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.