Deepfake: “when seeing isn’t believing anymore”

(di Elisabetta Stringhi)

Il tema del c.d. *deepfake* è stato trattato con crescente preoccupazione, negli ultimi due anni, da informatici e giuristi in virtù del potenziale “disruttivo” (“*disruptive*”) di questa nuova tecnologia.

Con il termine *deepfake* ci si riferisce a una «*Tecnica che utilizza l’intelligenza artificiale per combinare e sovrapporre immagini o video originali, ritraenti una persona, con quelli ritraenti qualcun altro, o per generare immagini o video completamente falsi e difficilmente riconoscibili come falsi*» (G. Ziccardi, P. Perri).

Finora, la maggior parte di accademici, esperti di sicurezza e giuristi si sono preoccupati delle conseguenze del *deepfake* in ambito politico, per finalità distorsive del dibattito pubblico e di diffusione di misinformazione. Tuttavia, fin dalle sue origini, il fenomeno è stato connotato da una netta prospettiva di genere. Infatti, secondo una rilevazione di settembre 2019 (AI, Sensity), il 96% del materiale video *deepfake* circolante su internet è diffusione non consensuale di immagini intime di donne e ragazze.

Prima di esaminare le implicazioni e prospettive giuridiche del fenomeno, è necessaria una premessa circa le sue origini sociologiche e tecnologiche.

In breve, Ian Goodfellow e un gruppo di ricercatori nel campo dell’Intelligenza Artificiale sviluppano le Reti Antagoniste Generative (*Generative Adversarial Networks*, diminutivo *GANs*), una classe di *Machine Learning* in cui due reti neurali sono sottoposte a *training* simultaneo per stimare dei modelli generativi mediante un processo competitivo. Le *GANs* hanno notevoli potenzialità per scopi fotografici, artistici e satirici. Ad esempio,

consentono di scambiare i volti di due persone ritratte in un video, di apporre il volto di una persona sul corpo di un'altra, così come di utilizzare foto, audio e video di un determinato soggetto per simularne movimenti e discorsi.

L'origine di tale fenomeno è da rintracciarsi nella comunità online di Reddit. Difatti, nel 2017, un utente di Reddit, con nickname "deepfakes" appunto, decide di mettere a disposizione della comunità online suggerimenti e consigli su come utilizzare le *GANs* per apporre il volto di donne target sul corpo di attrici pornografiche, così da realizzare e diffondere immagini sessualmente esplicite, senza il loro consenso. Il fenomeno viene scoperto e monitorato da una giornalista di Wired, Samantha Cole.

Presto, sempre più utenti richiedono informazioni su come creare video *deepfake* non soltanto ritraenti attrici, cantanti, *gamers,* lavoratrici dello spettacolo ma anche ex fidanzate, ex mogli, colleghe, insegnanti. Un altro utente, 'deepfakeapp', sviluppa una versione app per telefono del software. La tecnologia per creare immagini e video *deepfake* diventa sempre più accessibile e mercificata, tramite *repositories* online, distribuzioni di app e software e, persino, di marketplace.

Alla luce della democratizzazione dei contenuti *deepfake* e della tecnologia che ne consente la realizzazione, è fondamentale interrogarsi sul potenziale impatto sui diritti e sulle libertà fondamentali delle persone fisiche e giuridiche.

Si ravvisano, in particolare, 4 aree di rischio:

1. Violazione dei diritti della personalità (c.d. *personality rights*),
2. Diffusione non consensuale di immagini intime (c.d. Image-based sexual abuse),
3. Minacce e rischi da una prospettiva di sicurezza informatica e, infine,
4. Guerra dell'informazione (sponsorizzata dallo Stato o da soggetti privati).

1. La diffusione di materiale *deepfake* riguardante persone fisiche potrebbe comportare un danno all'immagine, alla reputazione e all'identità digitale della vittima. Il *deepfake* può senz'altro costituire un mezzo e, al contempo un contenuto idoneo a distorcere notevolmente l'identità, definibile nel suo complesso come un bene-valore giuridico meritevole di protezione da possibili travisamenti del proprio patrimonio intellettuale, ideologico, etico, religioso e professionale.

A tal proposito, è interessante il caso del tentativo di sabotaggio reputazionale operato da una ex coniuge ai danni del marito ai fini di capovolgere l'esito di una controversia per la custodia del figlio.

2. Come anticipato, il fenomeno del *deepfake* ha assunto una particolare connotazione di genere, come confermato dalle rilevazioni di Sensity AI, nonché da episodi di inaudita gravità come quello di Rana Ayubb, una giornalista investigativa indiana nota per le sue inchieste indipendenti. Dopo aver fermamente condannato il partito nazionalista, attualmente al Governo, Bharatiya

Janata Party (BJP) durante due dirette televisive sulla BBC e su Al Jazeera per il loro incondizionato supporto a un uomo indagato per abuso sessuale di minore, senza alcun riguardo nei confronti della bambina coinvolta, Rana è stata presa di mira da una violentissima campagna di impersonificazione, misinformazione, *cyberstalking* e abuso sessuale di immagine con un video *deepfake* realizzato e diffuso viralmente. Il caso di Rana mostra come qualsiasi voce femminile che non tema di farsi sentire sia potenzialmente a rischio di essere silenziata. La stessa racconta:

"*E' stato devastante. Non potevo mostrare il mio volto. Puoi definirti giornalista, puoi definirti femminista ma, in quel momento, non riuscivo a vedere al di là dell'umiliazione*".

Trattasi di un'ulteriore forma di diffusione non consensuale di immagini intime, peraltro stigmatizzata anche dall'Autorità Garante, la quale ha instaurato un procedimento a carico di Telegram, a seguito della diffusione del bot *Deepnude* diffusosi viralmente in gruppi e canali della piattaforma. Questa particolare fattispecie di *deepfake* sembra pertanto colpire prevalentemente categorie di soggetti vulnerabili, quali donne e minorenni.

3. La creazione di materiale audiovisivo comporta rischi di sicurezza informatica. Recentemente, l'Autorità Garante ha segnalato i potenziali rischi di attacchi di tipo *phishing*, *vishing*, nonché di furto di identità. Non si può fare a meno di aggiungere come il *deepfake* possa consentire anche la violazione di misure di sicurezza fondate sul riconoscimento biometrico, oltre che la profilazione dei soggetti interessati. È alquanto significativo il caso di un attacco di tipo *vishing* e *CEO fraud* realizzato ai danni di un amministratore delegato di una celebre società, fornitrice di energia elettrica. Inoltre, sono stati numerosi i casi di tentato spionaggio industriale operato mediante profili social (Linkedin nella specie) creati con il ricorso alle GANs e, naturalmente, all'ingegneria sociale. Il panorama del rischio, dunque, si caratterizza per particolare complessità e mutevolezza degli attacchi di tipo *deepfake*.
4. Nell'era della post-verità e dell'utilizzo politico dei social media, nonché della crescente preoccupazione dovuta ai rischi della misinformazione nel pubblico, è doveroso, infine, fare una riflessione sul fatto che la tecnologia *deepfake* può costituire uno strumento sofisticato ai fini della c.d. "*guerra dell'informazione*". In tal senso, oltre a ricordare la vicenda del c.d. *shallowfake* di Nancy Pelosi, si segnala come si siano già verificati due casi preoccupanti di instabilità politica ed alterazione dei processi democratici dovuti anche alla diffusione di contenuti *deepfake* (Gabon e Malesia).

La tecnologia *deepfake,* con il suo livello di complessità e di sofisticatezza in continua crescita ed evoluzione, costituisce una minaccia per l'attendibilità dei contenuti audiovisivi e, dunque, un rischio di alterazione, modificazione e vera e propria manipolazione della realtà. L'acuta espressione "*liar's dividend*" riassume efficacemente la prospettiva delineata. Con tale termine, Robert Chesney, Danielle Citron e Hany Farid fanno riferimento allo scenario paradossale in cui non soltanto saremmo indotti a

ritenere veri contenuti *deepfake* ma, al contrario, saremmo persino inclini a considerare falsi i contenuti reali. A parere di chi scrive, in conclusione, le criticità più profonde che il *deepfake* pone potrebbero non essere risolte tramite soluzioni tecnologiche e di *digital forensics*, ma sono necessarie la parallela organizzazione di iniziative di sensibilizzazione nel pubblico e di formazione degli operatori del diritto e della protezione dei dati, oltre che l'introduzione di appropriate norme a tutela delle parti offese dall'utilizzo malevolo dell'IA.