Digeat N.6 - 19 Giugno 2025

L’inautenticità: fake news, deepfake e hate speech.

Di Paolo Assirelli



Abstract

Questo articolo propone un’analisi di tre fenomeni distinti: fake news, deepfake e hate speech, mettendone in luce le loro caratteristiche distintive e le loro differenze essenziali. Vengono inoltre presentate alcune strategie di contrasto recentemente adottate sul piano giuridico, tecnologico e culturale.

Indice

- Introduzione: l’inattesa piega degli eventi
- Fake news, deepfake e hate speech: definizioni e differenze
- Strategie di contrasto
- Conclusioni e prospettive future

Introduzione: l’inattesa piega degli eventi

L’innegabile democratizzazione nella produzione di contenuti della digital age, sebbene abbia ampliato enormemente la possibilità di accesso all’informazione, ha tuttavia dato luogo alla **proliferazione di fake news, deepfake e hate speech, che oggi rappresentano una vera sfida per le democrazie contemporanee**, tanto da essere stati identificati come il principale rischio globale nel World Economic Forum Global Risks Report del 2025[1]. Per l’UNESCO, infatti, **tali pratiche possono normalizzare la violenza, danneggiando la sicurezza delle comunità**[2].

La gravità del fenomeno è evidente: la pandemia da COVID-19 ha dimostrato quanto le bufale sui vaccini e le “cure miracolose” possano propagarsi in modo virale sui social network, generando diffidenza nei confronti delle istituzioni sanitarie. Il conflitto in Ucraina ha prodotto un’ondata di informazioni manipolate (video deepfake di leader militari, fake news dai fronti di guerra), evidenziando il carattere transnazionale del problema. Nelle elezioni presidenziali statunitensi del 2024, l’impiego massiccio di deepfake ha sollevato seri interrogativi circa la legittimità del processo democratico e la percezione della realtà da parte dell’elettorato.

> Questo quadro ha quindi unito esperti di diritto e sociologia, epistemologi, psicologi sociali ed esperti di etica dell’informazione, nella ricerca di strumenti di tutela dei diritti individuali e della verità nell’era digitale, poiché la questione ormai impatta le fondamenta stesse delle democrazie contemporanee.

Fake news, deepfake e hate speech: definizioni e differenze

Per **fake news** si intendono **informazioni deliberatamente false, create per influenzare l'opinione pubblica o oscurare la verità con scopi ingannevoli**, distinguendo tra *misinformation* (informazione falsa diffusa per errore o ignoranza) e *disinformation* (informazione deliberatamente falsa).[3] In Europa, un'indagine dell'Eurobarometer[4], iniziata nel 2023, rileva che circa il 68% dei cittadini europei dichiara di essere stato esposto a contenuti di disinformazione.[5] **Le fake news riescono a contaminare anche i sistemi informativi tradizionalmente considerati affidabili**: nel 2024, diverse testate giornalistiche europee hanno inconsapevolmente riportato notizie false generate da farm di disinformazione operanti dal sud-est asiatico.

Le fake news possono spaziare dai complotti politici alle bufale in campo sanitario[6], aumentando la polarizzazione: una ricerca di psicologia condotta nel 2024 ha evidenziato che **le persone sono generalmente più influenzate dalla coerenza politica di una notizia rispetto alla sua verità oggettiva**[7]. Recenti studi della neuroscienza cognitiva dimostrano come la fruizione ripetuta di informazioni false attivi nel cervello circuiti di ricompensa simili a quelli delle dipendenze comportamentali: la conferma delle proprie credenze genera un piacere neurobiologico che sovrasta il beneficio cognitivo dell'acquisizione di informazioni corrette.

> I **deepfake** rappresentano una forma avanzata di fake news basata sull'Intelligenza Artificiale. Sono video, immagini o audio generati tramite algoritmi di IA che manipolano materiale autentico per creare contenuti iperrealistici ma completamente falsi[8], e gli odierni strumenti di IA ne permettono la facile creazione anche agli utenti non esperti; dal 2025, basta uno smartphone di fascia media. La democratizzazione della tecnologia ha reso la manipolazione dell'informazione visiva alla portata di tutti, con implicazioni profonde sul concetto stesso di evidenza empirica.

Già nell'aprile 2024, negli Stati Uniti, un audio deepfake diffuso su Telegram – attribuito a una preside di un'istituzione scolastica di Baltimora – conteneva insulti apertamente razzisti e violenti; la polizia locale ha poi accertato che il filmato era totalmente artefatto[9]. In Australia, a giugno 2024, sono stati individuati centinaia di video deepfake pornografici raffiguranti ragazzine minorenni, un fenomeno di molestie digitali che ha suscitato grande allarme sociale[10].

Il Parlamento Europeo, nella relazione annuale allegata alla Proposta Di Risoluzione sulla *"tutela degli interessi finanziari dell'Unione europea"* (2024/2083 INI), ha segnalato un incremento del 340% dei casi di frode finanziaria basata su deepfake nell'ultimo biennio, con perdite stimate intorno ai 2,7 miliardi di euro. Inoltre, rendendo plausibili i contenuti falsi, i **deepfake rischiano di screditare anche l'affidabilità dei media autentici (*liar's dividend*), rendendo arduo per gli utenti distinguere tra realtà e simulazione**, e minando alla base il concetto di "verità fattuale".

> L'**hate speech** attiene all'espressioni verbali, scritte o iconiche che denigrano, odiano o incitano alla violenza contro persone in quanto appartenenti a gruppi protetti (per razza, nazionalità, religione, genere, orientamento sessuale, disabilità, ecc.). L'elemento chiave è il pregiudizio: il discorso d'odio non si definisce in base alla veridicità del messaggio, ma in base al bersaglio e all'ideologia sottesa. Anche un'affermazione vera, se presentata per incentivare pregiudizi o azioni discriminatorie, è considerata hate speech.

I social network hanno inizialmente adottato regole proprie contro l'hate speech, ma la moderazione è complessa e spesso criticata con un'inversione di tendenza ormai manifesta: a gennaio 2025, Marc Zuckerberg ha annunciato la fine delle politiche di moderazione su Facebook; Elon Musk, ancora, ha abolito ogni forma di controllo e censura nella piattaforma X in favore delle *"community notes"*, che favorirebbero maggiore libertà di espressione. Nel 2024, l'analisi pubblicata dall'Università di Oxford nella sua piattaforma *"Monitoring Online Hate Speech"*[11], ha rilevato che gli algoritmi di rilevamento dell'hate speech dei cinque principali social network identificano correttamente solo il 67% dei

contenuti problematici in inglese, con percentuali ancora inferiori per lingue diverse o per varianti dialettali.

Nella pratica i tre fenomeni spesso si sovrappongono: una fake news a sfondo razziale può contenere elementi di hate speech; un deepfake può essere prodotto con finalità di incitamento all'odio contro un gruppo specifico, dando origine a strategie coordinate di manipolazione dell'opinione pubblica; fake news, deepfake e hate speech vengono orchestrati in modo sequenziale e complementare in "campagne ibride": prima si diffonde una fake news su un presunto crimine commesso da membri di una minoranza, poi si crea un deepfake che "documenta" visivamente l'accaduto, infine si alimenta l'hate speech contro quel gruppo attraverso commenti coordinati e bot sui social media.

Nel Regno Unito, un omicidio attribuito ad un rifugiato è stato seguito dalla diffusione di fake news che lo collegavano alla criminalità organizzata, scatenando sommosse anti-immigrati. La Nuova Zelanda, nell'agosto 2024, ha registrato la forte reazione civile dei portalettere di Wellington che si sono rifiutati di consegnare 80.000 volantini contenenti false accuse contro la comunità musulmana[12]. Il quotidiano "The Guardian", afferma che in Italia, nel 2024, la "Lega" ha diffuso sui suoi canali social immagini generate con sistemi di IA di donne e bambini intenti a mangiare insetti.

> Il rapporto *"The changing DNA of serious and organised crime (EU-SOCTA)",* pubblicato dall'Europol a marzo 2025, cita almeno sette *"disinformation farms"* operanti a livello globale, con sedi distribuite tra Russia, Cina, Medio Oriente e Sud-Est asiatico, in grado di produrre e diffondere contenuti ingannevoli in oltre 30 lingue, spesso per conto di attori statali. Questa industrializzazione della disinformazione rappresenta una sfida cruciale per la sicurezza nazionale e la stabilità democratica dei paesi occidentali.

Strategie di contrasto

L'ambito giuridico

Sul piano **giuridico**, diverse giurisdizioni stanno aggiornando le normative esistenti; **nell'UE è in vigore dal 2024 il *Digital Services Act (DSA)*, che impone ai grandi servizi online la segnalazione e tempestiva rimozione dei contenuti illegali e l'adozione di procedure di verifica durante le campagne elettorali** per prevenire la manipolazione dell'opinione pubblica. Il 13 febbraio 2025, la Commissione Europea e il Comitato Europeo per i Servizi Digitali hanno approvato l'integrazione ufficiale nel DSA del *"Codice di Condotta sulla Disinformazione"* ***[13]*** rendendo obbligatorie e misurabili, a partire dal 1° luglio 2025, le pratiche di contrasto[14], con pesanti sanzioni in caso di inosservanza.

In Italia, l'AGCOM (Autorità per le Garanzie nelle Comunicazioni) può intervenire durante le campagne elettorali per imporre rettifiche a emittenti radiofoniche o televisive che diffondano notizie palesemente false.

> Sul fronte penale, gli articoli 604-bis e 604-ter del Codice Penale italiano puniscono chi propaganda *"idee fondate sulla superiorità o sull'odio razziale o etnico, ovvero istiga a commettere o commette atti di discriminazione per motivi razziali, etnici, nazionali o religiosi"*, nonché *"l'incitamento alla discriminazione o alla violenza per motivi razziali, etnici, nazionali o religiosi."*[15], colpendo direttamente la propaganda d'odio e l'istigazione alla violenza.[16]

Il disegno di legge n. 1146 del 20 maggio 2024 (*Disposizioni e delega al Governo in materia di intelligenza artificiale)* prevede all'art. 23: che *"Qualunque contenuto informativo diffuso …. tramite*

qualsiasi piattaforma in qualsiasi modalità … che, … sia stato, attraverso l'utilizzo di sistemi di intelligenza artificiale, completamente generato ovvero, anche parzialmente, modificato o alterato in modo tale da presentare come reali dati, fatti e informazioni che non lo sono, deve essere reso, …, chiaramente visibile e riconoscibile da parte degli utenti mediante inserimento di un elemento o segno identificativo, anche in filigrana o marcatura incorporata … con l'acronimo "IA" ovvero, nel caso audio, attraverso annunci audio ovvero con tecnologie adatte a consentire il riconoscimento".

La Corte di Cassazione ricomprende la diffusione consapevole di deepfake diffamatori nell'ambito del reato di diffamazione aggravata (art. 595 c.p.), e di sostituzione di persona (art. 494 c.p.).

Negli Stati Uniti, è in fase di approvazione finale un disegno di legge – il "*Take It Down Act*" – con l'obbligo per le piattaforme social di rimuovere – entro 48 ore – qualsiasi immagine deepfake intima di una persona senza il consenso dell'interessato.[17] La California ha approvato una legge che punisce con l'arresto fino a un anno la diffusione di immagini deepfake sessualmente esplicite senza consenso.[18].

> Sul piano internazionale, si delinea una tendenza verso standard comuni: l'UE ha ampliato la tutela del GDPR e del DSA con l'*AI Act*, il cui approccio risk-based classifica le applicazioni di IA in quattro categorie di rischio e impone requisiti progressivamente più stringenti in base al loro potenziale impatto negativo; in tale schema, i deepfake possono essere potenzialmente ricompresi tra le applicazioni IA "ad alto rischio".

Il considerando n. 133 richiama la *"necessità di nuovi metodi e tecniche per risalire all'origine delle informazioni"*, e reputa opportuno imporre ai fornitori l'integrazione di *"soluzioni tecniche che consentano agli output di essere marcati in un formato leggibile meccanicamente e di essere rilevabili come generati o manipolati da un sistema di IA e non da esseri umani"*. Puntualmente, l'articolo 50, 2° comma, del Regolamento impone ai fornitori di sistemi che generano contenuti audio, immagine, video o testuali di operare sugli output generati in modo che *"siano marcati in un formato leggibile meccanicamente e rilevabili come generati o manipolati artificialmente"*.

Nel 2024, l'Assemblea Generale delle Nazioni Unite ha adottato la risoluzione non vincolante *"Integrità dell'informazione nell'era digitale"*, che esorta gli Stati membri a sviluppare strategie nazionali contro la disinformazione e a promuovere la resilienza sociale attraverso l'educazione ai media. Al G20 di Rio de Janeiro nel 2024, è stata presentata una proposta di un *"Global Compact against Digital Disinformation"* che impegna i firmatari a creare una task force internazionale per la condivisione di best practices e informazioni sui gruppi organizzati di disinformazione.

> Deepfake e hate speech pongono sfide significative al diritto e alla società nell'equilibrare il sostegno all'innovazione tecnologica e alla libertà di espressione con la tutela dei diritti individuali e collettivi e la prevenzione di danni. Anzi, sul piano giuridico sta emergendo il concetto di *"diritto all'autenticità informativa"*, come sviluppo del diritto all'informazione.

Il riconoscimento e il contrasto ai predetti fenomeni sottolineano la centralità dei modelli educativi e della consapevolezza. L'approccio educativo, tuttavia, deve confrontarsi con la velocità dell'evoluzione tecnologica, che spesso supera la capacità di adattamento dei sistemi formativi tradizionali. Quindi, **l'asse giuridico si sta muovendo su due fronti: ampliamento delle sanzioni per l'hate speech e costruzione di un'architettura normativa flessibile per gestire la rapida evoluzione delle tecnologie IA.**

L’ambito sociale: le strategie di mitigazione

L’adozione di misure **sociali** e **tecnologiche** di prevenzione è comunque cruciale: **l’UNESCO ha dedicato l’anno 2024 al tema dell’istruzione**, organizzando corsi di formazione per migliaia di insegnanti sulle tecniche di riconoscimento e smontaggio dei messaggi d’odio. In Italia, scuole e università hanno introdotto specifici moduli educativi su come individuare le fake news e valutare l’attendibilità delle fonti informative. Il Ministero dell’Istruzione italiano, nel corso del 2024, ha previsto diversi programmi per le scuole secondarie dedicati all’alfabetizzazione mediatica, con particolare attenzione all’identificazione di contenuti manipolatori e alla comprensione dei meccanismi di funzionamento degli algoritmi di raccomandazione dei social media.

La campagna paneuropea *“Think Before You Share”*, lanciata dalla Commissione Europea nell’ottobre 2024, ha coinvolto influencer e celebrità nella promozione di pratiche responsabili di condivisione dei contenuti. *FactCheckEU.info* è un progetto diretto a contrastare la disinformazione nell’UE nelle future elezioni del Parlamento Europeo, pubblicando documenti provenienti da 19 diverse testate giornalistiche e reti di rilettura collettiva (*crowdsourced fact-checking*) che cercano di smascherare notizie false appena emergono.

Le aziende tecnologiche hanno implementato algoritmi di machine learning che analizzano il testo dei post per segnalare discorsi d’odio o informazioni potenzialmente false; hanno introdotto meccanismi per inserire avvisi di contestazione fact-checking quando viene mostrata un’affermazione sensibile; hanno sviluppato software di *«deepfake detection»* che individuano incoerenze nei video generati artificialmente. Tali sistemi sono in continua sfida con l’evoluzione della IA. Alcuni esperti propongono soluzioni preventive, come l’inserimento obbligatorio di watermark digitali in tutti i contenuti generati da IA. Ma la sofisticazione delle tecniche di manipolazione cresce: i ricercatori dell’Università di Stanford hanno dimostrato come i deepfake di ultima generazione siano in grado di superare i test di verifica biometrica utilizzati da molte banche e istituzioni pubbliche. Alcune aziende tecnologiche stanno sperimentando approcci innovativi, come l’utilizzo di reti neurali generative antagoniste (GAN) per simulare attacchi e creare una sorta di “vaccino digitale”.

> Gli studiosi insistono sull’approccio cosiddetto *“whole-of-society”*: tutte le componenti devono collaborare attivamente, con una serie di misure integrate[19]: prevenzione (educazione digitale nelle scuole); rilevamento (politiche di audit delle IA); contrasto legale (diffusione di guide legali per vittime di deepfake); riparazione (supporto psicologico per chi subisce attacchi d’odio).[20] La comunità scientifica e la società scommettono sulle reti di fact-checker indipendenti; codici etici aziendali; piattaforme interne di segnalazione veloce; partnership pubblico/privato (la Commissione UE nel 2024 ha collaborato con Facebook e Google per sviluppare un *“Election Integrity Toolkit”* ***[21]***).

Conclusioni e prospettive future

Fake news, deepfake e hate speech costituiscono un panorama complesso, alimentato tanto dall’innovazione tecnologica quanto dalle dinamiche umane di pregiudizio e partigianeria. **Le democrazie reagiscono estendendo le norme già esistenti (penali o civili) alle nuove fattispecie digitali, ovvero mediante la creazione di nuovi reati.**

La sociologia sottolinea che gli interventi devono guardare anche al contesto in cui le informazioni si diffondono: è essenziale mantenere viva la fiducia nel giornalismo di qualità e incentivare una cittadinanza attiva che abbia un atteggiamento critico su ciò che legge online.

> Le tecnologie IA continueranno a evolversi, rendendo più credibili le simulazioni; la risposta deve quindi essere dinamica e richiede una cooperazione internazionale più stretta con accordi transnazionali sulla *"responsabilità algoritmica"* delle piattaforme.

La società deve innalzare il proprio livello di resilienza culturale tramite l'educazione ai media e la trasparenza dei canali informativi, per preservare il pluralismo informativo e i diritti fondamentali e garantire che la tecnologia informi anziché ingannare. **La sfida maggiore rimane quella culturale: senza l'alfabetizzazione digitale, le tecnologie più avanzate rischiano di essere inaffrontabili.**

NOTE

[1] World Economic Forum, Global Risks Report 2025.

[2] ClareO'Hagan, UNESCO dedicates the International Day of Education 2024 to countering hate speech, UNESCO, 28 giugno 2024.

[3] L'Electoral Commission australiana definisce la disinformation come "informazione sapientemente falsa progettata per ingannare e influenzare l'opinione pubblica. Cfr. "Disinformation in 2024 was rife, and it's likely to bring more risks in 2025, The University of Melbourne, 2025.

[4] A partire dal 1973, l'Eurobarometer raccoglie una serie di sondaggi di opinione pubblica condotti regolarmente per conto della Commissione europea.

[5] *"Giovani e fake news"* realizzata da YouTrend in collaborazione con la Rappresentanza in Italia della Commissione Europea, Comunicato stampa 20 aprile 2024.

[6] Una fake news diffusa nel 2024 affermava che la Commissione Europea puntasse a ottenere entro il 2030 un 15% di farina derivante dai grilli. Cfr. Disinformation in 2024 was rife, and it's likely to bring more risks in 2025, The University of Melbourne, 2025.

[7] Disinformation in 2024 was rife, and it's likely to bring more risks in 2025, The University of Melbourne, 2025.

[8] Nora Nahae Kim della Columbia University definisce i deepfake come «filmati, immagini o audio creati con IA che manipolano riprese reali per creare contenuti altamente realistici ma del tutto ingannevoli». Cfr. Nora Nahae Kim, Deepfakes and Democracy: Free Speech vs. Election Integrity, Columbia University Libraries

Published in partnership with Columbia University Libraries and Columbia Law School.

[9] Disinformation in 2024 was rife, and it's likely to bring more risks in 2025, The University of Melbourne, 2025.

[10] Disinformation in 2024 was rife, and it's likely to bring more risks in 2025, The University of Melbourne, 2025.

[11] Riferimento a questo link.

[12] Ben Quinn e Dan Milmo, How the far right is weaponising AI-generated content in Europe, The Guardian, 26 novembre 2024.

[13] Riferimento a questo link.

[14] EU Digital Service Act (DSA).

[15] Mariella Spata, Hate speech, cyberbullismo e i limiti alla libertà di espressione sul web, Diritto.it, 19 dicembre 2024.

[16] Mariella Spata, Hate speech, cyberbullismo e i limiti alla libertà di espressione sul web, Diritto.it, 19 dicembre 2024.

[17] Will Oremus, Congress passes bill to fight deepfake nudes, revenge porn, Washington Post 28 aprile 2025.

[18] Anna Merod, Congress passes bill criminalizing illicit deepfakes as students are targeted, in K-12 Dive, 29 aprile 2025.

[19] Anuragini Shirish e Shobana Komal, A socio-legal inquiry on deepfakes, in Dialnet, California Western In-ternational Law Journal, Febbraio 2024.

[20] Anuragini Shirish e Shobana Komal, *A socio-legal inquiry on deepfakes*, in Dialnet, California Western International Law Journal, Febbraio 2024.

[21] Riferimento a questo link.