

Intelligenza Artificiale e la ridefinizione del “Vero”.

Deepfake e sintesi della realtà

Per molto tempo siamo stati abituati a pensare che vedere o ascoltare qualcosa rappresentasse una prova particolarmente forte della sua autenticità. Una fotografia poteva essere ritoccata, un video montato e una frase estrapolata dal proprio contesto, ma produrre immagini, voci o filmati realistici di eventi mai accaduti richiedeva competenze, tempo e strumenti professionali. L'intelligenza artificiale generativa sta modificando profondamente questa situazione.

Oggi strumenti accessibili a un numero crescente di persone permettono di generare testi, fotografie, illustrazioni, voci e video partendo da semplici istruzioni. È possibile creare il volto di una persona che non esiste, modificare una fotografia reale, clonare una voce, far pronunciare a qualcuno parole mai dette o ricostruire in video una situazione mai avvenuta. NIST, il National Institute of Standards and Technology statunitense, sottolinea come la produzione di deepfake realistici sia diventata un processo sempre più economico e accessibile e come, parallelamente, la loro individuazione costituisca una sfida crescente anche per i sistemi automatici di rilevamento.

Questo non significa che qualsiasi contenuto generato dall'intelligenza artificiale sia falso o pericoloso. Un'immagine sintetica può essere utilizzata dichiaratamente per illustrare un racconto, un video può ricostruire un avvenimento storico a scopo didattico e una voce artificiale può rendere più accessibile un servizio. Il problema nasce quando il contenuto sintetico viene presentato come documentazione autentica di qualcosa che sarebbe realmente accaduto.

L'arrivo dell'intelligenza artificiale generativa modifica quindi una delle domande fondamentali della media literacy. Non dobbiamo più chiederci soltanto **“è vera questa informazione?”**, ma anche **“da dove proviene ciò che sto vedendo o ascoltando e posso dimostrarne l'autenticità?”**.

1. Dai contenuti manipolati ai contenuti sintetici

La manipolazione delle informazioni non nasce con l'intelligenza artificiale. Fotografie ritoccate, documenti falsificati, montaggi audiovisivi e citazioni inventate esistono da molto tempo. In molti casi, inoltre, non è nemmeno necessario modificare tecnicamente un contenuto: una fotografia autentica di un evento avvenuto dieci anni prima può essere presentata come immagine di qualcosa accaduto ieri e diventare così uno strumento di disinformazione.

L'intelligenza artificiale generativa introduce però una possibilità ulteriore: **creare direttamente il contenuto che serve alla narrazione**, invece di limitarsi a modificare materiale già esistente.

Possiamo distinguere, a scopo didattico, alcune modalità principali:

a) **Generazione completa**: il testo, l'immagine, l'audio o il video vengono prodotti artificialmente. Può trattarsi, ad esempio, della fotografia realistica di una manifestazione che non si è mai verificata.

b) **Modifica generativa**: un contenuto reale viene trasformato attraverso l'intelligenza artificiale. Una persona può essere aggiunta o eliminata da una fotografia, un ambiente modificato oppure una parte di un video sostituita.

c) **Clonazione o sintesi della voce**: partendo da registrazioni di una persona è possibile produrre nuovi contenuti audio che ne imitano voce, intonazione e caratteristiche espressive, facendo apparentemente pronunciare frasi mai dette.

d) **Deepfake audiovisivo**: volto, voce o movimenti possono essere generati o manipolati in modo da rappresentare una persona reale mentre compie azioni o pronuncia parole che non appartengono alla registrazione originale.

La normativa europea utilizza una definizione precisa di *deepfake*: contenuti di immagini, audio o video generati o manipolati dall'IA che assomigliano a persone, oggetti, luoghi, entità o eventi esistenti e che potrebbero apparire falsamente autentici o veritieri a chi li osserva.

Non tutti i contenuti sintetici sono quindi deepfake e non tutti i deepfake vengono utilizzati necessariamente per disinformare. La stessa tecnica può essere impiegata in un film, in un'opera satirica o in una ricostruzione storica. A determinare il rischio sono soprattutto il **contesto, l'intenzione e il modo in cui il contenuto viene presentato al pubblico**.

2. Immagini, video, audio e testi: come cambia la manipolazione

2.1. Immagini sintetiche

Le immagini generate dall'intelligenza artificiale possono rappresentare persone, oggetti e luoghi mai esistiti oppure eventi mai accaduti. Gli errori visivi che caratterizzano le prime generazioni — mani deformate, scritte incomprensibili, oggetti impossibili — stanno diventando progressivamente meno evidenti. Basare la verifica soltanto sulla ricerca di un dettaglio "strano" è quindi un metodo sempre meno affidabile.

Un'immagine dovrebbe invece essere considerata nel suo contesto: chi l'ha pubblicata per primo? È possibile trovare fotografie dello stesso evento realizzate da altre persone? Il luogo rappresentato corrisponde a quello dichiarato? Esistono fonti giornalistiche o istituzionali che documentano l'avvenimento?

2.2. Video e deepfake

Il video viene spesso percepito come una prova particolarmente forte perché sembra mostrarci direttamente ciò che è successo. L'intelligenza artificiale permette però di modificare volti, movimenti e labiale oppure di produrre sequenze completamente sintetiche.

Anche un video autentico può inoltre essere manipolatorio senza essere tecnicamente modificato: può essere tagliato eliminando ciò che avviene prima e dopo, rallentato, accompagnato da un audio differente oppure presentato come registrazione di un luogo o di un momento diversi.

Per questo **“il video è vero?” non è una domanda sufficiente**. Dobbiamo chiederci anche quando sia stato registrato, dove, da chi e se il contesto attribuito al filmato corrisponda a quello originale.

2.3. Audio e clonazione della voce

La voce è un altro elemento che siamo abituati ad associare fortemente all'identità personale. La sintesi vocale permette invece di creare registrazioni nelle quali una persona sembra pronunciare frasi mai dette.

In questo caso la verifica visiva non può aiutarci. Diventa quindi ancora più importante risalire alla fonte: l'audio compare su un canale ufficiale? Esiste la registrazione completa? Altri giornalisti presenti hanno ascoltato la stessa dichiarazione? La persona coinvolta o la sua organizzazione ne confermano l'autenticità?

2.4. Testi generati dall'IA

Anche un testo fluido, ben scritto e apparentemente autorevole non costituisce una prova della correttezza delle informazioni che contiene. I sistemi generativi possono infatti produrre affermazioni plausibili ma inesatte e attribuire informazioni a fonti che devono essere controllate separatamente.

La regola fondamentale resta quindi la stessa: **l'autorevolezza apparente della forma non sostituisce la verifica delle fonti**.

3. Autenticità, verità e contesto

L'intelligenza artificiale ci obbliga a distinguere tre domande che spesso confondiamo:

- a) **Il contenuto è autentico?** Ossia: la fotografia, il video o l'audio provengono realmente dalla fonte e dall'evento dichiarati?
- b) **Il contenuto è stato modificato?** Ossia: sono intervenute manipolazioni, generative o tradizionali, che ne cambiano il significato?
- c) **L'informazione che accompagna il contenuto è vera?** Anche una fotografia completamente autentica può diventare disinformazione se viene collocata nel contesto sbagliato.

Questo ultimo caso è particolarmente importante. Davanti a un'immagine sospetta potremmo concentrarci esclusivamente sulla domanda **“È stata generata dall'IA?”**. Ma una vecchia fotografia autentica presentata come prova di un nuovo evento può ingannarci esattamente come un'immagine sintetica. Allo stesso modo, un'immagine generata dall'IA non è automaticamente disinformazione. Se un quotidiano utilizza un'illustrazione artificiale chiaramente indicata come tale per accompagnare un articolo, non sta necessariamente cercando di convincere il lettore che la scena rappresentata sia realmente avvenuta.

L'obiettivo della verifica non è quindi **scoprire l'intelligenza artificiale ovunque**, ma ricostruire correttamente provenienza, trasformazioni e contesto del contenuto.

4. Verificare un contenuto multimediale

Il fact-checking professionale offre un metodo utile per affrontare questo nuovo ambiente informativo. AFP spiega che i propri fact-checker non utilizzano un rilevatore di contenuti generati dall'IA come prova sufficiente: l'esito di uno strumento automatico deve sempre essere accompagnato da altre verifiche. L'agenzia utilizza inoltre almeno due fonti indipendenti per verificare l'affermazione centrale di un fact-check.

Possiamo tradurre questo approccio in alcuni principi di verifica:

- a) **Individuare l'affermazione da verificare:** prima di cercare qualcosa dobbiamo capire esattamente che cosa il contenuto pretende di dimostrare. “Questa fotografia mostra davvero l’evento X avvenuto nel luogo Y e nella data Z?” è una domanda verificabile; “questa immagine mi sembra strana” non lo è.
- b) **Risalire alla fonte originaria:** un contenuto condiviso migliaia di volte può aver perso completamente il proprio contesto. Occorre cercare chi lo abbia pubblicato per primo, quando e con quale descrizione.
- c) **Effettuare una ricerca inversa:** per fotografie e fotogrammi di video è possibile utilizzare motori di ricerca inversa per verificare se lo stesso contenuto sia apparso precedentemente online. AFP utilizza questo metodo anche per scoprire immagini autentiche pubblicate con contesti falsi. Per i video è possibile estrarre fotogrammi significativi e cercarli separatamente.
- d) **Verificare il contesto:** elementi come insegne, edifici, targhe, condizioni meteorologiche, abbigliamento, ombre, lingua utilizzata e caratteristiche del territorio possono essere confrontati con altre fonti per verificare luogo e data.
- e) **Cercare conferme indipendenti:** se un video mostra un evento eccezionale avvenuto in una grande città, è ragionevole aspettarsi altre immagini, testimonianze, comunicati o articoli. Una singola pubblicazione virale non diventa più affidabile soltanto perché è stata ricondivisa da migliaia di account.
- f) **Utilizzare con cautela i rilevatori automatici:** un software può fornire un indizio, non un verdetto. NIST sta sviluppando sistemi e benchmark specifici proprio perché l'individuazione automatica dei deepfake rimane un problema tecnico complesso e le prestazioni possono diminuire significativamente quando gli strumenti vengono utilizzati in situazioni reali differenti da quelle dei test.

Un esempio concreto viene dal lavoro quotidiano dei fact-checker. Nell'agosto 2026 AFP ha verificato un'immagine diffusa come documentazione delle violenze durante proteste nel Kashmir amministrato dal Pakistan. La verifica non si è basata su un singolo indizio: i giornalisti hanno analizzato incongruenze visive, confrontato il contenuto con altre fonti e utilizzato anche strumenti di rilevazione dell'IA come elemento aggiuntivo, arrivando alla conclusione che l'immagine fosse stata generata artificialmente.

5. Provenienza: sapere da dove arriva un contenuto

Accanto agli strumenti che cercano di individuare una manipolazione esiste un approccio differente: conservare informazioni verificabili sulla **provenienza** di un contenuto.

La Coalition for Content Provenance and Authenticity (C2PA) ha sviluppato uno standard denominato *Content Credentials* che permette di associare ai contenuti informazioni sulla loro origine e sulle trasformazioni subite. L'obiettivo è consentire a chi osserva un'immagine o un altro contenuto digitale di conoscere meglio come sia stato prodotto e modificato.

È importante però comprendere il limite di questo strumento. Conoscere la provenienza non significa conoscere automaticamente la verità. Una fotografia può avere una provenienza perfettamente documentata ed essere comunque utilizzata con una didascalia falsa. Viceversa, l'assenza di credenziali di provenienza non dimostra che un'immagine sia falsa.

La provenienza costituisce quindi **un elemento della verifica, non la verifica stessa**.

Anche la normativa europea si sta muovendo verso una maggiore trasparenza. Dal 2 agosto 2026 sono applicabili gli obblighi dell'articolo 50 dell'AI Act relativi alla riconoscibilità di determinati contenuti generati o manipolati dall'IA e all'etichettatura dei deepfake. L'obiettivo è ridurre il rischio che contenuti sintetici vengano scambiati per autentici, senza vietare in generale l'utilizzo dell'intelligenza artificiale nella produzione di contenuti.

6. Dal “credere ai propri occhi” al dubbio metodico

L'implementazione dei contenuti sintetici potrebbe portarci alla conclusione opposta: se tutto può essere manipolato, allora non possiamo più credere a niente. Anche questa reazione sarebbe problematica.

Il pensiero critico non coincide con il sospetto indiscriminato. **Dubitare metodicamente significa sospendere il giudizio finché non abbiamo raccolto elementi sufficienti**, non decidere in partenza che qualsiasi fotografia, video o notizia sia falsa.

La competenza fondamentale diventa quindi la capacità di passare dalla percezione alla verifica. Un contenuto realistico non è necessariamente autentico, così come un contenuto sorprendente non è necessariamente falso. Prima di condividerlo dobbiamo ricostruire la fonte, il contesto e le conferme disponibili.

Questo passaggio è particolarmente importante perché le tecnologie cambieranno. Gli errori visivi che oggi permettono talvolta di riconoscere un'immagine sintetica potrebbero scomparire domani. Un metodo basato esclusivamente sul riconoscimento di “difetti dell'IA” rischia quindi di diventare rapidamente obsoleto. Un metodo fondato su **fonti, provenienza, contesto e validazione incrociata** rimane invece valido anche quando gli strumenti di generazione migliorano.

Il nuovo problema della media literacy non è dunque imparare a distinguere perfettamente con gli occhi ciò che è umano da ciò che è artificiale. È imparare a non confondere **plausibilità e prova**, e a trasformare il dubbio iniziale in una procedura di verifica.

Riferimenti bibliografici e sitografici

National Institute of Standards and Technology – NIST (2024), *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency*.

National Institute of Standards and Technology – NIST, *GenAI: Deepfakes 2026 e Open Media Forensics Challenge*.

Agence France-Presse – AFP, *How we work* e *AFP Fact-Checking Stylebook*.

Coalition for Content Provenance and Authenticity – C2PA, *C2PA Specifications – Content Credentials*.

Commissione europea, *Norme di trasparenza per i sistemi di IA e Codice di buone pratiche sulla trasparenza dei contenuti generati dall'IA*.

Traccia per l'attività in classe

Attività 1: Come i quotidiani raccontano l'intelligenza artificiale e la crisi del "vero"

Scopo di questa attività è stimolare gli allievi a comprendere come il giornalismo affronti il problema dei contenuti sintetici e quali elementi utilizzi per distinguere tra possibilità tecniche, rischi reali e scenari soltanto ipotetici.

Il docente seleziona dalle sezioni dedicate a tecnologia, cultura, società o media di uno o più quotidiani alcuni articoli relativi all'intelligenza artificiale generativa, ai deepfake, alla clonazione delle voci o alla manipolazione di immagini e video. È preferibile scegliere articoli che presentino anche casi concreti e che permettano agli studenti di individuare le fonti utilizzate dal giornalista.

Gli studenti, lavorando in coppia, analizzano uno degli articoli compilando il modulo seguente.

Analisi di un articolo sull'IA e sui contenuti sintetici

Descrizione

Articolo analizzato:

1. Quotidiano, autore e data di pubblicazione
2. Tecnologia o fenomeno descritto
3. Tipologia di contenuto coinvolto: testo, immagine, audio o video
4. Caso concreto utilizzato nell'articolo
5. Qual è il possibile rischio informativo evidenziato?
6. Quali fonti vengono citate dal giornalista?
7. Sono presenti fonti primarie? Quali?
8. Sono consultati esperti? Con quali competenze?
9. Viene spiegato come sia stata verificata l'autenticità dei contenuti?
10. L'articolo distingue tra contenuto sintetico e disinformazione?
11. Vengono indicate eventuali incertezze o limiti degli strumenti di verifica?
12. Il titolo rappresenta correttamente ciò che emerge dall'articolo?

13. Quali informazioni ulteriori sarebbero necessarie per comprendere meglio il caso?

14. Quali indicazioni utili offre l'articolo a chi incontra contenuti simili online?

Una volta compilati i moduli, le coppie presentano la propria analisi alla classe. Il confronto dovrà concentrarsi soprattutto sul modo in cui i diversi quotidiani costruiscono l'affidabilità della notizia: quali prove mostrano, quali fonti utilizzano e quali aspetti lasciano invece alla fiducia del lettore.

In una seconda sessione gli studenti potranno confrontare due articoli dedicati allo stesso fenomeno, verificando se utilizzano le stesse fonti, se raggiungono conclusioni differenti e quale dei due rende più trasparente il proprio percorso di verifica.

Attività 2: Fact-checking di un contenuto sintetico o manipolato

Scopo di questa attività è fornire agli allievi un **metodo replicabile di verifica delle fonti multimediali**, sviluppando l'abitudine a sospendere la condivisione finché non siano state raccolte prove sufficienti attraverso fonti indipendenti.

Il docente seleziona un contenuto recente diventato virale — fotografia, video o audio — successivamente verificato da una redazione giornalistica o da un'organizzazione professionale di fact-checking. Inizialmente agli studenti viene mostrato soltanto il contenuto originale e il messaggio con cui è stato diffuso, senza rivelare il risultato del fact-check.

Le coppie dovranno comportarsi come una piccola redazione di verifica e compilare progressivamente la seguente scheda.

Protocollo di verifica di un contenuto multimediale

Risultati della verifica

Contenuto analizzato:

- 1. Qual è esattamente l'affermazione che il contenuto pretende di dimostrare?**
- 2. Chi lo ha pubblicato?**
- 3. È possibile individuare la pubblicazione originaria?**
- 4. Quando è comparso per la prima volta?**
- 5. La fonte è identificabile e attendibile?**
- 6. La ricerca inversa produce risultati precedenti o simili?**
- 7. Luogo e data dichiarati sono verificabili?**
- 8. Esistono dettagli visivi o sonori incoerenti?**

9. Esistono altre immagini o registrazioni indipendenti dello stesso evento?
10. Almeno due fonti indipendenti confermano l'avvenimento?
11. Sono disponibili comunicazioni ufficiali o testimonianze dirette?
12. Sono presenti informazioni sulla provenienza o Content Credentials?
13. Eventuali strumenti di rilevazione dell'IA che risultato forniscono?
14. Quali sono le prove più solide raccolte?
15. Conclusione: autentico, sintetico, manipolato, fuori contesto oppure non verificabile?

Il principio fondamentale dell'attività è che **nessun singolo elemento deve essere considerato sufficiente**. Un dito deformato non dimostra che una fotografia sia artificiale, così come il risultato positivo o negativo di un rilevatore automatico non conclude la verifica. Gli studenti devono costruire la propria valutazione attraverso la convergenza di più prove.

Terminata l'indagine, il docente mostra il fact-check professionale relativo allo stesso contenuto. Gli studenti confrontano quindi il proprio procedimento con quello utilizzato dai giornalisti: quali passaggi avevano individuato? Quali prove erano sfuggite? Il fact-checker ha utilizzato strumenti diversi? Quale elemento è stato decisivo?

Come modello può essere utilizzata anche la metodologia pubblicata da AFP, che prevede la ricerca della fonte originale, l'analisi inversa delle immagini, la verifica del contesto, il confronto con fonti indipendenti e l'uso prudente degli strumenti automatici di riconoscimento dell'IA. Un esempio recente è il fact-check pubblicato da AFP il 12 agosto 2026 su un'immagine generata con l'IA e falsamente collegata alle violenze elettorali nel Kashmir amministrato dal Pakistan.

In una seconda sessione di lavoro potranno essere gli studenti stessi, utilizzando quotidiani e fonti giornalistiche affidabili, a individuare un nuovo caso di immagine, audio o video controverso e ad applicare autonomamente il protocollo. Il risultato finale dovrà assumere la forma di un breve **fact-check**, nel quale non sarà sufficiente scrivere "vero" o "falso": dovranno essere illustrate le fonti consultate, le verifiche effettuate e le prove che giustificano la conclusione.