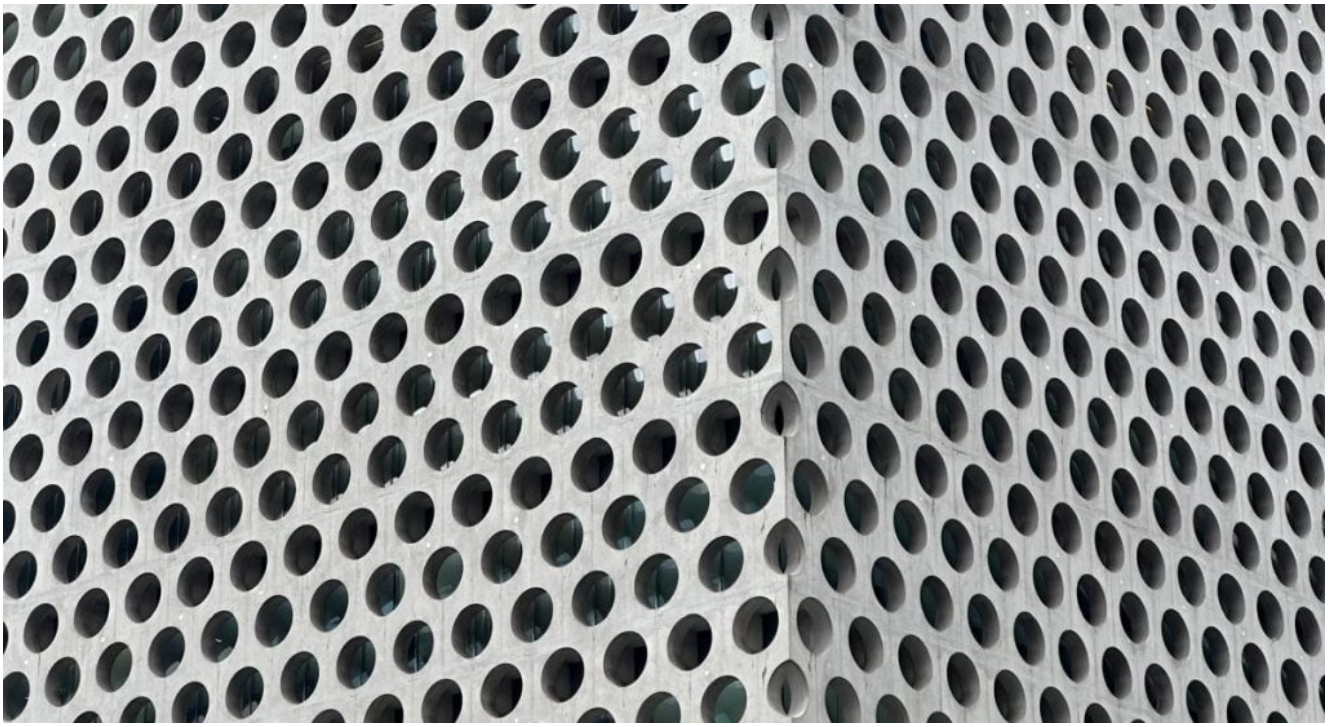
Chieti, un candidato fa campagna elettorale su Pornhub, Tinder e Hinge e abbandona i canali più tradizionali della propaganda politica

Nel filmato che precede i video a luci rosse su Pornhub, Paride Paci adotta apertamente il linguaggio dell’allusione. Sulle app di incontri, invece, la promozione passa attraverso immagini e call to action pensate per mimare l’estetica tipica di quelle piattaforme

Social, non è dipendenza. È ingegneria cognitiva.

Il verdetto di Los Angeles contro Meta e Google non chiude una causa. Apre una domanda che il mondo della comunicazione non può più fingere di non sentire e per la quale la risposta sta anche nell’internazionalizzazione della cultura della comunicazione e FERPI ne è protagonista.

La tela delle bugie: distopie cognitive e manipolazione



Nello spazio digitale contemporaneo, il falso non è più un’anomalia: è una tecnica di influenza, un fattore economico e uno strumento di potere. Fake news, recensioni manipolate, contenuti sintetici e deepfake agiscono in modo sempre più pervasivo sulla formazione dell’opinione pubblica, sulle scelte dei consumatori e sulla stessa affidabilità dell’ecosistema informativo. La menzogna online non si limita a distorcere il vero, ma costruisce realtà alternative credibili, capaci di orientare consenso, reputazione e mercato. In questo scenario, l’intelligenza artificiale amplifica quantità, velocità e sofisticazione della manipolazione, rendendo più sottile il confine tra autenticità e artificio. La posta in gioco non riguarda soltanto la correttezza dell’informazione o la lealtà commerciale, ma la tenuta di valori essenziali come trasparenza, fiducia e libertà di scelta. Contrastare questi fenomeni significa allora impegnarsi per realizzare forme di tutela giuridica, tecnologica e culturale: difendere la verità come bene pubblico in un ambiente in cui l’inganno, ormai, può essere prodotto, diffuso e monetizzato su scala industriale.

1. I dati: dimensione quantitativa e percezione sociale del fenomeno

La disinformazione digitale ha assunto una dimensione quantitativa e qualitativa senza precedenti. I dati statistici evidenziano infatti delle tendenze convergenti: la crescente sofisticazione delle fake news, l'aumento dei contenuti manipolati mediante strumenti di intelligenza artificiale, la difficoltà per gli utenti di distinguere tra contenuti autentici e sintetici, l'elevata esposizione delle fasce più giovani della popolazione ai flussi informativi digitali.

Secondo il Digital News Report 2025 del Reuters Institute for the Study of Journalism, il 58% degli intervistati a livello globale ritiene che distinguere tra informazioni vere e false online sia diventato sempre più difficile. Nello stesso rapporto la fiducia complessiva nei media si attesta mediamente intorno al 32%, con forti differenze tra Paesi, mentre la preoccupazione per la diffusione della disinformazione online rimane elevata nella maggior parte delle aree analizzate, superando il 50% in diversi contesti europei e negli Stati Uniti[1].

Analoghe evidenze emergono da un'indagine internazionale condotta nel 2023 da UNESCO in collaborazione con Ipsos, che mostra come una larga maggioranza degli utenti Internet a livello globale dichiari di essere entrata in contatto con fake news sui social network, con impatti percepiti in particolare nell'ambito dell'informazione politica e sanitaria. Lo studio rileva che oltre l'85% degli intervistati teme l'impatto politico della disinformazione, mentre il 53% identifica nell'intelligenza artificiale un fattore di amplificazione del fenomeno. Una quota significativa degli utenti riconosce inoltre di avere difficoltà nel distinguere contenuti autentici da contenuti manipolati, soprattutto quando questi sono generati mediante tecniche di intelligenza artificiale[2].

Nel contesto europeo, diversi studi segnalano un aumento dell'esposizione percepita alla disinformazione, con alcuni Paesi particolarmente vulnerabili alle campagne di manipolazione informativa. Tra questi viene frequentemente indicata anche l'Italia, dove la diffusione dei contenuti disinformativi appare particolarmente elevata rispetto alla media europea[3].

Secondo un Report 2025 dell'Unione Europea, l'82% del campione europeo (86% di quello italiano) ritiene di essere stato esposto disinformazione e fake news spesso o almeno qualche volta negli ultimi sette giorni. E sebbene poco più di sei intervistati su dieci si sentano sicuri di poter riconoscere la disinformazione quando la incontrano (61%), circa tre intervistati su dieci non hanno fiducia nella propria capacità di riconoscere la disinformazione[4].

Quanto ai giovani, i dati nazionali confermano tale tendenza. L'indagine IPSOS 2025 "*Alfabetizzazione digitale e fake news*" ha rilevato che l'85% dei ragazzi (studenti di scuola secondaria) ritiene che le fake news sui social influenzino opinioni e comportamenti delle persone. Ciò nonostante, proprio i più digitalmente competenti dimostrano di essere quelli che vi mettono più "like".[5]

Il Rapporto Censis–Ital Communications 2023 evidenzia, inoltre, che il 76,5% degli intervistati ritiene che tali contenuti siano diventati sempre più sofisticati e difficili da riconoscere. Un'indagine condotta dall'Italian National Council (INC) rileva, inoltre, che l'83% degli intervistati ammette di aver creduto almeno una volta ad una notizia falsa diffusa online[6].

Particolarmente significativo è anche il dato relativo alla percezione del fenomeno. Secondo Infodata – Il Sole 24 Ore, circa tre persone su dieci ritengono che la disinformazione non rappresenti un problema rilevante. Questo scarto tra l'elevato livello di esposizione alle fake news e la

percezione relativamente bassa della loro pericolosità costituisce uno degli elementi che contribuiscono a rendere il fenomeno strutturalmente persistente nell'ecosistema informativo digitale[7].

Sul piano globale, numerose analisi evidenziano una crescita significativa della disinformazione prodotta mediante strumenti di intelligenza artificiale generativa: l'uso di modelli di generazione automatica del linguaggio e delle immagini ha abbattuto i costi di produzione dei contenuti manipolati, rendendo possibile la creazione su larga scala di articoli, immagini e video falsi ma altamente verosimili[8].

Dati riportati da Agência Brasil nel 2026 indicano, ad esempio, che nel contesto brasiliano i contenuti disinformativi generati mediante intelligenza artificiale sono più che triplicati tra il 2024 e il 2025[9]. Ulteriori analisi sono state condotte da organizzazioni indipendenti come NewsGuard, che ha documentato la proliferazione di centinaia di siti web generati automaticamente mediante strumenti di intelligenza artificiale e progettati per imitare graficamente testate giornalistiche affidabili. Tali siti pubblicano contenuti generati automaticamente o semi-automaticamente, contribuendo ad amplificare la diffusione di informazioni manipolate o non verificate nello spazio digitale[10].

Parallelamente, un numero crescente di studi economici evidenzia la rilevanza del fenomeno nel contesto del mercato digitale e della reputazione online delle imprese. Analisi empiriche dimostrano che anche variazioni minime nei rating delle piattaforme di recensione possono produrre effetti economicamente rilevanti sui ricavi delle imprese, confermando il ruolo della reputazione digitale come asset strategico nell'economia delle piattaforme[11].

Secondo alcune stime internazionali, tra il 10% e il 30% delle recensioni online potrebbe essere manipolato o non autentico,

con effetti significativi sul comportamento dei consumatori. Alcuni, inoltre, indicano che un incremento artificiale del rating di un prodotto o servizio può generare aumenti di ricavi fino al 15 – 20%, evidenziando come la manipolazione reputazionale possa alterare non soltanto il flusso informativo, ma anche l'equilibrio concorrenziale del mercato[12].

Queste dinamiche sono ulteriormente amplificate dalla crescente diffusione di strumenti di automazione e generazione di contenuti mediante intelligenza artificiale nel settore del marketing digitale; si è registrata una crescita esponenziale dei sistemi automatizzati di generazione di recensioni e commenti, spesso utilizzati per simulare consenso o per alterare artificialmente l'appetibilità di prodotti e servizi[13].

Il salto qualitativo del fenomeno emerge con particolare evidenza nella sua crescente dimensione geopolitica: negli ultimi anni, infatti, numerose indagini hanno documentato l'esistenza di operazioni coordinate di disinformazione finalizzate ad influenzare processi elettorali e dibattiti politici in diversi Paesi. Tra queste, è stata ampiamente analizzata la cosiddetta "Operazione Doppelgänger", attribuita a reti di propaganda filorusse e basata sulla replicazione grafica di siti di informazione europei, sulla diffusione coordinata di contenuti manipolati relativi al conflitto in Ucraina ed a consultazioni elettorali in diversi Paesi dell'Unione europea[14].

L'Istituto per gli Studi di Politica Internazionale (ISPI) ha sottolineato come fake news, intelligenza artificiale generativa ed interferenze straniere costituiscano oggi forme di minaccia ibrida per le democrazie liberali, capaci di incidere simultaneamente sulla sicurezza informativa, sul dibattito pubblico e sulla stabilità istituzionale[15].

In questo contesto, episodi recenti – come le campagne di

disinformazione individuate tra il 2024 e il 2025 in Moldavia in occasione di importanti consultazioni elettorali – mostrano come la manipolazione informativa possa essere utilizzata come strumento di pressione geopolitica e di influenza strategica[16]. Secondo l'IDMO (European Digital Media Observatory) e il Microsoft Threat Analysis Center, l'impiego dell'intelligenza artificiale generativa sta ulteriormente ampliando la capacità di produzione e diffusione di contenuti propagandistici e manipolati, rendendo le operazioni di influenza sempre più sofisticate e difficili da individuare[17].

2. Dalla disinformazione episodica alla disinformazione strutturale: la disinformazione come fenomeno sistemico

Nella fase iniziale dello sviluppo di Internet, la disinformazione veniva generalmente interpretata come una distorsione episodica del flusso informativo, riconducibile a singoli attori o a specifiche campagne di propaganda. Tuttavia, l'evoluzione delle piattaforme digitali e dei modelli economici basati sull'attenzione ha progressivamente trasformato questo fenomeno in una componente strutturale dell'ecosistema comunicativo contemporaneo.

Oggi la disinformazione non costituisce più un'anomalia marginale del sistema informativo, bensì una dinamica sistemica strettamente connessa al funzionamento delle infrastrutture digitali. L'architettura delle piattaforme online si fonda, infatti, su algoritmi di raccomandazione, sistemi di profilazione degli utenti e modelli economici orientati alla massimizzazione dell'*engagement*. In tale contesto, i contenuti che suscitano reazioni emotive intense – indignazione, paura, conflitto – tendono ad essere maggiormente amplificati dai meccanismi di distribuzione algoritmica, favorendo la diffusione di narrazioni polarizzanti o manipolative[18].

A tale dinamica si aggiunge l’impatto crescente dell’intelligenza artificiale generativa, che ha ulteriormente ridotto i costi di produzione e distribuzione dei contenuti. I sistemi di generazione automatica consentono oggi di produrre su larga scala testi, immagini e prodotti audiovisivi plausibili, spesso difficilmente distinguibili da quelli autentici. La disinformazione diventa così replicabile, adattabile e personalizzabile, potendo essere modellata in funzione del destinatario e diffusa attraverso campagne altamente mirate[19].

Ne deriva una mutazione qualitativa dell’ecosistema informativo dove l’asimmetria informativa non riguarda più soltanto singole relazioni tra emittente e destinatario, ma tende a investire l’intero spazio pubblico digitale, modificando le condizioni nelle quali si formano opinioni, decisioni economiche e scelte politiche. È così, dunque, che il fenomeno contribuisce all’erosione della fiducia nelle istituzioni e nelle fonti autorevoli di informazione e può incidere profondamente sulla qualità del dibattito pubblico e sulla stabilità dei sistemi democratici.

Da questa duplice incidenza emerge progressivamente un bene giuridico composito, che può essere individuato nell’integrità dell’ecosistema informativo. La tutela non può essere affidata esclusivamente agli strumenti tradizionali del diritto dell’informazione o del diritto della concorrenza. Essa richiede piuttosto un approccio regolatorio multilivello, capace di integrare interventi normativi, responsabilità delle piattaforme digitali, strumenti di trasparenza algoritmica e politiche di “alfabetizzazione digitale”. In questo contesto, la regolazione dell’ecosistema informativo appare destinata a diventare uno dei principali terreni di evoluzione del diritto dell’economia digitale e della governance delle piattaforme online.

3. Intelligenza artificiale generativa, deepfake e nuove frontiere della disinformazione

Le evoluzioni più recenti, in questo ambito e non solo, sono strettamente connesse allo sviluppo dei sistemi di intelligenza artificiale generativa, capaci di produrre contenuti testuali, visivi e audiovisivi sempre più realistici.

I modelli generativi basati su tecniche di deep learning, come i *large language models* e i sistemi di sintesi audiovisiva, consentono oggi la produzione automatizzata di articoli, immagini, registrazioni vocali e video manipolati con un grado di verosimiglianza tale da rendere estremamente complessa la distinzione tra contenuti autentici e contenuti artificialmente generati. In questo contesto si colloca il fenomeno dei cosiddetti "deepfake", ossia contenuti audiovisivi sintetici realizzati mediante tecniche di intelligenza artificiale che permettono di alterare o ricreare l'immagine e la voce di una persona in modo estremamente realistico[20].

La diffusione di tali tecnologie introduce una nuova dimensione: se nelle fasi precedenti la manipolazione informativa si basava prevalentemente sulla selezione e reinterpretazione di contenuti reali, l'intelligenza artificiale generativa consente oggi la creazione di eventi informativi completamente artificiali, rendendo possibile la produzione su larga scala di narrazioni plausibili ma interamente fittizie. Studi recenti evidenziano come tali strumenti possano essere impiegati anche nell'ambito di vere e proprie campagne di manipolazione informativa, amplificando ulteriormente la capacità di diffusione della disinformazione nell'ambiente digitale[21].

In particolare, la possibilità di utilizzare l'IA per

introdurre in rete, in maniera massificata, immagini verosimili capaci di imprimersi nella memoria degli utenti e di diventare virali, può dare luogo al c.d. “effetto Mandela”: una falsa memoria collettiva in cui diverse persone ricordano gli stessi dettagli sbagliati riguardanti accadimenti, immagini o affermazioni. Un esempio di questo effetto, in relazione all’intelligenza artificiale, è la finta foto di Papa Francesco che indossava un enorme cappotto bianco di Balenciaga, diventata virale ed entrata nell’immaginario (falso) collettivo [57]. Un effetto che muove dalla constatazione empirica e neuropsicologica che la memoria (anche quella autobiografica, insieme di ricordi semantici ed episodici) mantiene traccia del percepito e del vissuto emotivo e dimentica progressivamente l’origine e la fonte: il vero psicologico che prevale sul vero oggettivo. Dinamica amplificata drammaticamente dall’intelligenza artificiale e dalla bulimia compulsiva verso contenuti digitali alimentata dagli algoritmi.

Di fronte a tali rischi, il legislatore europeo ha iniziato a introdurre specifici obblighi normativi volti a garantire una maggiore trasparenza dei contenuti generati artificialmente. Il Regolamento europeo sull’intelligenza artificiale (AI Act) prevede, ad esempio, che i sistemi di IA capaci di generare contenuti sintetici debbano adottare misure idonee a rendere tali contenuti chiaramente identificabili come artificiali. In particolare, i fornitori di sistemi di IA generativa sono tenuti a informare gli utenti quando interagiscono con contenuti prodotti o manipolati mediante intelligenza artificiale[22].

Tali interventi normativi riflettono una crescente consapevolezza: nell’era dell’intelligenza artificiale generativa, la tutela dell’ecosistema informativo richiede non soltanto strumenti di contrasto alla disinformazione tradizionale, ma anche meccanismi capaci di garantire la tracciabilità e la riconoscibilità dei contenuti sintetici.

La sfida per il diritto contemporaneo consiste dunque nell’elaborare forme di regolazione che consentano di valorizzare le potenzialità innovative dell’intelligenza artificiale senza compromettere l’affidabilità dell’ambiente informativo.

4. Fake news: misinformation, disinformation e propaganda

Nel dibattito scientifico ed istituzionale contemporaneo il fenomeno delle fake news viene sempre più frequentemente analizzato attraverso una distinzione concettuale tra misinformation, disinformation e propaganda. Categorie che, pur condividendo l’elemento della non veridicità o della manipolazione dell’informazione, si differenziano per struttura, finalità e grado di intenzionalità.

La misinformation consiste nella diffusione di informazioni false o inesatte che circolano nello spazio pubblico senza una specifica intenzione manipolativa. In tali casi la falsità del contenuto deriva spesso da errori interpretativi, incomprensioni o dalla riproduzione inconsapevole di notizie non verificate. L’elemento soggettivo che caratterizza la misinformation è dunque l’assenza di dolo: chi diffonde il contenuto non agisce con l’intento di alterare deliberatamente il processo informativo, ma contribuisce comunque alla circolazione di contenuti inaccurati o fuorvianti all’interno dell’ecosistema digitale[23].

Diverso è il caso della disinformation, che si configura quando la diffusione di informazioni false avviene in modo consapevole e deliberato, con l’obiettivo di influenzare percezioni, opinioni o comportamenti collettivi. L’elemento qualificante della disinformation può essere individuato nella volontà intenzionale di alterare la percezione della realtà attraverso la manipolazione del flusso informativo. In questo senso, la disinformazione rappresenta non soltanto una

deviazione patologica del processo comunicativo, ma una vera e propria strategia di influenza, capace di incidere sul funzionamento dei mercati, sul dibattito pubblico e, nei casi più estremi, sui processi democratici[24].

Accanto a tali fenomeni si colloca la propaganda digitale, che può essere definita come l'utilizzo sistematico e organizzato della disinformazione per finalità politiche, ideologiche o strategiche. A differenza delle forme episodiche di manipolazione informativa, la propaganda si caratterizza per la presenza di strutture organizzate, campagne coordinate e tecniche di amplificazione artificiale dei contenuti, spesso realizzate attraverso reti di account automatizzati, influencer coordinati o strategie di micro-targeting. Nel contesto delle piattaforme digitali, tali dinamiche assumono una portata particolarmente rilevante poiché l'architettura algoritmica dei sistemi di raccomandazione tende a privilegiare contenuti ad elevata capacità di generare engagement, ossia interazioni emotive intense, polarizzazione e conflitto[25].

Un ambito nel quale tali dinamiche emergono con particolare evidenza è quello dei conflitti armati e delle guerre ibride, dove le fake news non si esauriscono nella diffusione episodica di contenuti falsi, ma si inseriscono in strategie più ampie di influenza e manipolazione narrativa. Si pensi, ad esempio, alle già citate campagne filorusse legate all'operazione "Doppelgänger", fondate sulla clonazione grafica di testate giornalistiche europee e sulla diffusione di contenuti costruiti per alterare la percezione del conflitto in Ucraina e orientare il dibattito pubblico nei Paesi dell'Unione. Oppure si guardi, nel contesto della guerra tra Israele e Hamas, alla circolazione, soprattutto nei giorni immediatamente successivi al 7 ottobre 2023, di immagini e video di repertorio, tratti da altri teatri di guerra o da eventi precedenti, rilanciati come se documentassero in tempo reale bombardamenti, attacchi o atrocità del conflitto in

corso, con l’effetto di massimizzare l’impatto emotivo delle notizie e di orientare in senso polarizzato la reazione dell’opinione pubblica. Così anche le campagne di delegittimazione rivolte contro attori umanitari, come nel caso di UNRWA, che ha denunciato la diffusione sistematica di informazioni false e manipolate utilizzate come strumento di guerra per screditarne l’operato e ostacolarne indirettamente l’azione umanitaria. In questa prospettiva, la disinformazione diventa parte integrante dell’arsenale comunicativo del conflitto, poiché non mira soltanto a falsare singoli fatti, ma a condizionare il quadro interpretativo complessivo entro il quale opinione pubblica, istituzioni e comunità internazionali valutano gli eventi bellici[26].

Analoga funzione può assumere la disinformazione quando venga impiegata come strumento di delegittimazione nei confronti di magistrati e organi di polizia giudiziaria. In tali ipotesi, la diffusione di notizie false, contenuti decontestualizzati, accuse infondate o narrazioni manipolative non mira soltanto a colpire la reputazione individuale dei singoli funzionari, ma tende a incrinare la fiducia collettiva nell’imparzialità della giurisdizione e nella credibilità dell’attività investigativa, creando un ambiente ostile idoneo a condizionare indirettamente anche il sereno esercizio della funzione[27]. È in questa prospettiva che la giurisprudenza europea – pur riconoscendo l’ampio spazio che in una società democratica deve essere riservato alla critica dell’operato giudiziario – distingue nettamente tra il dissenso, anche severo, e l’attribuzione di fatti specifici privi di un’adeguata base fattuale, soprattutto quando tali affermazioni sono suscettibili di compromettere l’autorevolezza della giustizia[28]. Su un piano comparato, il fenomeno ha assunto forme particolarmente evidenti tanto in Europa orientale – ove non sono mancate campagne di discredito organizzato contro giudici percepiti come ostili al potere politico – quanto negli Stati Uniti, dove il Chief Justice Roberts ha incluso espressamente disinformazione e

intimidazione tra le principali minacce contemporanee all'indipendenza giudiziaria[29]. Per quanto concerne la polizia giudiziaria, il salto di qualità impresso dai deepfake rende oggi possibile fabbricare audio e video sintetici idonei non solo a frodi o sostituzioni di persona, ma anche a screditare investigatori, insinuare dubbi sulla genuinità delle attività di indagine o contaminare il contesto informativo nel quale la prova digitale viene percepita e valutata, tanto che Europol considera la capacità di prevenzione e di rilevazione di tali contenuti una priorità strategica per le forze dell'ordine[30].

Peraltro, gli algoritmi che regolano la visibilità dei contenuti sulle principali piattaforme social operano sulla base di logiche di ottimizzazione dell'attenzione e del tempo di permanenza degli utenti. Come sopra accennato, tale modello economico, fondato sulla cosiddetta "economia dell'attenzione"[31], tende inevitabilmente a favorire la diffusione di contenuti sensazionalistici, emotivamente polarizzanti o controversi, che risultano statisticamente più capaci di generare condivisioni, commenti e reazioni[32].

Una manifestazione ulteriore di tali dinamiche si osserva quando la disinformazione si innesta in competizioni elettorali già esposte a condizionamenti criminali, clientelari o corruttivi: in questi contesti, infatti, le fake news cessano di essere meri contenuti falsi e diventano uno strumento attraverso il quale il controllo del territorio, l'intermediazione opaca e lo scambio di utilità si trasferiscono sul piano simbolico e reputazionale, alterando la percezione dei candidati, delegittimando gli avversari e restringendo progressivamente lo spazio del confronto pubblico[33]. In questa prospettiva, le campagne delegittimanti e quelle di sostegno non appaiono più come fenomeni distinti, ma come due versanti di una medesima strategia di influenza: le prime mirano a isolare candidati, amministratori, giornalisti o istituzioni di garanzia

attraverso accuse infondate, insinuazioni o narrazioni di brogli; le seconde costruiscono attorno a determinati candidati un consenso artificiale, simulando radicamento sociale, affidabilità e inevitabilità dell'esito elettorale[34]. Nel contesto italiano il fenomeno emerge soprattutto come vulnerabilità strutturale dell'ambiente informativo locale: AGCOM ha descritto la disinformazione come fenomeno misurabile del sistema informativo nazionale. Nel 2024 il Ministero dell'interno ha registrato 630 atti intimidatori nei confronti degli amministratori locali, in aumento del 13,9 per cento rispetto all'anno precedente, e la letteratura sul rapporto fra mafia e competizione elettorale ha mostrato, da un lato, il sostegno elettorale assicurato in alcuni contesti dalle organizzazioni mafiose e, dall'altro, l'incremento della violenza contro figure politiche in prossimità del voto[35].

Fuori dal contesto nazionale, l'Operazione "Teatro Invisível" in Brasile ha riguardato un'organizzazione criminale dedita alla diffusione di notizie false su candidati a sindaco in più di dieci municipi dello Stato di Rio de Janeiro, mentre la successiva "Teatro Invisível II" ha collegato quel circuito disinformativo a ipotesi di riciclaggio, frodi in appalti e distruzione di prove digitali; in Moldavia diverse indagini hanno fatto emergere ingenti finanziamenti illeciti russi destinati alla manipolazione elettorale, anche tramite l'utilizzo strategico e la diffusione organizzata di fake news su Facebook, TikTok e Telegram.

Il risultato complessivo di tale dinamica è una crisi dell'autorità epistemica nell'ecosistema informativo contemporaneo. Ne deriva una crescente difficoltà per cittadini e consumatori nel distinguere informazioni affidabili da contenuti manipolati, con effetti rilevanti non solo sul piano culturale e sociale, ma anche su quello giuridico ed economico[36].

5. Recensioni false, strategie di marketing e pratiche ingannevoli

Un fenomeno particolarmente significativo della disinformazione nell'economia digitale è rappresentato dalle recensioni false online, che costituisce una forma specifica di manipolazione informativa capace di incidere direttamente sui meccanismi di funzionamento del mercato.

Nel contesto delle piattaforme digitali, le recensioni degli utenti hanno assunto una funzione sempre più rilevante nella costruzione della reputazione commerciale di imprese, prodotti e servizi. Esse rappresentano oggi uno dei principali strumenti attraverso cui i consumatori orientano le proprie scelte economiche, svolgendo una funzione informativa che, di fatto, integra o sostituisce le tradizionali forme di pubblicità e comunicazione commerciale. Secondo numerosi studi, una quota molto elevata di consumatori consulta sistematicamente le recensioni online prima di effettuare acquisti o scegliere servizi, attribuendo a tali contenuti un elevato grado di affidabilità[37].

Proprio la centralità economica della reputazione digitale ha favorito la diffusione di pratiche manipolative volte ad alterare artificialmente la percezione pubblica di prodotti e servizi. Tra le tecniche più diffuse si collocano la pubblicazione di recensioni positive artificiali, finalizzate a migliorare la reputazione di un'impresa, e la diffusione di recensioni negative coordinate, dirette a danneggiare concorrenti. Tali fenomeni rientrano nelle strategie di "*astroturfing*", ossia nella creazione artificiale di consenso apparente attraverso contenuti presentati come spontanei ma in realtà orchestrati da soggetti interessati[38].

Accanto all'*astroturfing* si è diffusa anche la pratica del cosiddetto "*review bombing*", consistente nella pubblicazione

massiva di recensioni negative coordinate, spesso motivate non da effettive esperienze di consumo ma da campagne organizzate con finalità reputazionali, ideologiche o commerciali. Tali dinamiche possono incidere significativamente sulla visibilità dei prodotti nelle piattaforme digitali e, di conseguenza, sulle scelte dei consumatori[39].

Il problema delle recensioni false assume una rilevanza non soltanto etica o informativa, ma anche giuridica ed economica, in quanto incide su diversi interessi tutelati dall'ordinamento. In primo luogo, la manipolazione delle recensioni altera il corretto funzionamento del mercato digitale, producendo effetti distorsivi sulla concorrenza tra operatori economici. In secondo luogo, essa compromette il diritto dei consumatori a ricevere informazioni veritiere e trasparenti, elemento essenziale per l'esercizio di scelte di acquisto consapevoli.

Proprio per tali ragioni, il diritto europeo ha progressivamente riconosciuto la rilevanza giuridica delle recensioni online all'interno della disciplina delle pratiche commerciali scorrette. La direttiva (UE) 2019/2161 – nota come Omnibus Directive – ha introdotto specifici obblighi di trasparenza per le piattaforme digitali, imponendo ai professionisti che pubblicano recensioni di indicare se e in che modo sia stato verificato che le stesse provengano da consumatori che hanno effettivamente acquistato o utilizzato il prodotto o servizio recensito[40].

Parallelamente, il Digital Services Act ha imposto alle piattaforme online obblighi più stringenti in materia di trasparenza algoritmica, gestione dei contenuti illegali e responsabilità nella moderazione dei contenuti. Sebbene il regolamento non disciplini specificamente le recensioni false, esso introduce strumenti normativi destinati ad incidere indirettamente anche su tali fenomeni, in particolare attraverso obblighi di gestione del rischio sistemico e di maggiore accountability delle piattaforme digitali[41].

6. Il quadro normativo italiano

Nel sistema giuridico italiano qualunque intervento giuridico in materia di disinformazione deve necessariamente confrontarsi con il principio fondamentale sancito dall'articolo 21 della Costituzione, che riconosce a tutti il diritto di manifestare liberamente il proprio pensiero con la parola, lo scritto e ogni altro mezzo di diffusione.

Tuttavia, tale libertà non assume carattere assoluto, ma incontra limiti derivanti dalla necessità di tutelare altri valori costituzionali di pari rilievo, tra cui la dignità della persona, la reputazione individuale e l'ordine pubblico.

L'esercizio della libertà di informazione deve tuttavia rispettare alcuni criteri fondamentali, tra cui la verità dei fatti narrati, la pertinenza dell'informazione e la continenza espressiva. Il bilanciamento tra libertà di espressione e tutela della reputazione rappresenta, dunque, uno degli snodi centrali del diritto dell'informazione.

In Italia la normativa che mira a disciplinare il contrasto ai fenomeni di disinformazione in senso lato non è organica e unitaria. La tutela si articola piuttosto attraverso una pluralità di strumenti normativi appartenenti a diversi settori dell'ordinamento, che operano su livelli tra loro complementari.

Da un lato, sicuramente assumono rilievo le tutele del Codice del Consumo (D.Lgs. 206/2005), per quanto attiene alle pratiche commerciali ingannevoli ai sensi degli articoli 20 e seguenti[42] , nonché della normativa in materia di concorrenza sleale, prevista dall'art. 2598 c.c., o della responsabilità civile per danno reputazionale. Dall'altro, il nostro ordinamento prevede anche diverse ipotesi di potenziale rilevanza penale delle condotte di disinformazione.

Una delle ipotesi più frequenti è rappresentata

dalla diffamazione aggravata mediante mezzo di pubblicità, prevista dall'art. 595, comma 3, c.p. Tale fattispecie ricorre quando la diffusione di contenuti diffamatori avviene attraverso strumenti idonei a raggiungere un numero indeterminato di persone, categoria nella quale la giurisprudenza ha incluso anche le piattaforme digitali ed i social network[43].

In presenza di notizie false idonee a turbare l'ordine pubblico o a generare allarme nella collettività può inoltre configurarsi il reato di procurato allarme presso l'autorità, previsto dall'art. 658 c.p. Questa disposizione è stata talvolta richiamata in relazione alla diffusione di informazioni false riguardanti emergenze sanitarie, eventi catastrofici o situazioni di pericolo per la sicurezza pubblica.

La disinformazione può assumere rilievo penale anche in ambito economico-finanziario. L'articolo 185 del Testo Unico della Finanza (TUF) punisce, infatti, le condotte di manipolazione del mercato, che possono includere la diffusione di notizie false o fuorvianti idonee ad influenzare il prezzo degli strumenti finanziari.

Nei casi in cui le campagne di disinformazione siano realizzate mediante specifiche tecniche informatiche, possono essere integrate la truffa (art. 640 c.p.), la sostituzione di persona (art. 494 c.p.), frequentemente associata alla creazione di profili falsi sui social network, nonché i reati di accesso abusivo a sistemi informatici (art. 615 *ter* c.p.) ed interferenze illecite nelle comunicazioni digitali (art. 617 *quater* e *sexies* c.p.). Il reato di truffa, poi, potrebbe essere ipotizzato anche rispetto alle recensioni false online, quando tali condotte siano strumentali a trarre un ingiusto profitto.

Tuttavia, nessuna di queste fattispecie è in grado di garantire una tutela unitaria (quantomeno *ex post*) rispetto al

più vasto e sfaccettato fenomeno sinora descritto. La mancata introduzione di una specifica disciplina sanzionatoria è forse il sintomo di una scarsa percezione dell'effettiva portata lesiva di questo insieme di condotte, che di volta in volta vengono analizzate e raccontate singolarmente nella cronaca e nel dibattito pubblico.

Negli ultimi anni, inoltre, è iniziato un percorso di ripensamento della disciplina relativa alle recensioni online. In particolare, alcune proposte normative – tra cui il disegno di legge sulle piccole e medie imprese, definitivamente approvato dal Senato il 4 marzo 2026 ed in attesa di pubblicazione[44] – hanno previsto l'introduzione di specifici obblighi di trasparenza e tracciabilità delle recensioni digitali, nonché il riconoscimento di strumenti di tutela come il diritto di replica e l'autenticazione delle recensioni pubblicate sulle piattaforme.

7. L'evoluzione del quadro normativo europeo: Digital Services Act, Digital Markets Act, Omnibus Directive e AI Act

La crescente attenzione istituzionale verso il fenomeno della disinformazione riflette una consapevolezza ormai diffusa a livello europeo ed internazionale. Nel 2024 la vicepresidente della Commissione europea Věra Jourová ha definito la diffusione sistemica delle fake news come "un pericolo reale per la democrazia", sottolineando come la manipolazione informativa rappresenti una delle principali sfide per la tutela del pluralismo informativo e per il corretto funzionamento dei sistemi democratici contemporanei[45].

Negli ultimi anni l'Unione europea ha adottato una serie di strumenti regolatori che, pur perseguendo finalità differenti, concorrono a delineare un sistema normativo sempre più articolato di governo delle piattaforme digitali.

Uno dei pilastri di tale architettura è rappresentato dal Digital Services Act (DSA), Regolamento (UE) 2022/2065, entrato pienamente in vigore nel 2024. Il DSA introduce un insieme di obblighi progressivi per i fornitori di servizi digitali, con particolare attenzione alle Very Large Online Platforms (VLOPs) ed ai motori di ricerca di grandi dimensioni. Tra gli elementi più innovativi del regolamento figurano gli obblighi di trasparenza degli algoritmi di raccomandazione, i sistemi di segnalazione e rimozione dei contenuti illegali, nonché la previsione di valutazioni periodiche dei rischi sistemici connessi al funzionamento delle piattaforme. Tra tali rischi il regolamento include espressamente anche la diffusione di disinformazione e la manipolazione dei processi democratici[46].

Accanto al DSA si colloca il Digital Markets Act (DMA), Regolamento (UE) 2022/1925, che affronta il problema della concentrazione del potere economico nel settore delle piattaforme digitali. Il DMA introduce specifici obblighi per i cosiddetti gatekeeper, ossia le grandi piattaforme che controllano infrastrutture digitali essenziali per l'accesso ai mercati online. Sebbene il regolamento sia principalmente orientato alla tutela della concorrenza, esso incide indirettamente anche sulla circolazione delle informazioni nell'ambiente digitale, imponendo limiti alle pratiche di auto-preferenziazione e rafforzando l'interoperabilità tra servizi digitali[47].

Parallelamente, l'Unione europea ha rafforzato la tutela dei consumatori nel contesto del commercio digitale attraverso la Direttiva (UE) 2019/2161, nota come Omnibus Directive. Tale atto, nel modificare la disciplina delle pratiche commerciali scorrette e dei diritti dei consumatori, introduce specifici obblighi di trasparenza relativi alle recensioni online. In particolare, i professionisti che pubblicano recensioni devono indicare se e in che modo sia stato verificato che tali recensioni provengano da consumatori che hanno effettivamente

utilizzato o acquistato il prodotto o servizio recensito. La diffusione di recensioni false o manipolate può quindi integrare una pratica commerciale ingannevole, con conseguenti responsabilità giuridiche[48].

Più recentemente, il quadro regolatorio europeo si è ulteriormente ampliato con l'adozione del Regolamento sull'intelligenza artificiale (AI Act), approvato nel 2024. Pur non essendo specificamente dedicato al fenomeno della disinformazione, l'AI Act introduce importanti obblighi di trasparenza per i sistemi di intelligenza artificiale capaci di generare contenuti sintetici, tra cui immagini, audio e video realistici comunemente definiti deepfake. Il regolamento prevede, infatti, che tali contenuti debbano essere chiaramente identificabili come generati artificialmente, al fine di evitare che possano essere utilizzati per manipolare l'opinione pubblica o diffondere informazioni ingannevoli[49].

Nel loro insieme, questi strumenti normativi delineano una trasformazione significativa del diritto europeo delle piattaforme digitali. Se nella fase iniziale dello sviluppo di Internet la regolazione si era concentrata prevalentemente sulla responsabilità degli intermediari e sulla libertà di circolazione delle informazioni, il nuovo quadro normativo europeo appare sempre più orientato alla tutela dell'integrità dell'ecosistema informativo digitale.

In questa prospettiva, la regolazione delle piattaforme non riguarda più soltanto la rimozione dei contenuti illegali, ma si estende alla gestione dei rischi sistemici connessi al funzionamento delle infrastrutture digitali, tra cui la disinformazione, la manipolazione algoritmica e le distorsioni del mercato digitale. L'Unione europea sembra quindi perseguire un modello regolatorio che mira a bilanciare l'innovazione tecnologica con la protezione dei valori fondamentali dell'ordinamento europeo, tra cui la libertà di espressione, la tutela dei consumatori, la concorrenza leale ed il corretto funzionamento dei processi democratici.

8. Profili probatori e difficoltà applicative

Nonostante la presenza di strumenti normativi civili, amministrativi e penali idonei a reprimere alcune forme manipolazione informativa, l'effettività dell'azione giudiziaria spesso si ferma di fronte ad ostacoli strutturali di natura probatoria e applicativa.

Uno dei principali fattori critici, infatti, è rappresentato dall'anonimato o dalla pseudonimizzazione degli utenti online. Le piattaforme digitali consentono infatti la creazione di account attraverso identità non verificabili o facilmente falsificabili, rendendo più difficile individuare con certezza i soggetti responsabili della diffusione di contenuti illeciti. L'utilizzo di identità digitali fittizie, profili coordinati o account temporanei consente spesso agli autori delle campagne di disinformazione di operare senza lasciare tracce facilmente riconducibili alla propria identità reale[50].

Un ulteriore elemento di complessità è rappresentato dall'impiego crescente di bot e sistemi automatizzati di diffusione dei contenuti. In questi casi la ricostruzione delle responsabilità individuali diventa particolarmente complessa, poiché la diffusione dei contenuti avviene attraverso infrastrutture tecnologiche distribuite e talvolta difficilmente riconducibili a specifici soggetti giuridici[51].

Alle difficoltà tecniche si aggiungono, inoltre, problemi di natura territoriale e giurisdizionale. Le piattaforme digitali operano su scala globale ed i contenuti possono essere generati, ospitati e diffusi attraverso server collocati in Stati diversi. Questa dimensione transnazionale rende più complesso individuare la giurisdizione competente ed applicare efficacemente gli strumenti di cooperazione giudiziaria

internazionale, soprattutto nei casi in cui i soggetti coinvolti operino da Paesi con normative differenti o con limitati meccanismi di collaborazione giudiziaria[52].

9. Prospettive di riforma

L'analisi dei profili giuridici, tecnologici ed economici della disinformazione evidenzia come il contrasto a tali fenomeni non possa essere affidato esclusivamente agli strumenti repressivi tradizionali. La natura sistemica dell'ecosistema informativo digitale, caratterizzato da intermediazione algoritmica, globalità delle piattaforme e rapidità di diffusione dei contenuti, richiede un approccio strutturale e multilivello, capace di integrare strumenti normativi, tecnologici e culturali.

In questa prospettiva, il rafforzamento delle politiche di contrasto alla disinformazione deve necessariamente muoversi lungo diverse direttrici complementari.

Un primo ambito di intervento riguarda il rafforzamento degli obblighi di trasparenza delle piattaforme digitali. La crescente centralità degli algoritmi nella selezione e distribuzione dei contenuti rende infatti necessario sviluppare meccanismi più avanzati di trasparenza e responsabilità rispetto ai criteri che determinano la visibilità delle informazioni online. In questa direzione si collocano le iniziative delle normative europee che introducono obblighi di *algorithmic accountability*, imponendo alle grandi piattaforme digitali di valutare e mitigare i rischi sistemici connessi alla diffusione di contenuti disinformativi[53].

Un secondo profilo riguarda la maggiore affidabilità delle recensioni online, che rappresentano oggi uno degli strumenti più influenti nella formazione delle decisioni di consumo. L'introduzione di sistemi di Step-up Authentication o di altri meccanismi avanzati di verifica dell'identità degli utenti

potrebbe contribuire a ridurre la diffusione di recensioni false o manipolate e ciò consentirebbe di rafforzare il legame tra identità digitale e contenuto pubblicato, aumentando la tracciabilità delle interazioni online.

Accanto a tali strumenti, possono essere accompagnati anche sistemi di certificazione delle fonti informative, capaci di fornire agli utenti indicatori affidabili sull'origine e sull'affidabilità dei contenuti. Un ulteriore elemento strategico consiste nella promozione di sistemi indipendenti di verifica delle informazioni (*fact-checking*), che possano operare in modo trasparente e verificabile anche rispetto alle recensioni online ed ai contenuti generati dagli utenti. Tali meccanismi, se adeguatamente strutturati ed indipendenti, possono contribuire a rafforzare la qualità del dibattito pubblico e a ridurre la diffusione di informazioni false o fuorvianti[54].

Parallelamente, assume un ruolo fondamentale il rafforzamento delle politiche di educazione digitale e *media literacy*dei singoli cittadini. La capacità di riconoscere contenuti manipolati, verificare le fonti e comprendere il funzionamento degli algoritmi rappresenta, infatti, uno degli strumenti più efficaci di prevenzione della disinformazione. Numerose istituzioni internazionali hanno evidenziato come la promozione della cultura digitale costituisca un elemento essenziale per la resilienza delle società democratiche rispetto alle campagne di manipolazione informativa[55].

Infine, un ruolo centrale nelle prospettive di riforma è rappresentato dalla crescente attenzione verso i principi di accountability algoritmica. Gli algoritmi che regolano la distribuzione dei contenuti sulle piattaforme digitali influenzano in modo significativo la visibilità delle informazioni e la formazione delle opinioni pubbliche. La possibilità di sottoporre tali sistemi a forme di audit indipendente e di verifica regolatoria rappresenta uno degli strumenti più promettenti per garantire maggiore trasparenza e

responsabilità nell'ecosistema informativo digitale[56].

Più che la diffusione del falso in sé, preoccupa oggi l'inadeguatezza dell'ordinamento (italiano ed europeo) a governarlo in modo organico. Non esiste ancora una disciplina unitaria che sappia prevenire *ex ante* la manipolazione informativa e, insieme, contrastarne *ex post* gli effetti con strumenti efficaci e coerenti con la natura del fenomeno. Il sistema attuale resta frammentato, episodico, affidato a regole sparse e a rimedi spesso tardivi rispetto alla velocità con cui il contenuto ingannevole si propaga, si consolida e produce danno. Così, mentre la tecnologia industrializza l'inganno, il diritto continua troppo spesso a inseguirla senza comprenderla appieno. È proprio in questa asimmetria che si misura la debolezza della risposta pubblica: non manca solo una sanzione adeguata, manca una visione normativa complessiva. Finché tale lacuna persisterà, fake news e manipolazioni continueranno a prosperare nelle fenditure di un sistema che, più che prevenire il fenomeno, ne subisce l'evoluzione.

Riferimenti

[1] Reuters Institute, *Digital News Report 2025*, <https://www.digitalnewsreport.org>

[2] UNESCO – Ipsos, *Survey on the Impact of Online Disinformation and hate speech*, 2023, <https://www.unesco.org/sites/default/files/medias/fichiers/2023/11/unesco_ipsos_survey.pdf>

[3] European Commission, *Tackling Online Disinformation: A European Approach*, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236>

[4] https://europa.eu/eurobarometer/surveys/detail/3592

[5] <https://drive.google.com/file/d/1KwGqR4cB8Yg03YPbfSTXBIe-

U_EaJ2s6/view> e

<https://www.rapportogiovani.it/wp-content/uploads/2025/02/Alfabetizzazione-digitale-e-fake-news-4-compresso_1.pdf>

[6] Ital Communications – Censis, *Disinformazione e fake news in Italia*, 2023, <https://italcommunications.it/wp-content/uploads/2025/01/Rapporto-ItalCommunications-Censis-2023.pdf>;

[7] Infodata – Il Sole 24 Ore, <https://www.infodata.ilsole24ore.com/2025/12/11/per-tre-persone-su-dieci-la-disinformazione-non-e-un-grosso-problema-ed-e-questo-il-problema/?refresh_ce=1>

[8] European Parliament Research Service, *Artificial Intelligence and Disinformation*, <https://www.europarl.europa.eu>

[9] Agência Brasil, analisi sull'aumento di contenuti generati da IA nel 2026, <https://agenciabrasil.ebc.com.br>

[10] NewsGuard, *AI-Generated News Sites Report*, <https://www.newsguardtech.com>

[11] Luca M., Zervas G., *Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud*, Harvard Business School.

[12] OECD, *Online Consumer Reviews: The Case of Fake Reviews and Consumer Protection*, <https://www.oecd.org>

[13] Duivenvoorde, *Generative AI and the future of marketing: A consumer protection perspective*, Computer Law & Security Review 57, luglio 2025, <https://www.sciencedirect.com/science/article/pii/S2212473X25000148>

[14] EUvsDisinfo – Doppelgänger Campaign Analysis, <https://euvsdisinfo.eu/tag/doppelganger/?numberposts=20>

[15] ISPI, *Disinformazione e minacce ibride alle democrazie*, <https://www.ispionline.it>

[16] Investigative reports on disinformation campaigns in Moldova (2024–2025).

[17] IDMO – Microsoft Threat Analysis Center, *Generative AI and Information Manipulation*.

[18] S. Vosoughi, D. Roy, S. Aral, *The Spread of True and False News Online*, **Science**, 2018,

<https://www.science.org/doi/10.1126/science.aap9559>

[19] European Parliamentary Research Service, *Information manipulation in the age of generative artificial intelligence* (2025), <https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/779259/EPRS_BRI(2025)779259_EN.pdf>

[20] Chesney, D. Citron,*Deepfakes and the New Disinformation War*, Foreign Affairs (2019), <https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war>

[21] European Parliamentary Research Service, *Tackling deepfakes in European policy* (2021), <https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039>

[22] Parlamento europeo, Artificial Intelligence Act – normativa europea sull'intelligenza artificiale, <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

[23] UNESCO, *Journalism, Fake News and Disinformation. Handbook for Journalism Education and Training* (2018), <https://unesdoc.unesco.org/ark:/48223/pf0000265552>

[24] European Commission, *Tackling Online Disinformation: A*

European Approach (2018), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236>

[25] European Commission, *Code of Practice on Disinformation* (2022), <https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation>

[26] Cfr. EEAS, "Doppelganger strikes back: FIMI activities in the context of the EE24", 2024, <https://euvsdisinfo.eu/uploads/2024/06/EEAS-TechnicalReport-DoppelgangerEE24_June2024.pdf>; EDMO, "Disinformation about Israel/Hamas conflict flooded the EU in October", 2023, <https://edmo.eu/wp-content/uploads/2023/09/EDMO-29-Horizontal-FCB-updated.pdf>; European Commission, "Commission opens formal proceedings against X under the Digital Services Act", 18 dicembre 2023, <https://digital-strategy.ec.europa.eu/en/news/commission-opens-formal-proceedings-against-x-under-digital-services-act>; UNRWA, "The spread of misinformation & disinformation continues to be used as a weapon of war in Gaza", 2024, https://www.unrwa.org/newsroom/official-statements/spread-misinformation-disinformation-continues-be-used-weapon-war-gaza>.

[27] Cfr. OHCHR, "A/75/172: Disciplinary measures against judges and the use of "disguised" sanctions" (2020), <https://www.ohchr.org/en/documents/thematic-reports/a75172-disciplinary-measures-against-judges-and-use-disguised-sanctions>; v. anche ENCJ, "Statement by the Executive Board of the ENCJ On Pressure and Intimidation of Judges through Media", 17 dicembre 2025, <https://obt-jud.hu/en/statement-executive-board-encj-pressure-and-intimidation-judges-through-media>

[28] Corte EDU, Grande Camera, "Morice v. France", 23 aprile 2015, <https://hudoc.echr.coe.int/eng?i=001-154265>

[29] OKO.press, “Smear campaign coordinated by the Ministry of Justice, aimed to discredit Polish judges, discovered” (2019), <https://oko.press/why-did-the-polish-deputy-minister-of-justice-resign-everything-you-need-to-know-about-the-piebiak-scandal> e Supreme Court of the United States, “2024 Year-End Report on the Federal Judiciary”, <https://www.supremecourt.gov/publicinfo/year-end/2024year-endreport.pdf>

[30] Europol, “Facing reality? Law enforcement and the challenge of deepfakes”, <https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes>; v. anche Europol, “Europol report finds deepfake technology could become staple tool for organised crime”, 28 aprile 2022, <https://www.europol.europa.eu/media-press/newsroom/news/europol-report-finds-deepfake-technology-could-become-staple-tool-for-organised-crime>

[31] L’economia dell’attenzione e il paradosso che sta uccidendo i giornali, <https://www.feltrinellieducation.it/magazine/l-economia-dell-attenzione-e-il-paradosso-che-sta-uccidendo-i-giornali>

[32] Shoshana Zuboff, *The Age of Surveillance Capitalism*, PublicAffairs, 2019; v. anche Reuters Institute, *Digital News Report 2024*, <https://www.digitalnewsreport.org>

[33] Council of Europe, “Information Disorder: Toward an interdisciplinary framework for research and policy making”, <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html>); International IDEA – Clingendael Institute, “Protecting Politics: Deterring the Influence of Organized Crime on Elections”, <https://www.idea.int/publications/catalogue/protecting-politics-deterring-influence-organized-crime-elections>

[34] Council of Europe, “Information Disorder”, cit.; International IDEA – Clingendael Institute, “Protecting Politics”, cit., che descrive le modalità attraverso cui la criminalità organizzata interferisce nei processi elettorali.

[35] AGCOM, “News vs. Fake nel sistema dell’informazione”, <https://www.agcom.it/pubblicazioni/rapporti/news-vs-fake-nel-sistema-dellinformazione-interim-report-indagine>; Ministero dell’interno, “Atti intimidatori nei confronti degli amministratori locali”, <https://www.interno.gov.it/it/stampa-e-comunicazione/dati-e-statistiche/atti-intimidatori-nei-confronti-amministratori-locali> e “Report anno 2024”, <https://www.interno.gov.it/sites/default/files/2025-04/report_atti_intimidatori_amm_locali_anno_2024.pdf>; G. De Feo, G. De Luca, “Mafia in the Ballot Box”, in *American Economic Journal: Economic Policy*, https://www.aeaweb.org/articles?id=10.1257%2Fpol.20150551; A. Alesina, S. Piccolo, P. Pinotti, “Organized Crime, Violence, and Politics”, in *The Review of Economic Studies*, <https://academic.oup.com/restud/article/86/2/457/5060718>

[36] C. Wardle, H. Derakhshan, *Information Disorder: Toward an interdisciplinary framework for research and policy making*, Council of Europe (2017), <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html>

[37] OECD, *Online Consumer Reviews: The Case of Fake Reviews and Consumer Protection* (2019), <https://www.oecd.org/competition/consumer-policy/fake-reviews.htm>

[38] European Commission, *Behavioural Study on Misleading Online Reviews* (2020), <https://op.europa.eu/en/publication-detail/-/publication/9f9b6d2a-0f0e-11ea-8c1f-01aa75ed71a1>

[39] M. Luca, G. Zervas, *Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud*, Harvard Business School Working Paper (2016), <https://www.hbs.edu/faculty/Pages/item.aspx?num=51974>

[40] Direttiva (UE) 2019/2161 del Parlamento europeo e del Consiglio (Omnibus Directive), <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32019L2161>

[41] Regolamento (UE) 2022/2065 – Digital Services Act, <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32022R2065>

[42] Sulle quali è competente l'Autorità Garante della Concorrenza e del Mercato (AGCM). Parallelamente, le autorità indipendenti svolgono un ruolo sempre più rilevante nel contrasto alle pratiche di disinformazione online. In particolare, l'Autorità per le Garanzie nelle Comunicazioni (AGCOM) esercita funzioni di vigilanza in materia di pluralismo informativo e regolazione delle piattaforme digitali, intervenendo anche con linee guida e strumenti di monitoraggio relativi alla diffusione di contenuti disinformativi nello spazio digitale.

[43] Cass. Pen., Sez. V, 8 giugno 2015, n. 24431.

[44] Atto S. 1484-B definitivamente approvato il 4 marzo 2026 ed in attesa di pubblicazione, <https://www.senato.it/service/PDF/PDFServer/BGT/01496644.pdf> <https://www.senato.it/export/ddl/full/59139>

[45] Commissione europea – dichiarazioni della vicepresidente Věra Jourová sulla disinformazione, <https://digital-strategy.ec.europa.eu/it/news/vice-president-jourova-mission-canada-discuss-issues-related-disinformation-and-foreign>

[46] Regolamento (UE) 2022/2065 del Parlamento europeo e del

Consiglio – Digital Services Act, <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32022R2065>

[47] Regolamento (UE) 2022/1925 del Parlamento europeo e del Consiglio – Digital Markets Act, <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32022R1925>

[48] Direttiva (UE) 2019/2161 del Parlamento europeo e del Consiglio – Omnibus Directive, <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32019L2161>

[49] Parlamento europeo – Artificial Intelligence Act (AI Act), <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

[50] Council of Europe, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making* (Wardle, Derakhshan, 2017), <https://rm.coe.int/information-disorder-report/1680762772>

[51] Ferrara E. e al., *The Rise of Social Bots*, Communications of the ACM, 2016, <https://cacm.acm.org/research/the-rise-of-social-bots/>

[52] Council of Europe, Budapest Convention on Cybercrime, <https://www.coe.int/en/web/cybercrime/the-budapest-convention>

[53] Regolamento (UE) 2022/2065 – Digital Services Act, <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32022R2065>

[54] European Commission, Code of Practice on Disinformation (2022), <https://digital-strategy.ec.europa.eu/en/policies/code-practi

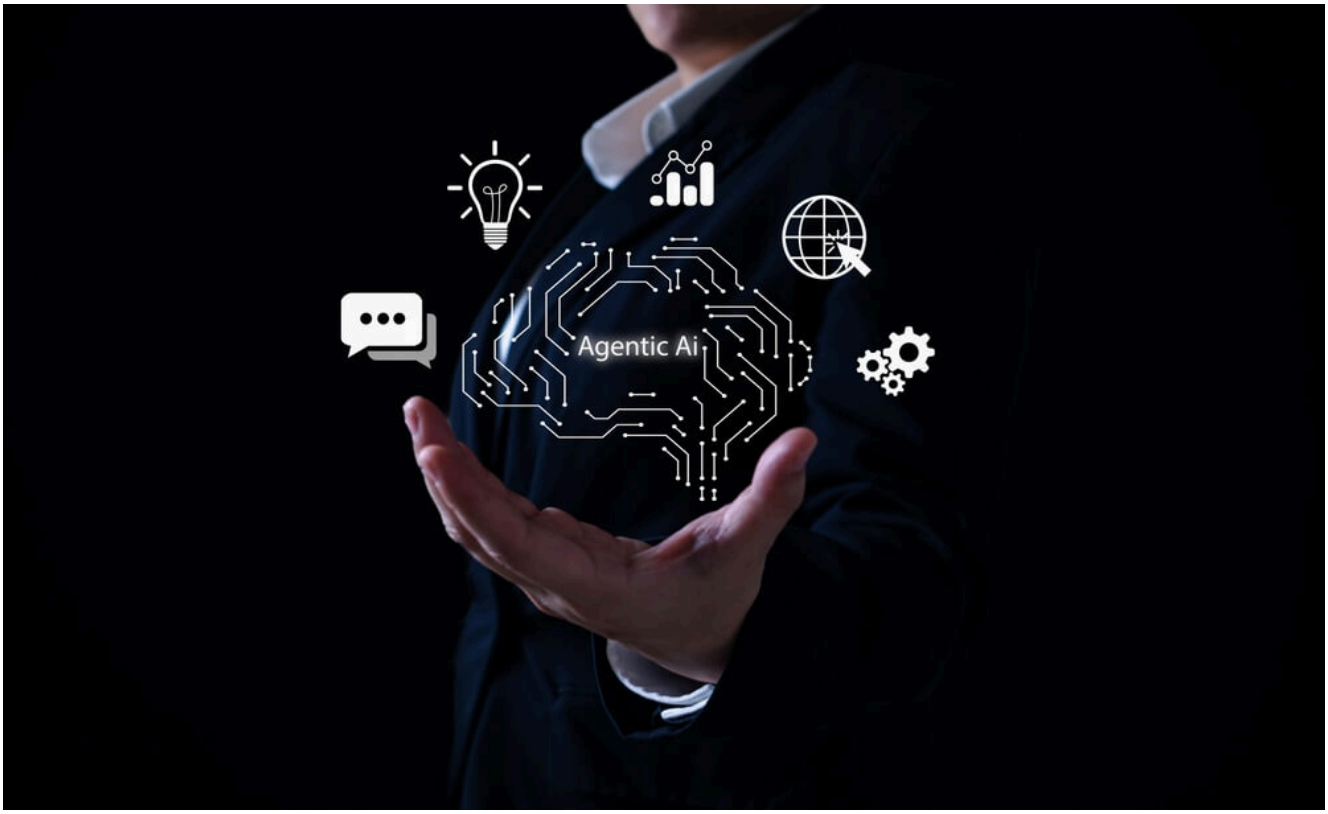
ce-disinformation>

[55] UNESCO, *Media and Information Literacy: Policy and Strategy Guidelines*, <https://www.unesco.org/en/media-information-literacy>

[56] European Parliamentary Research Service, *A governance framework for algorithmic accountability and transparency*, <https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262>

[57] S. Wells, Tutto ciò che sappiamo sull'Effetto Mandela: dai falsi ricordi all'intelligenza artificiale, National Geographic, 04.08.2025, <https://www.nationalgeographic.it/tutto-cio-che-sappiamo-sull-effetto-mandela-dai-falsi-ricordi-all-impatto-dell-intelligenza-artificiale>

Cognitive injection, la nuova frontiera della guerra psicologica AI



La cognitive injection trasforma la cybersecurity in uno scontro che prende di mira la percezione umana. Agenti AI autonomi, alimentati da OSINT e leak, non cercano più solo il click ma alterano convinzioni, decisioni e fiducia collettiva.

La verità sintetica sta divorando indagini e processo penale

Una volta si discuteva del rapporto epistemologico tra verità e verità processuale. Di questa se ne accettavano i limiti perché destinata a regolare funzionalmente i rapporti tra consociati nel bilanciamento tra certezza e ordine. La verità prodotta dall'intelligenza artificiale è costante falsificazione creativa e percettiva, che modifica anche la memoria autobiografica delle persone e la materialità dell' "oggettivo" e altera la valutazione della prova, rendendo ormai la categoria della verità processuale non più parte della realtà, della verità e della sua conoscibilità. Il leviatano è divenuto demiurgo. Proprio nel momento in cui la prova digitale diventa il baricentro, diretto o indiretto, di ogni indagine penale (perché l'uomo è parte di un tessuto digitale dalle maglie sempre più strette), i crimini legati all'AI generativa condizionano e subornano anche la prova analogica. Quello a cui assistiamo (per lo più inconsapevolmente) con i crimini legati alla AI e, in particolare deepfake, phishing, sextortion è solo sineddoche

della crisi della prova nell’era dei contenuti sintetici.

L’intelligenza artificiale generativa non ha soltanto ampliato la capacità di produrre contenuti, ma ha incrinato – forse definitivamente – uno dei presupposti più fragili della convivenza digitale: la fiducia nell’autenticità di ciò che vediamo, ascoltiamo e leggiamo. Deepfake, phishing evoluto, sextortion e pornografia artificiale mostrano come l’inganno non dipenda più soltanto dalla falsità del contenuto, ma dalla sua plausibilità, dalla sua rapidissima diffusione e dalla capacità di colpire emozioni, reputazione, patrimonio e autodeterminazione. In questo scenario, la minaccia non riguarda solo la commissione di nuovi reati o il potenziamento di quelli tradizionali, ma investe anche direttamente la tenuta della prova audiovisiva, la credibilità delle relazioni digitali, la tenuta del sistema processuale e l’effettività delle risposte sanzionatorie. L’AI generativa, infatti, non ha soltanto moltiplicato la capacità di produrre contenuti, ma li ha resi altamente credibili, difficilmente verificabili, praticamente gratuiti ed alla portata di tutti. Il punto non è più stabilire se l’intelligenza artificiale possa essere usata per commettere reati, ma comprendere come stia trasformando le modalità per commetterli.

La statistica dei crimini legati specificamente all’AI generativa è ancora incompleta, perché le principali banche dati classificano i fatti per tipologia – truffe, estorsioni, sfruttamento sessuale, manipolazioni informative – e non ancora per tecnologia impiegata. Eppure, i dati disponibili sono ormai troppo convergenti per essere liquidati come passeggeri segnali di cambiamento. Europol, nel SOCTA 2025, descrive l’AI come fattore di trasformazione della criminalità organizzata, un moltiplicatore strutturale capace di accrescerne efficienza, ordine di grandezza e adattabilità[1]. Nel rapporto Threat Landscape 2025, ENISA rileva che i cybercriminali sfruttano sempre più l’intelligenza artificiale per aumentare produttività e capacità operative, e conferma

che il phishing resta il mezzo di compromissione più diffuso, rappresentando circa il 60% dei casi osservati[2]. Nel 2024 Internet Crime Report, l'FBI ha registrato 859.532 denunce e 16,6 miliardi di dollari di perdite; il solo phishing tramite "spoofing"[3] (attacco informatico con uso di false identità) conta 193.407 segnalazioni, mentre il "Business Email Compromise" genera perdite per oltre 2,77 miliardi di dollari[4].

Il quadro si fa ancora più allarmante rispetto ai reati che colpiscono i minori. Il National Center for Missing & Exploited Children (NCMEC) segnala che nel 2024 la CyberTipline[5] ha ricevuto 20,5 milioni di report, corrispondenti a 29,2 milioni di incidenti distinti; le segnalazioni di "online enticement" (adescamento online), categoria che include la sextortion, superano le 546.000, con un incremento del 192% rispetto al 2023[6]. Nello stesso anno, le segnalazioni che coinvolgono l'uso di Generative AI sono cresciute del 1.325%, passando da 4.700 a 67.000[7]. Nel 2024, infine, l'Internet Watch Foundation ha individuato 245 report contenenti immagini di abuso sessuale su minori generate dall'AI, in aumento del 380% rispetto all'anno precedente[8].

1. La Gen AI: un'overview

L'intelligenza artificiale generativa (o generative AI) è una branca dell'AI che consente ai sistemi informatici di creare nuovi contenuti originali, come testi, immagini, musica, video o codice, partendo da dati di addestramento esistenti. A differenza dei modelli tradizionali di AI, progettati per classificare o riconoscere dati, i sistemi generativi sono in grado di produrre contenuti nuovi, che spesso risultano indistinguibili da quelli creati da esseri umani.

Il funzionamento dell'AI generativa si basa su reti neurali profonde (*deep neural networks*), in particolare su un'architettura chiamata trasformatore (*transformer*),

introdotta nel 2017 da un gruppo di ricercatori di Google[9]. Questa architettura è alla base di molti modelli di linguaggio avanzati come GPT (Generative Pre-trained Transformer) di OpenAI, PaLM di Google, Claude di Anthropic o LLaMA di Meta.

I modelli generativi sono solitamente pre-addestrati su enormi quantità di dati, spesso prelevati dal web, e successivamente ottimizzati per svolgere compiti specifici (ad esempio la scrittura automatica, la generazione di immagini o la creazione di codice). Durante l'addestramento, il modello apprende le strutture statistiche e semantiche del linguaggio o dei dati visivi, in modo da poterli riprodurre in maniera credibile.

Un esempio molto noto di AI generativa è ChatGPT, sviluppato da OpenAI, che può sostenere conversazioni complesse, scrivere saggi, rispondere a domande tecniche e persino imitare stili letterari. Altri strumenti come DALL-E o Midjourney generano immagini realistiche partendo da una semplice descrizione testuale (prompt), mentre altri, come Synthesia, permettono la creazione di video a partire da input linguistici.

Dal punto di vista tecnico, il cuore della generazione consiste nell'autocompletamento predittivo: il modello calcola la probabilità che una parola (o un pixel, o una nota musicale) segua un'altra, e genera sequenze coerenti sulla base di questo calcolo. Per i modelli linguistici, ad esempio, ciò avviene parola per parola o token per token, scegliendo quelli statisticamente più plausibili in relazione al contesto fornito dall'utente.

Per quanto riguarda i deepfake, questa tecnologia implica spesso l'uso di una "rete neurale" per l'apprendimento automatico[10]. La rete neurale inizia come una sorta di tabula rasa, caratterizzata da una rete nodale controllata da un insieme di parametri numerici impostati casualmente[11]. Proprio come l'esperienza affina i nodi neurali del cervello, gli esempi addestrano il sistema della rete neurale[12]. Se la

rete elabora un’ampia gamma di esempi di addestramento, è poi in grado di creare modelli sempre più accurati[13]. È attraverso questo processo che le reti neurali classificano audio, video o immagini e generano imitazioni o alterazioni realistiche.

Di per sé, l’emergere dell’apprendimento automatico tramite metodi basati su reti neurali preannuncia un aumento significativo della capacità di creare immagini, video e audio falsi. Ma la storia non finisce qui. Entrano in scena le “reti generative avversarie”, GAN (*Generative Adversarial Networks*)[14]. Queste ultime, inventate dal ricercatore di Google Ian Goodfellow, utilizzano simultaneamente due reti neurali: una, detta generatore, attinge da un insieme di dati per produrre un campione che imiti tali dati; l’altra, il discriminatore, valuta quanto il generatore abbia avuto successo. In modo iterativo, le valutazioni del discriminatore perfezionano il lavoro del generatore. Il risultato supera di gran lunga la velocità, la portata e la finezza che potrebbero ottenere i revisori umani, di talché la crescente sofisticazione dell’approccio GAN porterà inevitabilmente alla produzione di deepfake sempre più convincenti[15].

2. Dalla generazione del contenuto alla generazione dell’inganno

La GenAI è spesso descritta come una tecnologia “creativa”. Sul piano giuridico, però, la sua qualità più rilevante è un’altra: la capacità di produrre output plausibili, adattabili al contesto e difficili da distinguere da quelli umani. I modelli generativi contemporanei discendono dall’evoluzione delle architetture di apprendimento profondo, in particolare dai transformer, introdotti nel 2017, che hanno reso possibile trattare enormi sequenze di dati testuali, visivi e audiovisivi in modo molto più efficiente dei modelli precedenti[16]. L’OCSE ha osservato che la GenAI impone ai governi una sfida nuova: governare una tecnologia capace di

trasformare informazione, mercato e relazioni sociali senza sacrificare diritti fondamentali, affidabilità e fiducia pubblica[17].

Se in passato la manipolazione sofisticata di immagini, audio e video restava in gran parte confinata a soggetti altamente specializzati, oggi strumenti commerciali, spesso intuitivi e relativamente economici, consentono a una platea molto più ampia di produrre contenuti sintetici più o meno pericolosi. È qui che la questione smette di essere esclusivamente tecnologica ed entra nella vita reale generando non solo fattispecie di reato del tutto nuove, ma anche facendo acquisire una potenza offensive diversa a reati già noti quali frode, estorsione, diffamazione, sostituzione di persona e manipolazione probatoria, alterando i presupposti di credibilità dell'inganno[18].

Un esempio divenuto ormai paradigmatico è l'esperimento dell'Università di Washington sul volto di Barack Obama. Alcuni ricercatori, utilizzando uno strumento basato su rete neurale da loro creato, hanno mostrato come, disponendo di materiale audiovisivo sufficiente, fosse possibile produrre un video in cui l'ex presidente pronunciava frasi mai dette[19]. Quel caso aveva una funzione dimostrativa. Oggi, però, il problema non è più stabilire se ciò sia possibile, ma riconoscere che la medesima logica è ormai spendibile in frodi affettive, falsi comunicati, video-ricatti, vocal clone o pseudo-prove. In altre parole, la GenAI non genera soltanto contenuti: crea situazioni verosimili e credibili[20].

3. La criminalità persuasiva: phishing, voice cloning e frodi deepfake

It can happen on an international and serious organised crime scale, and it can happen in someone's bedroom … You can think of any crime type and put it through an AI lens and say: "What is the opportunity here?" (Alex Murray, UK National Crime

Agency)[21]

Il primo terreno sul quale la GenAI ha mostrato in modo plastico la propria capacità criminogena è quello della frode persuasiva. Il phishing è una forma di ingegneria sociale in cui l'autore del reato si finge un contatto affidabile e induce la vittima a condividere informazioni sensibili, come password o dati bancari[22]. Il phishing tradizionale era spesso riconoscibile per errori grammaticali, formule standardizzate, registri maldestri. La GenAI riduce proprio questi segnali: produce testi corretti, imita stili comunicativi, costruisce identità digitali verosimili, personalizza messaggi su destinatari specifici e imita stili di scrittura aggirando i controlli, sfruttando le vulnerabilità dei filtri e-mail e delle abitudini degli utenti. L'effetto non è solo quantitativo, ma qualitativo: il falso appare più credibile perché somiglia sempre di più alla comunicazione ordinaria[23].

L'Anti-Phishing Working Group (APWG) ha osservato 1.130.393 attacchi nel secondo trimestre del 2025, il valore trimestrale più alto dalla metà del 2023[24]; nel quarto trimestre, inoltre, i bersagli principali sono risultati i servizi di social media e SaaS/Webmail, ciascuno pari al 20,3% degli attacchi osservati, mentre lo smishing (phishing via SMS) ha continuato a crescere del 30-40% trimestralmente[25]. Inoltre, ENISA rileva che all'inizio del 2025 le campagne di phishing supportate dall'AI rappresentavano già oltre l'80% dell'attività di social engineering osservata[26]. Microsoft aggiunge che le e-mail di phishing automatizzate con AI raggiungono tassi di click del 54%, contro il 12% dei tentativi standard[27]. Nella stessa direzione va l'FBI, secondo cui nei primi sette mesi del 2025 l'AI compare in oltre 9.000 segnalazioni, distribuite tra diverse forme di frode online[28].

L'FBI ha avvertito che i criminali usano la GenAI per facilitare frodi finanziarie su scala più ampia, sfruttando

immagini artificiali, testi convincenti, audio clonati e video sintetici per rendere l'inganno più plausibile[29]. La FTC, dal canto suo, ha segnalato il rischio crescente dei sistemi di voice cloning, capaci di riprodurre la voce di un familiare o di un superiore gerarchico e di attivare una risposta emotiva immediata nella vittima[30]. L'elemento interessante, dal punto di vista giuridico, è che l'AI non interviene soltanto nell'atto finale della frode, ma già nella costruzione del contesto fiduciario: crea la relazione apparente che induce la vittima a credere, reagire d'impulso e abbassare le proprie difese.

Non sorprende, allora, che gli esempi più recenti siano sempre meno "artigianali" e sempre più organizzati. Nel febbraio 2025 la Hong Kong Police ha annunciato di avere smantellato due organizzazioni criminali che impiegavano deepfake e identità sintetiche per frodi online, con 58 arresti e perdite complessive stimate in circa 400 milioni di dollari di Hong Kong[31]. Non esistono però solo frodi economiche in senso classico. L'FBI ha diffuso nel 2025 un'allerta sulle forme di c.d. "virtual kidnapping" (rapimento virtuale) in cui i criminali usano immagini o video alterati come falsa "proof of life" per rendere più credibili richieste estorsive per rapimenti inesistenti[32]. Il salto qui è evidente: la GenAI rende più facile simulare non solo un messaggio, ma un'intera scena emotiva[33].

In questo senso, il phishing generativo rappresenta qualcosa di più di un semplice aggiornamento tecnologico dell'ingegneria sociale. È l'emersione di una criminalità persuasiva, nella quale il contenuto falso vale non tanto per ciò che dice, quanto per la sua capacità di inserirsi in una relazione già plausibile. È precisamente questo elemento che spiega perché la risposta del diritto non possa esaurirsi nella sola repressione del fatto consumato: quando l'inganno è automatizzato, personalizzato e scalabile, la prevenzione e la trasparenza diventano parte della risposta giuridica tanto

quanto la sanzione[34].

4. Sextortion, pornografia sintetica e sessualità artificiale non consensuale

Se la frode mostra il lato economicamente più visibile della GenAI, il suo impatto più disturbante emerge nei reati che colpiscono la sessualità e la vulnerabilità delle persone. La sextortion, già nota prima dell'esplosione dell'AI generativa, cambia natura quando l'autore non si limita a minacciare la diffusione di materiale reale, ma può creare o alterare contenuti sessualmente espliciti a partire da fotografie innocue, immagini reperite online o semplici dati identificativi. In questo caso, la minaccia non riguarda più soltanto ciò che esiste, ma ciò che può essere reso credibile in pochi minuti[35].

Questa trasformazione non si esaurisce, tuttavia, nella singola condotta estorsiva o nella manipolazione di un contenuto. Come ha osservato recentemente *Internazionale*[36], riprendendo un'analisi di *The Economist*, l'intelligenza artificiale sta rimodellando l'intera economia della pornografia digitale: dai generatori di immagini esplicite alle app di *nudify*, fino alle piattaforme che devono decidere se ammettere, monetizzare o almeno etichettare contenuti sintetici apparentemente autentici.

In questo contesto, sextortion, deepfake pornografici e sfruttamento dell'immagine altrui non appaiono più come deviazioni episodiche, ma come lo scenario pervasivo ed in crescita esponenziale che sta rimodellando i confini stessi della pornografia, della sua fruizione e dell'immaginario ad essa collegati, con inquietanti e potenziali normalizzazioni delle devianze illecite nella vita reale, soprattutto per quanto riguarda i minori.

I dati inglesi del NCMEC, ad esempio, confermano la portata del fenomeno. Nel 2024 i report di "online

enticement” (adescamento online) superano le 546.000 unità, con una crescita del 192% rispetto all’anno precedente; la stessa organizzazione riferisce di ricevere quasi 100 segnalazioni al giorno di “financial sextortion” (ovvero quelle condotte estorsive, perpetrate attraverso la rete, caratterizzate dalla minaccia di diffondere immagini o video sessualmente espliciti che ritraggono la vittima, al fine di ottenere qualcosa da quest’ultima), e di essere a conoscenza, dal 2021, di almeno 36 adolescenti che si sono tolti la vita dopo essere stati vittime di sextortion[37].

Il 2024, inoltre, si è concluso con alcuni eclatanti casi giudiziari in materia di intelligenza artificiale. In particolare, un ventisettenne di Bolton (UK), Hugh Nelson, è stato condannato a 18 anni di reclusione per aver creato con l’intelligenza artificiale immagini di abusi su bambini, utilizzando come base delle foto di minori realmente esistenti[38]. Si è trattato del primo processo di questo genere in Inghilterra.

Nelson aveva utilizzato Daz 3D, un programma per computer con una funzione di intelligenza artificiale, per trasformare le immagini “normali” in immagini di abusi. In alcuni casi, i pedofili avevano commissionato le immagini, fornendo direttamente fotografie di minori con cui avevano avuto contatti nella vita reale.

Dalla vendita delle immagini su diverse chat online ha guadagnato circa cinquemila sterline durante un periodo di 18 mesi.

Sebbene ci siano state precedenti condanne per deepfake (nei quali spesso si sovrappone un volto alle azioni commesse da un altro corpo) la novità di questo caso è rappresentata proprio dalla creazione integrale di personaggi 3D da fotografie innocue[39].

In questa direzione l’Internet Watch Foundation, nel 2024 ha

segnalato un aumento del 380% dei report relativi a immagini di abuso sessuale su minori generate dall'AI[40].

Il danno non consiste soltanto nell'eventuale acquisizione abusiva di immagini autentiche, ma nella costruzione artificiale di una sessualità artefatta e non voluta, nella mercificazione del corpo rappresentato, nella minaccia reputazionale e nella capacità del contenuto artificiale di circolare come se fosse vero. È per questo che il dibattito normativo si sta spostando verso categorie più ampie, come quella dei contenuti intimi non consensuali, idonee a ricomprendere anche immagini e video generati o alterati artificialmente[41].

La crescita del ricorso a immagini generate dall'AI per soddisfare il mercato nero degli abusanti porta con sé una sfida alla tenuta delle norme in materia, concepite in tutt'altra epoca. Assumono sempre maggiore rilevanza, dunque, quelle pronunce della Corte di Cassazione italiana che, già da tempo, hanno iniziato a qualificare come materiale pedopornografico fumetti e illustrazioni di racconti erotici raffiguranti minorenni, ritenendo di dover includere in tale nozione tutto ciò che sia idoneo a dare allo spettatore l'idea che oggetto della rappresentazione pornografica sia un minore[42].

5. Le difficoltà del riconoscimento

Le diverse tipologie di deepfake, da quelle più semplici, come il face-swapping, fino al lip-syncing e al puppet-mastery (rispettivamente lo scambio di volti, la sincronizzazione di un video di una persona che parla con un diverso audio, l'animazione di un soggetto inesistente), possono essere tanto complesse da realizzare in maniera realistica quanto poi ostiche da distinguere da immagini reali.

Sebbene nel tempo, di pari passo con l'evoluzione della tecnologia generativa, siano stati compiuti notevoli progressi

nelle prestazioni dei rilevatori di deepfake, esistono tuttora notevoli limiti che impediscono una completa ed efficace individuazione.

In primo luogo, come sempre avviene per le tecnologie AI che necessitano di apprendere per migliorare, l'accessibilità a grandi database di deepfake rappresenta un fattore determinante per lo sviluppo di tecniche di rilevamento efficaci. Tuttavia, analizzando la qualità dei video presenti in questi dataset emergono diverse ambiguità rispetto ai contenuti manipolati realmente presenti su Internet. Le più comuni tipologie in questi database sono lo sfarfallio temporale in alcuni momenti durante il discorso, la sfocatura attorno alle regioni facciali, l'eccessiva levigatezza nella texture del viso o la mancanza di dettagli, l'assenza di movimenti o rotazioni della testa, l'assenza di oggetti che ostruiscono il volto (occhiali, effetti di luce e simili), la sensibilità alle variazioni nella postura o nello sguardo, ad esempio con incoerenze nel colore della pelle. Le ambiguità citate derivano da imperfezioni nei passaggi delle tecniche di manipolazione. Inoltre, contenuti di bassa qualità difficilmente risultano convincenti o in grado di creare una reale impressione. Pertanto, anche se i metodi di rilevamento si dimostrano efficaci su tali video, non è garantito che mantengano le stesse prestazioni in situazioni reali, con foto e video creati con tecniche più avanzate.

Inoltre, i metodi di rilevamento dei deepfake sono formulati come un problema di classificazione binaria, in cui ogni campione è etichettato come vero o falso. Questo tipo di classificazione è più semplice in ambienti controllati, dove si sviluppano e testano le tecniche utilizzando contenuti audio-visivi di cui è nota l'origine – originali o falsificati. Tuttavia, nell'applicazione sul campo, i video possono essere alterati in modi diversi dai deepfake, per cui un contenuto non rilevato come manipolato non è necessariamente autentico (es. Photoshop). Inoltre, un

contenuto deepfake può presentare più tipi di alterazione, sia audio che visiva, rendendo poco accurata una singola etichetta. In contenuti con più volti, di solito uno o più di essi vengono manipolati con deepfake solo in alcune sequenze. Pertanto, lo schema di classificazione binaria dovrebbe essere ampliato a classificazioni multicategoria/multi-etichetta e a una classificazione/rilevamento locale a livello di fotogramma, per rendere più efficace l'analisi.

I metodi più elementari di rilevamento dei deepfake, dunque, sono generalmente progettati per analisi in batch su grandi dataset. Tuttavia, quando tali tecniche vengono impiegate da giornalisti o forze dell'ordine, spesso è disponibile solo un numero limitato di video. Inoltre, un punteggio numerico che rappresenta la probabilità che un contenuto sia reale o falso ha scarso valore se non accompagnato da una spiegazione che giustifichi tale punteggio, ma la maggior parte dei metodi di rilevamento, soprattutto quelli basati su tecniche di deep learning, non offre tali spiegazioni a causa della loro natura di "black-box", cioè di scatola nera di cui – dall'esterno – non si può conoscere il funzionamento[43].

6. Deepfake e giustizia: la crisi della prova audiovisiva

Una riflessione giuridica completa sul tema deve necessariamente confrontarsi con gli impatti dell'AI Generativa sul rapporto tra verità e accertamento processuale: gli effetti dei deepfake, infatti, non sono solo limitati alla diffusione del falso, ma sono tali da minare la fiducia nel vero. La prova audiovisiva ha tradizionalmente beneficiato, in molti contesti, di una sorta di affidamento pratico: un video "mostra", una fotografia "documenta", un audio "registra". I deepfake stravolgono questi assunti e incrinano proprio questa fiducia di base, dando luogo ad una delle più grandi minacce della società contemporanea[44].

La difficoltà, però, non è solo culturale ma anche tecnica. I limiti strutturali dei sistemi di rilevamento sono noti: dataset non pienamente rappresentativi, classificazione binaria troppo rigida, scarsa trasparenza dei modelli, vulnerabilità a *bias* di razza e genere, difficoltà a trattare contenuti multimodali e alterazioni locali. La valutazione forense dei deepfake, peraltro, si deve oggi confrontare con l'utilizzo di tecniche di *anti-forensics*[45], utilizzate proprio per rendere difficile la ricostruzione forense e volte a nascondere gli elementi utili all'accertamento dei falsi. Ad esempio, uno strumento di rilevazione di contenuti AI può rivelarsi molto meno affidabile quando gli stessi sono stati caricati, compressi, riconvertiti e diffusi attraverso piattaforme social[46].

Le piattaforme social come Twitter, Facebook o Instagram, che sono i principali canali online utilizzati per diffondere contenuti audiovisivi, per risparmiare banda di rete privano tali contenuti dei metadati, li sottopongono a *downsampling* (processo di riduzione della definizione) e li comprimono in modo significativo prima del caricamento. Queste manipolazioni, comunemente conosciute come *social media laundering*, eliminano indizi relativi a eventuali falsificazioni sottostanti e aumentano il tasso di falsi positivi nei rilevamenti. I metodi di rilevamento basati su punti chiave a livello di segnale sono particolarmente vulnerabili a questo tipo di distorsione. Una misura efficace per migliorare l'accuratezza dei rilevatori su contenuti "ripuliti" dai social media è includere tali effetti nei dati di addestramento, ampliando al contempo i database di valutazione con contenuti visivi manipolati tramite social network.

7. L'ingresso dei deepfake nella giustizia

Tra le preoccupazioni per l'utilizzo malevolo dei deepfake

spicca il possibile ruolo di questa tecnologia per la manipolazione della formazione della prova nei processi, penali e non solo. Già nel 2021, in uno studio della Commissione europea si osservava come i deepfake sollevassero "*serie preoccupazioni riguardo alla fondamentale credibilità e ammissibilità delle registrazioni audiovisive come prove elettroniche nei tribunali*"[47].

Ed infatti, la prima conseguenza nell'opinione pubblica del diffuso utilizzo dei deepfake è l'erosione della fiducia nelle istituzioni e nel concetto di verità, con un danno evidente alla credibilità del sistema giudiziario.

Questa tecnologia, infatti, costringe lo spettatore a rivalutare il livello di fiducia riposto nei video e nelle immagini con cui interagisce, in particolare in ambito giudiziario, proprio perché possono apparire estremamente realistici ma essere in realtà del tutto falsi; ed anche quando lo spettatore è consapevole della falsità del contenuto, la percezione può comunque influenzare il subconscio. La loro mera esistenza mina la fiducia "intrinseca" che viene normalmente attribuita a prove video e fotografiche[48].

Nella maggior parte degli ordinamenti, inoltre, il vuoto normativo derivante dalla novità della materia rende difficile scegliere in quale fattispecie penale inquadrare l'uso dei deepfake. All'estero, si è suggerito di ricorrere alle discipline – civili e penali – riguardanti il diritto d'autore[49], che tuttavia è spesso limitante. Sarebbe preferibile, al contrario, l'introduzione diffusa di nuove discipline specifiche, sulla scorta (in UE) della strada già tracciata dall'AI Act, come si vedrà più avanti[50].

L'aspetto più critico, in ogni caso, riguarda il valore che normalmente viene attribuito a prove documentali quali immagini e video[51], spesso ammessi senza alcun genere di accertamento informatico sulla loro genuinità, provenienza e

originalità. Talvolta, l'unica garanzia sulla provenienza è la mera integrità della catena di custodia dei file e dei dispositivi in cui sono contenuti, che tuttavia non danno alcuna garanzia sulla genuinità del contenuto *ab origine*[52]. Questa fiducia "intrinseca" è ulteriormente dimostrata dal fatto che – nella maggior parte degli ordinamenti – questi accertamenti sono compiuti fuori dall'aula, prima del dibattimento, solitamente durante le indagini[53].

È qui che la tecnologia produce il suo effetto forse più profondo sul diritto: la crisi della prova audiovisiva non consiste soltanto nel rischio che un falso entri nel processo, ma nel fatto che anche il vero smette di apparire immediatamente affidabile. Il danno, in altri termini, non è solo epistemico, ma istituzionale[54].

Consideriamo, ad esempio, uno dei mezzi di prova ritenuto di maggior rilevanza nell'ambito delle indagini e dei processi: le intercettazioni, a prescindere che siano telefoniche "tradizionali", ambientali o mediante captatori informatici. Allo stato, la veridicità e l'affidabilità tecnica sono valutate tramite la correttezza del metodo di utilizzo dello strumento, ma non valutano l'originalità del contenuto dal punto di vista tecnico-informatico. Aggiungiamo a questa carenza la progressiva diffusione dei *deepfake* vocali sempre più realistici e meno riconoscibili: il risultato è un dialogo correttamente acquisito e valutato sul piano procedimentale, ma che può avere come sottostante due o più voci di cui alcune sono prodotte dall'intelligenza artificiale. L'intercettazione, quindi, rischia di attribuire certezza – con un metodo formalmente corretto – ad elementi che in realtà sono frutto di una attività artefatta.

8. L'AI Act: non un codice penale dell'AI, ma un perimetro europeo

dell’illiceità

In questo scenario, il diritto europeo e quello italiano stanno iniziando a reagire non solo con categorie classiche, ma anche con obblighi di trasparenza, strumenti di rimozione, poteri di vigilanza e nuove fattispecie mirate[55].

L’AI Act attua una governance del rischio, della trasparenza e del controllo della filiera tecnica. Proprio per questo, tuttavia, esso è centrale anche per chi si occupa di crimini legati all’intelligenza artificiale: perché definisce il perimetro europeo di ciò che il legislatore considera non tollerabile o, quantomeno, bisognoso di presidio rafforzato[56].

Da un lato, l’art. 5 vieta pratiche manipolative, sfruttatrici o incompatibili con i diritti fondamentali, compresi alcuni usi particolarmente invasivi o lesivi della dignità e dell’autonomia individuale[57]. Il suo valore è soprattutto sistematico: chiarisce che l’Unione non guarda all’AI come a uno strumento neutro, ma come a una tecnologia capace di amplificare inganno, dominio e sfruttamento.

Dall’altro lato, l’art. 50 è la disposizione più direttamente collegata ai deepfake e ai contenuti sintetici. Il regolamento impone obblighi di trasparenza sui contenuti generati o manipolati artificialmente, chiedendo che deepfake e taluni contenuti testuali diffusi al pubblico siano dichiarati come tali, salvo eccezioni specifiche, ad esempio per attività di prevenzione o repressione dei reati. Il significato culturale e giuridico della norma è notevole: l’Unione non vieta in sé il contenuto sintetico, ma ritiene inaccettabile che esso possa circolare come autentico senza che il destinatario sia posto in grado di riconoscerne la natura artificiale[58].

Su questo sfondo acquista rilievo anche il dibattito, oggi aperto in sede europea, sull’art. 4 dell’AI Act in materia di AI Literacy: “*I fornitori e i deployer dei sistemi di IA*

adottano misure per garantire nella misura del possibile un livello sufficiente di alfabetizzazione in materia di IA del loro personale nonché di qualsiasi altra persona che si occupa del funzionamento e dell'utilizzo dei sistemi di IA per loro conto, prendendo in considerazione le loro conoscenze tecniche, la loro esperienza, istruzione e formazione, nonché il contesto in cui i sistemi di IA devono essere utilizzati, e tenendo conto delle persone o dei gruppi di persone su cui i sistemi di IA devono essere utilizzati". Nei fenomeni qui considerati – phishing generativo, deepfake, sextortion e pseudo-prove – la vulnerabilità dell'ordinamento nasce infatti anche da un'asimmetria cognitiva: tra chi, da un lato, genera i contenuti e chi, dall'altro, li riceve, li diffonde o è chiamato a valutarli. È perciò significativo che, nel confronto avviato con il c.d. Pacchetto Digital Omnibus[59], si discuta se mantenere in capo a fornitori e deployer un obbligo attivo di adozione di misure oppure degradarlo a semplice incoraggiamento[60].

Ancora più rilevante è il disposto degli artt. 53 e 55, dedicato ai modelli di AI per finalità generali e a quelli con rischio sistemico. Qui il Regolamento esce dalla logica del singolo utilizzatore malevolo e guarda alla responsabilità organizzativa di chi sviluppa e mette in circolazione modelli potenti: documentazione tecnica, informazioni per chi integra il modello, policy sul copyright, sintesi pubblica dei dati di training, valutazione e mitigazione dei rischi sistemici, reporting degli incidenti e obblighi di cybersicurezza[61]. In altri termini, l'AI Act sposta l'attenzione a monte del fatto criminale, verso la progettazione e la distribuzione dell'infrastruttura che rende possibili determinate forme di offesa[62].

9. Il DSA: dalla trasparenza del contenuto alla responsabilità della

piattaforma

Se l’AI Act presidia soprattutto la trasparenza del contenuto non autentico e la governance della filiera tecnica, il Digital Services Act[63] interviene sul diverso piano della circolazione dei contenuti illeciti sulle piattaforme. Ed è proprio questo il profilo che casi recenti, come quello di Grok integrato in X, rendono particolarmente evidente: il danno non deriva soltanto dalla generazione del deepfake, ma dalla sua immediata pubblicazione, dalla visibilità algoritmica e dalla rapidità della sua propagazione. In tale prospettiva assumono rilievo il meccanismo di “notice and action” di cui all’art. 16 DSA, gli ordini di rimozione previsti dall’art. 9, il ruolo dei “trusted flaggers” ex art. 22 e, per le Very Large Online Platforms, gli obblighi di valutazione e mitigazione dei rischi sistemici di cui agli artt. 34 e 35. Ne consegue che, nel contesto dei deepfake sessuali e delle immagini intime non consensuali, la tutela non può essere affidata soltanto alla sanzione penale successiva o agli obblighi di trasparenza del provider del modello, ma deve articolarsi anche in procedure rapide di segnalazione, rimozione, disabilitazione dell’accesso e contenimento della diffusione.

10. La legge italiana n. 132/2025: dal raccordo regolatorio alla risposta penale

In Italia la legge 23 settembre 2025, n. 132 non si limita a “seguire” l’AI Act. Lo assume come cornice, ma compie un passo ulteriore, innestando nel diritto interno princìpi, autorità, deleghe e alcune scelte penalistiche di forte rilievo. L’art. 3 stabilisce che sviluppo e utilizzo di sistemi e modelli di AI per finalità generali devono avvenire nel rispetto di diritti fondamentali, trasparenza, proporzionalità, sicurezza, protezione dei dati, accuratezza, non discriminazione, sorveglianza e intervento umano; lo stesso articolo richiede, inoltre, che l’uso dell’AI non pregiudichi il dibattito

democratico con interferenze illecite e che sia assicurata la cybersicurezza lungo tutto il ciclo di vita dei sistemi e dei modelli[64]. L'art. 4 aggiunge, per il settore dell'informazione, i princìpi di obiettività, completezza, imparzialità e lealtà, oltre alla trasparenza del trattamento dei dati personali[65]. Sono disposizioni che, lette insieme, forniscono già una chiave interpretativa robusta per fenomeni come deepfake politici, manipolazione informativa e impersonificazione.

L'art. 15 interviene poi direttamente sull'attività giudiziaria, stabilendo che, anche nei casi di impiego di sistemi di AI, resta sempre riservata al magistrato ogni decisione sull'interpretazione e applicazione della legge, sulla valutazione dei fatti e delle prove e sull'adozione dei provvedimenti[66]. È una clausola di valore non solo simbolico, ma sistematico: riconosce che l'AI può entrare nell'organizzazione della giustizia, ma non può sostituire il giudizio umano nel punto più delicato, quello dell'accertamento. Gli artt. 20 e 24, poi, designano l'Agenzia per l'Italia Digitale (AgID) e l'Agenzia per la Cybersicurezza Nazionale (ACN) quali autorità nazionali per l'intelligenza artificiale, attribuendo a queste ultime specifici poteri di vigilanza, ispezione e sanzione, nonché il compito di raccordare l'ordinamento nazionale con il sistema europeo, compresa l'applicazione delle misure e delle sanzioni amministrative richiamate dall'art. 99 AI Act[67]. Proprio l'art. 24, inoltre, delega il Governo a disciplinare i casi di realizzazione e impiego illeciti di sistemi di AI, prevedendo strumenti anche cautelari per inibire la diffusione e rimuovere contenuti generati illecitamente, nonché la possibilità di introdurre autonome fattispecie di reato[68].

Ma il dato forse più significativo è che la legge italiana non si ferma alla delega futura. Con l'art. 26 interviene già sul codice penale, introducendo (all'art. 61 n. 11-undecies)

un’aggravante comune per i reati commessi mediante sistemi di AI quando questi abbiano costituito mezzo insidioso o abbiano ostacolato la difesa, e soprattutto inserendo il nuovo art. 612-quater c.p., rubricato “Illecita diffusione di contenuti generati o alterati con sistemi di intelligenza artificiale”[69]. La fattispecie punisce chi cagiona un danno ingiusto diffondendo, senza consenso, immagini, video o voci falsificati o alterati mediante AI e idonei a indurre in inganno sulla loro genuinità. Si tratta di un passaggio di grande rilievo: il legislatore italiano riconosce, per la prima volta in modo esplicito, che il danno da contenuto sintetico non è solo una variante tecnologica della menzogna, ma un’offesa autonoma alla persona, alla reputazione e all’autodeterminazione.

La nuova fattispecie penale, peraltro, non esaurisce il quadro dei rimedi. Restano infatti applicabili, ricorrendone i presupposti, anche fattispecie quali diffamazione, sostituzione di persona e trattamento illecito di dati; ma, soprattutto, la protezione effettiva della vittima dipende dalla capacità di coordinare la risposta penale con gli strumenti europei di rimozione e contenimento della circolazione. Nei casi in cui immagini, video o voci manipolati siano veicolati tramite grandi piattaforme online, la tutela non coinvolge soltanto l’accertamento della responsabilità dell’autore materiale, bensì anche la tempestiva attivazione di procedure di segnalazione, disabilitazione dell’accesso e mitigazione del rischio sistemico. La risposta ai deepfake sessuali appare così non soltanto repressiva ma sempre più multilivello: penale, regolatoria e procedurale, distribuita tra autore dell’offesa, fornitore del sistema e piattaforma che ne rende possibile la propagazione.

Tuttavia, l’intervento normativo appare ancora parziale e con ampi margini di miglioramento. Infatti, si limita a considerare le condotte che implicano la creazione e la

diffusione di materiale prodotto con intelligenza artificiale, senza contemplare le ulteriori possibilità che quest'ultima offre. Infatti, ulteriori condotte criminali potrebbero essere integrate dal mero uso dell'AI quale strumento di ricerca, da un lato, poiché consente l'elaborazione di enormi quantità di dati a fini di catalogazione, anche di contenuti in rete con modalità che esondano dal lecito (ad esempio, le capacità di profilazione di utenti dei social per determinare la vulnerabilità alle truffe). Dall'altro, non viene presa in considerazione la possibilità che le AI di pubblico accesso vengano addestrate dagli utenti con informazioni false, al fine di condizionarne le risposte ai futuri fruitori.

Con la GenAI, dunque, non esiste più solo un discrimine tra vero e falso, ma tra contenuti credibili e contenuti verificabili. La frode si fa più persuasiva, la sextortion più devastante, la prova audiovisiva più fragile. In questo scenario, l'AI Act costruisce il primo grande perimetro europeo di trasparenza, governance e controllo dei modelli.

Il diritto, tuttavia, non può limitarsi a inseguire la tecnologia, soprattutto quando la sua evoluzione segue accelerazioni esponenziali. Deve imparare a prevenire l'inganno dove oggi davvero si forma: nella progettazione del contenuto, nella sua diffusione, nella capacità di sembrare autentico e nell'idoneità a produrre danno prima ancora che sia possibile smentirlo. È qui che si pone la vera sfida dei crimini legati all'AI generativa: non nel fascino del falso perfetto, ma nella tenuta dell'ordinamento davanti a contenuti che, sempre più spesso, chiedono di essere creduti prima ancora di essere verificati.