

L'IA e i social media. L'uso dei deepfake e di contenuti generati con l'intelligenza artificiale deve essere bandito dai social?

SCHEDA

Immaginiamo di aprire un social network e di imbatterci nel video di un personaggio famoso che pronuncia una frase sorprendente. Il volto è il suo, la voce sembra esattamente la sua e perfino i movimenti delle labbra coincidono con le parole che ascoltiamo. Il video potrebbe essere autentico. Oppure potrebbe non esserlo affatto: quelle parole potrebbero non essere mai state pronunciate e quel video potrebbe essere stato realizzato, in pochi minuti, utilizzando un sistema di intelligenza artificiale.

Fino a pochi anni fa, creare una manipolazione audiovisiva sufficientemente realistica da ingannare una parte consistente degli spettatori richiedeva competenze tecniche, software sofisticati e molto tempo. Oggi gli strumenti di intelligenza artificiale generativa permettono a un numero sempre maggiore di persone di creare immagini, audio, testi e video realistici partendo da semplici istruzioni. È possibile generare il volto di una persona che non esiste, clonare una voce, modificare una fotografia, ricostruire un ambiente mai esistito oppure mostrare una persona reale mentre compie azioni che non ha mai compiuto.

Tra questi contenuti troviamo i **deepfake**, termine utilizzato per indicare immagini, audio o video prodotti o manipolati attraverso sistemi di intelligenza artificiale in modo da rappresentare persone, luoghi, oggetti o avvenimenti reali facendo apparire autentico qualcosa che invece non lo è. La definizione adottata dall'Unione europea insiste proprio su questo aspetto: ciò che caratterizza il deepfake non è semplicemente il fatto che un contenuto sia stato modificato, ma la possibilità che venga erroneamente percepito come autentico o veritiero.

Il problema, quindi, non è necessariamente l'utilizzo dell'intelligenza artificiale. La stessa tecnologia può essere impiegata per creare un video comico, realizzare gli effetti speciali di un film, tradurre automaticamente un contenuto in un'altra lingua, creare un'illustrazione oppure ricostruire virtualmente un luogo storico. L'IA generativa offre possibilità creative enormi e sta diventando uno degli strumenti utilizzati quotidianamente da designer, studenti, aziende, artisti e content creator. La questione cambia quando chi guarda un contenuto non ha gli strumenti per capire che ciò che sta osservando è stato creato o manipolato artificialmente.

Ed è qui che entra in gioco la disinformazione.

Un video falso può mostrare un politico mentre annuncia una decisione mai presa; una fotografia artificiale può rappresentare una catastrofe mai avvenuta; la voce clonata di un personaggio conosciuto può essere utilizzata per promuovere un prodotto che quella persona non ha mai utilizzato. Potremmo persino ricevere un messaggio vocale apparentemente inviato da qualcuno che conosciamo, ma in realtà generato artificialmente. La stessa tecnologia che consente di creare contenuti divertenti o creativi può quindi diventare uno strumento di manipolazione, truffa, impersonificazione o propaganda.

Il problema diventa ancora più significativo sui social network, dove le informazioni viaggiano rapidamente e vengono spesso osservate per pochi secondi prima di decidere se mettere un “Mi piace”, commentarle oppure condividerle. Un contenuto sorprendente, scandaloso o emotivamente coinvolgente può raggiungere migliaia o milioni di persone prima che qualcuno abbia avuto il tempo di verificarne l'autenticità. E più il contenuto sembra realistico, più entra in gioco una convinzione molto semplice e profondamente radicata: **“l’ho visto con i miei occhi, quindi deve essere vero”**.

Proprio questa convinzione rischia di diventare sempre meno affidabile.

Per molto tempo, infatti, fotografie e video sono stati considerati una sorta di prova della realtà. Anche quando sapevamo che un'immagine poteva essere ritoccata, la possibilità di produrre facilmente un video completamente artificiale, con volti, ambienti e voci estremamente realistici, apparteneva più alla fantascienza che alla nostra esperienza quotidiana. L'intelligenza artificiale generativa modifica questo rapporto tra immagini e realtà e ci costringe ad aggiungere una nuova domanda al nostro modo di informarci: non soltanto **“quello che sto leggendo è vero?”**, ma anche **“quello che sto vedendo o ascoltando è realmente accaduto?”**.

Questo non significa, però, che possiamo riconoscere sempre un deepfake semplicemente guardandolo con attenzione. Alcuni consigli diventati popolari negli ultimi anni – controllare il numero delle dita, osservare gli occhi, cercare movimenti innaturali delle labbra o imperfezioni nello sfondo – possono talvolta aiutarci, ma rischiano rapidamente di diventare obsoleti. I sistemi di generazione migliorano continuamente e gli errori più evidenti tendono a diminuire. La media literacy non può quindi trasformarsi in una specie di gara tra esseri umani e algoritmi per scoprire il pixel sbagliato.

Dobbiamo piuttosto tornare alle regole fondamentali della verifica delle informazioni.

Chi ha pubblicato il contenuto? Possiamo risalire alla fonte originale? La stessa notizia viene riportata anche da altre fonti attendibili? Il video compare sul profilo ufficiale della persona rappresentata? Un quotidiano, un'agenzia di stampa o un ente istituzionale confermano ciò che vediamo? Il contenuto viene presentato come reale oppure come satira, ricostruzione o creazione artistica? Esistono informazioni sulla sua provenienza e sulle eventuali modifiche che ha subito?

Quest'ultima domanda sta diventando sempre più importante. Una delle strade seguite per affrontare il problema consiste infatti nel creare strumenti che consentano di conservare informazioni sulla **provenienza digitale** di immagini e video. La Coalition for Content Provenance and Authenticity (C2PA), ad esempio, ha sviluppato uno standard denominato *Content Credentials*, attraverso il quale è possibile associare ai contenuti informazioni verificabili sulla loro origine e sulle trasformazioni che hanno subito. L'obiettivo non è stabilire automaticamente se un contenuto sia “vero” o “falso”, ma fornire agli utenti maggiori informazioni sulla sua storia.

Ma allora perché non risolvere il problema alla radice e vietare semplicemente tutti i contenuti generati con l'intelligenza artificiale sui social?

La risposta è meno semplice di quanto possa sembrare. Un divieto generale eliminerebbe infatti, insieme agli utilizzi dannosi, anche moltissimi utilizzi legittimi e creativi dell'IA. Dovremmo vietare un'illustrazione creata per accompagnare una storia? Un video satirico dichiaratamente artificiale? Un'animazione? La ricostruzione di una città antica? Un doppiaggio realizzato con l'IA? E che cosa accadrebbe quando l'intelligenza artificiale viene utilizzata soltanto per modificare una parte di un contenuto reale?

Il confine tra contenuto umano e contenuto artificiale sta diventando sempre più difficile da tracciare. Per questo motivo le politiche pubbliche e quelle adottate dalle piattaforme tendono oggi a distinguere tra **uso dell'intelligenza artificiale** e **uso ingannevole dell'intelligenza artificiale**.

Nell'Unione europea, ad esempio, dal 2 agosto 2026 sono applicabili gli obblighi di trasparenza previsti dall'articolo 50 dell'AI Act. Tra le misure previste vi sono strumenti per rendere riconoscibili i contenuti generati o manipolati artificialmente e specifici obblighi di informazione nel caso dei deepfake. Per alcuni contenuti artistici, satirici o di fantasia sono previste modalità di informazione compatibili con la natura dell'opera. L'obiettivo, dunque, non è vietare in generale la produzione di contenuti attraverso l'intelligenza artificiale, ma fare in modo che una persona possa sapere quando si trova di fronte a qualcosa di artificiale che potrebbe essere scambiato per autentico.

Anche le grandi piattaforme stanno sperimentando soluzioni simili. YouTube richiede ai creator di dichiarare i contenuti realistici significativamente generati o modificati con l'IA e, nel maggio 2026, ha annunciato etichette più visibili e sistemi automatici per individuare alcuni contenuti artificiali. Meta utilizza l'etichetta "AI info" per determinati contenuti riconosciuti come generati artificialmente e richiede la dichiarazione di video fotorealistici o audio realistici creati o modificati digitalmente in determinate circostanze. TikTok, nel luglio 2026, ha annunciato ulteriori interventi dedicati alla trasparenza dei contenuti generati dall'IA e al contrasto degli account che pubblicano grandi quantità di spam artificiale.

Naturalmente anche le etichette hanno dei limiti. Un utente potrebbe ignorarle, un contenuto potrebbe essere pubblicato senza dichiarazione oppure i sistemi automatici potrebbero non riconoscerlo. Inoltre, sapere che un'immagine è stata generata dall'IA non ci dice necessariamente se il messaggio che comunica sia vero o falso. Un'immagine artificiale potrebbe illustrare correttamente una notizia vera; un video completamente autentico potrebbe invece essere pubblicato con una descrizione falsa o essere estratto dal suo contesto originale.

È quindi importante non trasformare la discussione in una semplice contrapposizione tra "contenuti veri" e "contenuti creati dall'IA". La vera questione riguarda **trasparenza, intenzione e contesto**.

L'UNESCO considera proprio la capacità di comprendere criticamente l'intelligenza artificiale una parte sempre più importante della media and information literacy. In un ambiente digitale nel quale strumenti automatici possono partecipare alla produzione di testi, immagini, video e informazioni, non basta sapere utilizzare la tecnologia: bisogna essere in grado di interrogarsi sulla sua origine, sui suoi obiettivi e sulle conseguenze della sua diffusione.

Forse, quindi, la domanda da porci non è soltanto se i contenuti generati dall'IA debbano essere banditi dai social. Potremmo chiederci piuttosto: **in quali casi dovrebbero essere consentiti? Quando devono essere chiaramente segnalati? Quando invece diventano così ingannevoli o pericolosi da dover essere rimossi? E chi dovrebbe assumersi la responsabilità di questa decisione: chi crea il contenuto, chi lo pubblica, la piattaforma oppure le istituzioni?**

Proviamo a discuterne insieme.

ATTIVITÀ IN CLASSE

Proviamo a costruire insieme una possibile politica per la gestione dei contenuti generati dall'intelligenza artificiale all'interno di un social network.

1. Dividere la classe in gruppi di lavoro.

2. Distribuire a ogni gruppo uno dei seguenti casi:

A) Un creator pubblica un video comico nel quale utilizza l'IA per far cantare una canzone a un personaggio storico. Nel video è indicato chiaramente che si tratta di un contenuto artificiale.

B) Circola sui social un video realistico nel quale un politico sembra annunciare una decisione molto importante. Il video è stato generato con l'IA, ma non contiene nessuna indicazione che lo segnali.

C) Un profilo pubblicitario utilizza la voce e il volto artificialmente ricostruiti di un personaggio famoso per fargli consigliare un prodotto che non ha mai utilizzato.

D) Un utente pubblica un'immagine generata dall'IA che rappresenta una grande alluvione in una città italiana. L'immagine viene accompagnata dalla frase "Sta succedendo proprio adesso", anche se l'evento non è mai avvenuto.

E) Un museo pubblica sui propri social una ricostruzione realistica realizzata con l'IA per mostrare come poteva apparire una città romana duemila anni fa. La natura della ricostruzione è indicata nella descrizione.

3. Ogni gruppo dovrà decidere quale delle seguenti azioni dovrebbe adottare il social network:

- lasciare il contenuto online senza interventi;
- lasciare il contenuto online aggiungendo un'etichetta che segnali l'utilizzo dell'IA;
- ridurre la diffusione;
- rimuoverlo;
- applicare una sanzione all'account che lo ha pubblicato.

La decisione dovrà essere motivata rispondendo alle seguenti domande:

1. **Il contenuto potrebbe essere scambiato per autentico? Perché?**
2. **Chi potrebbe essere danneggiato dalla sua diffusione?**
3. **L'autore ha dichiarato in modo sufficientemente chiaro l'utilizzo dell'intelligenza artificiale?**
4. **L'intenzione con cui è stato realizzato il contenuto è importante per decidere come intervenire?**
5. **È sufficiente aggiungere un'etichetta oppure, in questo caso, il contenuto dovrebbe essere rimosso?**
6. **Quale regola generale potrebbe adottare il social network per gestire in futuro casi simili?**

4. Presentazione e confronto

Ogni gruppo presenterà alla classe il proprio caso, la decisione adottata e la regola che propone di introdurre.

Al termine delle presentazioni, la classe proverà a redigere insieme un breve **“Regolamento dei contenuti generati dall’IA”**, composto da cinque regole condivise.

L’obiettivo non è necessariamente raggiungere una posizione unanime, ma comprendere quanto sia complesso trovare un equilibrio tra **libertà di espressione, creatività, diritto all’informazione, tutela delle persone e contrasto alla disinformazione**.

MATERIALI DI APPROFONDIMENTO (FACOLTATIVO)

- Per comprendere come l’Unione europea sta affrontando il problema dei deepfake e dei contenuti generati artificialmente, è possibile consultare i materiali della **Commissione europea dedicati all’articolo 50 dell’AI Act**. Dal 2 agosto 2026 sono applicabili specifici obblighi di trasparenza relativi ai sistemi di IA e ai contenuti generati o manipolati artificialmente. Le pagine della Commissione contengono anche una definizione di deepfake ed esempi delle situazioni nelle quali è richiesta una chiara informazione agli utenti.
- Sul rapporto tra **intelligenza artificiale, social media e media literacy**, UNESCO mette a disposizione il modulo *MIL Competencies in the Age of AI and Social Media*, che propone strumenti e attività per comprendere criticamente il funzionamento dell’intelligenza artificiale nell’ambiente informativo contemporaneo.
- Per approfondire il tema della provenienza delle immagini e dei contenuti digitali è possibile esplorare il progetto **C2PA – Coalition for Content Provenance and Authenticity** e il sistema dei *Content Credentials*. L’obiettivo è creare standard che permettano di associare a fotografie, immagini e altri contenuti informazioni verificabili sulla loro origine e sulle modifiche effettuate.
- Può essere interessante, infine, confrontare direttamente le politiche adottate dalle piattaforme. **YouTube, Meta e TikTok** hanno introdotto sistemi di dichiarazione, etichettatura o rilevamento dei contenuti generati con l’intelligenza artificiale. Confrontare le diverse soluzioni può essere un ulteriore esercizio per chiedersi quale sia il modo più efficace per garantire trasparenza senza rinunciare alle possibilità creative offerte dall’IA.