Digeat N.6 - 19 Giugno 2025

Alla ricerca dell’autenticità perduta: la tecnologia ci salverà dalla tecnologia?

Di Serena Deplano

Abstract

L’idea che gli algoritmi di Intelligenza Artificiale possano generare contenuti così verosimili da farli apparire autentici a un osservatore ignaro può indurre un senso di vertigine: è solo l’inizio di un vortice nel quale il confine tra ciò che è autentico e ciò che non lo è rischia di diventare sempre più labile e difficile da cogliere? Che strumenti abbiamo per far fronte a questo scenario?

Indice

Tempi bui per i Neal Caffrey: **l’Intelligenza Artificiale, infatti, sta velocemente rubando il lavoro ai falsari**! Se, un tempo, lo sforzo di riproduzione di un’opera richiedeva perizia, tecnica, dedizione e tempo, oggi l’Intelligenza Artificiale è in grado di produrre in qualche manciata di minuti contenuti che all’occhio umano sono perlopiù indistinguibili da quelli reali. Il vorticoso sviluppo dell’IA sta sprigionando, infatti, una straordinaria capacità generativa carica di potenzialità, ma portatrice di non pochi rischi: parliamo di *deepfake*.

Conoscere il fenomeno: alla radice del deepfake

Il neologismo *deepfake* deriva dall’insieme di *deep learning* – i sistemi di apprendimento profondo alla base dei meccanismi di funzionamento dell’Intelligenza Artificiale generativa – e *fake*, cioè falso[1], e deve il suo nome a un omonimo utente di Reddit che, nel 2017, postò video pornografici dopo aver sostituito in modo molto realistico – attraverso algoritmi di *machine learning* – i volti di attori famosi a quelli dei reali protagonisti dei video[2].

> Il termine *deepfake* è così diventato, nel tempo, etichetta di una categoria che indica tutti quei contenuti – video, audio, immagini – generati o manipolati dall’IA attraverso algoritmi di *deep learning* in modo tale da farli apparire autentici.

Per farlo, si sfruttano perlopiù le GAN (generative adversarial network) – reti avversarie generative[3] – e i recenti modelli di diffusione[4]. Le GAN sono modelli di apprendimento che sfruttano due reti neurali dette, rispettivamente, generatore e discriminatore: la prima ha il compito di generare contenuti; la seconda, invece, quello di verificarli per determinare quali siano veri o meno. Durante il periodo di addestramento, entrambe le reti evolvono, migliorando le loro prestazioni, fino a raggiungere il punto di equilibrio: quando, cioè, **il discriminatore non riesce più a distinguere tra contenuto artificiale e reale**.

> Il modello GAN è, quindi, un sistema intrinsecamente competitivo, orientato a un risultato specifico: far sì che la stessa rete di controllo non sia in grado comprendere se un contenuto sia autentico – reale! – o artificiale.

I modelli di diffusione invece funzionano grazie a un processo di *machine learning* in due fasi: progressiva aggiunta di rumore (*noise*) rispetto ai dati di addestramento disponibili e, successivamente, riduzione del rumore. In questa seconda fase, il modello, attraverso algoritmi predittivi, impara a identificare il rumore aggiunto nella fase precedente per poterlo rimuovere in modo efficace. Indirettamente, quindi, impara anche come generare contenuti.

Deepfake per ogni occasione

> Fatta questa premessa, non è difficile comprendere come questi modelli consentano di produrre i contenuti più disparati[5]. Qualche esempio può certamente aiutare a comprendere il fenomeno.

Un grande classico è il ***lip sync***: la manipolazione di video con cui si adegua il movimento delle labbra di una persona a un contenuto audio modificato o addirittura radicalmente sintetico, cioè prodotto artificialmente. In questo modo, di fatto, è possibile attribuire a chiunque qualsiasi affermazione, violando il controllo che le persone hanno sulla manifestazione dei propri pensieri.

Un altro esempio è quello del ***face swap***: una tecnica che consente di scambiare digitalmente i volti presenti in immagini o video. In buona sostanza, l'intelligenza artificiale identifica i tratti di un volto, mappandoli su quello di un'altra persona e generando così lo scambio, con risultati altamente realistici. Si tratta di una tecnica molto utile in campo artistico e cinematografico, ma altrettanto pericolosa se, per esempio, utilizzata allo scopo di diffondere contenuti espliciti con il volto di un target ignaro, come recentemente accaduto alla cantante americana Taylor Swift[6].

Sul fronte del *deepfake* vocale, vale la pena citare il ***voice cloning***, una tecnica che consente di ottenere la riproduzione molto precisa di una voce umana a partire dai suoi tratti più caratteristici, come il timbro e l'intonazione. Anche il *voice cloning* trova applicazione nel campo cinematografico, per esempio per il doppiaggio di film; tuttavia, esso costituisce anche un potenziale pericolo: se la voce sintetica e quella umana non sono più distinguibili, chi può dirci che quella che sentiamo al telefono è davvero la voce di una persona cara che ci chiede aiuto e non quella di un suo clone artificiale che punta a raggiraci per ottenere un vantaggio?

Implicazioni, potenzialità, rischi

> È chiaro dunque: l'IA è in grado di generare contenuti estremamente sofisticati e, soprattutto, realistici[7], con una vasta gamma di applicazioni utili in diversi settori[8]. Al tempo stesso, tuttavia, merita riflessioni per le implicazioni e rischi di cui è portatrice

> [9]. **Un primo, potente, risvolto è quello della diffusione della disinformazione**, intesa – nel lessico della Commissione Europea – come *"un'informazione rivelatasi falsa o fuorviante concepita, presentata e diffusa a scopo di lucro o per ingannare intenzionalmente il pubblico, e che può arrecare un pregiudizio pubblico"* ***[10]***.

Non è difficile ricordare i già numerosi casi in cui contenuti artificiali, grazie alla cassa di risonanza dei social media, si sono diffusi in modo virale, con lo scopo di ingannare i destinatari, ad esempio, influenzando l'opinione pubblica nei contesti sociali più sensibili, quando i target del *deepfake* sono personaggi di rilievo politico[11]. Emblematico, fra tutti, è il caso del video falso in cui il presidente ucraino Volodymyr Zelensky, a poche settimane dall'inizio della guerra nel 2022, invitava i suoi connazionali ad arrendersi[12].

I rischi del *deepfake*, tuttavia, vanno ben oltre, dischiudendo orizzonti controversi in numerosi e diversi contesti[13]. Basti pensare alle violazioni di identità perpetrabili attraverso il *face swap* o alle drammatiche esperienze collegate ai reati collegati con la sfera sessuale e in particolare al ***deep nude*** *:* casi nei quali gli algoritmi di *deep learning* vengono utilizzati per realizzare immagini o video totalmente sintetici a partire da una foto. Basta un'unica, semplice, foto postata in un profilo social, come riporta la cronaca di una cittadina del sud della Spagna, dove nel 2023 molte ragazze hanno dolorosamente scoperto di essere ignare protagoniste di video hard creati in modo totalmente artificiale, poi diffusi in rete[14].

> Altrettanto pericoloso è il rischio che il *deepfake* possa essere utilizzato come **misura ritorsiva e intimidatoria nei confronti dei giornalisti** per limitare, nei fatti, l'esercizio del diritto di cronaca, come accaduto alla giornalista investigativa Rana Ayyub, vittima di *fake porn* intimidatorio per le sue prese di posizione contro casi di stupro a danni di minori in India[15].

Non meno preoccupante, anche per i suoi risvolti economici, è il fenomeno delle **truffe** che sfruttano i *deepfake* per indurre le vittime a pagare ingenti somme di denaro. Il fenomeno può riguardare i contesti privati ma anche quelli lavorativi, come accaduto a una dipendente di una società di Hong Kong che, a seguito di una video call con il CFO dell'azienda e vari altri partecipanti a lei noti, aveva disposto pagamenti per circa 25 milioni di euro a favore di alcune società, per poi scoprire che le persone presenti nella call non erano altro che *deepfake* eseguiti in tempo reale[16].

Ultimo ma non ultimo, un effetto indiretto del *deepfake* è quello di produrre una **progressiva erosione della fiducia collettiva verso i contenuti digitali**[17].

Argini normativi

> Quali sono le difese disponibili? Una prima risposta arriva dal legislatore europeo, che ha affrontato il tema secondo diverse prospettive e angolature.

Partiamo dal Regolamento (UE) 2024/1689 (più noto come **AI Act**). In primo luogo, esso fornisce una definizione di *deepfake*: *"un'immagine o un contenuto audio o video generato o manipolato dall'IA che assomiglia a persone, oggetti, luoghi, entità o eventi esistenti e che apparirebbe falsamente autentico o veritiero a una persona* (art. 3 n. 60)". Si tratta di una definizione ampia, non limitata alle persone, ma estesa anche a tutela di entità materiali e immateriali che potrebbero essere falsamente rappresentate, e subire così danni reputazionali a causa di *deepfake*.

L'AI Act – non senza critiche[18] – colloca il *deepfake* nella categoria dei sistemi di IA a rischio limitato: sistemi, cioè, che, pur interagendo con esseri umani o generando contenuti, non pongono

rischi elevati per i diritti fondamentali. L'art. 50 prevede obblighi di trasparenza in capo ai fornitori e ai deployer. Nello specifico, è stabilito che i deployer debbano *"rendere nota la circostanza che il contenuto è stato generato o manipolato artificialmente"* (art. 50 co. 4). In aggiunta, lo stesso art. 50, al comma 2, obbliga *"i fornitori di sistemi di Intelligenza Artificiale che generano contenuti audio, immagine, video o testuali sintetici, a garantire che gli output del sistema di IA siano marcati in un formato leggibile meccanicamente e rilevabili come generati o manipolati artificialmente"*.

> Il Regolamento, dunque, promuove l'utilizzo di strumenti di etichettatura, lasciando però ai fornitori l'individuazione delle tecnologie da utilizzare, il che crea un dibattito sull'efficacia delle misure che è particolarmente vivace[19]. L'indirizzo del Regolamento sul *deepfake,* in ogni caso, è quello di garantire la maggiore trasparenza possibile affinché i destinatari possano essere consapevoli della natura del contenuto.

Un'altra angolatura di regolazione del *deepfake* è quella della **protezione dei dati personali**, spesso utilizzati sia per la generazione dei contenuti, sia per l'addestramento dei modelli. Se il *deepfake* riguarda o ha ad oggetto la rappresentazione di una persona, può essere considerato un trattamento di un dato personale, perché relativo a un individuo identificato o identificabile (art. 4 del Regolamento (UE) 2016/679 – GDPR). Il concetto di trattamento è piuttosto ampio, comunque, e comprende dunque tutti i possibili usi dei dati personali e, quindi, tutti i possibili step di un *deepfake*: dalla generazione alla diffusione[20].

> Il GDPR prevede, poi, che il trattamento del dato sia sorretto da una base giuridica tra quelle indicate dall'art. 6 e, più nello specifico, il legittimo interesse o il consenso informato. In mancanza, il trattamento è considerato illecito. Il Regolamento, quindi, offre un insieme di strumenti per affrontare le conseguenze del *deepfake*. Va detto, però, che **la tutela delle vittime risulta spesso piuttosto complessa perché è particolarmente difficile individuare i responsabili**[21].

Argini tecnologici

A questo punto la domanda è lecita: **la tecnologia può difenderci dalla tecnologia?** I fronti sono tanti e sono aperti. Alcune soluzioni perseguono l'obiettivo di rilevare il *deepfake* a valle della sua diffusione, secondo una logica che potremmo definire reattiva. Si tratta delle tecnologie che sfruttano i modelli di rilevazione dei *deepfake* basandosi su specifici algoritmi[22]. In pratica: Intelligenza Artificiale contro Intelligenza Artificiale. Pur molto promettente, è un po' come un inseguimento tra il gatto e il topo, dove il topo ha una rapidissima capacità di correggere i propri errori una volta individuati: alla rilevazione, infatti, segue, tipicamente, il miglioramento nella successiva generazione del *deepfake****[23]***.

> Un limite importante degli algoritmi di *deepfake detection* è imputabile al modello di classificazione basato su logica binaria (falso/reale), perché questi algoritmi sono intrinsecamente non spiegabili e non è dunque possibile comprendere in base a quali elementi sia stata realizzata la classificazione, anche al fine di migliorare i modelli e renderli più robusti [24]. Altre soluzioni tecnologiche, invece, si basano su un approccio preventivo, orientato a fornire al destinatario informazioni e garantire trasparenza.

Un primo esempio è quello dei **sistemi di etichettatura (*labelling*)** – promossi anche dall'AI Act – che aggiungono informazioni (metadati) al contenuto digitale. Un esempio su tutti è il **watermark digitale**: in sostanza, si tratta di un segno univoco – visibile o invisibile – che viene associato al contenuto ed è

rilevabile da specifici software. Il *watermark* può essere associato anche a contenuti prodotti attraverso l'Intelligenza Artificiale, creando una traccia rilevabile solo da appositi algoritmi che consentono di risalire al modello utilizzato. Tuttavia, le soluzioni di *watermarking* presentano alcuni problemi di robustezza e accuratezza[25].

Tra le altre tecnologie di tipo preventivo, utilizzate per contrastare il fenomeno del *deepfake*, c'è poi la ***blockchain*****,** nata per garantire la certificazione – o, meglio, la notarizzazione – di dati, grazie alla decentralizzazione e all'uso della crittografia[26]. In estrema sintesi, **la *blockchain* è un registro distribuito**: una rete di nodi in cui ognuno di essi contiene una trascrizione integrale di tutte le transazioni registrate nella *blockchain*, la cui affidabilità è garantita da un processo – definito consenso – attraverso il quale tutti i nodi del registro convergono sull'ordine e la validità delle transazioni garantendo una versione unica della *blockchain* stessa. Le transazioni sono identificate univocamente attraverso un'impronta digitale – hash crittografico – che ne garantisce l'univocità. Grazie, poi, alla firma digitale – che attribuisce la certezza dell'identità dell'autore – la blockchain traccia la provenienza di un determinato contenuto e ne viene così certificata l'esistenza in un dato momento storico, fornendo una garanzia di trasparenza e di integrità dei contenuti. Tuttavia, per garantire l'efficacia della *blockchain* nella difesa contro il *deepfake*, è necessario garantire scalabilità e diffusione tra gli utenti, oltre a creare uno standard per l'adozione della tecnologia.

> In fondo al tunnel c'è una luce? Probabilmente sì: benché nessuna tecnologia da sola possa essere decisiva nella difesa contro l'uso malevolo delle stesse tecnologie, **la strategia più funzionale sul lungo periodo potrebbe forse essere quella di generare sinergie tra le diverse soluzioni e adattarle in funzione dei contesti, combinandole tra loro.**

Tuttavia, a prescindere dallo sviluppo tecnologico, la strategia più efficace rimane sempre quella di **investire nel fattore umano**: esercitare il senso critico, alimentare il dubbio, valutare la credibilità delle informazioni, aumentare le conoscenze, non abdicare acriticamente davanti alla tecnologia. **La strategia più efficace – forse – è proprio quella di coltivare la natura autentica dell'essere umano.**

NOTE

[1] Abbas F. e altri, 2025

[2] Somers M., 2020

[3] Goodfellow I. e altri, 2014

[4] Dhariwal P., Nichol A., 2021

[5] Masood e altri, 2023

[6] BBC News, 27 gennaio 2024

[7] Noreen I. e altri, 2022

[8] A. Thanh Thi Nguyen e altri, 2022

[9] Alanazi S. e altri, 2025

[10] Commissione Europea, comunicazione 26 aprile 2018

[11] Tremolada L., 2024

[12] BBC Technology, 18 marzo 2022

[13] Moreno F. R., 2024

[14] BBC News, 24 settembre 2023

[15] Ayyub R., 2018

[16] CNN, 24 maggio 2024

[17] Singh J., 2022

[18] Moreno F. R., 2024

[19] Proietti G., 2024

[20] Huijstee, M. v., 2024

[21] Huijstee, M. v., 2024

[22] Seow J. W. e altri, 2022

[23] Bernaciak C., Ross D. A., 2022

[24] Kundu R. e altri, 2025

[25] Madiega T., 2023

[26] Calderini B., 2020

Bibliografia e sitografia

1. Abbas F., Taeihagh A., Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence, Expert Systems with Applications, Volume 252, Part B, 2024.
2. Alanazi, S., Asif, S., Caird-daley, A., Unmasking deepfakes: a multidisciplinary examination of social impacts and regulatory responses, Hum.-Intell. Syst. Integr., 2025.
3. Ayyub R., I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me, in Huffington Post UK, 201.
4. Bernaciak, C., Ross, D., How Easy Is It to Make and Detect a Deepfake?, Carnegie Mellon University, Software Engineering Institute's Insights (blog), 2025.
5. Calderini, B., La blockchain ci salverà dalle fake news, in ZeroUno, Rivista di Digital360, 2020.
6. Coccomini, D.A., Messina, N., Gennaro, C., Falchi, F., Combining Efficient Net and Vision Transformers for Video Deepfake Detection, In Sclaroff, S., Distante, C., Leo, M., Farinella, G.M. Tombari, F. (eds) *Image Analysis and Processing* – Lecture Notes in Computer Science, vol 13233, 2022.
7. Dhariwal P., Nichol A., Diffusion models beat GANs on image synthesis, in Proceedings of the 35th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, Article 672, 2021.
8. Comunicazione "Contrastare la disinformazione online: un approccio europeo", COM(2018) 236, del 26 aprile 2018.
9. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Bengio Y., Generative adversarial nets, in Advances in neural information processing system, 2014.

10. Huijstee, M. v., Boheemen, P. v., Das, D., Nierling, L. et al., Tackling deepfakes in European policy, European Parliament, 2021.
11. Kundu R., Balachandran A., Roy-Chowdhury A. K., TruthLens: Explainable DeepFake Detection for Face Manipulated and Fully Synthetic Data, arXiv:2503.15867, 2025.
12. Lisinka J., Castro D., Why AI-Generated Content Labeling Mandates Fall Short, Reports of the Center for Data Innovation, 2024.
13. Madiega T., Generative AI and watermarking, EPRS, 2023.
14. Masood M., Nawaz, M., Malik, K.M., Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward, Appl Intell 53, 2023.
15. Moreno F.R., Generative AI and deepfakes: a human rights approach to tackling harmful content, in International Review of Law, Computers & Technology, 38(3), 2024.
16. Noreen I., Muneer MS., Gillani S., Deepfake attack prevention using steganography GANs, PeerJ Comput Sci., 2022.
17. Proietti G., L'impianto regolatorio della società dell'informazione tra vecchi e nuovi equilibri. Il fenomeno del deep fake, in Media Laws – Rivista di diritto dei media, I-2024.
18. Seow J. W, Lim M. K., Phan R.C.W., Liu J. K., A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities, Neurocomputing, Volume 513, 2022.
19. Singh J., Deepfakes: The Threat to Data Authenticity and Public Trust in the Age of AI-Driven Manipulation of Visual and Audio Content, in Journal of AI-Assisted Scientific Discovery, Vol. 2 No. 1, 2022.
20. Somers M., Deepfakes, explained, MIT Sloan School of Management, 2020.
21. Thanh Thi Nguyen A., Quoc Viet Hung Nguyen B., Dung Tien Nguyen A., Duc Thanh Nguyen A., Thien Huynh-The C., Saeid Nahavandi D., Thanh Tam Nguyen E., Quoc-Viet Pham F., M. Nguyen C., Deep learning for deepfakes creation and detection: A survey, in Computer Vision and Image Understanding, 2022.
22. Tremolada L., Dai deepfake di Taylor Swift alle fake news con ChatGpt: come difendersi dalla disinformazione elettorale?, Il sole 24 ore, 2024