Il nuovo reato di deep fake: riflessioni tecnico – giuridiche sull'articolo 612 - quater del codice penale

*di Enrica Priolo**

Sommario: 1. Inquadramento normativo e dimensione socio-tecnologica del fenomeno – 2. Profili sostanziali – 3. Criticità tecnico-informatiche nella definizione normativa e problemi di determinatezza – 4. Profili di accertamento forense digitale dei deep fake – 5. La rapidità dell'obsolescenza tecnologica e l'adattabilità normativa – 6. Considerazioni conclusive

1. Inquadramento normativo e dimensione socio-tecnologica del fenomeno

L'art. **26 della Legge 23 settembre 2025, n. 132** (per una panoramica più ampia sulla nuova legge in materia, si rimanda a "*Approvata la legge italiana sull'IA. Verso un primato europeo*?") ha introdotto nel codice penale l'art. 612-quater, rubricato "*Illecita diffusione di contenuti generati o alterati con sistemi di intelligenza artificiale*". La fattispecie incriminatrice sanziona con la reclusione da uno a cinque anni chiunque cagioni un danno ingiusto ad una persona, cedendo, pubblicando o altrimenti diffondendo, senza il suo consenso, immagini, video o voci falsificati o alterati mediante l'impiego di sistemi di intelligenza artificiale e idonei a indurre in inganno sulla loro genuinità.

L'intervento normativo si colloca nell'alveo dell'adeguamento al **Regolamento (UE) 2024/1689** (cd. *AI Act*), che ha istituito un sistema di governance multilivello dell'intelligenza artificiale basato su una classificazione del rischio. Tuttavia, a differenza dell'approccio regolatorio dell'AI Act, che privilegia misure preventive e obblighi di conformità per gli operatori economici, la norma italiana introduce una tutela di tipo repressivo-sanzionatorio, tipica del diritto penale.

La collocazione sistematica della fattispecie tra i delitti contro la persona e, specificamente, contro la libertà morale, successiva agli atti persecutori (art. 612-bis c.p.) e al revenge porn (art. 612-ter c.p.), evidenzia la natura del bene giuridico tutelato che attiene alla sfera più intima della personalità individuale, alla libertà di autodeterminazione e al pieno svolgimento della personalità umana ex art. 2 Cost. Tale scelta topografica non è dunque casuale: il legislatore riconosce che la manipolazione digitale dell'identità personale mediante sistemi di IA generativa rappresenta una forma di violenza psicologica e di lesione dell'integrità morale comparabile, per gravità, alle condotte persecutorie e alle violazioni della privacy sessuale.

La ratio della criminalizzazione si fonda, in sintesi, su tre pilastri concettuali:

- l'esigenza di tutela anticipata rispetto alle tradizionali fattispecie di evento, attesa l'elevatissima idoneità lesiva dei contenuti manipolati mediante IA;
- la necessità di sanzionare specificamente la dimensione tecnologica della condotta, riconoscendo l'IA quale moltiplicatore di offensività;
- la valorizzazione del consenso quale cardine della disponibilità della propria immagine digitale, secondo una logica di autodeterminazione informativa coerente con la normativa sulla protezione dei dati personali[1].

*Borsista di ricerca in SSD GIUR-17/A presso l'Università degli Studi di Cagliari, Dipartimento di Ingegneria Elettrica ed Elettronica (DIEE).

[1] Sul nesso tra autodeterminazione informativa e tutela dell'immagine digitale, PIZZETTI, *Privacy e il diritto europeo alla protezione dei dati personali*, Torino, 2016, vol. I; RODOTÀ, *Il mondo nella rete. Quali i diritti, quali i vincoli*, Bari, 2014.

Peraltro, è opportuno precisare che prima dell'entrata in vigore della legge 132/2025, l'ordinamento presentava una lacuna di tutela difficilmente colmabile mediante il ricorso alle fattispecie incriminatrici tradizionali, ad eccezione del caso di *revenge porn* a contenuto sessuale. Le **ipotesi astrattamente applicabili** - diffamazione (artt. 594-595 c.p.), sostituzione di persona (art. 494 c.p.), trattamento illecito di dati personali (art. 167-bis cod. privacy) - risultavano **inadeguate** a cogliere la peculiare offensività delle condotte realizzate mediante *deep fake*.

La diffamazione, infatti, postula la lesione della reputazione attraverso comunicazioni di fatti specifici, mentre il *deep fake* opera mediante la creazione di una realtà parallela apparentemente autentica; la sostituzione di persona richiede l'attribuzione a sé o ad altri di un'identità altrui con finalità specifiche, mentre il *deep fake* può limitarsi a manipolare l'immagine altrui senza assumerne l'identità; il trattamento illecito di dati personali, infine, presuppone operazioni su dati esistenti, non la creazione ex novo di contenuti sintetici[2].

Sebbene il legislatore nazionale abbia introdotto una peculiare fattispecie delittuosa all'articolo 612 – quater c.p., sul piano definitorio risulta ancora del tutto **assente**, invece, una **definizione ufficiale** di deep fake. Il primo riferimento risale al 2017, quando un utente di Reddit con lo pseudonimo " deep fakes" diede il nome a questa tecnologia, utilizzando architetture "*Generative Adversarial Network - GAN*" per creare contenuti pornografici non consensuali (meglio descritti come contenuti multimediali generati mediante software di intelligenza artificiale che, utilizzando materiali autentici come immagini o suoni reali, sono in grado di alterare o ricostruire in modo estremamente realistico l'aspetto, i movimenti e persino la voce di una persona).
Da allora, la diffusione di strumenti sempre più accessibili ha determinato una proliferazione esponenziale del fenomeno:

- nel 2017-2018, la creazione di un deep fake convincente richiedeva competenze tecniche avanzate, hardware specializzato (GPU di fascia alta), dataset consistenti (centinaia di immagini del soggetto target) e tempi di elaborazione prolungati (giorni o settimane);
- nel 2020-2021, l'avvento di framework open-source come DeepFaceLab e Faceswap ha democratizzato la tecnologia, rendendo possibile la creazione di deep fake con strumenti accessibili a utenti non specializzati;
- nel 2023-2025, l'emergere di applicazioni *consumer* e servizi *cloud* ha ulteriormente abbassato le barriere di accesso, permettendo la generazione di *deep fake* di qualità elevata mediante semplici app mobile in pochi minuti.

L'etimologia del termine, che fonde "deep learning" e "fake", rivela la matrice tecnologica del fenomeno: si tratta di artefatti prodotti mediante reti neurali profonde, segnatamente attraverso architetture basate su Generative Adversarial Networks (GAN), Variational Autoencoders (VAE) o, più recentemente, modelli di diffusione (diffusion models).

Dal punto di vista tecnico, il processo di generazione di un deep fake può essere scomposto in diverse fasi tecniche: (i) raccolta e preprocessing dei dati di training, che consiste nell'acquisizione di un corpus di immagini, video o registrazioni audio del soggetto target e nella loro elaborazione mediante tecniche di normalizzazione, allineamento facciale, segmentazione e augmentation; (ii) addestramento del modello generativo, mediante il quale una rete neurale apprende le caratteristiche statistiche e morfologiche del soggetto attraverso l'ottimizzazione iterativa di funzioni di loss complesse; (iii) sintesi del contenuto manipolato, in cui il modello addestrato genera nuovi contenuti sintetici applicando le caratteristiche apprese a materiali di input diversi; (iv) post-processing e raffinamento, che include operazioni di color grading, blending, compositing e temporal smoothing per aumentare il realismo del risultato finale.

[2] Sulla distinzione tra manipolazione di dati esistenti e generazione sintetica ex novo, HILDEBRANDT, *Law for Computer Scientists and Other Folk*, Oxford, 2020.

Le manifestazioni del fenomeno deep fake sono molteplici e di crescente pericolosità sociale, con impatti che trascendono la sfera individuale per investire la dimensione collettiva della fiducia informativa e dell'integrità del discorso pubblico.

Sul piano delle lesioni individuali, a titolo esemplificativo, si registrano la pornografia non consensuale (***deep fake porn***) che costituisce la forma più diffusa e lesiva, con stime che indicano oltre il 96% dei deep fake online come contenuti pornografici non consensuali; frodi e truffe sofisticate, realizzate mediante sintesi vocale e video per impersonare dirigenti aziendali, familiari o autorità; danneggiamento reputazionale mediante la creazione di contenuti compromettenti falsi; ricatti e estorsioni basati sulla minaccia di diffusione di deep fake lesivi.

Sul piano sociale e istituzionale, gli effetti più severi si riverberano nei fenomeni di disinformazione politica mediante la creazione di dichiarazioni false attribuite a figure pubbliche, ovvero di manipolazione dell'opinione pubblica attraverso campagne coordinate di diffusione di contenuti sintetici. Altri esempi riguardano l'erosione della fiducia nei media e nelle istituzioni derivante dall'incertezza sulla veridicità dei contenuti, nonché l'interferenza nei processi democratici attraverso la diffusione strategica di deep fake in periodi elettorali.

2. Profili sostanziali

Alla luce delle premesse finora svolte, si può affermare che il neo-delitto evidenzia un'esigenza di tutela innovativa della persona rispetto ai pericoli delle tecnologie di intelligenza artificiale. La collocazione sistematica dopo l'art. 612-ter c.p. (revenge porn) manifesta l'espansione dei rischi nella società tecnologica, ma il bene giuridico tutelato appare più ampio e non riducibile alla sola riservatezza sessuale.

Secondo autorevole dottrina[3] si tratterebbe di un diritto nuovo: il diritto a non essere ingannati o condizionati sfavorevolmente dall'intelligenza artificiale.Tale interesse trova fondamento nell'art. 5, lett. a), AI Act, che vieta le pratiche manipolative o ingannevoli aventi lo scopo di distorcere il comportamento umano, configurando rischi inaccettabili già sanzionati amministrativamente con le pene più severe.

Il bene giuridico fa capo a qualunque persona destinataria della comunicazione o diffusione dei contenuti manipolati, restando però disponibile attraverso il consenso. L'offesa è penalmente rilevante solo se produce un danno ingiusto che costituisce l'evento consumativo e deve essere oggetto del dolo dell'agente.

Nondimeno, la fattispecie richiede l'impiego tassativo di sistemi di intelligenza artificiale e tale nozione extrapenale, sebbene ampia e criticata, coglie le caratteristiche distintive rispetto ai tradizionali sistemi informatici: l'autonomia, l'adattabilità e la capacità di generare output (previsioni, raccomandazioni, decisioni) che possano influenzare ambienti fisici o virtuali. Questi requisiti distintivi, pur al di là delle singole tecniche informatiche, possono e devono essere accertati giudizialmente.

Le condotte tipizzate (cessione, pubblicazione, diffusione) richiamano quelle di fattispecie vigenti in materia di pornografia minorile. In particolare, ciò trova conferma nelle nozioni di alterazione e falsificazione: la prima indica la modificazione non autorizzata di un contenuto genuino; la seconda include la creazione integrale di contenuti difformi dall'originale o dal vero oggettivo. Per alcuni[4] va

[3] PICOTTI, *I contenuti penali della legge sull'intelligenza artificiale*, in Sistema Penale, pubblicato il 2 dicembre 2025, p. 4.

[4] PICOTTI, cit., p. 7.

IA, materialmente o ideologicamente falsi. In ogni caso, il fenomeno supera il porn deep fake, estendendosi alle manipolazioni di dichiarazioni politiche per influenzare l'opinione pubblica o distorcere competizioni elettorali. Il requisito selettivo fondamentale è l'idoneità oggettiva ad indurre in inganno sulla genuinità, valutata rispetto alla comune capacità di discernimento dei destinatari.

L'elemento del danno ingiusto come evento consumativo appare distonico rispetto alle fattispecie analoghe che si consumano con le mere condotte di diffusione e crea problemi di accertamento del nesso eziologico data l'autonomia e imprevedibilità del sistema di IA.

Il regime di procedibilità è a querela, con procedibilità d'ufficio nei casi di connessione con delitti procedibili d'ufficio, di vittime incapaci per età o infermità, e di offesa a pubblica autorità nell'esercizio delle funzioni. Quest'ultima previsione estende la nozione di persona offesa oltre le persone fisiche, coinvolgendo enti pubblici, con incertezze sull'individuazione del soggetto danneggiato.

3. Criticità tecnico-informatiche nella definizione normativa e problemi di determinatezza

Stando alla lettera della legge, la fattispecie punisce condotte realizzate "mediante l'impiego di sistemi di intelligenza artificiale", senza, tuttavia, fornire criteri tecnici operativi per discriminare ciò che integra intelligenza artificiale da ciò che ne resta escluso. Autorevole dottrina ha evidenziato che la condotta risulta eccessivamente generica in quanto riferita a contenuti generati o manipolati con IA, sottolineando l'assenza di parametri tecnici identificativi. Tale lacuna genera problemi applicativi concreti che investono tanto la fase investigativa quanto quella processuale.

Dal punto di vista dell'informatica giuridica, però, si pongono ulteriori interrogativi fondamentali sulla demarcazione tra sistemi che integrano intelligenza artificiale ai sensi della norma e sistemi che, pur realizzando risultati analoghi, operano mediante paradigmi computazionali differenti.

Si pensi a software come Instagram, Snapchat o ad applicazioni professionali come Adobe Photoshop, i quali applicano filtri che modificano l'aspetto del volto mediante algoritmi di elaborazione digitale delle immagini. Tali algoritmi operano attraverso operazioni matematiche deterministiche (convoluzione, trasformazioni affini, filtering nel dominio delle frequenze) codificate esplicitamente dai programmatori, ma senza coinvolgere processi di apprendimento automatico da dati. Un filtro che leviga la pelle, rimuove imperfezioni, modifica le proporzioni facciali, altera il colore degli occhi, opera mediante l'applicazione di funzioni matematiche predefinite ai pixel dell'immagine.

Al riguardo la questione giuridica è se tali manipolazioni integrino "impiego di sistemi di intelligenza artificiale" ai sensi dell'art. 612-quater c.p. Dal punto di vista tecnico, la risposta appare negativa: gli algoritmi di image processing tradizionali non apprendono da dati, non presentano autonomia decisionale, non operano mediante reti neurali o modelli statistici complessi. Tuttavia, dal punto di vista fenomenologico, producono risultati analoghi ai sistemi di IA, immagini alterate che possono indurre in inganno sulla genuinità. La definizione normativa non fornisce criteri per risolvere tale ambiguità.

Un operatore esperto che utilizzi strumenti professionali di editing video (Adobe After Effects, Nuke, Blender) può manipolare frame per frame un video per sovrapporre il volto di una persona, mediante tecniche di rotoscoping, chroma keying, tracking manuale dei punti caratteristici, color matching e compositing. Il risultato finale può essere indistinguibile da un deep fake generato mediante GAN, ma il processo di creazione è radicalmente diverso: non coinvolge sistemi di IA, ma competenze artistiche e tecniche applicate manualmente.

L'operatore, quindi, non impiega sistemi di intelligenza artificiale in senso tecnico, ma utilizza software che automatizzano operazioni specifiche (tracking, stabilizzazione) mediante algoritmi

manipolati idonei a ingannare) suggerirebbe una risposta affermativa. Tale tensione tra elemento letterale e ratio normativa genera incertezza applicativa.

Ancora un esempio di condotta dagli esiti incerti è data dal fatto che nella prassi operativa, molti deep fake sono prodotti mediante workflow ibridi che combinano l'utilizzo di algoritmi di IA per alcune fasi specifiche (es. face detection mediante CNN, face-swapping mediante GAN, sintesi vocale mediante modelli text-to-speech) ed interventi manuali per fasi di rifinitura (color grading, compositing finale, correzione manuale di artefatti) oppure impiego di algoritmi tradizionali per operazioni ausiliarie (stabilizzazione, noise reduction, encoding video).

In tali ipotesi, il risultato finale è il prodotto di una catena operativa complessa in cui l'IA costituisce solo uno dei componenti. Dal punto di vista giuridico, sorge il problema di capire se la fattispecie si integra anche quando l'IA è impiegata solo parzialmente nel processo di creazione.

E anche qui la norma non fornisce criteri risolutivi, rimettendo all'interprete valutazioni caso per caso con alto rischio di arbitrarietà.

Richiamando le citate difficoltà di accertare processualmente l'elemento normativo "impiego di sistemi di intelligenza artificiale", in uno con l'indeterminatezza tecnica della norma, introduce potenziali disparità di trattamento: infatti, condotte identiche dal punto di vista fenomenologico potrebbero ricevere trattamenti processuali differenti in funzione della sofisticazione tecnologica impiegata, sollevando questioni di uguaglianza sostanziale nell'applicazione della norma penale.

4. Profili di accertamento forense digitale dei deep fake

L'attribuzione causale tra impiego di IA e produzione del contenuto costituisce un fatto costitutivo del reato, la cui dimostrazione, come noto, grava sull'accusa in applicazione del principio generale dell'onere probatorio nel processo penale. Pare opportuno, a questo punto, analizzare i ragionamenti e gli strumenti di cui ci si avvale, in ambito tecnico, per l'esecuzione delle perizie digitali.

L'accertamento peritale deve rispondere a tre quesiti fondamentali:

- identificazione della manipolazione,
- attribuzione tecnologica e
- caratterizzazione forense.

Per poter rispondere a tali domande, l'analisi forense dei deep fake si avvale di un complesso di metodologie tecnico-scientifiche che possono essere classificate secondo la tassonomia proposta dal NIST (National Institute of Standards and Technology) nel framework Digital Evidence. Tale framework, sviluppato inizialmente per l'analisi forense generale di evidenze digitali, è stato adattato e specializzato per rispondere alle peculiarità dei contenuti sintetici generati mediante IA.

La fase di **Collection** costituisce il momento critico in cui viene raccolta l'evidenza digitale, preservandone l'integrità attraverso procedure che garantiscano la catena di custodia (chain of custody). Nel contesto dei deep fake, questa fase presenta peculiarità tecniche.

- *Calcolo di hash crittografici.* Immediatamente dopo l'acquisizione del file sospetto, deve essere calcolato un hash crittografico (tipicamente SHA-256, eventualmente integrato con MD5 per retrocompatibilità con sistemi legacy) del file originale. L'hash costituisce un'impronta digitale univoca del contenuto: qualunque modifica, anche minimale, del file produce un hash completamente differente. L'hash deve essere documentato in un verbale di acquisizione firmato digitalmente e conservato separatamente dall'evidenza. Nei procedimenti giudiziari, la corrispondenza tra l'hash calcolato al momento dell'acquisizione e l'hash del file presentato in dibattimento costituisce prova dell'integrità dell'evidenza.

framerate, risoluzione, informazioni sulle tracce audio, timestamp di creazione e modifica; per i file audio, i metadati includono formato di encoding, bitrate, sample rate, metadata ID3 per file musicali. Tali metadati devono essere estratti mediante strumenti forensi validati (ExifTool, MediaInfo) e documentati integralmente. Particolare attenzione deve essere posta alla rilevazione di anomalie nei metadati: incongruenze temporali (timestamp di creazione posteriori a timestamp di modifica), incompatibilità tra metadati dichiarati e caratteristiche effettive del file, assenza di metadati tipicamente presenti, presenza di metadati associati a software di manipolazione.

- *Creazione di copie forensi bit-a-bit.* L'evidenza originale non deve mai essere analizzata direttamente, per evitare qualunque rischio di alterazione. Deve essere creata una copia forense bit-a-bit (immagine forense) su supporti write-protected. Nel contesto dei deep fake diffusi online, l'acquisizione presenta sfide aggiuntive: necessità di documentare l'URL di origine, lo screenshot della pagina web, le informazioni sul server hosting, i log di accesso se disponibili; gestione dei contenuti effimeri su piattaforme social (Stories di Instagram, Snap di Snapchat) che si auto-distruggono dopo un periodo predefinito, richiedendo acquisizione rapida; acquisizione di contenuti su piattaforme cifrate end-to-end (Telegram, Signal, WhatsApp) che presentano protezioni tecniche contro l'estrazione forense.

La fase di **Examination** costituisce il cuore dell'analisi forense, in cui vengono applicate tecniche specialistiche per identificare indicatori di manipolazione. Per i deep fake, tale analisi si articola su diversi livelli.

a) Analisi degli artefatti algoritmici specifici delle GAN

I sistemi di deep learning, in particolare le GAN, lasciano tracce caratteristiche nei contenuti generati, definite artefatti algoritmici. Tali artefatti derivano dalle limitazioni intrinseche dell'architettura e dal processo di addestramento[5].

b) Analisi delle inconsistenze fisiche e illuminotecniche

I contenuti manipolati mediante IA spesso presentano violazioni delle leggi fisiche dell'ottica e dell'illuminazione che possono essere rilevate mediante analisi computazionale[6].

[5] *Pattern nella distribuzione spettrale.* Le immagini naturali presentano una distribuzione caratteristica delle frequenze: le componenti a bassa frequenza (variazioni graduali di colore e luminosità) sono predominanti, mentre le componenti ad alta frequenza (dettagli fini, bordi netti) decrescono secondo una legge di potenza predicibile. Le immagini generate mediante GAN spesso deviano da tale distribuzione, presentano eccesso di energia in specifiche bande di frequenza, mostrando pattern periodici anomali rilevabili mediante trasformata di Fourier bidimensionale; mostrano discontinuità nella distribuzione spettrale tra diverse regioni dell'immagine. L'analisi spettrale si effettua mediante: conversione dell'immagine nel dominio delle frequenze mediante FFT 2D, calcolo dello spettro di potenza radiale, confronto con modelli statistici di immagini naturali, identificazione di anomalie mediante test statistici. *Artefatti nei pattern di rumore ad alta frequenza.* Le fotocamere digitali introducono pattern di rumore caratteristici derivanti dalle imperfezioni del sensore e dell'elettronica di acquisizione. Tale rumore presenta proprietà statistiche specifiche: distribuzione gaussiana, correlazione spaziale limitata, dipendenza dall'ISO e dalle condizioni di illuminazione. Le GAN tendono a produrre pattern di rumore differenti: rumore eccessivamente uniforme o, viceversa, con variazioni anomale; correlazioni spaziali inusuali; assenza del Pattern Response Non-Uniformity tipico dei sensori reali. L'analisi del rumore si effettua mediante: estrazione del rumore mediante filtering passa-alto, analisi statistica delle proprietà del rumore (media, varianza, curtosi, asimmetria), confronto con modelli di rumore di sensori reali, ricerca di pattern periodici o correlazioni anomale. *GAN fingerprinting.* Ricerche recenti hanno dimostrato che diverse architetture GAN lasciano "impronte digitali" distintive nei contenuti generati. Ogni modello GAN specifico (StyleGAN, StyleGAN2, ProGAN) produce artefatti caratteristici nei pattern di upsampling, nelle modalità di generazione delle texture, nelle proprietà statistiche degli strati latenti. Mediante tecniche di GAN fingerprinting è possibile non solo identificare che un contenuto è stato generato mediante GAN, ma anche determinare quale architettura specifica è stata impiegata, e in alcuni casi persino identificare il modello pre-addestrato specifico utilizzato.

[6] *Analisi fotometrica della coerenza dell'illuminazione.* In una scena reale, tutte le superfici sono illuminate dalle stesse sorgenti luminose, producendo ombre e highlights coerenti. I deep fake spesso violano tale coerenza: volti face-swapped possono presentare direzioni di illuminazione incompatibili con il resto della scena; le ombre proiettate possono essere incoerenti con la posizione apparente delle sorgenti luminose; i riflessi speculari (negli occhi, su superfici lucide) possono mancare o essere fisicamente impossibili. L'analisi si effettua mediante: stima computazionale della direzione della sorgente luminosa per diverse regioni dell'immagine mediante inversione del modello di illuminazione lambertiana; verifica della consistenza delle direzioni stimate; analisi dei riflessi speculari mediante tecniche di specular decomposition; simulazione fisica delle ombre e confronto con quelle osservate. *Analisi dei riflessi negli occhi.* Gli occhi umani agiscono come sfere riflettenti che catturano l'immagine dell'ambiente circostante. In un video autentico, i riflessi negli occhi devono essere coerenti con l'ambiente visibile nel frame, devono cambiare dinamicamente quando il soggetto muove lo sguardo o la testa, devono presentare le corrette proprietà geometriche di riflessione speculare su superficie sferica. I deep fake spesso falliscono nel replicare correttamente tali riflessi: i riflessi possono essere assenti, statici quando dovrebbero cambiare, geometricamente impossibili, inconsistenti tra occhio destro e sinistro. L'analisi si effettua mediante: estrazione computazionale delle regioni oculari, enhancement dei riflessi corneali, analisi della coerenza temporale nei video, confronto con modelli fisici di riflessione su superfici sferiche.

d) Analisi della compressione e dei pattern di manipolazione[8]

e) Analisi mediante reti neurali avversarie per il rilevamento

Paradossalmente, le stesse tecnologie di IA impiegate per creare deep fake vengono utilizzate per rilevarli. Sono stati sviluppati modelli di deep learning specificamente addestrati sul rilevamento di contenuti sintetici[9].

Dopo l'applicazione delle tecniche di examination, la fase di **Analysis** consiste nella valutazione complessiva delle evidenze raccolte per formulare conclusioni forensi solide e scientificamente fondate. Ciascuna delle tecniche applicate produce indicatori che devono essere valutati statisticamente. Non è sufficiente rilevare anomalie, occorre quantificare la loro significatività statistica. Per esempio se l'analisi spettrale rileva anomalie nella distribuzione delle frequenze, occorre calcolare la probabilità che tali anomalie si verifichino per caso in un'immagine autentica; se l'analisi del blinking rileva frequenze atipiche, occorre confrontarle con la distribuzione statistica del blinking naturale. La valutazione impiega test statistici appropriati (test chi-quadrato, test di Kolmogorov-Smirnov, likelihood ratio tests) per quantificare la significatività.

Le conclusioni forensi così ottenute devono essere accompagnate da una stima del grado di confidenza, espressa secondo scale standardizzate. Una formulazione tipica potrebbe essere: "Con alta confidenza (probabilità superiore al 90%), il video analizzato è stato manipolato mediante tecniche di intelligenza artificiale". Il grado di confidenza dipende da numero e coerenza degli indicatori rilevati, significatività statistica di ciascun indicatore, robustezza delle tecniche impiegate, qualità dell'evidenza analizzata.

---

[7] *Blinking pattern analysis.* L'ammiccamento palpebrale umano segue pattern statistici ben caratterizzati: frequenza media di 15-20 ammiccamenti al minuto in condizioni normali, durata tipica di 100-400 millisecondi, distribuzione temporale che segue processi stocastici specifici (processo di Poisson modulato). I primi deep fake mostravano assenza quasi totale di ammiccamento, problema ora parzialmente risolto ma che lascia comunque anomalie rilevabili: frequenza anomala, durata atipica, pattern temporali non naturali, asimmetrie tra occhio destro e sinistro. L'analisi si effettua mediante: tracking automatico delle palpebre frame-by-frame mediante facial landmark detection, estrazione della serie temporale degli eventi di ammiccamento, analisi statistica della distribuzione temporale, confronto con modelli di blinking umano naturale. *Analisi della sincronizzazione audio-visiva.* Nel parlato umano naturale, esiste una sincronizzazione precisa tra il segnale audio (fonetica) e il movimento delle labbra (visemes). I deep fake che manipolano il parlato spesso presentano: disallineamenti temporali micrometrici tra audio e movimento labiale, movimenti labiali che non corrispondono esattamente ai fonemi pronunciati, mancanza dei movimenti preparatori che precedono l'articolazione di specifici fonemi, anomalie nei movimenti della mascella e della lingua visibili in video ad alta risoluzione. L'analisi si effettua mediante: estrazione dei landmark facciali relativi a bocca e labbra, estrazione delle caratteristiche acustiche del segnale audio (MFCC, spectrogrammi), analisi cross-modale della sincronizzazione mediante tecniche di correlazione, confronto con modelli di sincronizzazione audio-visiva derivati da dataset di parlato naturale.

[8] *Analisi del Photo Response Non-Uniformity (PRNU).* Ogni sensore fotografico digitale presenta imperfezioni di fabbricazione uniche che introducono un pattern di rumore caratteristico, il PRNU, che agisce come una "impronta digitale" del dispositivo. Quando un'immagine è acquisita da un sensore reale, tale pattern è presente in modo coerente su tutta l'immagine. Quando un'immagine è manipolata o sintetizzata: il PRNU può essere assente nelle regioni sintetiche, può presentare discontinuità tra regioni autentiche e manipolate, può mostrare inconsistenze statistiche. L'analisi PRNU si effettua mediante: estrazione del pattern di rumore dell'immagine mediante filtering avanzato, confronto con il PRNU di riferimento del presunto dispositivo di acquisizione se disponibile, analisi delle discontinuità spaziali nel PRNU che indicano manipolazioni localizzate, test statistici di correlazione per verificare la presenza del PRNU atteso. *Error Level Analysis (ELA).* Le immagini digitali sono tipicamente compresse mediante algoritmi lossy come JPEG. Ogni ciclo di compressione/decompressione introduce artefatti specifici. Quando un'immagine è manipolata localmente, le regioni modificate presentano livelli di compressione differenti rispetto alle regioni originali, creando discontinuità rilevabili. L'ELA opera mediante: ricompressione dell'immagine a un livello di qualità noto, calcolo della differenza tra immagine originale e ricompressa, visualizzazione della differenza come heatmap dove regioni con error level alto indicano possibili manipolazioni. Tale tecnica è particolarmente efficace per identificare manipolazioni grossolane, ma può produrre falsi positivi su immagini con contenuto eterogeneo.

[9] *XceptionNet e architetture derivate.* XceptionNet è una rete convoluzionale profonda che impiega depthwise separable convolutions per estrarre feature discriminative. Addestrata su vasti dataset di deep fake annotati (come FaceForensics++, che contiene oltre 1.8 milioni di frame manipolati con diverse tecniche), XceptionNet raggiunge accuratezze superiori al 95% nel rilevamento di deep fake di qualità media. Tuttavia, le performance degradano significativamente su deep fake di alta qualità o su architetture non presenti nel training set. Varianti migliorate includono: EfficientNet-B4, che bilancia accuratezza e efficienza computazionale; Capsule Networks, che apprendono relazioni gerarchiche nelle feature facciali risultando più robuste a variazioni; ensemble di modelli multipli, che combinano le predizioni di diversi detector per aumentare robustezza e accuratezza. *Transformer multimodali.* Per i deep fake video con componente audio manipolata, sono stati sviluppati modelli che analizzano congiuntamente: il dominio visivo (frame video), il dominio audio (segnale acustico), la sincronizzazione cross-modale tra i due. Tali modelli, basati su architetture Transformer che permettono l'attenzione cross-modale, sono particolarmente efficaci nell'identificare inconsistenze di sincronizzazione audio-video che sfuggono all'analisi separata dei singoli domini.

compressioni successive; anomalie nei pattern di blinking potrebbero derivare da condizioni fisiologiche particolari del soggetto; inconsistenze nell'illuminazione potrebbero derivare da setup di illuminazione complessi in studio. La perizia deve argomentare perché tali spiegazioni alternative sono implausibili nel caso specifico.

La relazione peritale deve tradurre l'analisi tecnica complessa in un formato comprensibile per i soggetti processuali non esperti (giudici, avvocati, giurati), mantenendo, allo stesso tempo, grande rigore scientifico. Ciò significa che le tecniche impiegate devono essere descritte evitando eccessivo tecnicismo, utilizzando analogie e spiegazioni intuitive ove possibile (invece di descrivere matematicamente la trasformata di Fourier, si può spiegare che *permette di scomporre un'immagine nelle sue componenti di frequenza, analogamente a come un prisma scompone la luce nei colori dell'arcobaleno*). La chiarezza è richiesta anche quando si devono presentare le evidenze, le visualizzazioni devono essere autoesplicative e corredate di legende comprensibili (immagini side-by-side che evidenziano anomalie, heatmap che mostrano regioni sospette, grafici che illustrano deviazioni statistiche, timeline che ricostruiscono la sequenza di manipolazioni).

Nonostante i progressi metodologici e strumentali descritti, l'accertamento forense dei deep fake presenta criticità operative significative che impattano sull'effettività della tutela penale e sulla certezza del diritto.

Intanto, i laboratori di polizia scientifica italiani dispongono di competenze consolidate nell'informatica forense tradizionale (analisi di hard disk, recupero dati, analisi di log). Tuttavia, l'analisi forense dei deep fake richiede competenze ultra-specialistiche che intersecano: machine learning e deep learning (comprensione delle architetture neurali, familiarità con framework come TensorFlow e PyTorch); computer vision avanzata (conoscenza di algoritmi di image processing, facial recognition, object detection); signal processing (analisi nel dominio delle frequenze, elaborazione di segnali audio-video); statistica computazionale (test statistici, analisi bayesiana, stima di confidenza). Tale profilo di competenze è raro e richiede formazione lunga e specializzata. Il rischio è che molti casi non possano essere adeguatamente analizzati per carenza di personale qualificato, traducendosi in impunità de facto.

Inoltre, esiste una dinamica di adversarial learning tra tecniche di generazione e tecniche di rilevamento. Gli sviluppatori di GAN possono addestrare i loro modelli specificamente per eludere i detector esistenti: tecnica nota come adversarial training, in cui il generatore è addestrato non solo a produrre immagini realistiche, ma anche a evitare il rilevamento da parte di specifici detector. Il risultato è che, appena viene sviluppato un nuovo detector efficace, emergono GAN capaci di evaderlo. Le ricerche dimostrano tassi di evasion attack superiori al 90% contro detector non specificamente hardened. Questo implica che le metodologie forensi devono essere continuamente aggiornate, richiedendo investimenti costanti in ricerca e sviluppo.

Ancora, si tenga conto che anche le tecniche più avanzate non garantiscono certezza assoluta. Gli studi sperimentali evidenziano: falsi negativi (deep fake non rilevati) con tassi del 5-20% per deep fake di alta qualità; falsi positivi (contenuti autentici classificati erroneamente come deep fake) con tassi del 2-10% a seconda delle soglie di confidenza impiegate. Nel processo penale, che richiede una prova oltre ogni ragionevole dubbio, sarà difficile capire quale soglia di confidenza statistica è sufficiente e l'assenza di standard consolidati rischia di generare valutazioni difformi tra diversi giudici e diverse perizie.

Un'analisi forense completa di un video deep fake, poi, può richiedere settimane per casi di media complessità (video brevi, qualità standard), mesi per casi complessi (video lunghi, qualità elevata, necessità di analisi cross-referenziate). I costi per consulenze tecniche ultra-specializzate possono raggiungere decine di migliaia di euro per caso. Ciò genera problemi di ragionevole durata del processo, disparità di trattamento tra imputati abbienti (che possono permettersi consulenze tecniche di parte sofisticate) e non abbienti ed inefficienze sistemiche con accumulo di arretrati.

Infine, due aspetti meritano di essere menzionati.

I contenuti digitali presentano peculiarità che complicano la chain of custody: volatilità (facilmente cancellabili, modificabili senza lasciare tracce evidenti); duplicabilità perfetta (impossibile distinguere originale da copia); dipendenza da metadati (che possono essere facilmente alterati); natura distribuita (contenuti hosted su server esteri, giurisdizioni multiple). La dimostrazione dell'integrità dell'evidenza digitale richiede procedure rigorose spesso non implementate correttamente nelle fasi iniziali dell'indagine, quando non vi è ancora consapevolezza della rilevanza penale. Il rischio è di evidence exclusion per vizi procedurali.

In secondo luogo, i contenuti sulle piattaforme come Instagram Stories, Snapchat, Telegram (detti *ephemeral content)* con il loro messaggi autodistruttivi implementano meccanismi di cancellazione automatica. I deep fake diffusi su tali piattaforme possono scomparire prima che sia stata avviata l'indagine. Inoltre, gli autori consapevoli possono procedere a cancellazioni massive: mediante scraping automatizzato dei propri post, utilizzo di servizi di anonimizzazione (VPN, Tor) per ostacolare tracciamento, impiego di tecniche anti-forensi (metadati stripping, encryption, steganografia). Ciò comporterebbe che in molti casi l'evidenza digitale potrebbe essere irrecuperabile al momento in cui l'autorità giudiziaria interviene.

5. La rapidità dell'obsolescenza tecnologica e l'adattabilità normativa

Come si intuisce da quanto detto finora, lo sviluppo dei sistemi di IA generativa segue una traiettoria di crescita esponenziale che solleva dubbi sulla sostenibilità temporale della disciplina normativa. La legge 132/2025 codifica una fattispecie modellata sulle tecnologie disponibili nel 2025, ma l'evoluzione tecnologica potrebbe rendere obsolete le categorie concettuali della norma nel breve-medio periodo.

Le proiezioni per il 2027-2030 suggeriscono che esisteranno deep fake perfetti indistinguibili dalla realtà anche per detector specializzati, avremo una democratizzazione completa della tecnologia con strumenti accessibili a chiunque mediante smartphone, si farà sintesi in tempo reale durante videoconferenze, vi sarà integrazione con realtà aumentata e virtuale. In tale scenario, l'elemento costitutivo del reato (impiego di sistemi di IA) potrebbe divenire impossibile da provare processualmente, rendendo la fattispecie sostanzialmente inapplicabile.

Dinanzi alla rapidità dell'evoluzione tecnologica e alla limitatezza della tecnica legislativa tradizionale, si propongono approcci alternativi.

Risulterebbe opportuno formulare norme che rinviino a parametri tecnici elaborati da autorità specializzate (AgID, Agenzia per la Cybersicurezza Nazionale, enti di standardizzazione internazionali come ISO), aggiornabili mediante atti amministrativi senza necessità di intervento legislativo. Però, sebbene si potrebbe guadagnare in flessibilità ed aggiornabilità rapida, potrebbero esserci tensioni con il principio di riserva di legge in materia penale (necessità di garantire pubblicità e conoscibilità degli standard tecnici).

In secondo luogo, si potrebbero prevedere ***sunset clauses*** che impongano al legislatore di rivedere periodicamente (es. ogni 3 anni) la disciplina alla luce dell'evoluzione tecnologica, con eventuale decadenza automatica in assenza di aggiornamento. Questo assicurerebbe che la normativa non diventi obsoleta per inerzia legislativa.

In modo più generico, si potrebbero formulare norme che sanzionino il risultato lesivo (diffusione di contenuti manipolati idonei a ingannare) indipendentemente dalla specifica tecnologia impiegata,

evitando riferimenti a categorie tecnologiche destinate a obsolescenza. Ciò richiederebbe, però, meccanismi per delimitare l'ambito di applicazione ed evitare indeterminatezza.

Il modello attualmente adottato dalla legge 132/2025 (che prevede decreti attuativi da adottare entro 12 mesi) rappresenta un passo verso maggiore adattabilità, ma potrebbe risultare insufficiente per tutti i motivi esposti.

6. Considerazioni conclusive

L'introduzione dell'art. 612-quater c.p. segna una cesura paradigmatica nel rapporto tra ordinamento penale e tecnologie algoritmiche generative, ponendo sul banco di prova la tenuta epistemologica delle categorie dogmatiche tradizionali dinanzi all'irruzione di artefatti computazionali dotati di autonomia operativa e capacità sintetica. Tuttavia, l'indagine condotta rivela come tale intervento normativo manifesti aporie strutturali che trascendono la mera tecnica di formulazione legislativa, investendo piuttosto la stessa ontologia giuridica del fatto tecnologicamente mediato.

Sul piano della tassatività tecnico-normativa, la fattispecie sconta un'indeterminatezza che non attiene alla dimensione semantico-linguistica del precetto, quanto alla sua sostanza epistemologica: il concetto di "sistema di intelligenza artificiale" – lungi dal configurare una nozione descrittiva di contorni definiti – si atteggia quale categoria fluida e porosa, irriducibile a parametri tecnici univoci e stabili. Siffatta evanescenza concettuale non costituisce mera imperfezione redazionale emendabile mediante l'intervento ermeneutico della giurisprudenza o della dottrina, ma riflette una frattura più profonda tra il ritmo incrementale dell'innovazione tecnologica e la vocazione stabilizzatrice del diritto penale.

La lacuna non può essere colmata mediante l'affinamento interpretativo; essa esige, al contrario, una riformulazione normativa che ancori la fattispecie a criteri operativi verificabili, capaci di segnare demarcazioni nette tra condotte penalmente rilevanti e lecite manipolazioni digitali.

Sul versante dell'accertamento forense digitale, l'analisi ha evidenziato come le metodologie investigative disponibili – per quanto tecnicamente raffinate – scontino deficit di affidabilità epistemica e asimmetrie applicative che ne minano la sostenibilità processuale. La digital forensics applicata ai contenuti sintetici generati mediante architetture neurali profonde si configura quale disciplina ancora in fase di consolidamento scientifico, priva di protocolli standardizzati, di *threshold* di confidenza condivisi dalla comunità scientifica, di validazione inter-laboratorio. L'accertamento giudiziale dell'elemento costitutivo del reato – l'impiego di sistemi di IA – risulta, così, subordinato alla contingente disponibilità di competenze ultra-specialistiche e di infrastrutture computazionali avanzate, generando il paradosso per cui l'an della punibilità dipende de facto dalle asimmetrie di risorse tra diversi uffici giudiziari, con evidenti ricadute sul principio di uguaglianza sostanziale nella giurisdizione penale.

Quanto alla dimensione diacronica dell'efficacia normativa, emerge un'aporia sistemica di più ampia portata: il diritto penale, ancorato a paradigmi epistemologici che postulano la stabilità e la predicibilità delle fattispecie incriminatrici, rivela una costitutiva inadeguatezza a governare fenomeni tecnologici caratterizzati da evoluzione esponenziale e discontinuità paradigmatiche. La fattispecie ex art. 612-quater c.p., calibrata sullo stato dell'arte tecnologico del 2025, rischia l'obsolescenza precoce già nel breve-medio periodo, allorché l'avanzamento delle architetture generative renderà i contenuti sintetici forensicamente indistinguibili dalla realtà empirica, vanificando l'accertabilità processuale dell'elemento normativo.

Di fronte a tali criticità costitutive, si impone una riconfigurazione metodologica dell'approccio regolatorio alle tecnologie algoritmiche emergenti. Ciò postula, in prima istanza, l'instaurazione di un dialogo interdisciplinare autentico e paritario tra penalisti, informatici teorici, esperti di machine learning e data scientists, finalizzato alla co-costruzione di definizioni tecniche che siano al contempo rigorose sul piano scientifico e operazionali sul piano giuridico. Non è sufficiente il mero recepimento

acritico di nozioni elaborate in contesti regolatori extra-penali, occorre, invece, una ridefinizione concettuale che tenga conto delle specificità garantiste proprie del diritto punitivo e delle esigenze di tassatività che ne conformano la legittimazione costituzionale.

In secondo luogo, si rende indifferibile un investimento strategico nella **formazione di personale giudiziario** e **investigativo** dotato di competenze tecnico-forensi avanzate. La risposta istituzionale non può limitarsi alla consulenza tecnica occasionale, ma deve strutturarsi mediante: l'istituzione di corsi di alta formazione permanente in digital forensics applicata ai contenuti IA-generated per magistrati inquirenti e giudicanti; la creazione di laboratori forensi ultra-specializzati presso le Procure generali, dotati di infrastrutture computazionali e personale tecnico-scientifico stabilmente dedicato; l'elaborazione e validazione di protocolli metodologici certificati secondo standard internazionali; l'implementazione di meccanismi di quality assurance e peer review per le perizie tecniche, sul modello delle best practices consolidate nelle scienze forensi tradizionali.

Sul piano della **tecnica legislativa**, si palesa l'esigenza di sperimentare meccanismi normativi adattivi che consentano l'aggiornamento dinamico della disciplina penale senza compromettere la certezza del diritto. Il modello delle deleghe legislative ex art. 76 Cost., per quanto apprezzabile, si rivela insufficiente. Forse servirebbe articolare un sistema ibrido che integri la costituzione di osservatori tecnico-scientifici permanenti, composti pariteticamente da giuristi e tecnologi, con mandato di monitoraggio continuo dell'evoluzione tecnologica e delle sue implicazioni giuridiche; l'introduzione di clausole di revisione obbligatoria a scadenze predeterminate (sunset clauses) che impongano al legislatore la riconsiderazione periodica della disciplina alla luce del mutato contesto tecnologico; il ricorso a rinvii normativi dinamici a standard tecnici elaborati da autorità indipendenti (AgID, ACN), aggiornabili mediante atti regolamentari secondo procedure semplificate ma garantite; il coordinamento sistematico con le istituzioni europee e gli organismi internazionali di standardizzazione, onde assicurare l'armonizzazione delle definizioni tecniche e l'interoperabilità giurisdizionale.

Infine – e qui si annida forse la questione più radicale – urge riconsiderare ab imis l'approccio stesso alla tutela penale nel dominio delle tecnologie algoritmiche. La proliferazione di fattispecie incriminatrici ad hoc, ancorate a specifiche modalità tecnologiche destinate a rapida obsolescenza, genera un diritto penale frammentario e instabile, incapace di assicurare orientamento ex ante ai consociati. Potrebbe rivelarsi più efficace e garantista l'elaborazione di fattispecie "aperte" orientate al risultato lesivo – l'offesa dell'autodeterminazione informativa, la manipolazione fraudolenta del consenso, la lesione dell'integrità cognitiva – anzichè alla modalità tecnica di realizzazione. Tale approccio andrebbe coniugato con meccanismi di tutela anticipata e preventiva, operanti a monte della catena produttiva: obblighi di due diligence tecnologica per sviluppatori e distributori di sistemi di IA ad alto rischio; forme di responsabilità oggettiva attenuata (strict liability) per chi immette sul mercato tecnologie dual-use facilmente strumentalizzabili a fini illeciti; sistemi di certificazione ex ante e tracciabilità crittografica dei contenuti digitali, mediante blockchain permissioned o protocolli di content authentication che consentano la verificabilità della provenienza e dell'integrità dei contenuti multimediali sin dal momento della loro generazione.

In definitiva, l'art. 612-quater c.p. costituisce un primo tentativo – meritorio nelle intenzioni, ma ancora embrionale nell'articolazione tecnica – di presidiare penalmente i rischi derivanti dall'IA generativa. La vera sfida per l'informatico giuridico oggi consiste nella rifondazione epistemologica dell'approccio regolatorio; dallo *ius conditum*, ancorato a tipizzazioni rigide e statiche, verso uno *ius condendum* caratterizzato da plasticità adattativa, apertura interdisciplinare e capacità di autoregolazione dinamica. Solo mediante tale mutamento paradigmatico – che postula un ripensamento profondo del rapporto tra norma penale e realtà tecnologica – sarà possibile elaborare un sistema di tutela penale al contempo efficace nella prevenzione delle condotte lesive e rispettoso delle garanzie costituzionali che fondano la legittimazione democratica dello *ius puniendi* nell'ordinamento liberal-democratico.