivo all'impiego di tali sistemi e dunque un *bonus* di cui tenere imprescindibilmente conto in sede di esame degli aspetti positivi della tecnologia in questione.

L'essere umano è, peraltro, un animale computazionalmente limitato, con scarse capacità di memorizzazione, di calcolo, di analisi: le sue scelte, per quanto si sforzi, saranno sempre compiute tenendo in considerazione solo una rosa limitata di fattori[82]. L'apporto offerto dai prodotti dell'AI permette effettivamente alla persona di compiere un notevole passo avanti nel processo di evoluzione tecnologica.

Eppure, se è vero che l'essere umano ha bisogno della macchina per supplire alle proprie carenze, egli è al contempo emozionalmente più evoluto di qualunque macchinario potrà mai sviluppare. È dunque compito dell'uomo inquadrare e determinare principi e regole che governino l'utilizzo della tecnologia, anche stabilendo quando e in che modo quest'ultima possa validamente sopperire alle lacune che egli riconosce di avere[83].

Questa direzione è condivisa anche dalle istituzioni a livello sovranazionale: già nel 2018 l'EDPS sottolineava che, alla radice del problema delle conseguenze malevole della tecnologia AI, vi era (in parte) un utilizzo irresponsabile, illegale o non etico delle informazioni personali gestite nell'ecosistema digitale[84]. La trasparenza rappresentava e rappresenta una via necessaria ma non sufficiente; allo stesso modo, un'opera di gestione dei contenuti può rivelarsi utile, ma deve essere realizzata nella misura in cui non comprometta i diritti fondamentali. Già allora, dunque, parte della soluzione è

[81] N. Bontridder, Y. Poullet, ivi, 10.
[82] Ivi*, par. 3.*
[83] Ivi, *par. 3.*
[84] EDPS*, Opinion 3/2018, par. 7,* 22.

stata individuata nell'applicare le regole esistenti (specialmente il GDPR, specificava l'EDPS) con rigore e congiuntamente ad altre norme[85].

È dunque sottile l'equilibrio ideale che permette una proficua "convivenza" con i fenomeni deepfake e DeepNude: da un lato occorrono norme capaci di comprendere e analizzare queste applicazioni; dall'altro, il legislatore deve comprendere il funzionamento della tecnologia ed essere disponibile a uno scambio di mutua integrazione, di sostegno e reciproco completamento, sviluppato in un clima di fiducia per quanto riguarda l'utilizzo delle nuove tecnologie, che non dimentichi la centralità della figura dell'essere umano nel processo e all'esito del processo di regolamentazione[86]. L'obiettivo cui mirano le norme tese a disciplinare deepfake, infatti, non può essere realizzato se non con un supporto tecnologico adeguato.

A questo approccio legislativo si deve affiancare poi un certo grado di consapevolezza che deve essere maturata dai destinatari delle norme e delle tecnologie: vanno incentivate, pertanto, le *corporate policies* e i codici etici da un lato e, dall'altro, le campagne di sensibilizzazione e *training*[87]. Invero, queste misure sono già ad oggi adottate dalle Istituzioni[88] e auspicabilmente saranno indirizzate, nel tempo, a un pubblico sempre maggiore, fino a realizzare un progetto educativo che promuova l'analisi critica dei contenuti reperiti in rete nell'ottica di una ormai imprescindibile alfabetizzazione digitale[89].

---

[85] Ibid, *par. 7,* 22.
[86] G. D'Acquisto; C. A. Trovato; L. De Benedetti, *ibid.*
[87] Westerlund, *ibid*.
[88] In questo senso, si vedano ad esempio le campagne informative promosse dal Garante per la Protezione dei Dati Personali, come GPDP, *Deepfake: dal Garante una scheda informativa sui rischi dell'uso malevolo di questa nuova tecnologia*, Doc. Web 9512278, 28/12/20.
[89] Westerlund, *ibid*. Mika Westerlund, nell'accingersi a concludere tale contributo, dedica una riflessione agli obiettivi dell'alfabetizzazione digitale la cui promozione si pone come un tema imprescindibile con particolare riferimento al percorso educativo dei più giovani: «*In general, there is a need to raise public awareness about AI's potential for misuse. Whereas deepfakes provide cyber criminals new tools for social engineering, companies and organisations need to be on high alert and to establish cyber resilience plans. Governments, regulators, and individuals need to comprehend that video, contrary to appearances, may not provide an accurate representation of what happened, and know which perceptual cues can help to identify deepfakes. It is recommended that critical thinking and digital literacy be taught in schools as these traits contribute to children's ability to spot fake news and interact more respectfully with each other online».*

Nel lungo periodo, infatti, un approccio di questo genere sembra avere i presupposti per evitare – o quanto meno limitare – gli abusi compiuti a mezzo DeepNude, o la diffusione di fake news, o, ancora, le truffe ai danni di qualche azienda, come quella che coinvolse il CEO inglese salvato (si fa per dire) dalla Euler Hermes.