Dipartimento di Scienze Politiche

Corso di Laurea Magistrale in Governo, Amministrazione e Politica

Cattedra di Diritto dell'Informazione e della Comunicazione

Deepfake: implicazioni giuridiche e impatto sociale tra innovazione e minaccia digitale.

Relatore:

Prof. Pietro Falletta

Correlatore:

Prof. Gianluca Giansante

Candidato:

Antonia De Rosa

Matr. 656382

Anno Accademico 2024/2025

Indice

INTRODUZIONE

Nel corso degli ultimi anni, il termine *deepfake* ha acquisito una rilevanza sempre maggiore, diventando il simbolo di una delle frontiere più complesse e controverse dell'intelligenza artificiale. La manipolazione di immagini e video, così come la realizzazione di contenuti audio sintetici, ha portato questa tecnologia emergente al centro del dibattito pubblico globale. Collocato tra innovazione e rischio, l'uso dei *deepfake* solleva interrogativi etici, giuridici e sociali di crescente urgenza per la nostra società.

Il termine "*deepfake*" nasce dalla combinazione di "*deep learning"* – un insieme avanzato di tecniche di apprendimento automatico basato su reti neurali profonde – e "*fake*", ovvero falso. Originariamente diffusi in piattaforme come *Reddit*, i *deepfake* si sono rapidamente evoluti da semplici strumenti sperimentali a tecnologie capaci di alterare significativamente la percezione della realtà. Se da un lato il fenomeno ha impatti positivi su settori come l'intrattenimento, l'educazione e la ricerca medica, dall'altro, emergono rischi significativi legati alla diffusione di disinformazione, alla violazione della *privacy*, all'uso per fini criminali e alla manipolazione dell'opinione pubblica.

I progressi nell'intelligenza artificiale permettono oggi di produrre contenuti audiovisivi falsificati con un livello di realismo tale da rendere estremamente difficoltosa la loro individuazione, anche per gli esperti. Questo quadro è ulteriormente aggravato da un sistema comunicativo caratterizzato dalla rapida circolazione delle informazioni, spesso privo di un controllo approfondito delle fonti. L'interazione tra IA e mezzi di comunicazione di massa amplifica il potenziale impatto negativo dei *deepfake* su scala globale.

La tecnologia che alimenta i *deepfake* si basa principalmente sulle reti neurali artificiali, con particolare attenzione alle GAN (*Generative Adversarial Networks*). Queste operano attraverso un processo competitivo tra due reti – un generatore e un discriminatore – allo scopo di perfezionare progressivamente la qualità e la credibilità dei contenuti generati. Tuttavia, questa sofisticata capacità di simulazione solleva un quesito cruciale: come distinguere il vero dal falso in una realtà dove anche le prove visive possono essere falsificate?

L'analisi di tale fenomeno richiede un approccio multidisciplinare: appare necessario, da un lato, comprendere gli aspetti tecnici legati alla produzione dei *deepfake* e al funzionamento degli strumenti che lo rendono possibile; dall'altro, risulta indispensabile sviluppare un quadro normativo adeguato a contrastare gli usi illeciti di questa tecnologia.

A livello europeo, iniziative come l'*Artificial Intelligence Act* rappresentano un importante passo verso la regolamentazione di sistemi IA potenzialmente pericolosi. A ciò si aggiunga il Regolamento Generale sulla Protezione dei Dati (GDPR), quale punto di riferimento normativo per la protezione dei dati personali degli individui. Tuttavia, il rapido sviluppo della tecnologia rende necessaria una revisione continua delle normative per garantire una tutela adeguata della dignità, della reputazione e dell'identità digitale dei cittadini.

Sul piano sociale, i *deepfake* rappresentano una minaccia crescente per la veridicità dell'informazione. La diffusione di *fake news*, la creazione non consensuale di contenuti pornografici (*deep porn*), le frodi identitarie e la propaganda politica erodono la fiducia collettiva nelle fonti del diritto e nei governi. Avvenimenti recenti dimostrano quanto queste tecniche possano essere sfruttate per danneggiare la reputazione di individui o manipolare ampiamente l'opinione pubblica.

L'obiettivo di questa tesi è analizzare il fenomeno dei *deepfake* nella sua complessità: partendo dalla sua definizione tecnica e dagli ambiti applicativi, si esplorano poi le principali implicazioni giuridiche e sociali derivanti dal loro utilizzo. Più nello specifico, il primo capitolo affronta l'origine e lo sviluppo della tecnologia *deepfake*, evidenziando le principali categorie, applicazioni e rischi associati, oltre ad approfondire le basi tecnologiche con un *focus* sull'intelligenza artificiale e le *Generative Adversarial Networks* (GAN). Il secondo capitolo esamina il quadro normativo vigente a livello europeo e internazionale, prestando particolare attenzione al GDPR, all'AI Act e alle risposte legislative adottate in altre giurisdizioni come Stati Uniti, Cina e Regno Unito. Infine, il terzo capitolo raccoglie una serie di casi studio significativi – dalle campagne elettorali statunitensi, all'impiego dell'intelligenza artificiale come mezzo di censura da parte del governo cinese, e ancora, al fenomeno del *deep porn* che ha colpito celebrità internazionali – per evidenziare l'impatto concreto e multidimensionale dei *deepfake* sulla realtà contemporanea.

Attraverso un approccio critico e documentato, il presente studio si propone di offrire un contributo al dibattito sull'equilibrio tra innovazione tecnologica e tutela dei diritti fondamentali. In un mondo sempre più digitalizzato e interconnesso, comprendere la natura e le implicazioni dei *deepfake* non è solo un interesse di ricerca, ma un imperativo etico e civile.

CAPITOLO I

DEFINIZIONE, SVILUPPO E APPLICAZIONI DEI *DEEPFAKE*

1. Definizione e quadro introduttivo

Il *deepfake* è definito come un "*filmato che presenta immagini corporee e facciali catturate in Internet, rielaborate e adattate a un contesto diverso da quello originario tramite un sofisticato algoritmo*".[1] Il termine *deepfake* è un neologismo inglese che nasce dalla fusione di "*deep learning*" (l'insieme di tecniche che permettono all'intelligenza artificiale di imparare a riconoscere le forme) e "*fake*" (falso, notizia falsa)[2].

Un *deepfake* è un filmato manipolato digitalmente, in cui immagini facciali e corporee sono alterate e sovrapposte a contesti diversi da quelli originali attraverso algoritmi sofisticati.[3] Si tratta di foto, video e audio in grado di ingannare gli utenti del *web* grazie al loro alto livello di realismo.[4] I *deepfake* si basano, di fatto, sull'utilizzo di reti neurali che imitano la struttura dei neuroni cerebrali; tali reti vengono "addestrate" con una grande quantità di dati reali per riconoscere e imitare espressioni facciali, movimenti labiali, timbri di voce e altre caratteristiche umane per creare un contenuto falso che appaia, però, autentico.[5]

Se è vero che questa tecnologia si dimostra essere estremamente utile se impiegata per scopi leciti, è altrettanto vero che, quando utilizzata per finalità illecite, risulta fortemente dannosa. Nel settore dell'intrattenimento, i *deepfake* sono utilizzati per migliorare gli effetti speciali nei film o per il doppiaggio automatizzato[6], mentre nel campo della pubblicità le aziende possono creare *testimonial* virtuali realistici. Tuttavia, i *deepfake*

---

[1] "*Deepfake*" in Enciclopedia Treccani Online, Istituto della Enciclopedia Italiana, https://www.treccani.it/vocabolario/deepfake_(Neologismi)/ (consultato il: 15/02/2025).
[2] Garante per la Protezione dei Dati Personali, *Vademecum,* dicembre 2020, pp. 1-6.
[3] F. RAMOS, *Deepfake: Análisis de sus implicancias tecnológicas y jurídicas en la era de la Inteligencia Artificial. Derecho Global. Estudios sobre Derecho y Justicia, IX*, 2024, p. 365.
[4] Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020, pp. 1-6.
[5] R. SONG, *Faking It: A Proposed Solution to Counter Nonconsensual Pornographic Deepfakes, 31 Wash. & Lee J. Civ. Rts. & Soc. Just. 157*, 2025, pp. 161-163.
[6] S. TROZZI, *La dimensione costituzionale dell'intelligenza artificiale generativa. La tutela della dignità umana nell'era del deepfake, in Diritto Pubblico Europeo Rassegna online*, 1, 2024, pp. 226-228.

possono essere impiegati anche per scopi illeciti; basti pensare a chi ne usufruisce per diffondere disinformazione politica, diffamare, commettere frode, o per creare contenuti pornografici falsi senza il consenso delle persone coinvolte. Dunque, si tratta di uno strumento che ha il potere di manipolare l'opinione pubblica e innescare, così, gravi danni reputazionali.[7]

Il primo *deepfake* è stato pubblicato nel 2017 sulla piattaforma *Reddit* dall'utente "*Deepfakes*" a discapito di alcune *star* di Hollywood. Ed invero, le vittime più frequenti di *deepfake* sono personaggi noti e, nello specifico, politici, a causa dell'ampia quantità di dati necessari per creare dei contenuti tramite intelligenza artificiale che possano essere considerati degli originali. Oltretutto, i primi contenuti di questo tipo venivano sviluppati in ambito pornografico, sovrapponendo il volto di un personaggio famoso sul corpo di un attore del settore.[8]

In sostanza, l'evoluzione di sistemi di intelligenza artificiale ha reso sempre più difficile distinguere i contenuti autentici da quelli manipolati, sollevando numerose questioni di carattere giuridico e normativo.[9] In questo contesto, risulta essenziale un'adeguata regolamentazione per bilanciare l'innovazione tecnologica con la tutela dei diritti fondamentali degli individui.

1.1 La definizione data dal Garante per la Protezione dei Dati Personali

Come anticipato, il termine "*deepfake*" nasce dalla combinazione di "*deep learning*" e "*fake*", riferendosi a contenuti audiovisivi manipolati attraverso l'uso di tecnologie di intelligenza artificiale avanzata. Il Garante per la Protezione dei Dati Personali (GPDP) definisce i *deepfake* come "*foto, video e audio creati grazie a software di intelligenza artificiale (AI) che, partendo da contenuti reali (immagini e audio), riescono a modificare o ricreare, in modo estremamente realistico, le caratteristiche e i movimenti di un volto o di un corpo e a imitare fedelmente una determinata voce*".[10]

---

[7] *Ibidem.*
[8] *Ivi*, p. 240.
[9] F. GALVANO, L. BADIALI, *Analisi comportamentale applicata al Deepfake, Behaviour Analysis Team*, 2025, pp. 3, 6.
[10] Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020, pp. 1-6.

A differenza di altre forme di disinformazione, come le *fake news* testuali, i *deepfake* operano direttamente su elementi audiovisivi, creando una falsa percezione di autenticità e credibilità. Il Garante Privacy evidenzia come tali contenuti siano generati attraverso *software* di intelligenza artificiale che elaborano dati esistenti, trasformandoli in un nuovo *output* altamente realistico.[11] La condizione necessaria alla base della natura dei *deepfake* è dunque che questa categoria di contenuti debba necessariamente derivare da materiale audiovisivo preesistente: si vanno quindi ad impiegare immagini, video o registrazioni vocali di persone reali. Questa caratteristica è identificativa dei *deepfake* che si distinguono da altre forme di manipolazione, come le illustrazioni digitali o le creazioni *ex novo* di personaggi fittizi.[12]

Giova ripetere che la tecnologia dei *deepfake* si avvale principalmente di reti neurali artificiali e algoritmi di *deep learning*, che apprendono i *pattern* facciali, vocali e gestuali di un soggetto, per poi riprodurli con estrema fedeltà. Grazie alla capacità dell'IA di analizzare grandi quantità di dati, è possibile generare video in cui una persona appare dire o fare qualcosa che in realtà non ha mai detto o fatto, creando contenuti con un livello di sofisticazione tale da impedire talvolta anche agli osservatori più attenti di scindere un falso dalla realtà.[13]

Come sottolineato dal GPDP, la creazione dei *deepfake* è strettamente connessa all'innovazione dell'intelligenza artificiale. Nel caso specifico dei *deepfake*, i modelli di IA vengono addestrati su enormi *dataset* di immagini e suoni per apprendere le caratteristiche distintive di un volto o di una voce. Successivamente, attraverso l'uso di algoritmi avanzati, il *software* è in grado di sovrapporre il volto di un individuo su quello di un altro, oppure di modificare un audio per renderlo indistinguibile dalla voce originale.[14]

---

[11] *Ibidem.*

[12] Camera dei deputati, Introduzione dell'articolo 612-quater del Codice penale, in materia di manipolazione artificiale di immagini di persone reali allo scopo di ottenerne rappresentazioni nude, A.C. 2986, XVIII legislatura.

[13] F. V. VALENTI, *Il deep fake: la nuova sfida dell'intelligenza artificiale generativa, Derecom 37*, 2024, p. 11

[14] F. GALVANO, L. BADIALI, *op. cit.*, pp. 8-11.

È però importante rilevare che, come specificato anche dal Garante Privacy, non tutti i contenuti generati dall'IA possono essere considerati *deepfake*. Affinché un contenuto rientri in questa categoria, quest'ultimo deve essere il risultato di un'elaborazione avanzata da parte di un *software* specifico, il cui obiettivo è quello di ottenere un *output* estremamente realistico.[15] Ad esempio, l'uso di *software* di *editing* tradizionali, come Photoshop per la manipolazione di immagini o Premiere Pro per il montaggio video, non potrebbe in alcun modo essere categorizzato come *deepfake*: non ci si limita al semplice *editing* di una foto, ma si richiede l'intervento di un *software* in grado di imitare l'intelligenza umana.

Dunque, l'obiettivo principale dei *deepfake* è quello di creare contenuti ingannevoli, in grado di raggirare chi li visualizza. La manipolazione audiovisiva può essere utilizzata per molteplici scopi, alcuni leciti e altri illeciti.[16] Tuttavia, appare evidente la minaccia che questa tecnologia rappresenta, soprattutto quando viene impiegata per diffondere disinformazione, creare contenuti pornografici non consensuali o commettere frodi.[17] Il Garante per la Protezione dei Dati Personali sottolinea che i *deepfake* sono potenzialmente capaci di ingannare gli utenti medi del *web*, oltre ad evidenziare la loro capacità di compromettere la veridicità delle informazioni e di violare il diritto all'immagine e alla riservatezza delle persone coinvolte.[18]

Un ultimo elemento essenziale è il grado di realismo che un *deepfake* deve possedere: la qualità dei *deepfake* è aumentata esponenzialmente negli ultimi anni, al punto che alcuni video e audio generati tramite IA risultano quasi indistinguibili da quelli reali.[19] Questo pone problemi significativi non solo per la *privacy* delle persone coinvolte, ma anche per la sicurezza informatica e la tutela della verità nei media digitali.[20]

---

[15] Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020, pp. 1-6.
[16] A. ORLANDO, *La regolamentazione del deepfake in Europa, Stati Uniti e Cina, Medialaws*, 2024, pp. 308-309.
[17] Camera dei deputati, A.C. 2986, cit.
[18] Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020, pp. 1-6.
[19] B. DOLHANSKY, R. HOWES, B. PFLAUM, N. BARAM, C. CANTON FERRER, *The Deepfake Detection Challenge (DFDC) Preview Dataset, ArXiv*, 2019, pp. 1-4.
[20] L. TREMOLADA, *Autoregolamentazione, trasparenza e sorveglianza: i nodi da sciogliere dell'AI Act, Il Sole 24 Ore*, 2023.

Per questi motivi, dal punto di vista giuridico, appare necessario regolamentare l'uso dei *deepfake* per prevenire abusi e proteggere i diritti fondamentali degli individui. Il *General Data Protection Regulation* (cd. GDPR)[21] prevede già norme stringenti per il trattamento dei dati personali, compresi quelli biometrici come sancito dall'art. 9, ma l'evoluzione tecnologica dei *deepfake* richiede interventi normativi più specifici. In Europa, l'*Artificial Intelligence Act* (cd. AI Act)[22] rappresenta un passo importante per stabilire regole armonizzate sull'uso dell'intelligenza artificiale, compresi i sistemi in grado di generare *deepfake*.

1.2 Approcci internazionali: definizioni e classificazioni

Il fenomeno del *deepfake* ha sollevato questioni legali non trascurabili a livello nazionale, essendo spesso l'innovazione tecnologica troppo avanguardista rispetto all'assetto normativo, che fatica a starle dietro. Non esistono, infatti, ancora norme *ad hoc* che riguardino nello specifico i *deepfake*, ma l'Europa sembra muoversi verso una regolamentazione più precisa e concreta con l'adozione dell'AI Act.[23]

A ciò si aggiunga che ogni Stato adotta norme conformi alla propria società e soprattutto alle proprie priorità politiche. Se alcuni ordinamenti giuridici hanno, infatti, adottato misure stringenti per contrastare gli usi illeciti del *deepfake*, altri hanno preferito privilegiare la loro natura innovativa, richiamando alla tutela della libertà di espressione mentre si cerca il connubio perfetto tra tecnologia e sicurezza.[24]

Negli Stati Uniti, la regolamentazione dei *deepfake* si è sviluppata prevalentemente a livello statale. La California e il Texas hanno introdotto leggi che vietano l'uso dei *deepfake* per interferire con i processi elettorali o per creare contenuti pornografici non

[21] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 (GDPR).
[22] Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio del 13 giugno 2024 (AI Act).
[23] F. V. VALENTI, *op. cit.*, pp. 12-13.
[24] A. ORLANDO, *op. cit.*, p. 321.

consensuali. A livello federale, il *DEEPFAKES Accountability Act*[25] mira a rendere obbligatoria l'etichettatura di contenuti manipolati tramite AI.[26]

In Europa, abbiamo evidenziato come il GDPR imponga severe restrizioni sull'uso di dati biometrici, che sono spesso coinvolti nella creazione di *deepfake*, e come la Commissione Europea abbia proposto l'*Artificial Intelligence Act* con l'obiettivo di regolamentare la materia e promuoverne uno sviluppo sicuro ed etico.[27]

L'AI Act prevede regole specifiche per i sistemi di IA che creano un rischio alto per la salute e la sicurezza o per i diritti fondamentali delle persone fisiche. Il regolamento europeo mira a tutelare gli utenti del *web* garantendo certezza rispetto alla veridicità dei contenuti che vengono caricati e affinché gli stessi possano fare scelte più informate, in quanto consapevoli che ciò che stanno guardando o ascoltando sia effettivamente autentico o meno. Prevede, infatti, obblighi di trasparenza per i sistemi di intelligenza artificiale, in particolare quando vengono utilizzati *chatbot* o *deepfake*.[28]

I sistemi di intelligenza artificiale sono classificati in base a livelli di rischio puntualmente individuati: rischio inaccettabile, rischio elevato, rischio limitato e rischio minimo o nullo. I sistemi rientranti nel primo livello di rischio sono del tutto vietati, includendo usi intrusivi e discriminatori dell'AI (come, ad esempio, sistemi di identificazione biometrica remota in tempo reale in spazi accessibili al pubblico).[29]

L'AI Act adotta, dunque, un approccio *risk-based*, ossia si distinguono quattro livelli di rischio per l'intelligenza artificiale: rischio inaccettabile e, perciò, sistemi vietati; alto

---

[25] Questo disegno di legge stabilisce i requisiti per le registrazioni tecnologiche avanzate di falsa rappresentazione (cioè *deepfakes*) e stabilisce sanzioni penali per violazioni correlate.
In particolare, richiede che i produttori di *deepfakes* si conformino generalmente a determinati requisiti di filigrana digitale e di divulgazione (ad esempio, dichiarazioni verbali e scritte). Stabilisce nuovi reati relativi a (1) la produzione di *deepfakes* che non soddisfano i requisiti di filigrana digitale o di divulgazione, e (2) l'alterazione di *deepfakes* per rimuovere o significativamente oscurare tali divulgazioni richieste. Il trasgressore è passibile di una multa, fino a cinque anni di carcere o entrambi. Stabilisce inoltre sanzioni civili e consente ai singoli di intentare azioni civili per danni. Inoltre, rivede il reato di frode in relazione a determinati documenti di identificazione per includere *deepfakes*.

[26] A. ORLANDO, *op. cit.*, p. 318.

[27] A. LONGO, *Il Parlamento europeo approva l'AI Act, cosa cambierà per le nostre aziende?, Il Sole 24 Ore*, 2023.

[28] S. TROZZI, *op. cit.*, pp. 226-228.

[29] *Ivi*, pp. 229-230.

rischio, che necessitano di requisiti stringenti per i sistemi e di specifici obblighi per gli operatori; basso rischio, che si configurano come sistemi leciti con limitati requisiti di trasparenza; rischio minimo specifico per la trasparenza, che presuppone obblighi meramente informativi.[30] L'art. 5 dell'AI Act stabilisce i divieti connessi ai rischi considerati inaccettabili per la salute e sicurezza, oltre che per i diritti fondamentali. Dal 2 febbraio 2025 il divieto dei sistemi IA a rischio inaccettabile è diventato applicabile in tutta Europa. [31]

L'Italia ha recepito le fonti legislative e giurisprudenziali comunitarie, le quali hanno contribuito a definire la disciplina italiana del diritto d'autore che fa capo alla legge 633/1941. I cosiddetti "diritti connessi al diritto d'autore" tutelano l'esposizione, la riproduzione e la commercializzazione dell'immagine di una persona.[32] Affinché sia possibile usufruire dell'immagine di un individuo, è necessario ottenere il consenso della persona ritratta, salvo quanto previsto dall'art. 97 della legge 633/1941[33].

Tuttavia, non esiste ancora una legislazione specifica sui *deepfake*; in Italia, infatti, la normativa si basa principalmente sulle disposizioni del GDPR e del Codice penale.[34] Nello specifico, l'art. 612-quater del Codice penale, in materia di manipolazione artificiale di immagini di persone reali allo scopo di ottenerne rappresentazioni nude, tutela i cittadini contro l'uso illecito dell'immagine altrui e interviene per contrastare il fenomeno dei "*deepnude*"[35]. Il fulcro di questa disposizione risiede nella criminalizzazione di condotte di invio, cessione, pubblicazione e diffusione di immagini di persone reali e identificabili che siano state manipolate artificialmente. La

[30] *Ivi*, pp. 230, 239-240.
[31] M. CASADEI, *Intelligenza artificiale, moda al test dell'AI Act: aziende in ritardo. Servono policy e formazione, Il Sole 24 Ore*, febbraio 2025.
[32] G. CASSANO, B. TASSONE, C. GALLI, V. FRANCESCHELLI, *Diritto industriale e diritto d'autore nell'era digitale, Giuffrè*, 2022, pp. 767-797.
[33] Art. 97 legge 633/1941: "Non occorre il consenso della persona ritrattata quando la riproduzione dell'immagine è giustificata dalla notorietà o dall'ufficio pubblico coperto, da necessità di giustizia o di polizia, da scopi scientifici, didattici o culturali, o quando la riproduzione è collegata a fatti, avvenimenti, cerimonie di interesse pubblico o svoltisi in pubblico. Il ritratto non può tuttavia essere esposto o messo in commercio, quando l'esposizione o messa in commercio rechi pregiudizio all'onore, alla reputazione od anche al decoro della persona ritrattata".
[34] F. V. VALENTI, *op. cit.*, p. 17.
[35] I *deepnude* sono immagini generate artificialmente, spesso tramite intelligenza artificiale o tecniche di deep learning, che rimuovono digitalmente gli indumenti da foto di persone reali (generalmente donne), creando rappresentazioni false e sessualmente esplicite senza il consenso del soggetto ritratto.

manipolazione deve avvenire attraverso l'impiego di strumenti tecnologici o sistemi di intelligenza artificiale, e la sua finalità specifica deve essere quella di ottenere rappresentazioni nude delle persone ritratte, capaci di indurre in errore chi le osserva. L'elemento della finalità ingannevole della rappresentazione nuda è cruciale, in quanto distingue questa fattispecie da altre possibili forme di manipolazione di immagini.[36] La sanzione prevista per tale condotta, salvo che il fatto non costituisca un reato più grave, si attesta sulla reclusione da due a sette anni e una multa dai 6.000 ai 16.000 euro.

Un aspetto importante da sottolineare sono le circostanze aggravanti specificamente previste: la pena aumenta se il reato è commesso dal coniuge, anche in caso di separazione o divorzio, o da una persona che è o è stata legata da relazione affettiva alla persona offesa, riconoscendo la particolare vulnerabilità che può derivare da dinamiche relazionali pregresse o in corso. Ulteriormente, l'aggravante si applica se il reato è commesso attraverso strumenti informatici o telematici, essendo incontrollata e immediata la diffusione che tali mezzi possono favorire.[37]

L'ordinamento italiano, in realtà, già prevedeva tutele attraverso l'art. 10 del Codice civile[38], il quale vieta l'esposizione o la pubblicazione dell'immagine altrui fuori dai casi consentiti dalla legge o con pregiudizio al decoro o alla reputazione. L'art. 612-quater si distingue proprio per la sua specifica attenzione alla manipolazione artificiale finalizzata alla creazione di rappresentazioni nude ingannevoli. Questo elemento di "artificialità" e la specifica finalità di ottenere una nudità ingannevole, resa tecnicamente sempre più realistica dai *deepfake*, rappresentano il cuore della nuova fattispecie di reato. In tal senso, il legislatore sembra voler scongiurare in modo più diretto i rischi specifici connessi alla diffusione di *deepfake* a contenuto sessualmente esplicito, riconoscendo la loro particolare capacità di ledere la dignità, l'onore e la reputazione delle persone in maniera insidiosa e potenzialmente virale.[39]

---

[36] Camera dei deputati, A.C. 2986, cit.

[37] *Ibidem.*

[38] Art. 10 del Codice civile: "Qualora l'immagine di una persona o dei genitori, del coniuge o dei figli sia stata esposta o pubblicata fuori dei casi in cui l'esposizione o la pubblicazione è dalla legge consentita, ovvero con pregiudizio al decoro o alla reputazione della persona stessa o dei detti congiunti, l'autorità giudiziaria, su richiesta dell'interessato, può disporre che cessi l'abuso, salvo il risarcimento dei danni".

[39] Camera dei deputati, A.C. 2986, cit.

Sempre in Europa, nel Regno Unito, l'*Online Safety Act* rappresenta il principale strumento normativo per regolamentare i contenuti *online* dannosi, compresi i *deepfake*. Questa normativa impone alle piattaforme digitali di adottare misure per ridurre i rischi legati alla diffusione di contenuti manipolati e vieta specificamente la creazione e la condivisione di *deepfake* con intenti dannosi, come la diffamazione o la violazione della *privacy*.[40] Inoltre, sul territorio britannico, le vittime di *deepfake* possono intraprendere azioni legali basandosi su normative esistenti in materia di protezione dei dati, diffamazione e molestie *online*. L'Autorità per la regolamentazione dei media digitali (Ofcom)[41] è stata incaricata di supervisionare l'applicazione di queste norme per garantire che le piattaforme rispettino gli obblighi previsti.

In Spagna, un ruolo fondamentale in questo ambito è svolto dall'*Agencia Española de Supervisión de la Inteligencia Artificial* (AESIA), istituita per garantire che lo sviluppo e l'uso dell'intelligenza artificiale in territorio spagnolo siano conformi ai principi etici e normativi europei. L'AESIA si occupa di supervisionare l'implementazione delle tecnologie IA, compresi i *deepfake*, assicurando che il loro utilizzo rispetti la *privacy* e la sicurezza delle persone. L'agenzia promuove lo sviluppo di strumenti per il rilevamento e la prevenzione dell'abuso dei *deepfake*, collaborando con altre istituzioni europee per armonizzare le normative e migliorare la protezione dei dati.[42]

Al di fuori dell'Europa, la Cina è uno dei Paesi con le normative più restrittive sui *deepfake*: dal 2020, il governo ha imposto alle piattaforme digitali di segnalare in modo evidente i contenuti generati da IA, vietando l'uso dei *deepfake* per la disinformazione e altre attività illecite.[43] Tali video devono essere esplicitamente contrassegnati con un *disclaimer* o censurati se danneggiano gli "interessi nazionali" del Partito Comunista Cinese. Appare, quindi, evidente che le normative cinesi siano molto più stringenti di

[40] LATHAM & WATKINS, *UK Online Safety Act 2023: A primer on the new law for relevant service providers*, 2024, pp. 4-9.
[41] L'Ofcom (*Office of Communications*) è l'autorità regolatrice indipendente delle comunicazioni nel Regno Unito. È un'agenzia governativa che si occupa di regolazione e concorrenza nei settori delle telecomunicazioni, dell'emittenza, dell'internet e dei servizi postali.
[42] *Arriverà quest'anno l'agenzia spagnola per l'intelligenza artificiale, Notizie.AI*, 2023.
[43] A. ORLANDO, *op. cit.*, p. 322.

quelle europee, preferendo un approccio di censura preventiva rispetto all'atteggiamento più permissivo adottato nella sfera occidentale.[44]

In Giappone e Corea del Sud, invece, la regolamentazione è ancora in fase di sviluppo, con un'attenzione particolare ai rischi legati all'uso dei *deepfake* nel settore politico e nel *revenge porn*[45].

1.3 Le origini e gli sviluppi del *deepfake*

I *deepfake* sono nati negli anni '90 con lo sviluppo di tecnologie di manipolazione, guadagnando, tuttavia, popolarità solo negli ultimi anni.[46] Si possono far risalire i primi sviluppi del fenomeno già al 1997, quando Christoph Bregler, Michele Covell e Malcolm Slaney svilupparono il programma "*Video Rewrite*" che si occupava di modificare riprese video aggiungendo dettagli non presenti nel video originale. In sostanza, questa tecnologia modificava video esistenti implementando dettagli e particolari, ma non sfruttava propriamente l'intelligenza artificiale. Utilizzava, piuttosto, tecniche di *computer vision*, come il *tracking* per tracciare i movimenti effettuati dalla bocca, ed il *morphing* per trasformare questi movimenti e trasportarli nell'*output* finale.[47]

La terminologia "*deepfake*" è stata adottata per la prima volta nel 2017, quando un utente anonimo di *Reddit* con il nickname "*Deepfakes*" ha iniziato a pubblicare video falsi ritraenti attrici famose di Hollywood in contesti pornografici.[48] Questi video sono stati creati utilizzando un codice *open source*, rendendo la tecnologia accessibile a chiunque avesse una buona conoscenza di *machine learning* e programmazione.[49]

In poco tempo, altri utenti hanno iniziato a replicarne l'idea, facendo germogliare sentimenti contrastanti riguardo a questa nuova tecnologia: se, da un lato, la rapida

---

[44] F. V. VALENTI, *op. cit.*, p. 14.

[45] Con *revenge porn* si intende la diffusione, senza il consenso della persona ritratta, di immagini o video a contenuto sessualmente esplicito originariamente realizzati in un contesto privato. Tale condotta è finalizzata spesso a danneggiare o vendicarsi dell'ex *partner* e costituisce reato ai sensi dell'art. 612-ter del Codice penale.

[46] F. ARRUZZOLI, *Deepfake – Significato, Storia, evoluzione, ICT Security Magazine*, 2022.

[47] C. BREGLER, M. COVELL, M. SLANEY, *Video Rewrite: Visual Speech Synthesis from Video, ISCA Speech*, 1997, pp. 153-156.

[48] F. ARRUZZOLI, *Deepfake – Significato, Storia, evoluzione, ICT Security Magazine*, 2022.

[49] R. SONG, *op. cit.*, pp. 161, 165, 170.

diffusione dei video *deepfake* ha entusiasmato molte persone, dall'altro ha portato a nutrire una serie di preoccupazioni per la sicurezza dell'individuo e per il potenziale uso improprio della tecnologia. In particolare, la diffusione di video *deep porn* ha sollevato questioni legali e preoccupazioni per la reputazione e la dignità delle vittime.[50] *Reddit* ha, in seguito, bloccato l'account "*Deepfakes*" a causa della violazione delle *policy* della *community* per la pubblicazione di video *Not Safe For Work* (NSFW).

Tuttavia, il merito dell'esplosione reale dei video *deepfake* va all'attore e regista statunitense Jordan Peele, il quale nel 2018 ha creato un video che ritraeva protagonista l'*ex* presidente degli Stati Uniti, Barack Obama. Nel suddetto video Obama sembra pronunciare parole mai dette, rivolgendo insulti ad altri personaggi pubblici. Questo esperimento era volto a dimostrare come la tecnologia dei *deepfake* possa essere impiegata per diffondere false dichiarazioni attribuite a personaggi noti, con possibili conseguenze sul piano politico e sociale.[51] E, di fatto, è riuscita a destare preoccupazioni persino da parte di *intelligence* governative, le quali hanno colto il potenziale rischio dell'utilizzo della tecnologia in scenari di propaganda politica, disinformazione e *cyberwarfare*.[52]

Per contrastare l'eccessiva propagazione di *deepfake*, diverse aziende, tra cui *Facebook* (ora *Meta*), hanno avviato delle iniziative proprie: nel 2019, ad esempio, *Facebook* ha lanciato la *Deepfake Detection Challenge* (DFDC), una competizione per accelerare lo sviluppo di nuovi metodi per rilevare i video *deepfake*.[53]

In questi anni la tecnologia alla base dei *deepfake* si è evoluta rapidamente, passando da contenuti multimediali processati con lunghi tempi di montaggio a video prodotti in tempo reale, in grado di cambiare l'immagine del volto di un soggetto durante *video call* o *live streaming*. Questa evoluzione ha portato gli esperti di sicurezza delle informazioni a classificare la tecnologia *deepfake* tra le nuove e più pericolose *cyber* minacce.[54]

---

[50] D. SCOTT, *Deepfake Porn Nearly Ruined My Life, Elle*, 2020.
[51] F. RAMOS, *op. cit.,* p. 366.
[52] F. ARRUZZOLI, *Deepfake – Significato, Storia, evoluzione, ICT Security Magazine*, 2022.
[53] B. DOLHANSKY, R. HOWES, B. PFLAUM, N. BARAM, C. CANTON FERRER, *op. cit.*, pp. 1-4.
[54] F. GALVANO, L. BADIALI, *op. cit.*, pp. 8-14.

Inoltre, si sono diffusi su larga scala anche *deepfake* audio: nel 2019, ad esempio, è stato segnalato un caso in cui la voce dell'amministratore delegato di una società tedesca è stata ricreata con un programma di intelligenza artificiale per truffare un altro dirigente.[55]

Data la crescente preoccupazione per i rischi associati ai *deepfake*, diversi Paesi e organizzazioni stanno lavorando per regolamentare l'uso di questa tecnologia.[56] In Italia, il Garante della Protezione dei Dati ha emesso nel 2020 un *Vademecum* per informare i cittadini sui rischi dell'uso malevolo di questa tecnologia, sottolineando l'importanza di un controllo normativo per evitare abusi.[57]

Oggi, la tecnologia *deepfake* viene utilizzata in diversi settori, tra cui il cinema, la pubblicità e l'educazione, ma continua a rappresentare una sfida etica e legale. Mentre alcuni esperti vedono nei *deepfake* un'opportunità per innovare il settore dell'intrattenimento e della comunicazione[58], altri sottolineano i rischi associati alla disinformazione e alla manipolazione dell'identità digitale.[59]

2. Tipologie di *deepfake*: tecniche, applicazioni e rischi. Innovazione o minaccia digitale?

I *deepfake* si suddividono in diverse tipologie, a seconda delle tecniche utilizzate per la loro realizzazione e degli scopi per cui vengono impiegati.[60] Dal punto di vista tecnico, i *deepfake* possono essere suddivisi in tre principali categorie.

La prima categoria è la *Full Image Synthesis*, la quale comprende immagini e video generati *ex novo*, e che quindi non necessitano di *input* esterni. Si possono produrre contenuti del tutto nuovi e spesso indistinguibili da quelli reali attraverso le reti neurali, di cui la *Generative Adversarial Networks* (GAN) è emblematica. Le GAN sono, di fatto, tra le tecniche di intelligenza artificiale maggiormente utilizzate con lo scopo di creare *deepfake* verosimili. Risultano così efficienti poiché in grado di limitare gli errori

---

[55] G. ZHENG, J. SHU, K. LI, *Regulating deepfakes between Lex Lata and Lex ferenda - a comparative analysis of regulatory approaches in the U.S., the EU and China, Crime, Law and Social Change*, 2024, pp. 2-3.
[56] S. TROZZI, *op. cit.*, p. 236.
[57] Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020, pp. 1-6.
[58] A. ORLANDO, *op. cit.*, pp. 307-308.
[59] S. TROZZI, *op. cit.*, pp. 230-231.
[60] B. KIRA, *When non-consensual intimate deepfakes go viral: The insufficiency of the UK Online Safety Act, Computer Law & Security Review 54*, 2024, p. 2.

superficiali del *deep learning* mediante l'implementazione di modelli che in certo qual modo si sfidano fra loro.[61]

La seconda categoria è quella della *Conditional Image Manipulation*: i *deepfake* appartenenti a questa categoria scaturiscono dalla manipolazione di immagini e video esistenti, reali, originali. Questo sistema di intelligenza artificiale modifica il contenuto originale, alterando volti, espressioni o caratteristiche fisiche per generare nuovi *output*, distaccandosi totalmente da quella che era l'immagine veritiera utilizzata in primo luogo.

Infine, il *Face Swapping* è una delle applicazioni più diffuse, consistendo nel sovrapporre la faccia di un soggetto sul corpo di un altro, cambiando di fatto l'identità al corpo che si sta utilizzando.[62] Parallelamente, gli audio *deepfake* sfruttano il *machine learning* per clonare voci esistenti riproducendone fedelmente il tono e far loro pronunciare frasi mai dette realmente.[63]

Quando si pensa ad un *deepfake*, questo generalmente scaturisce due possibili reazioni: ilarità, se associato a video satirici, o timore, quando utilizzato per minare la reputazione altrui o a scopi pornografici.[64] Ma, a dire il vero, nonostante la cattiva reputazione che circonda i *deepfake*, questa tecnologia può essere impiegata in molteplici contesti positivi e legittimi. Uno di questi è senza dubbio l'industria cinematografica e dell'intrattenimento, ambiente in cui l'uso dei *deepfake* permette di ringiovanire gli attori, ricreare volti di celebrità scomparse e migliorare gli effetti speciali senza dover ricorrere a complesse e costose tecniche di post-produzione.[65]

Allo stesso modo, il *deepfake* viene frequentemente impiegato sia per il doppiaggio in diverse lingue, sia per la traduzione automatica. Per quanto riguarda il secondo caso, un esempio emblematico è quello di *Spotify*, la piattaforma musicale che sta già sperimentando l'uso di *deepfake* vocali per tradurre automaticamente i *podcast*, mantenendo intatta la voce originale del *creator*.

---

[61] F. GALVANO, L. BADIALI, *op. cit.*, pp. 8-14.
[62] F. RAMOS, *op. cit.,* p. 366.
[63] *What is Machine Learning?*, *IBM Cloud Education*, 2020.
[64] F. GALVANO, L. BADIALI, *op. cit.*, pp. 15-16.
[65] A. ORLANDO, *op. cit.*, p. 321.

Un'altra categoria interessata è quella dell'educazione e della formazione, dove i *deepfake* appaiono notevolmente utili per creare simulazioni realistiche in ambito accademico, migliorando la capacità di apprendimento attraverso la realtà aumentata e l'uso di intelligenza artificiale.

Se, dunque, da un lato i *deepfake* offrono molteplici opportunità innovative, dall'altro continuano a rappresentare una minaccia concreta per la sicurezza e l'integrità dell'informazione.[66]

All'apice di queste minacce vi sono la disinformazione e le *fake news*: la possibilità di manipolare video e audio con estrema facilità ha reso i *deepfake* un'arma potente per diffondere informazioni false, specialmente in ambito politico ed elettorale. Basti pensare alla propagazione di innumerevoli interviste false che diffondono dichiarazioni (generalmente a discapito di personaggi politici) mai realmente rilasciate. O ancora, ai video umoristici o cosiddetti "*meme*", video con finalità ironiche o parodistiche che spesso vedono come protagonisti i politici rivolgersi in termini scortesi e talvolta volgari alla propria controparte, termini mai utilizzati realmente.[67]

Appare evidente come i *deepfake* rischino di pregiudicare il processo democratico, diffondendo disinformazione politica o impersonando candidati, con l'obiettivo di influenzare l'opinione pubblica e ingannare gli elettori: nell'eventualità in cui il pubblico non riesca a contraddistinguere un falso dalla realtà, questo potrebbe gravare fortemente sulla scelta di voto, andando a sfalsare le preferenze politiche dell'elettore.[68] Come dichiara il Garante della Privacy: "*I deepfake possono riguardare politici o opinion leader, con lo scopo di influenzare l'opinione pubblica. Video deepfake possono ad esempio essere mostrati o inviati agli elettori che simpatizzano per un determinato personaggio politico, rappresentandolo mentre compie azioni poco lecite o mentre si trova in situazioni sconvenienti, allo scopo di screditarlo ed influenzare le opinioni o il*

[66] *Ibidem.*
[67] T. RAMLUCKAN, *Deepfakes: The Legal Implications*, *Proceedings of the 19th International Conference on Cyber Warfare and Security*, 2024, pp. 282-283, 286-287.
[68] *Ivi*, pp. 282-287.

voto. In questo modo, i deepfake possono purtroppo contribuire alla diffusione di fake news e alla disinformazione".[69]

Non è un caso, invero, che l'AI Act preveda la possibilità che i *deepfake* rientrino nella categoria ad alto rischio a causa dell'alto potenziale di manipolazione di contenuti politici o elettorali.[70]

Un'area particolarmente preoccupante è quella del cosiddetto "*deep porn*", ovvero l'uso di questa tecnologia per creare contenuti pornografici falsi, spesso senza il consenso delle persone coinvolte.[71] Questa fattispecie ha un'implicazione fortemente grave, che è quella del *revenge porn*: ossia la "*diffusione nella rete di immagini sessualmente esplicite, senza il consenso del soggetto ritratto, che di solito è una donna, da parte di individui che intendono così denigrare l'ex partner*". [72] A dire la verità, il fenomeno del *deepfake* si è generato originariamente proprio con l'intento di realizzare e propagare materiale pornografico.[73] In particolari tipologie di *deepfake*, dette "*deepnude*", persone ignare possono essere rappresentate nude, in pose discinte, in situazioni compromettenti (ad esempio, a letto con presunti amanti) o addirittura in contesti pornografici.[74]

I video *deepfake* possono altresì essere creati *ad hoc* per realizzare veri e propri atti di *cyber*bullismo, che hanno come vittime soprattutto i più giovani. Un *deepfake* può essere realizzato al fine di denigrare, irridere e screditare le persone coinvolte, o addirittura per ricattarle, chiedendo soldi o altro in cambio della mancata diffusione del video, oppure per la sua cancellazione, quando già stato diffuso in rete.[75]

Il *Cybercrime* riguarda, infine, le attività illecite *online*: è il caso di frodi e truffe digitali. Si definisce *spoofing* quella "*attività illecita su Internet che consiste nel furto di identità. Si esplicita attraverso la falsificazione di indirizzi di siti web e del contenuto di questi*

---

[69] Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020, pp. 1-6.
[70] A. RUFFO, *Il disordine informativo e l'Intelligenza Artificiale; tra insidie e possibili strumenti di contrasto, Medialaws*, 2024, pp. 419-420.
[71] T. RAMLUCKAN, *op. cit.*, p. 283.
[72] "*Revenge porn*" in Enciclopedia Treccani Online, Istituto della Enciclopedia Italiana, https://www.treccani.it/vocabolario/deepfake_(Neologismi)/ (consultato il: 26/02/2025).
[73] F. ARRUZZOLI, *Deepfake – Significato, Storia, evoluzione, ICT Security Magazine*, 2022.
[74] Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020, pp. 1-6.
[75] *Ibidem.*

ultimi, sovente clonati, per indurre il navigatore a credere di essere nel sito web cercato mentre invece si trova in un sito copiato e falso, nel tentativo di carpirgli con l'inganno informazioni e dati personali, quali per es. numeri di conti correnti bancari o password. Quest'ultima condotta può anche prendere il nome di phishing."[76]

Questa attività viene svolta tramite l'utilizzo di un *ransomware*, ossia un "*programma maligno che limita o impedisce l'accesso al dispositivo sul quale si installa a insaputa dell'utente, richiedendo un riscatto da pagare per ripristinare l'uso normale del dispositivo.*" [77]

Il furto di identità tramite *deepfake* è una forma particolarmente grave di furto di identità, per quanto semplice da attuare. Nel momento in cui una persona compare in un *deepfake* a sua insaputa deve automaticamente rinunciare non solo ad una perdita di controllo sulla propria immagine, ma viene privata anche del controllo che ha sulle proprie idee e pensieri, facili da travisare mediante i discorsi e i comportamenti falsi attuati nei video di cui è vittima. In questo contesto volti e voci artefatti possono essere utilizzati per ingannare i sistemi di sicurezza basati su dati biometrici vocali e facciali: è il caso in cui il *deepfake* può arrivare a privare le persone della cosiddetta "autodeterminazione informativa", che consiste nella capacità del singolo di poter decidere che informazioni riguardanti la sua persona voglia o meno diffondere.[78]

2.1 *Deepfake* per intrattenimento: *meme* e satira, tra creatività e controversie

Come anticipato, i *deepfake*, pur avendo sollevato numerose questioni etiche e legali, hanno trovato ampio utilizzo nel settore dell'intrattenimento. Nello specifico, uno degli ambiti in cui si sono maggiormente diffusi è quello dei *meme* e dei contenuti ironici: l'uso di questa tecnologia per scopi comici e satirici ha dato origine a un nuovo e distintivo genere di contenuti digitali, che sfruttano la manipolazione audiovisiva per creare effetti esilaranti o parodici. [79] I *deepfake* in questo contesto sono sempre video o immagini creati

[76] "*Spoofing*" in Enciclopedia Treccani Online, Istituto della Enciclopedia Italiana, https://www.treccani.it/vocabolario/deepfake_(Neologismi)/ (consultato il: 26/02/2025).
[77] "*Ransomware*" in Enciclopedia Treccani Online, Istituto della Enciclopedia Italiana, https://www.treccani.it/vocabolario/deepfake_(Neologismi)/ (consultato il: 26/02/2025).
[78] Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020, pp. 1-6.
[79] S. TROZZI, *op. cit.*, p. 240.

o modificati con tecniche di intelligenza artificiale per sovrapporre il volto di una persona su un altro corpo o per alterare espressioni facciali e discorsi, ma con la peculiarità di essere a carattere umoristico.[80]

Il neologismo "*meme*" è stato coniato dal biologo inglese Richard Dawkins nel 1976 nel suo libro "*The Selfish Gene*". Dawkins definisce il *meme* come un'unità di trasmissione culturale o un "gene culturale". Il biologo britannico fa riferimento ai "nuovi replicatori" culturali, ovvero elementi che si diffondono attraverso l'imitazione e la riproduzione.[81] Con l'avvento di internet, i *meme* si sono evoluti in una forma di espressione digitale che utilizza immagini, video e testi per trasmettere concetti a carattere ludico, umoristico, parodico o satirico, tramite manipolazione e rielaborazione semantica di testi preesistenti. I *deepfake* si inseriscono perfettamente in questo contesto, offrendo strumenti avanzati per alterare i volti e le voci dei protagonisti di video virali, film famosi o discorsi pubblici, generando nuovi contenuti con un impatto comico.[82] I *meme* di internet sono testi appartenenti a diverse forme espressive come immagini statiche o animate, porzioni di testo e video.

Uno degli esempi più celebri di *deepfake* a scopo ironico è il video "*Home Stallone*", in cui il volto di Sylvester Stallone viene sovrapposto a quello del giovane protagonista del film "*Home alone*" (Mamma ho perso l'aereo). Il risultato, reso ancora più divertente dal gioco di parole con il titolo, ha riscosso enorme successo *online*, con milioni di visualizzazioni su *YouTube*. Questo tipo di contenuti ottiene estrema viralità, e generalmente in tempi record, poiché sfrutta il paradosso visivo per generare ilarità, creando accostamenti assurdi e imprevedibili tra volti celebri e contesti inaspettati, che tendono ad attirare una grandissima fetta di pubblico.

Un altro esempio noto di *deepfake* umoristico, tornando in Italia, è rappresentato dai video creati dal programma televisivo "Striscia la Notizia". Tra gli innumerevoli video di *deepfake* caricati nell'apposita rubrica presente sul loro sito *web*, uno particolarmente emblematico risale al 2019 e ritrae l'*ex premier* italiano Matteo Renzi mentre pronuncia dichiarazioni ironiche e surreali; sebbene il video fosse chiaramente satirico, alcuni

[80] F. V. VALENTI, *op. cit.*, p. 16.
[81] R. DAWKINS, *The Selfish Gene, Oxford University Press*, 2006.
[82] B. KIRA, *When non-consensual intimate deepfakes go viral, op. cit.*, pp. 2-3.

spettatori hanno creduto che fosse reale, sollevando polemiche sulla necessità di etichettare chiaramente i contenuti *deepfake* per evitare eventuali fraintendimenti.

Se da un lato i *deepfake* hanno un grande potenziale creativo e possono contribuire a una nuova forma di espressione artistica e comica, dall'altro sollevano interrogativi sull'autenticità delle informazioni e sulla necessità di regolamentare il loro uso, soprattutto per combattere quanto più efficacemente possibile la diffusione di disinformazione. Anche se creati a scopo umoristico, tali contenuti possono, e talvolta mirano, ad essere scambiati per veri e diffondere notizie false: un video ironico può essere percepito come una notizia reale, portando a fraintendimenti e causando potenziali danni irreparabili all'immagine degli individui che ne sono vittime.[83] Seppur in grado di generare ilarità, non si può trascurare il rischio che l'utilizzo improprio dei *deepfake* possa ledere la reputazione e la dignità personale delle persone coinvolte, nonché far perdere il controllo della propria narrativa circa l'immagine che si vuole proiettare di sé al mondo.[84]

Per questo motivo, molte piattaforme *social* hanno iniziato a implementare strumenti per segnalare i contenuti *deepfake* e informare gli utenti sulla loro natura artificiale. L'AI Act prevede requisiti minimi di trasparenza per i *deepfake*, obbligando a indicare la natura artificiale dei contenuti: non vieta in alcun modo la creazione di questi, ma impone chiarezza a chi li realizza e chi li diffonde, essendo necessario etichettare esplicitamente i contenuti come artificiali o manipolati digitalmente.[85] È compito degli stessi utenti segnalare contenuti sospetti alle piattaforme che li ospitano e rivolgersi alle autorità competenti in caso di reati o violazioni della *privacy*.[86]

2.2 *Deepfake* e *fake news*: rischi per l'informazione pubblica

L'avvento dei *deepfake* ha avuto un impatto ancor più significativo sulla propagazione delle *fake news*. La possibilità di manipolare immagini, video e audio con un risultato via via più realistico ha reso ormai difficile distinguere tra contenuti autentici e contenuti

[83] F. V. VALENTI, *op. cit.*, p. 13.
[84] S. TROZZI, *op. cit.*, p. 242.
[85] *Ivi*, p. 240.
[86] Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020, pp. 1-6.

falsificati, compromettendo l'affidabilità delle fonti di informazione e la fiducia nei media.[87]

Il maggior rischio correlato alla diffusione di *deepfake* nel contesto delle *fake news* è chiaramente la disinformazione politica: video alterati possono essere utilizzati per diffondere dichiarazioni false attribuite a *leader* politici o altre figure pubbliche, con il potenziale di influenzare elezioni, creare tensioni internazionali o incitare al disordine sociale. Non è un caso che durante le campagne elettorali vengano diffusi numerosi video *deepfake*, in cui candidati vengono falsamente mostrati mentre pronunciano discorsi o compiono azioni compromettenti, alterando la percezione dell'opinione pubblica, e di conseguenza anche la scelta di voto.[88]

Un esempio particolarmente significativo è il *deepfake* del presidente ucraino Volodymyr Zelensky diffuso nel 2022, nel quale veniva mostrato mentre ordinava alle forze armate di deporre le armi di fronte ai militari russi. Nonostante fosse stato dichiarato falso, questo video è stato ampiamente condiviso, e in un momento storico così delicato per l'Ucraina, quel *deepfake* ha mostrato la capacità di avere un impatto diretto sulla geopolitica e sulla percezione degli eventi globali.[89]

Oltre alla politica, anche il giornalismo è fortemente colpito da questo fenomeno: la capacità di generare contenuti falsificati ha reso più complesso il lavoro di verifica delle fonti per le redazioni giornalistiche, aumentando il rischio di diffusione involontaria di informazioni errate. Le testate giornalistiche devono ora adottare strumenti di rilevamento avanzati e collaborare con esperti di intelligenza artificiale per garantire l'autenticità delle notizie pubblicate.

Non bisogna trascurare che con l'avvento del digitale, si sia dato avvio ad una proliferazione di contenuti informativi offerti dalla rete, creando un'informazione sempre più orizzontale con la possibilità per l'utente di esprimere il proprio pensiero.[90] Con regole sempre più destrutturate, le redazioni giornalistiche si ritrovano, dunque, a

[87] S. TROZZI, *op. cit.*, p. 246.
[88] T. RAMLUCKAN, *op. cit.*, p. 282.
[89] F. GALVANO, L. BADIALI, *op. cit.*, pp. 15-18.
[90] *Ivi*, pp. 17-18.

contrastare la disinformazione soltanto attraverso il *fact-checking*, a tal punto che la Commissione Europea ha evidenziato la necessità di creare una vera e propria rete di *fact-checkers*.[91]

L'impatto dei *deepfake* sull'informazione pubblica rappresenta una grande minaccia per la sicurezza sociale: la diffusione di contenuti manipolati può generare panico o creare falsi allarmi, anche in contesti in cui si può non essere sufficientemente preparati per gestire imminenti emergenze.[92] Un esempio lampante è rappresentato da quei video che annunciano catastrofi inesistenti o diffondono messaggi falsi attribuiti ad autorità sanitarie, creando un panico generale che mira a compromettere la risposta statale a situazioni critiche e a minare ulteriormente la fiducia nelle istituzioni.

Per contrastare i suddetti rischi, le piattaforme *social* (spesso il principale veicolo di diffusione di questi contenuti) stanno introducendo algoritmi di rilevamento e segnalazione per avvisare gli utenti della possibile manipolazione di immagini, video e audio.[93] Anche l'educazione digitale gioca un ruolo fondamentale nella lotta contro la disinformazione generata dai *deepfake*: sensibilizzare il pubblico sui rischi di questi contenuti e insegnare strategie per riconoscere le manipolazioni può contribuire a mitigare l'impatto delle *fake news* sulla società.[94]

2.3 *Deep porn*: problematiche etiche e violazioni della *privacy*

L'utilizzo improprio dei *deepfake* è ancora più pericoloso nel momento in cui sfocia in *deep porn*. Questa categoria di contenuti digitali si basa sulla manipolazione di immagini e video pornografici per inserirne i volti di persone reali (spesso celebrità o talvolta anche individui comuni) senza il loro consenso.[95] Secondo una ricerca di CNN Business[96], nel 2019 su internet erano presenti almeno 14.678 video *deepfake*, di cui il 96% erano video pornografici.

---

[91] A. RUFFO, *op. cit.*, pp. 408-409.
[92] S. TROZZI, *op. cit.*, p. 246.
[93] A. RUFFO, *op. cit.*, pp. 408-409.
[94] L. METSELAAR, *Framing Deepfake Technology in European Union Governance: Discursive strategies and regulatory responses to deepfake technology, Management Society and Technology Universiteit Twente*, 2025, pp. 17-18.
[95] B. KIRA, *When non-consensual intimate deepfakes go viral, op. cit.*, pp. 3-4.
[96] *The number of deepfake videos online is spiking. Most are porn, CNN Business*, 2019.

Le problematiche etiche legate al *deep porn* sono molteplici. Innanzitutto, il fenomeno si configura come una grave violazione della dignità personale e del diritto all'immagine: la diffusione di video *deep porn* può ledere la reputazione delle vittime, causando danni significativi alla loro vita privata e professionale, trovandosi queste esposte a umiliazioni pubbliche e discriminazioni. Le persone coinvolte non consensualmente in questi video falsificati spesso riportano conseguenze psicologiche devastanti, inducendo stati di ansia, depressione, vergogna, isolamento sociale e, nei casi più estremi, pensieri suicidari.[97] Oltretutto, frequentemente il *deep porn* serve da trampolino di lancio per l'attuazione del *revenge porn*: laddove non si hanno foto o video intimi reali da condividere con un fine vendicativo e diffamatorio, questi vengono generati mediante intelligenza artificiale.[98]

È bene sottolineare che i *deep porn* vengono creati e diffusi senza il consenso delle persone ritratte, privandole del controllo sulla propria immagine e identità, motivo per cui la produzione degli stessi costituisce una fattispecie particolarmente grave di furto d'identità.[99] Oltre alle implicazioni etiche, il *deep porn* pone, infatti, questioni giuridiche complesse: in molti ordinamenti, la creazione e la distribuzione di contenuti pornografici falsificati senza consenso sono già considerati reati, rientrando nelle fattispecie del *revenge porn* e della diffamazione.[100] Tuttavia, la rapidità con cui questi video si diffondono *online* rende difficile per le autorità intervenire tempestivamente e rimuovere il materiale dannoso. Un esempio emblematico è il caso dell'attivista australiana Noelle Martin, che ha combattuto per anni contro l'uso non consensuale della sua immagine in contenuti *deepfake* pornografici, contribuendo alla criminalizzazione esplicita di questa pratica in diverse giurisdizioni.[101] In realtà i *deep porn* ledono anche il settore pornografico stesso, violando il diritto d'autore dei video pornografici originali, che vengono scaricati illegalmente e modificati senza il permesso dei detentori dei diritti. Infine, il *deep porn* solleva questioni relative alla protezione dei dati personali: è importante notare come la capacità di generare video realistici a partire da semplici

[97] A. MIOTTI, A. WASIL, *Combatting deepfakes: Policies to address national security threats and rights violations, ArXiv*, 2024, p. 12.
[98] K. MANIA, *Legal Protection of Revenge and Deepfake Porn Victims in the European Union: Findings From a Comparative Legal Study, Trauma Violence & Abuse 25(1)*, 2022, p. 117.
[99] A. MIOTTI, A. WASIL, *op. cit.*, p. 9.
[100] K. MANIA, *op. cit.*, pp. 120-122.
[101] D. SCOTT, *Deepfake Porn Nearly Ruined My Life, Elle*, 2020.

immagini disponibili *online* renda questa prassi troppo semplice da attuare e talvolta anche a costi bassissimi.[102]

Questo evidenzia la necessità di una maggiore tutela della *privacy* e di normative più severe per impedire l'uso improprio dell'intelligenza artificiale a fini lesivi. Bisogna adottare un approccio multilivello che combini innovazioni tecnologiche, interventi normativi efficaci e una maggiore consapevolezza sociale. Alcune organizzazioni stanno promuovendo lo sviluppo di tecnologie *blockchain* per tracciare l'origine dei contenuti multimediali e garantire la loro autenticità. E se le piattaforme, dal loro canto, cercano di implementare sempre di più gli strumenti di rilevamento basati sull'intelligenza artificiale per individuare e rimuovere contenuti manipolati, tuttavia la continua evoluzione delle tecniche di *deepfake* rende questa sfida particolarmente ardua.[103] A suo modo, l'Unione Europea, con il già citato AI Act, sta cercando di regolamentare queste pratiche, imponendo restrizioni particolarmente rigide sulla creazione e diffusione di *deepfake* a contenuto sessuale.[104]

3. La tecnologia alla base dei *deepfake*: intelligenza artificiale, GAN e tecniche di manipolazione

La tecnologia alla base dei *deepfake* si fonda su avanzate tecniche di intelligenza artificiale, che sfruttano in particolare l'apprendimento automatico (*machine learning*) e, in misura ancora maggiore, l'apprendimento profondo (*deep learning*). Questi approcci permettono di creare contenuti altamente realistici, dove immagini, video e audio sembrano perfettamente autentici, nonostante siano in realtà sintetici.[105] L'intelligenza artificiale, grazie a modelli complessi di reti neurali, è in grado di analizzare, manipolare e ricreare dettagli tanto accurati da far sì che i *deepfake* risultino indistinguibili dalla realtà.[106]

Il *focus* di questa tecnologia è rappresentato dalle Reti Generative Avversarie (*Generative Adversarial Networks* – GAN), che costituiscono uno degli strumenti più potenti nel

[102] A. ORLANDO, *op. cit.*, p. 321.
[103] A. RUFFO, *op. cit.*, pp. 423-424.
[104] A. ORLANDO, *op. cit.*, p. 314.
[105] T. RAMLUCKAN, *op. cit.*, p. 282.
[106] R. SONG, *op. cit.*, pp. 161-163.

campo della creazione di contenuti falsi ma credibili. Le GAN operano attraverso due reti neurali che collaborano in modo competitivo: una rete (che viene chiamata generatore), ha il compito di produrre immagini o video sintetici, mentre l'altra (nota come discriminatore) valuta la veridicità dei contenuti generati, cercando di identificare se siano reali o falsi. Con il tempo e l'allenamento, il generatore diventa sempre più abile nel creare contenuti via via più realistici, mentre il discriminatore migliora nel rilevare le discrepanze tra il falso e il vero. Si tratta di un processo di continuo affinamento delle capacità delle reti che consente di ottenere *deepfake* così ben fatti da risultare difficili da distinguere dai contenuti reali, anche per un occhio esperto.[107]

Un altro aspetto cruciale nel *deepfake* è il *face-swapping*, una tecnica che si avvale dell'intelligenza artificiale per sostituire il volto di una persona con quello di un'altra in un contenuto multimediale, mantenendo però intatte le espressioni facciali e i movimenti, affinché risultino coerenti con il contesto. Questo è reso possibile tramite l'uso delle Reti Neurali Convoluzionali (*Convolutional Neural Networks* – CNN), che permettono di analizzare e riprodurre con grande precisione le caratteristiche facciali e i dettagli espressivi, rielaborando immagini o sequenze video in modo che il risultato sembri perfettamente naturale.[108]

Come già preannunciato diverse volte, i *deepfake* non si limitano solo alle immagini o ai video, ma si estendono anche al campo dell'audio, dove l'IA viene utilizzata per replicare voci con una sorprendente fedeltà. Tecnologie come il *text-to-speech* avanzato e il *voice cloning* sono in grado di riprodurre con precisione il tono, il ritmo, le inflessioni e le caratteristiche vocali di una persona, creando discorsi sintetici che sembrano pronunciati dalla stessa.[109] In pratica i modelli di IA, addestrati su enormi volumi di dati audio, riescono ad apprendere le caratteristiche specifiche della voce da riprodurre, potendo quindi generare registrazioni vocali indistinguibili da quelle reali.[110]

Proprio per questo motivo, oltre alla creazione, anche il rilevamento e la prevenzione degli abusi legati ai *deepfake* sono diventati temi di grande interesse per le aziende

[107] G. ZHENG, J. SHU, K. LI, *op. cit.*, p. 3.
[108] R. SONG, *op. cit.*, pp. 161-163.
[109] F. GALVANO, L. BADIALI, *op. cit.*, pp. 7-11.
[110] F. RAMOS, *op. cit.,* pp. 366, 370.

tecnologiche e le istituzioni.[111] Tra le soluzioni più avanzate, l'analisi delle anomalie nei video tramite l'intelligenza artificiale ha consentito di sviluppare algoritmi in grado di individuare segnali di falsificazione: un esempio emblematico è *FakeCatcher*, uno strumento sviluppato da *Intel*, che utilizza un'analisi avanzata del flusso sanguigno nei volti presenti nei video per rilevare eventuali manipolazioni con un'affidabilità garantita del 96%.[112]

Google ha implementato, a sua volta, la funzione di ricerca di *Google Lens*, la quale permette agli utenti di verificare la veridicità delle immagini, controllando se siano state pubblicate in precedenza e se siano state modificate tramite IA. Tali strumenti di rilevamento si propongono essenzialmente di contrastare la diffusione di contenuti falsi che potrebbero essere altrimenti utilizzati per scopi dannosi, come la manipolazione dell'opinione pubblica o la violazione della *privacy*.[113]

Dunque, per creare un *deepfake*, si deve acquisire un *dataset* di immagini o video del soggetto da imitare; questi stessi dati vengono poi utilizzati come *input* per un algoritmo di *machine learning*, che identifica i tratti somatici distintivi della persona e crea una maschera digitale. In seguito, questa maschera viene sovrapposta al volto di un'altra persona, risultando in un video che appare genuino, ma che è il frutto di un'elaborazione digitale.[114] La creazione di *deepfake* è resa possibile anche grazie all'uso di Modelli Autoregressivi, di *Variational Autoencoders* (VAE) e di GAN, che migliorano costantemente la qualità del risultato finale.[115]

Più nel dettaglio, i Modelli Autoregressivi lavorano direttamente sull'immagine reale, modellando la distribuzione di ogni *pixel* in base al *pixel* precedente, producendo in sostanza immagini di alta qualità;[116] le VAE, invece, sono modelli di rete neurale utilizzati per l'apprendimento non supervisionato, addestrati a copiare l'*input* nell'*output*. Essendo

---

[111] F. GALVANO, L. BADIALI, *op. cit.*, p. 7.
[112] D. BARBERA, *Il sistema Intel per riconoscere i deepfake con una precisione del 96%. FakeCatcher è una tecnologia sviluppata con la State University di New York che riconosce i falsi dal flusso sanguigno del viso, Wired*, 2022.
[113] F. V. VALENTI, *op. cit.*, p. 18.
[114] R. SONG, *op. cit.*, pp. 161-163, 166.
[115] B. DOLHANSKY, R. HOWES, B. PFLAUM, N. BARAM, C. CANTON FERRER, *op. cit.*, pp. 1-4.
[116] C. BREGLER, M. COVELL, M. SLANEY, *op. cit.*, pp. 153-156.

formati da un *coder* e un *decoder*, comprimono e codificano i dati in ingresso per poi ricostruirli in uscita. Vengono utilizzati principalmente per il rilevamento di anomalie, l'apprendimento di caratteristiche e la rimozione del rumore dall'*input*.[117]

3.1 *Machine learning* e *deep learning*: principi fondamentali

Negli ultimi anni, il progresso delle tecnologie di intelligenza artificiale ha rivoluzionato numerosi settori, compreso quello della manipolazione dei contenuti multimediali. In questo contesto, il fenomeno dei *deepfake* rappresenta una delle applicazioni più controverse e avanzate del *machine learning* e del *deep learning*. Sarà quindi essenziale analizzare e comprendere i meccanismi che si celano dietro queste discipline per poter essere in grado di esaminare come i *deepfake* vengono generati e quali sono le loro implicazioni tecniche ed etiche.[118]

Il *machine learning* (ML) è alla base della tecnologia *deepfake*: si tratta di una branca dell'intelligenza artificiale che si occupa di creare sistemi in grado di apprendere o migliorare le proprie prestazioni in base ai dati ricevuti. In questo modo, l'apprendimento automatico emula esattamente il modo in cui gli esseri umani apprendono, e man mano migliora le proprie *performance*.[119] Il *machine learning* si divide in tre principali categorie: l'apprendimento auto-supervisionato, l'apprendimento non supervisionato e l'apprendimento per rinforzo.[120]

L'apprendimento auto-supervisionato riguarda un modello che viene addestrato su un *dataset* etichettato e impara a generalizzare i dati per compiere previsioni su nuove informazioni; l'apprendimento non supervisionato, si occupa invece di identificare strutture e schemi nascosti all'interno di dati non etichettati; ed infine, l'apprendimento per rinforzo si basa su un sistema di ricompense e penalizzazioni per migliorare progressivamente le prestazioni di un agente autonomo.[121]

---

[117] A. SIAROHIN, S. LATHUILIÈRE, S. TULYAKOV, E. RICCI, N. SEBE, *First Order Motion Model for Image Animation, Advances in neural information processing systems 32*, 2019, p. 4.
[118] F. GALVANO, L. BADIALI, *op. cit.*, pp. 4-6.
[119] *What is Machine Learning?*, *IBM Cloud Education*, 2020.
[120] A. RUFFO, *op. cit.*, p. 416.
[121] F. V. VALENTI, *op. cit.*, pp. 15-16.

Il *machine learning* è a sua volta composto da un suo sottoinsieme, che prende il nome di *deep learning* e utilizza reti neurali artificiali composte da tre o più livelli per l'elaborazione dei dati, cercando, in sostanza, di imitare il funzionamento del cervello umano, apprendendo da grandi quantità di dati e migliorando le proprie prestazioni a ogni esecuzione dell'algoritmo.[122]

Gli algoritmi di *deep learning* hanno trovato ampio impiego nella creazione di *deepfake* grazie alla loro capacità di generare contenuti iperrealistici. Si sono peraltro diffuse varie tecniche che utilizzano questa tecnologia, di cui le principali sono essenzialmente due, già precedentemente citate e che giova ripetere: le Reti Generative Avversarie (GAN), costituite da due reti neurali che competono tra loro nelle figure di generatore e discriminatore, con il fine ultimo di creare immagini sempre più convincenti;[123] e le *Variational Autoencoders* (VAE), le quali invece si occupano della compressione di immagini in rappresentazioni latenti, consentendo modifiche e manipolazioni sofisticate dei contenuti multimediali.[124]

I primi *deepfake*, i cosiddetti "*face-swaps*", si basavano su un tipo di sistema di *deep learning* chiamato "*autoencoder*". Con il tempo, si è passati ad una tipologia più evoluta di sistemi di *deep learning*, scaturita dagli studi del ricercatore statunitense Ian J. Goodfellow: fu proprio Goodfellow a pensare e programmare nel 2014 il nuovo sistema di *deep learning* che ora conosciamo come GAN. Ci troviamo di fronte ad un sistema che si risolve nella contrapposizione di due diversi *deep learning networks*, i quali si sfidano in una sorta di gioco basato sulla generazione di volti umani credibili e coerenti. [125]

L'evoluzione delle tecnologie di *deep learning* ha reso la *creazione* di *deepfake* accessibile a un pubblico sempre più vasto, con strumenti *open-source* che permettono anche a utenti con conoscenze limitate di generare contenuti falsificati.[126] Basti pensare al più recente sviluppo di *software* come *DeepFaceLab* e *FakeApp*, che ha reso possibile

[122] F. GALVANO, L. BADIALI, *op. cit.*, p. 7.
[123] *Ivi*, p. 8.
[124] A. SIAROHIN, S. LATHUILIÈRE, S. TULYAKOV, E. RICCI, N. SEBE, *op. cit.*, pp. 1-2.
[125] I. GOODFELLOW, Y. BENGIO, A. COURVILLE, *Deep learning, The MIT Press,* 2016.
[126] F. ARRUZZOLI, *Deepfake – Significato, Storia, evoluzione, ICT Security Magazine*, 2022.

la creazione di video *deepfake* con una qualità sorprendente, aprendo nuove prospettive ma anche preoccupazioni legate alla diffusione incontrollata dei contenuti in esame.[127]

Per contrastare le implicazioni negative dei *deepfake*, la comunità scientifica e tecnologica sta sviluppando nuove tecniche di rilevazione, basate sullo studio di artefatti digitali e anomalie nei movimenti facciali.[128] In particolare, *Google* e *Facebook* hanno sviluppato algoritmi avanzati di *detection* che si rifanno ai modelli di *machine learning* per identificare *pattern* comuni nei *deepfake* e segnalare i contenuti manipolati.[129]

L'apprendimento automatico è fortemente impiegato nell'ambito dei *social media* dove il *machine learning* è utile a suggerire contenuti rilevanti ai propri utenti che siano affini ai loro interessi sulla base di comportamenti passati. Tuttavia, i *social* non sono l'unico settore in cui l'apprendimento automatico viene utilizzato; basti pensare a banche, siti di *shopping online*. Ad esempio, le banche utilizzano il *machine learning* per rilevare frodi e prevenire transazioni sospette, mentre i siti di *shopping online* ne usufruiscono per fornire raccomandazioni di prodotti personalizzati ai propri clienti.[130]

Tra le applicazioni positive di queste tecnologie, vi sono altresì il miglioramento della sintesi vocale e la ricostruzione di immagini in ambito medico, nonostante spesso il loro utilizzo improprio tenda ad offuscare le potenzialità che queste dispongono.[131]

3.2 Il processo di creazione di un *deepfake*: dalla raccolta dati alla post-produzione

Come anticipato, la creazione di un *deepfake* segue un processo articolato che si sviluppa in più fasi, dalla selezione dei dati iniziali fino alla produzione e post-produzione del video finale.[132] Più nel dettaglio, il primo passo fondamentale riguarda la selezione e la preparazione dei dati. In sostanza, devono essere raccolte immagini o video del soggetto da imitare.[133] Per ottenere un *deepfake* convincente, è, inoltre, necessario un *dataset* di

---

[127] S. TROZZI, *op. cit.*, p. 246.
[128] F. GALVANO, L. BADIALI, *op. cit.*, pp. 10-11.
[129] B. DOLHANSKY, R. HOWES, B. PFLAUM, N. BARAM, C. CANTON FERRER, *op. cit.*, pp. 1-4.
[130] *What is Machine Learning?*, *IBM Cloud Education*, 2020.
[131] J. WILSON, *Deepfake: Post the Bruce Willis Controversy What Disruption To Entertainment Could Be Caused, Forbes*, 2022.
[132] B. DOLHANSKY, R. HOWES, B. PFLAUM, N. BARAM, C. CANTON FERRER, *op. cit.*, pp. 1-4.
[133] Garante per la Protezione dei Dati Personali, *Vademecum,* dicembre 2020, pp. 1-6.

alta qualità, composto da un numero significativo di immagini che ritraggano il soggetto da diverse angolazioni e con varie espressioni facciali: maggiore è la varietà e la quantità di dati a disposizione, migliore sarà la qualità del *deepfake* prodotto.[134] Tali *dataset* di immagini o video del soggetto da imitare, verranno poi utilizzati come *input* per l'algoritmo di *machine learning*: l'algoritmo cercherà, sulla base di questi dati, di individuare i tratti somatici distintivi del volto del soggetto e di creare una maschera digitale in grado di riprodurli fedelmente.[135]

La qualità dei dati raccolti è un elemento fondamentale, da cui dipende la qualità del risultato finale stesso. È preferibile che nel video sia presente un solo soggetto o che comunque non siano ripresi più soggetti nella stessa scena; questo facilita il lavoro del *software* e previene paragoni che potrebbero portare a delle imperfezioni.[136] Il soggetto utilizzato deve avere un volto quanto più simile possibile a quello da sovrapporre. Questo perché il *deepfake* funziona come una via di mezzo tra una maschera e un trucco; se, perciò, il soggetto base ha caratteristiche molto diverse rispetto alla maschera, si potrebbero riscontrare problemi di coerenza dell'immagine nella sovrapposizione.[137] Appare evidente l'importanza di trovare quanti più dati efficienti possibili. Da qui, la ragione per cui le vittime di *deepfake* sono solitamente persone famose, in quanto risulta più semplice e immediato trovare foto, audio e video di politici o celebrità. Inoltre, ritrarre celebrità contribuisce a rendere il video non solo più interessante, ma generalmente virale.

Una volta conclusa la prima fase della raccolta del *dataset*, quest'ultimo viene elaborato da un algoritmo di *deep learning* (spesso una Rete Generativa Avversaria – GAN o un *autoencoder*). Il modello viene addestrato per apprendere le caratteristiche distintive del volto e dei movimenti del soggetto, trattasi di un processo di *training* reiterativo che può richiedere un tempo variabile in base alla potenza computazionale disponibile e alla complessità del *dataset*. Dopo l'addestramento, il modello generato viene applicato ad un cosiddetto "*video target*", sostituendo il volto originale del soggetto con quello sintetizzato. Questa operazione viene realizzata attraverso delle tecniche avanzate di *face-*

[134] B. DOLHANSKY, R. HOWES, B. PFLAUM, N. BARAM, C. CANTON FERRER, *op. cit.*, pp. 1-4.
[135] *What is Machine Learning?*, *IBM Cloud Education*, 2020.
[136] J. WILSON, *Deepfake: Post the Bruce Willis Controversy What Disruption To Entertainment Could Be Caused, Forbes*, 2022.
[137] C. BREGLER, M. COVELL, M. SLANEY, *op. cit.*, pp. 153-156.

swapping e *motion capture*. Come anticipato, il *face-swapping* è una tecnica che permette di sovrapporre il volto generato dal modello a quello di un attore o di un soggetto esistente nel video; mentre la *motion capture* consente di replicare espressioni facciali e movimenti in modo realistico, rendendo la manipolazione praticamente indistinguibile da un video originale. Durante questa fase, il *software* regola dettagli come illuminazione, ombreggiature e proporzioni del viso per garantire che vi sia una perfetta integrazione del volto sintetico nel *video target*.[138]

Una volta che la maschera digitale è stata creata, questa viene sovrapposta al volto della persona di base nel video, in modo da creare un video che possa apparire autentico, quanto più reale possibile, ma che in realtà è il risultato di un'efferata elaborazione digitale. Il volto cosiddetto "di base" definisce i movimenti e viene tracciato dal programma punto per punto. Il secondo volto, ossia quello che funge da "maschera", copre il primo volto e ne ricopia i movimenti. Al riguardo, è stato creato un programma apposito su *Google Colab*: trattasi del programma "*First Order Model*", che permette essenzialmente di analizzare i dati forniti e di rielaborarli in modo da dirigere un volto in base ai movimenti registrati da un altro soggetto.[139]

Nella prima fase, la funzione *display* è utilizzata per creare e visualizzare un'animazione a partire da un'immagine d'origine (*source_image*) e da una sequenza di immagini guida (*driving_video*). Il programma analizza *frame* per *frame* il video "*driver*", tracciandone i movimenti e riportandoli sulla "maschera". Il generatore crea, quindi, i *frame* di *output* basati sul *file* indicato al programma, mentre il *keypoint detector* si occupa di individuare i movimenti del volto nei *frame* del video *driver*, replicandoli sulla sorgente.[140]

First Order Model utilizza librerie come "*imageio"* e *"skimage"* per la lettura, riscrittura e manipolazione delle immagini. La funzione "*make_animation*" contiene sia l'immagine maschera (*source_image*) che la lista dei frame estratti dal video *driver* (*driving_video*), oltre ai modelli già addestrati. L'*output* che ne risulta sono *frame* rinvenibili dall'animazione che vengono riuniti e convertiti in un video e salvati come *file* MP4.

[138] F. V. VALENTI, *op. cit.*, pp. 15-16.
[139] A. SIAROHIN, S. LATHUILIÈRE, S. TULYAKOV, E. RICCI, N. SEBE, *op. cit.*, pp. 1-9.
[140] *Ibidem.*

L’animazione risultante viene visualizzata sullo schermo, ed è possibile scaricare il *file* video generato tramite la funzione di *download*.[141]

È bene specificare che la creazione *deepfake* richieda competenze di programmazione, o l’accesso a programmi già precedentemente codificati. Tuttavia, in un contesto sempre più avvezzo alle tecnologie di intelligenza artificiale come quello odierno, diventa via via più lampante la necessità per gli Stati di aggiornare e adeguare all’evoluzione tecnologica il proprio assetto normativo.

[141] *Ibidem.*

CAPITOLO II

IL QUADRO GIURIDICO: LEGGI E NORMATIVE SUI *DEEPFAKE*

1. La protezione dei dati personali nel contesto dei *deepfake*

Nel contesto dei *deepfake* – e più in generale con l'avvento di tutte le nuove tecnologie di intelligenza artificiale – la protezione dei dati personali è una questione di crescente importanza.[142] In particolare, il Regolamento Generale sulla Protezione dei Dati (GDPR), un corpo normativo europeo che mira a tutelare i diritti e le libertà fondamentali delle persone fisiche con riguardo al trattamento dei loro dati personali[143] e, più recentemente, l'AI Act, cercano di affrontare le sfide poste da questa tecnologia.

Il GDPR, come esplicitamente indicato nell'art. 1 del regolamento stesso, si prefigge l'obiettivo di stabilire norme armonizzate per la protezione dei dati all'interno dell'Unione Europea e di assicurare la libera circolazione di tali dati, per garantire il buon funzionamento del mercato interno. In sostanza, lo scopo primario del GDPR è quello di proteggere i diritti e le libertà fondamentali delle persone fisiche, in particolare il diritto alla protezione dei dati personali.[144] E proprio l'articolo 1, nel delineare l'oggetto e le finalità del regolamento, vi fa esplicito riferimento:

"1. Il presente regolamento stabilisce norme relative alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché norme relative alla libera circolazione di tali dati.

2. Il presente regolamento protegge i diritti e le libertà fondamentali delle persone fisiche, in particolare il diritto alla protezione dei dati personali.

[142] S. TROZZI, *op. cit.*, p. 234.
[143] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 (GDPR).
[144] GDPR, art. 1.

3. La libera circolazione dei dati personali nell'Unione non può essere limitata né vietata per motivi attinenti alla protezione delle persone fisiche con riguardo al trattamento dei dati personali." [145]

Questo regolamento si applica al trattamento interamente o parzialmente automatizzato di dati personali e al trattamento non automatizzato di dati personali contenuti in un archivio o destinati a figurarvi.[146] È fondamentale notare che il GDPR si applica al trattamento dei dati personali di interessati che si trovano nell'Unione, effettuato da un titolare del trattamento o da un responsabile del trattamento che non è stabilito nell'Unione, quando le attività di trattamento riguardano l'offerta di beni o la prestazione di servizi a tali interessati, indipendentemente dall'obbligatorietà di un pagamento, o il monitoraggio del loro comportamento nella misura in cui tale comportamento ha luogo all'interno dell'Unione.[147]

L'art. 4 del GDPR fornisce una definizione onnicomprensiva di dato personale, intendendo con ciò "qualsiasi informazione riguardante una persona fisica identificata o identificabile".[148] Questa definizione acquista particolare rilevanza nel contesto dei *deepfake*, in quanto le immagini del volto e la voce di una persona, elementi centrali nella creazione di tali contenuti manipolati, rientrano pienamente nella nozione di dato personale. Poiché la creazione di *deepfake* implica spesso la manipolazione o la generazione di immagini e video che riproducono fedelmente le sembianze di persone reali, si configura un trattamento di dati personali, inclusi i dati biometrici.[149]

In modo specifico, il GDPR definisce i dati biometrici come "dati personali ottenuti da un trattamento tecnico specifico relativi alle caratteristiche fisiche, fisiologiche o comportamentali di una persona fisica che ne consentono l'identificazione univoca", una categoria particolare di dati.[150] Il trattamento di queste categorie speciali di dati personali è disciplinato con maggiore rigore dall'art. 9 del GDPR, che in via generale ne statuisce il divieto, salvo il ricorrere di specifiche deroghe, come il consenso esplicito

[145] *Ibidem*.
[146] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 (GDPR).
[147] GDPR, art. 3.
[148] *Ivi*, art. 4, punto 1.
[149] S. TROZZI, *op. cit.*, pp. 230, 243.
[150] GDPR, art. 9.

dell'interessato prestato per una o più finalità determinate, o la necessità del trattamento per motivi di interesse pubblico rilevante, fondati sul diritto dell'Unione o degli Stati membri e proporzionati alla finalità perseguita, nel rispetto dell'essenza del diritto alla protezione dei dati e prevedendo misure appropriate e specifiche per tutelare i diritti fondamentali e gli interessi dell'interessato.[151]

Il GDPR sancisce una serie di principi che devono essere rispettati nel trattamento dei dati personali: questi includono la liceità, correttezza e trasparenza del trattamento.[152] La liceità implica che il trattamento deve basarsi su una delle basi giuridiche previste dall'art. 6 del GDPR, come il consenso dell'interessato, che deve essere manifestato in modo libero, specifico e informato tramite un atto positivo inequivocabile, e il legittimo interesse perseguito dal titolare del trattamento o da terzi, a condizione che su tale interesse non prevalgano gli interessi o i diritti e le libertà fondamentali dell'interessato, richiedendo in tal caso un'attenta valutazione di bilanciamento.[153] Nel contesto dei *deepfake*, ottenere il consenso esplicito delle persone raffigurate è spesso fondamentale per garantire la liceità del trattamento[154], ma allo stesso tempo altrettanto complicato, soprattutto quando l'immagine o la voce di una persona vengono manipolate all'insaputa di quest'ultima o per finalità distorte rispetto al consenso originariamente prestato per un diverso trattamento.[155]

Il GDPR enuncia inoltre una serie di principi cardine che permeano l'intero impianto normativo e che devono essere osservati in ogni fase del trattamento dei dati personali. Tra questi spiccano la trasparenza (art. 5, paragrafo 1, lettera a), che postula che l'interessato sia pienamente consapevole dell'esistenza del trattamento e delle sue finalità, nonché delle modalità con cui i suoi dati vengono trattati;[156] la minimizzazione dei dati (art. 5, paragrafo 1, lettera c), che impone che i dati personali raccolti e trattati siano pertinenti, adeguati e limitati a quanto strettamente necessario per il perseguimento delle finalità dichiarate;[157] e il principio di *accountability* (art. 5, paragrafo 2), che radica

[151] *Ibidem.*
[152] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 (GDPR).
[153] GDPR, art. 6.
[154] S. TROZZI, *op. cit.*, p. 239.
[155] F. V. VALENTI, *op. cit.*, p. 12.
[156] GDPR, art. 5, paragrafo 1, lettera a.
[157] *Ivi*, art. 5, paragrafo 1, lettera c.

in capo al titolare del trattamento la responsabilità di assicurare e dimostrare la conformità del trattamento al regolamento. [158]

Il GDPR riconosce, poi, agli interessati un robusto catalogo di diritti esercitabili nei confronti del titolare del trattamento, finalizzati a garantire il controllo sui propri dati personali.[159] Tra questi diritti figurano il diritto di accesso ai propri dati (art. 15), che consente all'interessato di ottenere conferma che sia in corso un trattamento di dati personali che lo riguardano e, in caso affermativo, di accedere a tali dati e a informazioni correlate;[160] il diritto di rettifica (art. 16), che permette all'interessato di ottenere la correzione di dati personali inesatti;[161] il diritto alla cancellazione (c.d. "diritto all'oblio", art. 17), che consente all'interessato di ottenere la cancellazione dei dati personali in determinate circostanze;[162] il diritto di limitazione del trattamento (art. 18), che permette all'interessato di ottenere la limitazione del trattamento in specifiche ipotesi;[163] il diritto alla portabilità dei dati (art. 20), che consente all'interessato di ricevere in un formato strutturato i dati personali forniti a un titolare del trattamento e di trasmetterli ad un altro;[164] e il diritto di opposizione (art. 21), che permette all'interessato di opporsi al trattamento dei propri dati personali in determinate situazioni.[165]

Questi diritti possono essere esercitati anche nel contesto dei *deepfake*, ad esempio richiedendo la cancellazione di contenuti manipolati illecitamente; tuttavia, si potrebbe incorrere in degli ostacoli pratici, legati in particolare alla difficoltà di identificare il titolare del trattamento effettivo, alla potenziale diffusione incontrollata e transnazionale dei contenuti manipolati e alla persistente incertezza sull'origine e le successive modifiche apportate ai *deepfake*.[166]

Il GDPR impone obblighi sia ai titolari del trattamento (coloro che determinano le finalità e i mezzi del trattamento) che ai responsabili del trattamento (coloro che trattano i dati

[158] *Ivi*, art. 5, paragrafo 2.
[159] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 (GDPR).
[160] GDPR, art. 15.
[161] *Ivi,* art. 16.
[162] *Ivi,* art. 17.
[163] *Ivi,* art. 18.
[164] *Ivi,* art. 20.
[165] *Ivi,* art. 21.
[166] S. TROZZI, *op. cit.*, p. 243.

per conto del titolare). Questi obblighi includono l'attuazione di misure tecniche e organizzative adeguate a garantire la sicurezza del trattamento (art. 32),[167] la notifica all'autorità di controllo delle violazioni dei dati personali (art. 33)[168] e, in alcuni casi, la comunicazione della violazione all'interessato (art. 34)[169]. Inoltre, quando il trattamento può presentare un rischio elevato per i diritti e le libertà delle persone fisiche, il titolare del trattamento è tenuto a effettuare una valutazione d'impatto sulla protezione dei dati (art. 35)[170] e, in alcuni casi, a consultare preventivamente l'autorità di controllo (art. 36)[171].

Infine, il GDPR istituisce, all'art. 51, un sistema di autorità di controllo indipendenti con il compito di vigilare sull'applicazione del regolamento e di garantire il rispetto dei diritti degli interessati, conferendo loro poteri investigativi, correttivi e sanzionatori, per assicurare l'effettività della tutela. [172] Nel contesto dei *deepfake*, le autorità di controllo possono svolgere un ruolo cruciale nell'indagare su possibili violazioni delle disposizioni del GDPR connesse alla raccolta, all'elaborazione e alla diffusione illecita di dati personali finalizzate alla creazione di tali contenuti manipolati, nonché nell'adottare misure correttive e nell'irrogare sanzioni amministrative pecuniarie effettive, proporzionate e dissuasive in caso di accertata violazione.[173]

Pertanto, la combinazione di misure legali come il GDPR e l'AI Act, insieme a contromisure tecnologiche e all'aumento della consapevolezza degli utenti e delle autorità di controllo, rappresenta un approccio necessario per affrontare efficacemente i rischi posti dai *deepfake* alla protezione dei dati personali.

[167] GDPR, art. 32.
[168] *Ivi*, art. 33.
[169] *Ivi*, art. 34.
[170] *Ivi*, art. 35.
[171] *Ivi*, art. 36.
[172] *Ivi*, art. 51.
[173] S. TROZZI, *op. cit.*, pp. 234-237.

2. Il Regolamento Europeo sull'Intelligenza Artificiale (AI Act) e il suo rapporto con il GDPR

Il Regolamento Europeo sull'Intelligenza Artificiale (AI Act) è il primo quadro giuridico completo dell'Unione Europea volto a regolamentare l'intelligenza artificiale.[174] Questo coesiste e stabilisce un rapporto di sinergia e complementarità con il GDPR, nonostante la netta prevalenza di quest'ultimo in materia di trattamento dei dati personali.[175] In sintesi, l'AI Act si concentra sulla regolamentazione dei rischi specifici derivanti dai sistemi di intelligenza artificiale, mentre il GDPR configura la normativa orizzontale di riferimento per la protezione dei dati personali, indipendentemente dalla tecnologia utilizzata per il trattamento.

L'art. 1, paragrafo 7 dell'AI Act chiarisce che il diritto dell'Unione in materia di protezione dei dati personali, della vita privata e della riservatezza delle comunicazioni si applica ai dati personali trattati in relazione ai diritti e agli obblighi stabiliti dall'AI Act. Il regolamento lascia però impregiudicati il GDPR, il regolamento (UE) 2018/1725 (sul trattamento dei dati personali da parte delle istituzioni UE), la direttiva 2002/58/CE (ePrivacy) e la direttiva (UE) 2016/680 (sul trattamento dei dati personali a fini di contrasto), fatte salve specifiche eccezioni previste dall'AI Act.[176]

L'AI Act, dunque, non si pone l'obiettivo di sostituire il GDPR, bensì quello di integrarlo. E lo fa introducendo requisiti specifici per i sistemi IA ad alto rischio, i quali possono facilitare l'osservanza dei principi del GDPR relativi alla qualità, all'esattezza e alla minimizzazione dei dati personali; o, ancora, stabilendo obblighi di trasparenza per determinati sistemi di IA (art. 13 e art. 50 AI Act), inclusi quelli destinati all'interazione diretta con le persone e quelli che utilizzano dati biometrici, che si mostrano coerenti con i principi di correttezza e trasparenza del GDPR e con gli obblighi di fornire informazioni agli interessati (art.12, 13 e 14 GDPR).[177] L'AI Act all'art. 27 prevede, poi, l'obbligo di effettuare una valutazione d'impatto sui diritti fondamentali per i *deployer* (utilizzatori)

[174] D. KRAUSE, *The EU AI Act and the Future of AI Governance: Implications for U.S. Firms and Policymakers*, 2025, p. 1.
[175] S. TROZZI, *op. cit.*, pp. 234-235.
[176] AI Act, art. 1, paragrafo 7.
[177] D. KRAUSE, *op. cit.*, pp. 6-15.

di sistemi di IA ad alto rischio, organismi pubblici o enti privati che forniscono servizi pubblici.[178] Tale valutazione deve considerare i potenziali rischi per i diritti delle persone, inclusa la protezione dei dati personali, e può essere integrata con la valutazione d'impatto sulla protezione dei dati prevista invece dall'art. 35 del GDPR.[179]

Un'altra novità introdotta dall'AI Act è il proprio sistema sanzionatorio (art. 99-101): in questo caso, tali sanzioni sono distinte e autonome rispetto a quelle previste dal GDPR per le violazioni della normativa sulla protezione dei dati.[180] Tuttavia, in molti casi, l'uso di un sistema di IA comporta anche il trattamento di dati personali, il che significa che sia l'AI Act che il GDPR potrebbero essere applicabili e potenzialmente sanzionabili in caso di violazione delle rispettive disposizioni.[181]

L'obiettivo principale dell'AI Act è promuovere lo sviluppo e l'adozione di un'IA affidabile, bilanciando l'innovazione con la protezione dei diritti fondamentali e la sicurezza, attraverso un approccio basato sul rischio che classifica i sistemi di IA in diverse categorie con obblighi proporzionati.[182]

L'*iter* legislativo che ha condotto all'approvazione di questo complesso *corpus* normativo è stato lungo e articolato: la proposta iniziale è stata, infatti, presentata dalla Commissione Europea nell'aprile del 2021[183], e solo dopo un'intensa fase di negoziati e discussioni tra le varie istituzioni europee, il Parlamento Europeo ha dato il suo via libera definitivo al testo nel marzo del 2024, seguito dall'approvazione formale del Consiglio dell'Unione Europea nel maggio dello stesso anno. Il regolamento è entrato in vigore il 1° agosto 2024, con diverse date di applicazione previste per le sue varie disposizioni.[184]

L'AI Act si configura come un Regolamento di grande importanza in quanto mira a posizionare l'UE come un *leader* globale nello sviluppo e nell'adozione di un'IA umano-centrica: il suo scopo primario è quello di migliorare il funzionamento del mercato interno, istituendo un quadro giuridico uniforme che disciplini lo sviluppo, l'immissione

[178] AI Act, art. 27.
[179] GDPR, art. 35.
[180] AI Act, artt. 99-101.
[181] *Ivi*, art. 4.
[182] A. RUFFO, *op. cit*., p. 412.
[183] F. RAMOS, *op. cit.,* p. 375.
[184] A. RUFFO, *op. cit*., pp. 419-422.

sul mercato, la messa in servizio e l'uso dei sistemi di intelligenza artificiale nell'Unione, nel pieno rispetto dei valori fondanti dell'UE e dei diritti fondamentali sanciti dalla Carta. Il regolamento intende promuovere l'innovazione e la competitività, garantendo al contempo un elevato livello di preservazione della salute, della sicurezza, della democrazia, dello stato di diritto e dell'ambiente, proteggendo i cittadini dagli effetti nocivi che i sistemi di IA potrebbero generare.[185]

Uno dei punti cardine dell'AI Act è l'adozione di un approccio basato sul rischio, che classifica i sistemi di intelligenza artificiale in quattro categorie distinte: rischio inaccettabile, rischio elevato, rischio limitato e rischio minimo, con obblighi proporzionati al livello di rischio identificato.[186] I sistemi considerati a rischio inaccettabile, in quanto lesivi dei valori e dei diritti fondamentali dell'UE (come i sistemi di categorizzazione biometrica basati su caratteristiche sensibili), sono espressamente vietati.[187] I sistemi ad alto rischio, che possono avere un impatto significativo sulla vita delle persone (come quelli utilizzati in settori critici quali la finanza, la sanità, l'istruzione o l'applicazione della legge), sono soggetti a requisiti rigorosi in termini di qualità dei dati, documentazione tecnica, trasparenza, sorveglianza umana, accuratezza, robustezza e *cybersecurity*, oltre ad una valutazione di conformità *ex ante*.[188] I sistemi a rischio limitato sono soggetti a obblighi di trasparenza, come quello di informare gli utenti che stanno interagendo con un sistema di IA o che i contenuti che stanno visualizzando sono stati generati o manipolati artificialmente. I sistemi a rischio minimo, infine, non sono soggetti a particolari obblighi.[189] Abbiamo già precedentemente chiarito come, dal 2 febbraio 2025 il divieto dei sistemi AI a rischio inaccettabile sia diventato applicabile in tutta Europa.

Con specifico riferimento ai *deepfake*, il regolamento prevede obblighi di trasparenza per i fornitori e i *deployer* di sistemi di IA che generano o manipolano immagini, audio o video in modo tale da assomigliare a persone, oggetti, luoghi o eventi esistenti e da apparire falsamente autentici o veritieri. È sancito l'obbligo di indicare chiaramente che

[185] AI Act, art. 1, paragrafo 1.
[186] A. RUFFO, *op. cit*., p. 419.
[187] S. TROZZI, *op. cit.*, p. 236.
[188] AI Act, art. 1, paragrafo 1.
[189] *Ibidem.*

il contenuto sia stato generato o manipolato artificialmente, attraverso un'etichettatura appropriata e la divulgazione della sua origine artificiale.[190] Tuttavia, tale obbligo è mitigato in caso di utilizzi autorizzati dalla legge per quel che concerne indagini e perseguimento di reati, o qualora il contenuto faccia parte di opere o programmi artistici.[191] È importante notare che, sebbene i *deepfake* siano spesso classificati come a rischio limitato, l'uso di tali tecnologie per influenzare l'esito di elezioni o *referendum* potrebbe farli rientrare nella categoria ad alto rischio, data la potenziale minaccia ai processi democratici.[192]

Per garantire l'efficace attuazione e il rispetto del regolamento, l'AI Act ha istituito una complessa struttura di *governance* a livello sia europeo che nazionale, dove, a livello europeo, un ruolo centrale è svolto dall'Ufficio Europeo per l'Intelligenza Artificiale (*AI Office*), incaricato di sviluppare competenze, fornire orientamenti e coordinare le autorità nazionali. Accanto ad esso operano l'*AI Board,* un *panel* scientifico di esperti indipendenti e un *forum* consultivo con la partecipazione di una vasta gamma di *stakeholder*. A livello nazionale, gli Stati membri sono tenuti a designare autorità competenti per la notifica degli organismi di valutazione della conformità e per la vigilanza del mercato.[193]

Alla luce di quanto detto, possiamo constatare che l'AI Act consideri l'intelligenza artificiale come uno strumento al servizio dell'uomo, e che non vada in sostituzione a questo, essendo sviluppata e utilizzata in modo responsabile ed etico. Dunque, il regolamento mira sì a promuovere l'innovazione tecnologica, ma al contempo garantisce e preserva i diritti fondamentali e la sicurezza dei cittadini.

2.1 L'impatto dell'AI Act sull'uso dei *deepfake*

L'*Artificial Intelligence Act* affronta anche la questione dell'uso dei *deepfake*, definendoli come "un'immagine o un contenuto audio o video generato o manipolato dall'IA che

[190] A. RUFFO, *op. cit*., p. 420.
[191] S. TROZZI, *op. cit.*, pp. 239-240.
[192] F. V. VALENTI, *op. cit.*, p. 10.
[193] P. FALLETTA, *Lezioni di diritto pubblico del digitale, Cedam*, 2024, pp. 153-156.

assomiglia a persone, oggetti, luoghi, entità o eventi esistenti e che apparirebbe falsamente autentico o veritiero a una persona".[194]

Sebbene il regolamento non imponga un divieto generalizzato sull'uso dei *deepfake*, riconosce il loro potenziale per la disinformazione, la manipolazione e la lesione dei diritti individuali, introducendo specifiche misure volte a mitigarne i rischi.[195] Un ruolo centrale nella gestione di questi rischi associati ai *deepfake* è rivestito dall'art. 50, il quale stabilisce obblighi di trasparenza per i fornitori e i *deployer* di determinati sistemi di IA.[196]

Il primo punto di questo articolo constata che: "I fornitori garantiscono che i sistemi di IA destinati a interagire direttamente con le persone fisiche siano progettati e sviluppati in modo tale che le persone fisiche interessate siano informate del fatto di stare interagendo con un sistema di IA, a meno che ciò non risulti evidente dal punto di vista di una persona fisica ragionevolmente informata, attenta e avveduta, tenendo conto delle circostanze e del contesto di utilizzo. Tale obbligo non si applica ai sistemi di IA autorizzati dalla legge per accertare, prevenire, indagare o perseguire reati, fatte salve le tutele adeguate per i diritti e le libertà dei terzi, a meno che tali sistemi non siano a disposizione del pubblico per segnalare un reato." [197] Insomma, l'articolo impone ai fornitori di sistemi IA di informare gli utenti che stiano interagendo con un sistema artificiale, a meno che non sia palese.

Inoltre, prevede che i fornitori di sistemi di IA, inclusi quelli per finalità generali, che generano contenuti sintetici audio, immagine, video o testuali, debbano garantire che tali *output* siano marcati in un formato leggibile meccanicamente e rilevabili come generati o manipolati artificialmente. Tale marcatura deve essere "efficace, interoperabile, solida e affidabile", tenendo però sempre conto delle specificità dei diversi tipi di contenuto.[198]

Un particolare riferimento ai *deepfake*, emerge nel paragrafo 4 dell'art. 50, il quale stabilisce che "I *deployer* di un sistema di IA che genera o manipola un testo pubblicato allo scopo di informare il pubblico su questioni di interesse pubblico devono rendere noto

[194] A. RUFFO, *op. cit*., p. 421.
[195] F. V. VALENTI, *op. cit.*, pp. 12-14.
[196] S. TROZZI, *op. cit.*, pp. 240-241.
[197] AI Act, art. 50, paragrafo 1.
[198] AI Act, art. 50, paragrafo 2.

che il testo è stato generato o manipolato artificialmente. Tale obbligo non si applica se l'uso è autorizzato dalla legge per accertare, prevenire, indagare o perseguire reati o se il contenuto generato dall'IA è stato sottoposto a un processo di revisione umana o di controllo editoriale e una persona fisica o giuridica detiene la responsabilità editoriale della pubblicazione del contenuto." [199]

Questo obbligo di trasparenza indicato al quarto paragrafo si applica attraverso l'etichettatura degli *output* dell'IA e la divulgazione della loro origine artificiale in modo chiaro e distinto al momento della prima interazione o esposizione.[200] Tuttavia, il Regolamento prevede delle eccezioni a tale obbligo, in particolare "se l'uso è autorizzato dalla legge per accertare, prevenire, indagare o perseguire reati. Qualora il contenuto faccia parte di un'opera analoga o di programmi manifestamente artistici, creativi, satirici o fittizi, gli obblighi di trasparenza di cui al presente paragrafo si limitano all'obbligo di rivelare l'esistenza di tali contenuti generati o manipolati in modo adeguato, senza ostacolare l'esposizione o il godimento dell'opera." [201] Tale obbligo è analogo a quello precedentemente citato e previsto per i *deployer* di sistemi di IA che generano o manipolano testo pubblicato allo scopo di informare il pubblico su questioni di interesse pubblico.

L'AI Act, trattandosi di un regolamento UE, è direttamente applicabile in tutti gli Stati membri a partire dalla data di entrata in vigore delle singole disposizioni, senza necessità di un formale recepimento attraverso leggi nazionali. Ciò significa che le norme stabilite dal regolamento diventano parte integrante dell'ordinamento giuridico di ciascun Stato membro UE nel momento in cui acquisiscono efficacia. Tuttavia, il regolamento prevede anche un certo margine di manovra per gli Stati membri, in particolare per quanto riguarda la designazione delle autorità competenti responsabili dell'applicazione del regolamento, nonché per l'adozione di eventuali disposizioni più restrittive in ambiti specifici, come l'uso dei sistemi di identificazione biometrica remota.[202]

---

[199] *Ivi*, art. 50, paragrafo 4.

[200] A. RUFFO, *op. cit*., p. 421.

[201] AI Act, art. 50, paragrafo 4.

[202] L'art. 52 del Regolamento (UE) 2024/1689 (AI Act) stabilisce che "Uno Stato membro può decidere di prevedere la possibilità di autorizzare in tutto o in parte l'uso di sistemi di identificazione biometrica remota «in tempo reale» in spazi accessibili al pubblico a fini di attività di contrasto, entro i limiti e alle condizioni di cui al paragrafo 1, primo comma, lettera h), e ai paragrafi 2 e 3. Gli Stati membri interessati stabiliscono

Il regolamento è entrato ufficialmente in vigore il 1° agosto 2024, mentre la data di applicazione di tutto il regolamento è prevista per agosto 2026; tuttavia, i divieti relativi alle pratiche di IA inaccettabili e alcune disposizioni generali, come la definizione di IA, sono stati applicati solo a partire dal 2 febbraio 2025.[203] Mentre le norme relative alle autorità di notifica e alla struttura di *governance*, inclusa la creazione del Comitato europeo per l'intelligenza artificiale, sono diventate operative a partire dal 2 agosto 2025. Sempre dal 2 agosto 2025, sono stati applicati gli obblighi per i fornitori di modelli di IA per finalità generali, inclusi quelli relativi al diritto d'autore e alla trasparenza dei contenuti utilizzati per l'addestramento.[204] La data definitiva di applicazione del regolamento, inclusi i requisiti per i sistemi di IA ad alto rischio diversi da quelli già disciplinati da normative di armonizzazione dell'Unione, è ora fissata al 2 agosto 2026.[205]

Nonostante l'AI Act rappresenti un passo avanti significativo nella regolamentazione dei *deepfake*, non è esente dal dibattito tra gli studiosi: ad esempio, la sola efficacia dell'obbligo di trasparenza, previsto dall'art. 50, nel contrastare gli usi dannosi dei *deepfake* è stata messa in discussione.[206] Infatti, qualora i creatori o i *deployer* di *deepfake* agiscano con intenti malevoli e omettano di indicare la natura artificiale dei contenuti, la norma potrebbe rivelarsi insufficiente a proteggere gli individui dalle conseguenze negative che ne scaturiscono.[207] Inoltre, l'ampia portata delle eccezioni all'obbligo di trasparenza, in particolare per ragioni legate alla libertà di espressione e artistica, potrebbe potenzialmente limitare l'efficacia della disposizione. In ogni caso, sebbene il regolamento non vieti in assoluto i *deepfake*, e cerchi di mitigare molto i divieti

nel proprio diritto nazionale le necessarie regole dettagliate per la richiesta, il rilascio, l'esercizio delle autorizzazioni di cui al paragrafo 3, nonché per le attività di controllo e comunicazione ad esse relative. Tali regole specificano inoltre per quali degli obiettivi elencati al paragrafo 1, primo comma, lettera h), compresi i reati di cui alla lettera h), punto iii), le autorità competenti possono essere autorizzate ad utilizzare tali sistemi a fini di attività di contrasto. Gli Stati membri notificano tali regole alla Commissione al più tardi 30 giorni dopo la loro adozione. Gli Stati membri possono introdurre, in conformità del diritto dell'Unione, disposizioni più restrittive sull'uso dei sistemi di identificazione biometrica remota".

[203] AI Act, art. 113, lettera a.

[204] *Ivi*, art. 113, lettera b.

[205] *Ivi*, art. 113, paragrafo 2.

[206] F. V. VALENTI, *op. cit.*, p. 13.

[207] A. RUFFO, *op. cit*., p. 421.

all'utilizzo di questo, mira perlopiù a rendere gli utenti consapevoli della natura artificiale di tali contenuti, consentendo loro di valutare criticamente le informazioni ricevute.[208]

L'obiettivo dell'AI Act è quello di proiettare l'Unione Europea verso una regolamentazione armonizzata per l'intelligenza artificiale e tutte le nuove fattispecie che ne discendono, al fine di bilanciare innovazione tecnologica e protezione dei diritti fondamentali.[209] Ad ogni modo, l'AI Act non esclude la possibilità che gli Stati membri mantengano o adottino ulteriori misure nazionali, purché compatibili con il regolamento europeo, per affrontare le specificità dei propri contesti giuridici e sociali.[210] E, di fatto, è importante menzionare che alcuni Stati membri avevano già effettivamente intrapreso iniziative a livello nazionale per affrontare specifici aspetti legati ai *deepfake*, come nell'eventualità di diffusione non consensuale di *deepfake* pornografici.[211]

La sua importanza, dunque, consiste sì nel fornire un quadro comune a livello europeo, ma non esclude la possibilità che gli Stati membri mantengano o adottino ulteriori misure nazionali per affrontare le specificità dei propri contesti giuridici e sociali, purché compatibili con il regolamento stesso.

3. Analisi comparata delle normative internazionali

L'avanzamento rapido e pervasivo dell'intelligenza artificiale e, in particolare, delle tecnologie che consentono la creazione di *deepfake*, ha suscitato un'intensa attività normativa a livello internazionale. La creazione di questi contenuti ha da subito rappresentato e continua a rappresentare una sfida per i quadri giuridici tradizionali, andandosi ad intersecare con questioni relative alla libertà di espressione, alla *privacy*, alla disinformazione e alla sicurezza nazionale. Di conseguenza, diverse giurisdizioni hanno adottato o stanno considerando approcci normativi distinti.[212] È stato già

[208] F. V. VALENTI, *op. cit.*, pp. 13-14.
[209] P. FALLETTA, *Lezioni di diritto pubblico del digitale, Cedam*, 2024, pp. 153-156.
[210] Il Senato italiano ha approvato il disegno di legge sull'intelligenza artificiale, che passa all'esame della Camera dei deputati. Il Ddl 1146/24 recante "Disposizioni e delega al Governo in materia di intelligenza artificiale" è articolato in 26 articoli e conferisce delega al Governo per adottare, entro un anno, uno o più decreti legislativi per allineare la normativa nazionale all'AI Act europeo. I principi fondamentali stabiliscono che i sistemi di intelligenza artificiale devono essere sviluppati e applicati nel rispetto dell'autonomia e del potere decisionale umano, garantendo la supervisione e l'intervento delle persone.
[211] A. ORLANDO, *op. cit.*, pp. 311-317.
[212] S. TROZZI, *op. cit.*, pp. 239, 243.

ampiamente analizzato come l'Unione Europea si sia posta all'avanguardia nella regolamentazione dell'IA con l'adozione dell'AI Act, applicando parallelamente anche il GDPR al trattamento dei dati personali utilizzati nella creazione e diffusione di *deepfake*.

Dal suo canto, anche il *Digital Services Act* (DSA)[213] svolge un ruolo cruciale nella regolamentazione della diffusione *online* di *deepfake*, imponendo obblighi di moderazione dei contenuti e gestione dei rischi alle piattaforme *online* di grandi dimensioni. Il suddetto Regolamento è entrato pienamente in applicazione nel febbraio 2024, e si configura come un regolamento europeo volto a modernizzare il quadro giuridico per i servizi digitali nel mercato unico: in sostanza, il suo obiettivo principale è quello di creare un ambiente *online* più sicuro e responsabile, stabilendo obblighi armonizzati per i prestatori di servizi intermediari, che includono piattaforme *online*, *social network* e motori di ricerca.[214] Il DSA introduce un sistema di responsabilità differenziata in base alle dimensioni e alla natura dei servizi offerti, andando a prevedere degli obblighi più stringenti per le piattaforme *online* di dimensioni molto grandi (c.d. VLOPs) e i motori di ricerca *online* di dimensioni molto grandi (c.d. VLOSEs)[215].

Più nello specifico, in relazione ai *deepfake*, il DSA affronta indirettamente la questione attraverso diverse disposizioni: innanzitutto, ad esempio, il regolamento impone agli intermediari l'obbligo di adottare misure per contrastare la diffusione di contenuti illegali. Sebbene il termine "*deepfake*" non sia esplicitamente menzionato nella definizione di contenuto illegale, quest'ultima è interpretata in modo così ampio da includere qualsiasi informazione che violi il diritto dell'Unione o degli Stati membri.[216] Pertanto, quei *deepfake* utilizzati per scopi illeciti, come la diffamazione, la violazione della *privacy*, la pornografia non consensuale o la manipolazione a fini di frode, potrebbero rientrare nella categoria di contenuti illegali che le piattaforme sono tenute a rimuovere o disabilitarne l'accesso.[217]

---

[213] Regolamento (UE) 2022/2065 del Parlamento Europeo e del Consiglio del 19 ottobre 2022 relativo a un mercato unico dei servizi digitali e che modifica la direttiva 2000/31/CE (DSA).
[214] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 10-11, 17.
[215] J. P. QUINTAIS, *Generative AI, Copyright and the AI Act, Computer Law & Security Review 56*, 2025, p. 7.
[216] A. ORLANDO, *op. cit.*, p. 313.
[217] L. METSELAAR, *op. cit.*, p. 10.

Il DSA prevede anche misure per aumentare la trasparenza *online*, inclusi gli obblighi per le piattaforme di informare gli utenti sulle ragioni per cui determinati contenuti vengano loro raccomandati e sulle politiche di moderazione dei contenuti adottate. Nonostante, anche in questo caso, non specificamente mirato ai *deepfake*, il DSA potrebbe indirettamente contribuire a promuovere una maggiore consapevolezza e capacità critica da parte degli utenti nei confronti di contenuti potenzialmente manipolati.[218]

In particolar modo, l'art. 24 paragrafo 1, lettera b menziona direttamente la risposta delle piattaforme ala fornitura di contenuti illegali, che principalmente sfociano nella diffusione di contenuti pornografici generati dagli utenti, imponendo in seguito una serie di requisiti tecnici e organizzativi per le piattaforme e i fornitori.[219]

Sebbene il DSA non sia stato preposto per questo, è innegabile il suo potenziale circa il contrasto alla diffusione di *deepfake* dannosi attraverso i suoi obblighi di moderazione dei contenuti e gestione dei rischi. Tuttavia, l'efficacia delle sue disposizioni dipende perlopiù dalla capacità delle piattaforme di rilevare e identificare i *deepfake*; inoltre, l'applicazione del DSA a contenuti specifici, come ad esempio la fattispecie di *deepfake* politici che non rientra chiaramente nella definizione di contenuti illegali, può essere complessa poiché richiede un bilanciamento delicato con la libertà di espressione.[220]

Ad oggi, l'ordinamento europeo sembra essere quello più preparato e pronto per affrontare le nuove sfide imposte dall'utilizzo delle tecnologie di intelligenza artificiale: il suo approccio normativo proattivo e basato sul rischio, come esemplificato dall'AI Act, emerge rispetto al *modus operandi* che stanno, invece, adottando altri Stati terzi.

Negli Stati Uniti, ad esempio, il panorama normativo in materia di *deepfake* e disinformazione generata dall'intelligenza artificiale appare molto più frammentato e meno centralizzato rispetto al modello europeo. Questo è dato in gran parte dalla tradizione costituzionale statunitense, fortemente ancorata al Primo Emendamento e alla tutela della libertà di espressione: quest'ultima esercita un'influenza considerevole sia socialmente che politicamente, portando ad una riluttanza a imporre restrizioni

[218] A. RUFFO, *op. cit*., pp. 411-412.
[219] DSA, art. 24 paragrafo 1, lettera b.
[220] A. ORLANDO, *op. cit.*, pp. 311-316.

generalizzate sui contenuti, anche se falsi.[221] Strettamente correlata alla rilevanza del Primo Emendamento, vi è dottrina del "*marketplace of ideas*", la quale si basa sull'idea che la libera circolazione delle idee e delle opinioni, senza indebite restrizioni da parte del governo, sia il modo migliore per scoprire la verità. In questa sorta di "mercato", diverse idee competono tra loro e si presume che, nel lungo periodo, le idee vere e valide prevarranno su quelle false o dannose grazie al dibattito pubblico e alla capacità di discernimento dei cittadini.[222] Questa forte enfasi sulla libertà di espressione e sulla fiducia nella capacità del pubblico di valutare criticamente le informazioni, porta inevitabilmente a una certa riluttanza nell'imporre divieti o restrizioni generalizzate sui *deepfake*, anche quando questi veicolano disinformazione o contenuti potenzialmente dannosi.[223] Ciò significa che, secondo la logica del *marketplace of ideas*, anche le idee false o manipolate, benché venga garantito il dibattito aperto, non ostacolano la possibilità di contro-informazione e la consapevolezza del pubblico. Tuttavia, è importante menzionare che anche negli Stati Uniti, questa dottrina non è assoluta e incontra dei limiti quando la libertà di espressione entra in conflitto con altri interessi pubblici fondamentali, come la sicurezza nazionale, la protezione dalla diffamazione o il diritto alla *privacy*.[224]

Inoltre, la crescente sofisticazione dei *deepfake*, che li rende sempre più difficilmente distinguibili dalla realtà, fa vacillare proprio la premessa alla base della dottrina del *marketplace of ideas*, ossia la capacità insindacabile del pubblico di riconoscere e respingere le falsità. Sostanzialmente, se i *deepfake* diventano indistinguibili dalla realtà, la fiducia nel mercato delle idee come meccanismo di auto-correzione potrebbe soccombere. Risulta chiaro che gli Stati Uniti propendano per un approccio più frammentato e basato su settori specifici circa la regolamentazione dei *deepfake*.[225] A livello federale, non esiste una legislazione onnicomprensiva sui *deepfake*, sebbene siano state presentate diverse proposte di legge, come il *DEEPFAKES Accountability Act* e il *No AI Fraud Act*.[226] Questi progetti di legge mirano a stabilire obblighi di trasparenza,

[221] *Ibidem.*
[222] G. DE GREGORIO, *The market place of ideas nell'era della post-verità: quali responsabilità per gli attori pubblici e privati* online?, *Medialaws*, 2019, pp. 95-96.
[223] A. ORLANDO, *op. cit.*, p. 321.
[224] G. DE GREGORIO, *op. cit.*, pp. 93-95.
[225] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 2, 13-14.
[226] A. ORLANDO, *op. cit.*, pp. 320-321.

responsabilità e sanzioni per la creazione e diffusione di *deepfake* dannosi, con una particolare attenzione alla pornografia non consensuale, alla frode e all'interferenza elettorale.[227]

A livello statale, invece, alcuni Stati come Texas, California e Virginia hanno approvato leggi che vietano l'uso di *deepfake* per influenzare le elezioni o per creare e diffondere pornografia non consensuale. In ognuno di questi casi, però, la regolamentazione dei *deepfake* risulta ardua da ideare in conformità con il Primo Emendamento, essendo il diritto alla libertà di espressione pressoché illimitato nella visione statunitense.[228]

In contrapposizione con la frammentazione normativa in materia di *deepfake* che sta caratterizzando gli Stati Uniti, la Cina ha invece puntato ad un approccio più rigoroso e centralizzato alla regolamentazione dei *deepfake* con l'introduzione delle *Deep Synthesis Provisions* nel 2022.[229] Queste disposizioni essenzialmente vietano la creazione di *deepfake* senza il consenso dell'utente e richiedono che i contenuti generati con l'IA siano chiaramente contrassegnati. L'approccio cinese, in materia di *deepfake,* così come accade in ambito politico, riflette una forte volontà di controllo governativo sullo spazio digitale e sui contenuti *online*[230]: la Cina pone, di fatto, una forte enfasi sulla responsabilità dei fornitori di servizi di *deep synthesis*, imponendo loro obblighi di verifica dell'identità degli utenti, moderazione dei contenuti e segnalazione di informazioni false.[231]

Tornando sul suolo occidentale, il Regno Unito ha introdotto nel 2023 l'*Online Safety Act*, che mira a stabilire un nuovo regime di responsabilità per le piattaforme *online* in relazione a contenuti dannosi, inclusi i *deepfake*. Tuttavia, più avanti noteremo che il Regolamento potrebbe non essere sufficiente per contrastare efficacemente la minaccia dei *deepfake* non consensuali.

La Spagna ha, invece, istituito l'Agenzia Spagnola per la Supervisione dell'Intelligenza Artificiale (AESIA) e sta valutando misure legislative per contrastare la diffusione di *deepfake* dannosi.

---

[227] F. RAMOS, *op. cit.,* pp 372-373, 377-379.
[228] A. ORLANDO, *op. cit.*, p. 320.
[229] F. V. VALENTI, *op. cit.*, pp. 14-15.
[230] A. ORLANDO, *op. cit.*, pp. 322-324.
[231] F. V. VALENTI, *op. cit.*, pp. 14-15.

Volendo attuare un'analisi comparata tra gli Stati sopracitati, emergono tre principali paradigmi normativi: un paradigma "basato sulla applicazione", adottato principalmente dagli Stati Uniti, che si concentra quindi sulla regolamentazione dei *deepfake* in specifici scenari d'uso; un paradigma "basato sul soggetto", adottato dalla Cina, che si concentra sulla responsabilità dei fornitori di servizi di *deep synthesis*; ed infine un paradigma "basato sul ciclo di vita", seguito dall'UE, che mira a regolamentare più meticolosamente le diverse fasi del processo di creazione e diffusione dei *deepfake*.[232]

È chiaro come i maggiori ordinamenti mondiali stiano adottando misure molto eterogenee tra loro circa il tema dei *deepfake* (e l'IA più in generale). In prospettiva futura, si auspica, però, ad un maggiore coordinamento internazionale e ad una crescente cooperazione tra le autorità politiche delle tre potenze, in modo da costruire un "ponte" per avvicinare le visioni sul futuro e sui rischi dell'IA[233]: non si può più trascurare l'alfabetizzazione mediatica, oltre che incentivare la consapevolezza del pubblico di riconoscere e valutare criticamente i *deepfake* così come previsto dal *marketplace of ideas*.

In ogni caso, sebbene vi siano approcci diversi e in evoluzione, vi è un consenso generale sulla necessità di intervenire per regolamentare l'uso di queste tecnologie in modo responsabile ed etico, garantendo che l'innovazione tecnologica proceda di pari passo con la tutela dei diritti fondamentali e la salvaguardia della fiducia pubblica nel contesto digitale.

3.1 Europa: *focus* su Spagna e Regno Unito (*Online Safety Act*)

Precedentemente sono state citate come esempio di modello europeo, oltre all'Italia, anche la Spagna ed il Regno Unito. Sebbene la Gran Bretagna si trovi in Europa, la *Brexit*, che ha determinato la sua "uscita" dall'Unione Europea, chiaramente la pone in una situazione molto diversa rispetto – ad esempio – a quella italiana o spagnola. Il Regno Unito non è soggetto alle regole, né tantomeno gode della protezione, del DSA, del GDPR, e più recentemente, dell'AI Act. Questo ha portato il Regno Unito a formulare una propria normativa per affrontare la problematica dei contenuti dannosi sempre più

[232] A. ORLANDO, *op. cit.*, pp. 322-324.
[233] *Ibidem.*

frequentemente distribuiti *online*: *l'Online Safety Act* (OSA) è il frutto di questa crescente necessità.

Promulgato con l'assenso reale il 26 ottobre 2023, l'*Online Safety Act* introduce un esteso quadro normativo che impone doveri di diligenza proattivi ai fornitori di servizi *online user-to-user* (che consentono agli utenti di pubblicare contenuti o interagire tra loro) e ai servizi di ricerca che hanno un legame con il Regno Unito[234]. Questa legge si pone come obiettivo primario quello di proteggere sia i minori che gli adulti *online*, cercando di prevenire la proliferazione di contenuti e attività illegali e di contenuti che possano danneggiare i minori, oltre che proteggere dalla pubblicità fraudolenta.[235]

Il processo legislativo dell'OSA è stato complesso e prolungato, e questo *iter* quadriennale può rintracciare la sua genesi nella pubblicazione dell'*Online Harms White Paper*[236] nell'aprile del 2019: questo documento iniziale ha posto le basi per un nuovo approccio normativo volto a stabilire un "*duty of care*" (dovere di diligenza) per le piattaforme *online* nei confronti dei propri utenti.[237] L'idea centrale era quella di rendere le aziende tecnologiche maggiormente responsabili per la sicurezza degli individui *online*, in particolare per quanto riguarda la diffusione di contenuti illeciti e dannosi.

Con il susseguirsi di varie evoluzioni e revisioni prima della sua adozione, si può considerare che le negoziazioni per l'OSA si siano sviluppate in gran parte parallelamente a quelle che hanno portato all'adozione del *Digital Services Act* (DSA) nell'Unione Europea. Nonostante, dunque, alcune similitudini tra le due normative, come l'introduzione di valutazioni del rischio e misure di mitigazione, esistono differenze chiave significative. Ad esempio, il DSA impone alle piattaforme di valutare e mitigare i "rischi sistemici" che vanno oltre il semplice contenuto illegale, mentre l'attenzione del Regno Unito con l'OSA si concentra maggiormente sulla rimozione dei contenuti illegali.[238] Questa divergenza di *focus* riflette differenti priorità e approcci alla regolamentazione dello spazio digitale, e ha dato vita ad un dibattito pubblico e politico

[234] LATHAM & WATKINS, *op. cit.*, 2024, pp. 2-3.
[235] *Ivi*, p. 2.
[236] Un libro bianco (dall'inglese *white paper*) è generalmente un rapporto ufficiale pubblicato da un governo nazionale o da un'organizzazione internazionale su un determinato argomento o settore di attività.
[237] B. KIRA, *When non-consensual intimate deepfakes go viral, op. cit.*, p. 5.
[238] *Ivi*, pp. 12-13.

circa l'obiettivo principale dell'OSA: questo ha portato a focalizzarsi sulla necessità di trovare un equilibrio tra la protezione degli utenti *online* e la salvaguardia della libertà di espressione. La complessità di definire cosa costituisca un "*online harm*" (danno *online*) e di stabilire obblighi proporzionati per le piattaforme ha rappresentato una sfida costante proprio nel processo di elaborazione della legge.

La necessità di dar vita ad una legge come l'OSA scaturisce *in primis* dalla constatazione che le normative preesistenti non fossero sufficienti a contrastare la proliferazione di contenuti illeciti e dannosi *online*, rendendo impellente l'introduzione di un quadro giuridico più completo e incisivo. Il governo britannico, ha di fatto, pubblicamente lodato questa legge indicandola come promotrice della "nuova era di sicurezza e scelta *online*", con l'ambizioso obiettivo di rendere la Gran Bretagna "il luogo più sicuro al mondo in cui essere *online*".[239] Questa aspirazione riflette senza dubbio una risposta alla crescente preoccupazione pubblica e politica riguardo ai rischi associati all'uso di internet, che spaziano dallo sfruttamento minorile alla diffusione di contenuti violenti, all'incitamento all'odio e, in modo sempre più rilevante, all'abuso sessuale basato su immagini, inclusi i *deepfake* intimi non consensuali (*Non-Consensual Intimate Deepfakes* – NCID).[240]

Le ragioni che hanno portato all'adozione dell'OSA sono molteplici e interconnesse: in primo luogo, sono da ricercare nella crescente pressione per proteggere i bambini *online* da una vasta gamma di minacce, tra cui contenuti che promuovono comportamenti dannosi e soprattutto materiale pedopornografico.[241] Inizialmente focalizzato sulla tutela dei minori, il campo di applicazione dell'OSA è stato solo successivamente ampliato per includere la protezione degli adulti da contenuti illegali. Tuttavia, è importante notare che, durante le fasi finali di negoziazione, le disposizioni relative a contenuti legali ma dannosi per gli adulti sono state rimosse, limitando la versione finale dell'OSA alla regolamentazione di contenuti illegali e dannosi per i bambini.[242]

L'OSA mira a stabilire un dovere di diligenza (*duty of care*) per i fornitori di servizi *online* e motori di ricerca, imponendo loro di integrare la sicurezza nella progettazione dei propri

[239] *Ivi*, p. 2.
[240] LATHAM & WATKINS, *op. cit.*, 2024, pp. 1-4.
[241] *Ivi*, p. 2.
[242] B. KIRA, *When non-consensual intimate deepfakes go viral, op. cit.*, p. 6.

servizi e di adottare misure proporzionate per prevenire e mitigare la diffusione di contenuti illeciti e dannosi. Questo approccio, ispirato alle normative britanniche in materia di salute e sicurezza, va a configurarsi come un modello di co-regolamentazione, in cui l'autorità di regolamentazione per le comunicazioni (Ofcom), è investita del potere di supervisionare i sistemi di *governance* privati delle piattaforme.[243]

L'atto normativo conferisce a Ofcom ampi poteri di esecuzione, inclusa la possibilità di imporre sanzioni pecuniarie significative, fino a 18 milioni di sterline o al 10% del fatturato mondiale qualificato, a seconda di quale sia il maggiore. Inoltre, in determinate circostanze, i dirigenti *senior* potrebbero incorrere in responsabilità penale qualora non adottino tutte le misure ragionevoli per garantire che la propria organizzazione ottemperi alle richieste di informazioni di Ofcom.[244]

L'OSA ha istituito anche nuovi reati penali relativi all'abuso sessuale basato su immagini, superando le disposizioni precedenti del *Criminal Justice and Courts Act* [245] del 2015. Questi nuovi reati criminalizzano la condivisione o la minaccia di condividere fotografie o filmati intimi senza consenso, riconoscendo diverse fattispecie di reato con differenti livelli di gravità e sanzioni, inclusi i casi in cui la condivisione avviene con l'intento di causare allarme, umiliazione e disagio, o per ottenere gratificazione sessuale.[246]

Per quanto concerne i meccanismi introdotti dall'OSA, la legge impone ai fornitori di servizi *online* (definiti come "servizi di parte 3", i quali includono servizi utente-a-utente e servizi di ricerca con collegamenti al Regno Unito) una serie di obblighi volti a garantire la sicurezza *online*. Tra questi spiccano maggiormente la conduzione di valutazioni del rischio di contenuti illegali e la preparazione di valutazioni sull'accesso dei bambini per

[243] *Ivi*, pp. 4-5.
[244] LATHAM & WATKINS, *op. cit.*, 2024, p. 14.
[245] Questa legge è stata prodotta per "stabilire le modalità di trattamento dei trasgressori prima e dopo la condanna; creare un reato per maltrattamenti o negligenza intenzionale da parte di una persona che presta assistenza sanitaria o sociale; creare un reato di corruzione o altro esercizio improprio dei poteri e privilegi della polizia; prevedere disposizioni in merito ai reati commessi da conducenti squalificati; creare un reato di divulgazione di fotografie o filmati sessuali privati con l'intento di causare sofferenza; modificare il reato di incontro con un minore a seguito di adescamento sessuale; modificare il reato di possesso di immagini pornografiche estreme; prevedere le procedure e i poteri dei tribunali; prevedere il controllo giurisdizionale; e per finalità connesse".
[246] A. R. HUBER, Z. WARD, *Non-consensual intimate image distribution: Nature, removal, and implications for the Online Safety Act, European Journal of Criminology, 22 (1)*, 2025, pp. 2, 7.

i servizi che potrebbero essere utilizzati dai minori. Queste valutazioni devono essere mantenute aggiornate, oltre che riflettere le linee guida e i profili di rischio pubblicati da Ofcom, nonché le pratiche commerciali delle piattaforme.[247]

L'OSA stabilisce anche doveri di sicurezza che impongono alle piattaforme di adottare misure proporzionate per prevenire la diffusione di contenuti illegali e dannosi, tenendo conto della progettazione delle funzionalità, degli algoritmi, delle politiche di moderazione dei contenuti e degli strumenti che consentono agli utenti di controllare ciò che incontrano *online*. In pratica, le piattaforme sono tenute a predisporre meccanismi di segnalazione dei contenuti che siano facilmente accessibili e utilizzabili dagli utenti e dalle persone interessate (coloro che sono soggetti al contenuto o membri di un gruppo targettizzato dal contenuto). Inoltre, devono mantenere registri scritti delle proprie valutazioni del rischio e delle misure adottate per conformarsi ai codici di condotta di Ofcom.[248]

L'OSA prevede, poi, diverse categorie di servizi regolamentati (c.d. "Categoria 1", "Categoria 2A" e "Categoria 2B"), a cui si applicano obblighi specifici aggiuntivi, in particolare per le piattaforme più grandi e a maggiore impatto. Ad esempio, i servizi di Categoria 1 sono soggetti a obblighi più stringenti, come la fornitura di strumenti di responsabilizzazione degli utenti che consentano loro di escludere contenuti ritenuti dannosi, anche se non illegali, e la protezione dei contenuti ritenuti importanti per non ostruire il processo democratico, garantendo un trattamento equo a una vasta gamma di opinioni politiche. Principalmente, l'OSA mira ad un approccio basato sui "sistemi e processi" piuttosto che sui risultati, il che significa che l'enfasi va posta sull'implementazione di procedure adeguate a prevenire e rimuovere contenuti illegali, piuttosto che su *standard* misurabili di sicurezza *online*.[249]

Per quanto concerne, invece, il rapporto tra *Online Safety Act* e *deepfake*, e più specificatamente i *deepfake* intimi non consensuali, la legge riconosce che le immagini create o alterate tramite *computer grafica*, come i *deepfake*, rientrano nell'ambito dei nuovi reati relativi alla condivisione di immagini intime. Questo aspetto è cruciale in

[247] LATHAM & WATKINS, *op. cit.*, 2024, pp. 2-7.
[248] *Ivi*, 5-7, 21.
[249] B. KIRA, *When non-consensual intimate deepfakes go viral, op. cit.*, pp. 6-8.

quanto la precedente legislazione[250] era stata criticata per non aver affrontato adeguatamente questa forma emergente di abuso. Con l'OSA, la distribuzione di *deepfake* intimi non consensuali viene identificata come un reato prioritario, il che implica che le piattaforme avranno maggiori obblighi in termini di prevenzione e rimozione rapida di tali contenuti.[251]

Tuttavia, l'efficacia dell'OSA nel contrastare la diffusione di *deepfake* intimi non consensuali è stata oggetto di dibattito, essendo questa legge basata perlopiù sulle politiche delle piattaforme per la moderazione dei contenuti, le quali spesso mancano di coerenza e specificità riguardo ai media sintetici. Inoltre, l'OSA si concentra principalmente sulla rimozione dei contenuti dopo che sono stati caricati e segnalati, offrendo un rimedio limitato alle vittime una volta che il danno si è già verificato.[252]

Gli esperti hanno, infatti, sottolineato la necessità di meccanismi di prevenzione più efficaci, che affrontino la creazione e la disseminazione di NCID alla fonte, coinvolgendo non solo le piattaforme *social* e i siti pornografici, ma anche le aziende che sviluppano strumenti di intelligenza artificiale. Sostanzialmente l'OSA si affida in modo significativo alle segnalazioni degli utenti per l'identificazione di contenuti illegali, ma questo approccio potrebbe risultare problematico nel caso di NCID in quanto le persone raffigurate potrebbero non essere a conoscenza della loro esistenza e quindi impossibilitate a segnalarli. Infatti, sebbene Ofcom chiarisca che i *deepfake* intimi non consensuali dovrebbero essere considerati nell'ambito dei reati sessuali rilevanti, non è detto che le linee guida offerte alle piattaforme su come affrontare i potenziali contenuti illegali e come interpretare gli obblighi dell'OSA siano sufficientemente specifiche per la gestione degli NCID.[253]

Una delle principali criticità risiede nel fatto che l'OSA non imponga direttamente obblighi specifici ai fornitori di strumenti di creazione di *deepfake* basati sull'intelligenza artificiale, in modo da impedire la generazione di contenuti intimi non consensuali.

[250] Sezione 33 del *Criminal Justice and Courts Act* del 2015.
[251] A. R. HUBER, Z. WARD, *Non-consensual intimate image distribution: Nature, removal, and implications for the Online Safety Act, European Journal of Criminology, 22 (1)*, 2025, pp. 30-50.
[252] B. KIRA, *When non-consensual intimate deepfakes go viral, op. cit.*, pp. 1-2.
[253] *Ibidem.*

Oltretutto, attualmente, le definizioni all'interno dell'*Online Safety Act* probabilmente escludono gli strumenti di IA generativa perché non soddisfano i criteri di servizi utente-a-utente o motori di ricerca: ciò crea un vuoto normativo che si auspica venga colmato con una futura regolamentazione sull'intelligenza artificiale, così da poter affrontare efficacemente la problematica degli NCID.[254] In questo contesto, nonostante gli NCID non siano esplicitamente citati né nel DSA né tantomeno nell'AI Act, la direttiva europea sulla lotta contro la violenza nei confronti delle donne e la violenza domestica[255] prevede come reato la diffusione non consensuale di contenuti intimi prodotti o manipolati, inclusa la fabbricazione di *deepfake*. Tuttavia, la produzione di contenuti sintetici non consensuali in sé non è considerata un reato.[256]

In definitiva, sebbene l'*Online Safety Act* rappresenti un passo importante nel tentativo di regolamentare la sicurezza *online* e includa i *deepfake* intimi non consensuali nella sua definizione di abuso sessuale basato su immagini, la sua efficacia nel contrastare la creazione e la diffusione di NCID è limitata dal suo approccio basato sui "sistemi e processi" delle piattaforme e dalla mancanza di obblighi diretti sui fornitori di strumenti di IA generativa.[257]

Parallelamente, la Spagna si inserisce nel più ampio contesto della regolamentazione europea sull'intelligenza artificiale, ma avendo, in qualche modo, anticipato lo stesso ordinamento europeo: la *Agencia Española de Supervisión de la Inteligencia Artificial* (AESIA) rappresenta, infatti, la prima autorità di controllo specificamente dedicata all'intelligenza artificiale a livello europeo.

La sua istituzione ha anticipato e preparato il Paese all'assunzione degli obblighi e delle responsabilità derivanti dall'AI Act, ponendosi come obiettivo primario quello di supervisionare i sistemi di intelligenza artificiale impiegati sia dal settore pubblico che da quello privato in Spagna. Tra le sue competenze principali rientrano la garanzia della corretta implementazione dell'AI Act a livello nazionale, il potenziamento dei sistemi di

[254] *Ibidem.*
[255] Direttiva (UE) 2024/1385 del Parlamento europeo e del Consiglio, del 14 maggio 2024, sulla lotta alla violenza contro le donne e alla violenza domestica.
[256] B. KIRA, *When non-consensual intimate deepfakes go viral, op. cit.*, pp. 12-13.
[257] *Ivi*, p. 14.

IA sviluppati o utilizzati in Spagna, l'incorporazione di principi etici e di valore nei sistemi di IA e l'esercizio di poteri sanzionatori in caso di mancata conformità con il regolamento europeo. In relazione alle nuove fattispecie di intelligenza artificiale, inclusi i *deepfake*, AESIA si configura, in pratica, come l'ente incaricato di assicurare il rispetto delle disposizioni stabilite dall'AI Act.[258]

Abbiamo visto come la normativa europea, così come delineata nell'AI Act, adotti un approccio basato sul rischio per la regolamentazione dell'IA: non vieta in modo imperativo i *deepfake*, ma introduce obblighi di trasparenza specifici per i fornitori e i *deployer*. Ciò significa che mentre i fornitori di tali servizi sono obbligati ad esplicitare che gli *output* siano stati generati o manipolati artificialmente, analogamente i *deployer* di sistemi di IA che generano o manipolano testi pubblicati allo scopo di informare il pubblico su questioni di interesse collettivo sono tenuti a rendere noto che il testo è stato generato o manipolato artificialmente.[259] L'AESIA si inserisce perfettamente in questo meccanismo, essendo essa designata a garantire che gli operatori del settore IA, inclusi quelli che sviluppano o utilizzano tecnologie *deepfake*, si conformino agli obblighi di trasparenza e alle altre disposizioni dell'AI Act.[260]

È importante notare che, parallelamente all'implementazione del regolamento europeo, in Spagna il Governo ha approvato una proposta di legge nazionale[261] volta a introdurre disposizioni normative per assicurare la trasparenza nella pubblicazione e diffusione di contenuti prodotti mediante sistemi di intelligenza artificiale, prevedendo l'etichettatura univoca di tali contenuti. Questa iniziativa nazionale si pone in linea con i risultati degli studi condotti a livello europeo sull'etichettatura dei *deepfake* e mira a complementare e specificare ulteriormente gli obblighi previsti dall'AI Act, con l'Autorità per le garanzie nelle comunicazioni designata come responsabile del monitoraggio e dell'applicazione

[258] A. ARTERO MUÑOZ, C. F. RUIZ DE TOLEDO RODRÍGUEZ, P. MAIRAL MEDINA, *Agencia Española de Supervisión de la Inteligencia Artificial, la clave para un desarrollo tecnológico ético, justo y sostenible, Revista Española de Control Externo, vol. XXV, n. º 74-75*, 2023, pp. 35, 39-42, 45.
[259] F. V. VALENTI, *op. cit.*, p. 13.
[260] A. ARTERO MUÑOZ, C. F. RUIZ DE TOLEDO RODRÍGUEZ, P. MAIRAL MEDINA, *op. cit.*, pp. 34, 39, 45.
[261] La norma sull'etichettatura di IA stabilisce che qualsiasi immagine, audio o video generato o manipolato da AI che rappresenti persone reali o inesistenti in situazioni che non hanno mai vissuto deve essere etichettato in modo chiaro e distinguibile. Questo deve avvenire "non oltre il momento della prima interazione o esposizione", in linea con i requisiti dell'AI Act.

della legge. Il ministro Óscar López[262] ha, inoltre, dichiarato che "Quando la normativa sarà sviluppata, la *Agencia Española de Supervisión de la Inteligencia Artificial* (AESIA) stabilirà le norme", e sarà la *Agencia Española de Protección de Datos* (AEPD)[263] a vigilare che queste vengano rispettate.[264]

In contrapposizione a quello del Regno Unito, dunque, l'approccio spagnolo si configura come un'adesione convinta al quadro normativo europeo, con un'autorità nazionale specificamente istituita per garantirne l'efficacia e con iniziative legislative interne volte a rafforzare ulteriormente la trasparenza e la protezione degli utenti di fronte alle nuove sfide poste dall'intelligenza artificiale generativa, inclusi i *deepfake*.

3.2 Stati Uniti: iniziative legislative e approcci regionali

Negli Stati Uniti, la regolamentazione del fenomeno dei *deepfake* si distingue significativamente da quella europea, caratterizzandosi per una mancanza di un intervento legislativo federale onnicomprensivo e per l'emergere di iniziative a livello statale che tentano di colmare il vuoto normativo. A differenza dell'Unione Europea, che ha intrapreso la strada di una regolamentazione olistica dell'intelligenza artificiale con l'AI Act, gli Stati Uniti mostrano una tendenza a interventi più settoriali e, a livello federale, si limitano prevalentemente a proposte di legge, che faticano ad essere approvate.[265]

Questa peculiarità dell'ordinamento statunitense riflette la forte enfasi sul Primo Emendamento della Costituzione, che garantisce la libertà di espressione, e che pone limiti significativi alla possibilità di introdurre restrizioni sui contenuti, inclusi potenzialmente i *deepfake*. La protezione accordata ai fornitori di servizi internet, ai motori di ricerca e alle piattaforme *online*, generalmente esentati da responsabilità per i contenuti pubblicati dagli utenti, contribuisce ulteriormente a questo scenario.[266]

---

[262] Ministro per la trasformazione digitale e la funzione pubblica della Spagna.
[263] La *Agencia Española de Protección de Datos* (AEPD) è l'agenzia spagnola per la protezione dei dati. Si tratta di un'autorità pubblica indipendente che si occupa di garantire il rispetto della normativa sulla *privacy* in Spagna.
[264] M. G. PASCUAL*, El Gobierno aprueba la norma para el buen uso de la IA, que obliga a etiquetar contenidos creados con esta tecnología, El País*, 2025.
[265] A. ORLANDO, *op. cit.*, pp. 317-318.
[266] *Ivi*, p. 320.

Nonostante l'assenza di una legge federale unitaria, diverse proposte di legge sono state presentate al Congresso nel tentativo di affrontare le sfide poste dai *deepfake*; tra queste spicca il *DEEPFAKES Accountability Act*, presentato alla Camera dei Rappresentanti nel settembre 2023. Questa proposta mira a stabilire obblighi per i creatori di *deepfake*, differenziandoli a seconda del tipo di contenuto (audio, video o entrambi). In linea generale, si imporrebbe l'inserimento di tecnologie di provenienza dei contenuti per identificare chiaramente il materiale come alterato o generato tramite IA; poi, a seconda della natura del contenuto, si richiederebbero dichiarazioni verbali, scritte, o icone da integrare nel *deepfake* per prevenire fraintendimenti. La violazione di tali obblighi comporterebbe sanzioni diverse a seconda dell'offensività del contenuto, con pene detentive fino a cinque anni previste in casi di *deepfake* volti ad arrecare molestie sessuali, interferire con procedimenti ufficiali (incluse le elezioni), perpetrare frodi o furti di identità, o influenzare il dibattito pubblico interno nell'interesse di potenze straniere. Sono previste eccezioni per contenuti utilizzati dalle forze dell'ordine per la sicurezza pubblica e per contenuti pubblicati in contesti tali da non indurre una persona ragionevole a confondere l'attività falsificata con quella reale, come parodie, rievocazioni storiche o opere di finzione manifesta. La proposta prevede anche l'istituzione di una *task force* all'interno del Dipartimento di Sicurezza Nazionale, oltre che degli obblighi per gli sviluppatori di tecnologie per la creazione di *deepfake*, come garantire la capacità tecnica di inserire la provenienza e informare gli utenti sugli obblighi, e la richiesta ai fornitori di piattaforme *online* di dotarsi di sistemi per il rilevamento dei *deepfake*.[267]

Un'altra proposta federale rilevante è il *DEFIANCE Act*[268], approvato dal Senato nel luglio 2022 con emendamenti. Questa legge mira a garantire tutele giudiziarie alle vittime di *deepfake* intimi diffusi senza il loro consenso, estendendo la protezione già prevista per la divulgazione non consensuale di immagini intime. Più recentemente è stata

[267] *Ivi*, p. 317-318.

[268] Il *DEFIANCE Act* introduce la possibilità di usufruire di un ricorso civile federale per le vittime che sono identificabili in una "contraffazione digitale", definita come una rappresentazione creata attraverso l'uso di *software*, *machine learning*, intelligenza artificiale o qualsiasi altro mezzo generato da computer o tecnologico per falsamente apparire autentico. La legge si applica alle falsificazioni digitali che raffigurano la vittima nuda o impegnata in comportamenti sessualmente espliciti o scenari sessuali. La legge è opponibile contro gli individui che hanno prodotto o posseduto la contraffazione con l'intenzione di distribuirla; o che hanno prodotto, distribuito o ricevuto la contraffazione, se l'individuo sapeva o ha trascurato sconsideratamente che la vittima non ha acconsentito alla condotta.

presentata la proposta *COPIED Act* (*Content Origin Protection and Integrity from Edited and Deepfaked Media Act*), dai senatori Maria Cantwell, Marsha Blackburn e Martin Heinrich. Questa proposta legislativa mira a regolamentare l'uso di contenuti generati dall'intelligenza artificiale e a proteggere il diritto d'autore, promuovendo trasparenza e tracciabilità per i contenuti sintetici, definiti come quelli generati o modificati da algoritmi, inclusi quelli basati sull'IA: questo ambizioso progetto richiede la presenza di elementi distintivi, come filigrane digitali, con lo scopo di identificare rapidamente e in modo standardizzato il lavoro dell'IA da quello umano.[269] Gli scopi principali del *COPIED Act* sono: combattere l'aumento dei *deepfake* dannosi attraverso l'introduzione di nuove linee guida federali sulla trasparenza per contrassegnare, autenticare e rilevare i contenuti generati dall'intelligenza artificiale; proteggere giornalisti, attori e artisti dai furti guidati dall'IA; rendere i trasgressori responsabili degli abusi; standardizzare le informazioni sulla provenienza dei contenuti, la filigrana e il rilevamento dei contenuti sintetici con l'intento di incrementare gli *standard* di trasparenza per identificare quali contenuti siano stati generati o manipolati dall'IA e accertarne la provenienza; vietare l'uso di opere protette da *copyright* per addestrare sistemi di IA o per creare nuovi contenuti sintetici senza il consenso esplicito e informato dei titolari dei diritti d'autore. Questo concederebbe agli autori consapevolezza sull'uso delle loro opere e il diritto di stabilirne le condizioni, incluso il compenso.[270]

Un'ulteriore proposta degna di nota è il *No AI Fraud Act* (*No Artificial Intelligence Fake Replicas And Unauthorized Duplications Act*)[271], che riconosce ufficialmente il diritto di proprietà di ogni individuo sulla propria immagine e voce, rendendo illegale l'uso dell'IA per la clonazione di una persona.

In assenza di una normativa federale organica, a livello regionale, si possono osservare due principali tendenze tra gli Stati: gli interventi focalizzati sui *deepfake* pornografici e quelli incentrati sui *deepfake* politici. In tale contesto, appare emblematico il caso della California, che ha approvato nel 2019 due leggi distinte per queste due specifiche

[269] A. ORLANDO, *op. cit.*, p. 318-319.
[270] F. V. VALENTI, *op. cit.*, p. 15.
[271] Il disegno di legge stabilisce un quadro federale per proteggere i diritti individuali degli statunitensi al loro aspetto fisico e la loro e voce contro i falsi generati da IA e le contraffazioni. Esso cerca altresì di impedire alla clonazione, all'impersonificazione e ai falsi generati con IA di minare l'espressione artistica.

applicazioni; la legge sui *deepfake* politici, tuttavia, era soggetta a una "*sunset clause*" (clausola di decadenza) ed è stata abrogata nel gennaio 2023. Essa vietava, nei sessanta giorni precedenti un'elezione, la diffusione di "supporti audio o visivi materialmente ingannevoli" con effettiva malizia e con l'intento di danneggiare la reputazione di un candidato o ingannare gli elettori, con eccezioni per contenuti satirici o per gli operatori dell'informazione che ne dichiaravano chiaramente la natura ingannevole. Rimane, invece, pienamente applicabile la legge californiana[272] sui *deepfake* pornografici, che prevede il diritto di agire in giudizio contro chi intenzionalmente distribuisce *deepfake* di foto o video intimi o sessuali senza il consenso della persona ritratta, con eccezioni difficili da realizzare, come interesse pubblico, valore politico o giornalistico, protezione costituzionale.[273]

Altri Stati hanno seguito l'esempio californiano, adottando normative in una o nell'altra direzione. Per quanto riguarda i *deepfake* di matrice pornografica, la Virginia ha approvato nel 2019 una legge[274] che sanziona penalmente la distribuzione di *deepfake* pornografici idonei a costringere, molestare o intimidire una persona.[275] Nel 2024 sempre la Virginia ha emanato un'altra legge sui *deepfake* pornografici (VA SB 731), che ha modificato la definizione di "pornografia infantile" includendo materiale visivo sessualmente esplicito che raffigura un minore in uno stato di nudità o impegnato in una condotta sessuale dove questa rappresentazione è considerata oscena e il minore rappresentato non deve effettivamente esistere, nel caso in cui sia generato con intelligenza artificiale. La Florida[276], dal 2022, ha esteso le sanzioni penali relative alla

[272] Il *California Assembly Bill 602* (AB 602) crea una causa privata di azione contro una persona che: (1) crea e divulga intenzionalmente materiale sessualmente esplicito quando la persona sa o avrebbe dovuto ragionevolmente sapere che l'individuo raffigurato non ha dato il suo consenso alla creazione o divulgazione; o (2) divulga intenzionalmente materiale sessualmente esplicito che la persona non ha creato e la persona sa che l'individuo raffigurato non ha acconsentito alla creazione del materiale. Un "individuo raffigurato" è un individuo che, come risultato della digitalizzazione, sembra dare una *performance* che non ha effettivamente eseguito o che si esibisce in una rappresentazione alterata.

[273] A. ORLANDO, *op. cit.*, p. 319.

[274] Nel 2019, la Virginia è diventata il primo stato ad affrontare la tematica dei *deepfakes* sessuali non consensuali (VA H.B. 2678), aggiungendo ad una legge già esistente sul *revenge porn* la disposizione che "una persona la cui immagine è stata utilizzata per creare, adattare o modificare un'immagine video o fissa con l'intento di raffigurare una persona reale e che è riconoscibile come una persona reale dal volto della persona, somiglianza, o altra caratteristica distintiva".

[275] A. ORLANDO, *op. cit.*, p. 320.

[276] Florida, S.B. 1798, 24 giugno 2022.

pedopornografia e alla pornografia non consensuale a chi "promuove" *deepfake* (seppur il termine non sia esplicitamente usato, è incluso nella definizione di immagini create, alterate o modificate elettronicamente), senza che eventuali filigrane o etichettature del contenuto abbiano rilevanza.[277] Normative simili sono state adottate anche in Louisiana[278], South Dakota[279] e Washington[280].

In relazione ai *deepfake* di natura politica, invece, il Texas ha approvato nel 2019 una legge (Texas, S.B. 751)[281] simile a quella californiana, impedendo la distribuzione di *deepfake* politici entro trenta giorni dalle elezioni.[282] Nel Mississippi, dal 2022, è previsto il reato di diffusione di "*digitization*"[283] (termine preferito a *deepfake*) nei novanta giorni precedenti le elezioni, qualora manchi il consenso della persona ritratta, per chi diffonde il materiale e ne conosca la natura e miri a influenzare il dibattito elettorale.[284] Il New Mexico, nel 2024, ha introdotto una legislazione[285] simile al *Deepfakes Accountability Act*, imponendo l'etichettatura dei contenuti e sanzionando penalmente la diffusione di "*materially deceptive media*".[286] Discipline simili sull'obbligo di etichettatura sono state approvate anche in Indiana[287] e in Oregon[288].

Degno di nota è anche l'*Ensuring Likeness, Voice and Image Security (ELVIS) Act*, recentemente approvato in Tennessee[289], che aggiorna il *Personal Rights Protection Act*

---

[277] A. ORLANDO, *op. cit.*, p. 320.
[278] Louisiana, S.B. 1, 2 giugno 2023.
[279] South Dakota, S.B. 9, 13 febbraio 2022.
[280] Washington, S.B. 1999, 6 giugno 2024.
[281] SEZIONE 1. La sezione 255.004, Codice elettorale, è stata modificata aggiungendo le sottosezioni (d) e (e) così redatte:
(d) Una persona commette un reato se la persona, con l'intento di danneggiare un candidato o influenzare il risultato di un'elezione:
(1) crea un video *deepfake*; e
(2) causa la pubblicazione del video *deepfake* o la sua distribuzione entro 30 giorni dalle elezioni.
(e) In questa sezione, "*deep fake video*" significa un video creato con intelligenza artificiale che, con l'intento di ingannare, sembra rappresentare una persona reale che esegue un'azione che non si verifica nella realtà.
[282] A. ORLANDO, *op. cit.*, p. 320.
[283] Mississippi, S.B. 2577, 30 aprile 2024.
[284] A. ORLANDO, *op. cit.*, p. 320.
[285] Nex Mexico, H.B. 182, 5 marzo 2024.
[286] A. ORLANDO, *op. cit.*, p. 320.
[287] Indiana, H.B. 1133, 12 marzo 2024.
[288] Oregon, S.B. 1571, 27 marzo 2024.
[289] Tennessee, H.B. 2091, 26 marzo 2024 (ELVIS Act).

del 1984 e prevede una sanzione civile per chi rende disponibile al pubblico "*voice or likeness*" senza autorizzazione. Seppur concepita per la tutela della proprietà intellettuale, questa legge ha un'influenza significativa sulla natura legittima dei *deepfake*, con importanti eccezioni per utilizzi a fini giornalistici, informativi, educativi, satirici e parodistici protetti dal Primo Emendamento.[290]

Il panorama normativo statunitense, sia a livello federale che statale, mostra, pertanto, una tendenza a regolamentare i *deepfake* abbastanza settoriale e spesso debole, con un'attenzione prevalente a contesti specifici come le elezioni e la pornografia non consensuale. Questa impostazione è fortemente influenzata dalla preminenza del Primo Emendamento, che rende controversi interventi normativi che limitino l'uso di tali contenuti, a meno che non mettano seriamente in pericolo beni di innegabile rilevanza come la sicurezza pubblica, la dignità umana e la democraticità delle elezioni. La frammentazione a livello statale rischia di creare un mosaico di normative che possono risultare difficili da applicare e da rispettare, soprattutto in un contesto transnazionale come quello del *web*. Nonostante ciò, il dibattito e gli interventi normativi sui *deepfake* negli Stati Uniti sono in crescita; ciò appare rappresentativo della crescente necessità di trovare un equilibrio tra la protezione della libertà di espressione e la tutela da potenziali danni provocati dai *deepfake*.[291]

3.3 Cina: regolamentazione severa o controllo governativo?

La Repubblica Popolare Cinese ha intrapreso un approccio distintivo e rigoroso nei confronti delle tecnologie di *deepfake*, inquadrato all'interno di una più ampia strategia governativa volta ad esercitare una solida sovranità informatica e a garantire che lo sviluppo e l'uso dell'intelligenza artificiale siano pienamente allineati con gli interessi nazionali e i valori sociali cinesi.[292]

Questo orientamento si traduce in un modello regolatorio che può essere definito come "basato sul soggetto" (*subject-based regulatory paradigm*), in quanto pone una forte enfasi sulle responsabilità dei fornitori di servizi che utilizzano tecnologie di *deep*

[290] A. ORLANDO, *op. cit.*, pp. 320-321.
[291] *Ivi.*, pp. 317-321.
[292] *Ivi*, p. 321.

synthesis: proprio questo termine nel contesto cinese appare più ampio e inclusivo rispetto al solo "*deepfake*", essendo in grado di comprendere anche ambienti virtuali e *chatbot*.[293]

L'attenzione delle autorità cinesi verso il fenomeno dei *deepfake* si è precocemente sviluppata a partire dal 2019, in concomitanza con la crescente popolarità di applicazioni che permettevano la creazione di contenuti manipolati. Questa sollecitudine si è concretizzata nell'emanazione di normative significative, come le *Administrative Regulations on Online Audio and Video Information Services*[294] del 2020, che hanno introdotto un divieto generalizzato circa l'uso di contenuti generati tramite *deep synthesis* per creare o diffondere notizie false, giustificato dalla necessità di preservare l'ordine sociale e proteggere gli interessi individuali e la sicurezza nazionale.

Sostanzialmente tali normative impongono alle piattaforme *online* l'obbligo di rafforzare l'autoregolamentazione, oltre che istituire sistemi di responsabilità editoriale, garantire la sicurezza informatica, verificare l'identità degli utenti e segnalare o eliminare contenuti non autentici.[295]

Un ulteriore e più specifico intervento si è avuto nel 2022 con l'introduzione dei *Provisions on the Administration of Deep Synthesis Internet Information Services* [296], entrati effettivamente in vigore a gennaio del 2023. Queste norme mirano a regolamentare le attività di *deep synthesis* in un'ottica di promozione dei "valori socialisti fondamentali", oltre che di tutela della sicurezza nazionale e dei diritti dei cittadini.[297] Sebbene non stabiliscano un divieto generalizzato alla creazione di *deepfake*, i confini dei contenuti non consentiti appaiono particolarmente ampi e definiti in maniera potenzialmente vaga.[298]

Essenzialmente è vietato utilizzare servizi di *deep synthesis* per produrre, riprodurre, pubblicare o trasmettere informazioni proibite da leggi o regolamenti amministrativi, o per intraprendere attività che mettono a repentaglio la sicurezza e gli interessi nazionali,

[293] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 11-13.
[294] *Administrative Regulations on Online Audio and Video Information Services,* 18 novembre 2019. Cfr. *China issues regulation for online audio, video services, in english.gov.cn,* 30 novembre 2019.
[295] A. ORLANDO, *op. cit.*, p. 322.
[296] *Provisions on the Administration of Deep Synthesis Internet Information Services*, 25 novembre 2022.
[297] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 11-12.
[298] A. ORLANDO, *op. cit.*, p. 323.

o che danneggiano l'immagine della nazione, l'interesse pubblico, l'ordine economico o sociale, o i diritti e gli interessi legittimi di altri. In aggiunta, è esplicitamente vietato l'uso di tali tecnologie per la creazione o la diffusione di notizie false.[299]

Coerentemente con l'approccio basato sul soggetto, la normativa cinese pone un forte accento sugli obblighi imposti ai fornitori di servizi[300]: questi sono tenuti a verificare l'identità degli utenti, ad etichettare adeguatamente i contenuti generati o manipolati per facilitarne la tracciabilità, a rafforzare la gestione dei dati e, soprattutto, a segnalare agli utenti l'obbligo di ottenere il consenso dalle persone interessate prima di utilizzare la loro immagine o voce per la creazione di contenuti di *deep synthesis*. Per i servizi che possono generare o alterare significativamente informazioni, come la simulazione di testo, voce e volto di un individuo, è richiesta una etichettatura ben visibile. Inoltre, per i *deepfake* dotati del potenziale di influenzare l'opinione pubblica o la mobilitazione sociale, i fornitori devono espletare formalità di deposito e condurre una valutazione di sicurezza prima del lancio di nuovi prodotti tecnologici di *deep synthesis* che abbiano tali capacità. È bene notare che il mancato rispetto di questi obblighi può comportare sanzioni sia a livello civile che penale.[301]

L'approccio cinese si pone in chiaro contrasto con quello degli Stati Uniti, che invece si configura come un modello basato sull'applicazione (*application-based regulatory paradigm*), piuttosto che sul soggetto: si tratta di un approccio caratterizzato da una tendenza a intervenire normativamente solo in specifici ambiti di utilizzo considerati particolarmente critici, ad esempio le elezioni e la pornografia non consensuale, come analizzato in precedenza. A livello federale, al netto di alcune proposte di legge, non si registra una volontà di adottare una regolamentazione forte e generalizzata, preferendo un approccio più mitigato nei confronti della tecnologia, in linea con la tradizionale importanza attribuita alla libertà di espressione sancita dal Primo Emendamento.[302] In merito a ciò, la Corte Suprema ha persino riconosciuto una certa protezione a dichiarazioni false, a meno che non provochino danni gravi legalmente riconoscibili.[303]

[299] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 11-13.
[300] A. ORLANDO, *op. cit.*, p. 322.
[301] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 11-13.
[302] *Ivi*, pp. 5-7.
[303] A. ORLANDO, *op. cit.*, pp. 320-321.

Di conseguenza, l'attenzione legislativa, sia a livello federale che statale, si concentra poco su casi specifici, tralasciando i dettagli individuali a favore di problematiche molto più generali e lampanti. Sostanzialmente l'approccio statunitense è prevalentemente di tipo *ex post*, intervenendo principalmente a valle della creazione e diffusione dei *deepfake*, con un'enfasi sui rimedi legali per le vittime e sulle sanzioni per utilizzi specificamente dannosi[304]; e di fatto, la responsabilità dei *provider* di servizi *online* è tendenzialmente limitata dall' Art. 230 del *Communications Decency Act*[305], dal Primo Emendamento e dalle leggi di *copyright*.[306]

Questi due approcci si distinguono, a loro volta, dall'approccio adottato dall'Unione Europea, che si definisce come "basato sul ciclo di vita" (*life cycle-based regulatory paradigm*), e si mostra caratterizzato da una visione olistica basata sul rischio che mira a regolamentare l'IA in generale, e i *deepfake* in particolare, in tutte le loro fasi, dallo sviluppo della tecnologia alla creazione e alla diffusione dei contenuti. Sebbene non esista un atto normativo specificamente dedicato ai *deepfake*, è stato analizzato come il fenomeno sia ampiamente discusso all'interno di normative come il GDPR, il DSA e, soprattutto, l'AI Act[307], dove i *deployer* di sistemi di IA che generano o manipolano immagini o contenuti audio o video che costituiscono un *deepfake* sono esplicitamente obbligati a rendere noto che il contenuto sia stato generato o manipolato artificialmente.[308]

In certo qual modo, dunque, l'UE riesce ad adottare un approccio che combina misure *ex ante* (approccio cinese), come la regolamentazione degli algoritmi e dei processi di creazione, ed *ex post* (approccio americano), come la disciplina sulla diffusione e le sanzioni: l'obiettivo finale dell'UE è quello di trovare un equilibrio tra la promozione dello sviluppo tecnologico e la tutela dei diritti fondamentali, dei valori democratici e della sicurezza individuale e collettiva, adottando un approccio "*human-centric*".[309]

---

[304] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 13-14.
[305] L'articolo 230 del *Communications Decency Act* del 1996 protegge i fornitori di servizi internet e gli utenti dalla responsabilità per i contenuti pubblicati da altri. È stato creato per affrontare il dilemma che le piattaforme *online* devono affrontare tra la moderazione dei contenuti e l'essere ritenuti responsabili.
[306] Е.А. СВИРИДОВА, *Rules for the Use of Deepfake Technologies in the Law of the USA and the People's Republic of China: Adaptation of Foreign Experience in Legal Regulation, Современное право*, 2024, pp. 119-123.
[307] G. ZHENG, J. SHU, K. LI, *op. cit.*, p. 8.
[308] A. ORLANDO, *op. cit.*, p. 316.
[309] G. ZHENG, J. SHU, K. LI, *op. cit.*, p. 14.

In definitiva, mentre la Cina adotta un approccio centralizzato e proattivo, focalizzato sulla responsabilità dei fornitori di servizi e sul controllo dei contenuti in linea con i valori sociali e la sicurezza nazionale, gli Stati Uniti prediligono un modello decentralizzato e reattivo, intervenendo in modo mirato su specifiche applicazioni dannose e ponendo un forte accento sulla libertà di espressione e sulla limitazione della responsabilità dei *provider*. L'Unione Europea si colloca tra i due in una posizione intermedia, cercando di stabilire un quadro regolatorio completo e basato sul rischio che copra l'intero ciclo di vita dei *deepfake*, imponendo obblighi di trasparenza e mirando a bilanciare l'innovazione tecnologica con la protezione dei diritti e dei valori fondamentali.

La Cina, con la sua enfasi sul controllo e sulla responsabilità dei fornitori, sembra adottare un approccio di "*deep control*" sul fenomeno, mentre gli Stati Uniti mostrano una maggiore tolleranza verso la tecnologia, a meno che non causi danni specifici e comprovati. L'UE cerca, dal suo canto, di costruire un mercato unico digitale sicuro e affidabile, in cui i *deepfake* siano gestiti attraverso una combinazione di obblighi di trasparenza, valutazione del rischio e responsabilità dei diversi attori coinvolti. Questa diversa impostazione riflette le peculiarità degli ordinamenti giuridici, le visioni politiche e strategiche in materia di sviluppo tecnologico e le priorità in termini di bilanciamento tra libertà individuali, sicurezza collettiva e progresso economico. Un elemento trasversale, tuttavia, è la crescente consapevolezza della necessità di contrastare la diffusione di contenuti *deepfake* dannosi, specialmente in settori sensibili come i processi democratici, sebbene le modalità e la portata degli interventi normativi differiscano significativamente tra le tre potenze.[310]

[310] A. ORLANDO, *op. cit.*, pp. 324-327.

CAPITOLO III

CASI STUDIO: L'IMPATTO CONCRETO DEI *DEEPFAKE*

1. Elezioni presidenziali USA 2024: il caso Trump

La politica non è estranea all'impiego di intelligenza artificiale, e nello specifico di *deepfake,* a scopo propagandistico: il miglior esempio a riguardo sono le elezioni presidenziali tenutesi nel 2024 che hanno visto vincitore Donald Trump. Durante le elezioni, sia Trump che la sua avversaria Kamala Harris hanno utilizzato attivamente i *social*, proponendo però campagne mediatiche molto diverse. Se Harris ha puntato ad una strategia più sobria e talvolta pedissequa, Trump, in linea con la sua personalità politica, ha fatto un uso di IA piuttosto considerevole e sfrontato.

Nello specifico, durante le ultime presidenziali statunitensi, Trump ha utilizzato attivamente i *social media*, e più in particolare la sua piattaforma *Truth Social*, per comunicare con i suoi sostenitori e diffondere messaggi politici.[311] Ma non si è limitato alla diffusione di contenuti sterili, e ha fatto uso attivo di *deepfake* a scopo propagandistico: l'esempio più emblematico è stato la pubblicazione di un'immagine artefatta in cui l'artista statunitense Taylor Swift appariva vestita come l'iconico Zio Sam in un contesto patriottico, sottintendendo falsamente il suo sostegno a Trump. Questa immagine, inizialmente condivisa da un sostenitore trumpiano su X, è stata poi ripubblicata da Trump stesso sul suo account *Truth Social*.[312] Questo episodio è particolarmente interessante poiché ci fa comprendere come persino una figura politica di alto profilo come Trump possa contribuire alla proliferazione di media sintetici falsi. È infatti notorio come Taylor Swift sia, in realtà, dichiaratamente schierata a sinistra, e altrettanto notorio è il suo burrascoso rapporto con l'attuale presidente USA.

---

[311] A. RUDNIEVA, A. O. РУДНЄВА, *Innovative information technologies in election political communications, Epistemological studies in Philosophy, Social and Political Sciences,* 2024, p. 178.
[312] *Ibidem.*

Le conversazioni tra i due, tendenzialmente a senso unico da parte di Trump, si sono dispiegate perlopiù su *Twitter* (poi, X) nel corso degli anni, con una sempre più crescente ostilità da parte di Trump verso la cantante. Se, infatti, nel corso del 2012 Trump scriveva diversi *tweet* di apprezzamento diretti a Taylor Swift, nel 2018, a seguito della prima presa di posizione politica della cantante, dove esortava i suoi *followers* in Tennessee a votare contro la repubblicana Marsha Blackburn per le elezioni di medio termine, Trump ha cambiato rotta, dichiarando che a quel punto la sua musica gli sarebbe piaciuta un 25% in meno. Molti utenti credono che Trump sia in qualche modo ossessionato da Taylor Swift a causa di tutte le dichiarazioni da lui rilasciate in merito nel corso degli anni, e avendo più volte sia apprezzato pubblicamente il suo aspetto fisico che commentato negativamente la sua attuale relazione con il giocatore di *football* Travis Kelce.[313] L'apice si è raggiunto nell'agosto del 2018, quando ha, come detto anticipatamente, pubblicato diversi *deepfake* figuranti Taylor Swift, tra cui anche un'immagine dove era ritratta con lo slogan "*Swifties for Trump*", oltre a riferimenti alla minaccia terroristica sventata per il concerto di Swift a Vienna. Trump, in seguito, ha dichiarato di non aver generato in prima persona le immagini, attribuendole ad "altre persone" e definendo l'IA "molto pericolosa in quel senso".

Dopo l'*endorsement* di Swift a Kamala Harris, Trump ha poi reagito dicendo di non essere mai stato un *fan* di Taylor Swift e che lei avrebbe "probabilmente pagato un prezzo per questo sul mercato", per poi pubblicare qualche giorno dopo sul suo profilo *Truth Social* un *post* in cui scriveva letteralmente: "*I Hate Taylor Swift*".

Questa vicenda è abbastanza indicativa di come la tecnologia odierna possa essere sfruttata erroneamente persino da capi di Stato del calibro di Donald Trump, oltre a sollevare interrogativi circa quale sia il confine insorpassabile tra la propaganda politica tradizionale e le nuove forme di manipolazione rese possibili dall'intelligenza artificiale. Il cosiddetto "*political deepfake*" può essere definito come una "campagna speciale che utilizza la tecnologia di intelligenza artificiale per minare la reputazione dei *leader* politici al fine di cambiare il corso della lotta elettorale o di screditare un politico in carica".[314]

---

[313] H. DAILEY, *"I Hate Taylor Swift": Everything Donald Trump Has Ever Said About the Pop Star, Billboard*, 2025.

[314] E. A. VINOGRADOVA, *Potential threats of unauthorized use of political deepfakes during political elections: international experience, Мировая политика*, 2024, pp. 44-45.

In questo contesto, la propaganda politica, tradizionalmente definita come la diffusione di informazioni distorte o esagerate per influenzare l'opinione pubblica, si intreccia perfettamente con l'IA quando quest'ultima viene utilizzata per creare e distribuire contenuti manipolatori come i *deepfake*: l'intelligenza artificiale può automatizzare la creazione e la diffusione di articoli di notizie false e *post* sui *social media*, che possono essere mirati a specifici segmenti di elettori con disinformazione personalizzata progettata per influenzare le loro percezioni e comportamenti.[315]

Le implicazioni dell'uso di *deepfake* nella propaganda politica minano, innanzitutto, la fiducia nelle informazioni e nei media: la capacità di creare video e audio iperrealistici di persone che dicono o fanno cose che non hanno mai fatto rende sempre più difficile per il pubblico distinguere tra realtà e finzione, e questo può portare ad un deterioramento dell'ambiente informativo.[316]

Inoltre, i *deepfake* possono essere utilizzati per screditare candidati, diffondere false informazioni o incitare disordini. Casi internazionali mostrano come questa tecnologia sia già stata impiegata in contesti elettorali: ad esempio, durante le elezioni generali indiane del 2019, si è verificata una rapida diffusione di notizie false e video manipolati, in seguito dimostratisi generati dall'IA, che hanno influenzato il *sentiment* degli elettori. Nelle elezioni brasiliane del 2020, invece, è stato affermato che *bot* automatizzati hanno inondato i *social media* con disinformazione, manipolando l'opinione pubblica.[317] Proprio in Brasile, il Tribunale Superiore Elettorale ha proibito l'uso dell'intelligenza artificiale per creare e propagare contenuti falsi nelle elezioni.

Nelle Filippine, invece, le elezioni presidenziali del 2022 sono state presumibilmente influenzate dall'uso di pubblicità mirata guidata dall'IA, che sfruttava i dati degli elettori per creare messaggi politici altamente personalizzati e spesso fuorvianti.[318]

---

[315] M. B. E. ISLAM, M. HASEEB, H. BATOOL, N. AHTASHAM, Z. MUHAMMAD, *AI Threats to Politics, Elections, and Democracy: A Blockchain-Based Deepfake Authenticity Verification Framework*, Blockchains 2, 2024, pp. 465-466.
[316] E. A. VINOGRADOVA, *op. cit.*, pp. 47-48.
[317] M. B. E. ISLAM, M. HASEEB, H. BATOOL, N. AHTASHAM, Z. MUHAMMAD, *op cit.*, pp. 459.
[318] *Ivi*, pp. 459, 461.

In Indonesia, è stato creato nel 2024 un *deepfake* ritraente un *leader* politico deceduto, in procinto di sostenere apparentemente il suo *ex* partito politico. Sempre in Indonesia il generale Subianto, candidato alle presidenziali, si è munito di IA per creare un *avatar*, una versione mitigata e non fotorealistica di sé stesso, al fine di distogliere l'attenzione da passate accuse di violazione dei diritti umani. Nel frattempo, in Pakistan, la tecnologia *deepfake* ha permesso all'*ex* Primo Ministro, incarcerato, Imran Khan di entrare in contatto con il pubblico: essenzialmente i suoi messaggi scritti dalla prigione sono stati convertiti in videomessaggi, permettendogli di interagire con il suo elettorato e mantenere una posizione in politica nonostante il suo confinamento.[319] Lo stesso Donald Trump non è stato risparmiato dalla tecnologia *deepfake*, essendo stato nel marzo 2024 protagonista di un video falso diventato estremamente virale, in cui sembrava essere stato arrestato.

Insomma, il confine tra propaganda politica e IA si sfuma quando le tecniche di IA vengono utilizzate per amplificare e sofisticare le strategie propagandistiche: l'intelligenza artificiale non solo facilita la creazione di contenuti manipolatori, ma può anche analizzare i dati degli elettori per indirizzare la propaganda in modo più efficace e personalizzato[320]; questa microtargettizzazione è molto proficua nel rendere la propaganda più persuasiva e difficile da contrastare.[321] E non è un caso che l'IA costituisca un nuovo ed apparentemente efficace mezzo di propaganda proprio in quei Paesi in cui scarseggiano normative organiche volte a regolamentare queste nuove tecnologie.

Tuttavia, è importante considerare che l'IA può anche avere effetti positivi in contesti politici. Alcuni scienziati brasiliani ritengono che l'uso dell'IA generativa nelle campagne politiche possa avere una duplice natura, capace di generare sia conseguenze positive che negative: da un lato, essa può servire come mezzo per incoraggiare l'attivismo e l'impegno degli elettori, mentre dall'altro, può portare a una maggiore polarizzazione politica e alla diffusione di disinformazione. Scienziati americani condividono questa visione, evidenziando come l'IA possa creare *deepfake* convincenti aumentando la

[319] M. PAWELEC, *Deepfakes: Manipulation on Demand? Evolution of Deepfake Technology, Societal Impact, and the Path Forward, Heinrich Böll Foundation Tel Aviv, Israel Public Policy Institute*, 2025, p. 14.
[320] M. B. E. ISLAM, M. HASEEB, H. BATOOL, N. AHTASHAM, Z. MUHAMMAD, *op cit.*, pp. 459, 462, 466.
[321] M. PAWELEC, *op. cit.*, p. 12.

probabilità di diffondere informazioni false durante le campagne elettorali, ma anche potenzialmente raggiungere nuovi gruppi *target* e coinvolgere i cittadini in modo più diretto attraverso i cosiddetti "*softfake*" (manipolazioni meno evidenti, ma che possono comunque fuorviare gli spettatori sulle capacità o le posizioni dei candidati).[322]

In definitiva, i mezzi di comunicazione politica necessitano per natura di adeguarsi ai tempi, malgrado la dicotomia apparente tra tecnologia e politica. Ciononostante, l'impiego dell'IA a scopo propagandistico, soprattutto se a discapito di altre persone, come nel caso dei *deepfake* ritraenti Taylor Swift diffusi da Trump, potrebbe varcare quel confine necessario a definire quali siano i limiti entro i quali muoversi rispetto a questo nuovo contesto tecnologico. Sebbene l'IA possa offrire opportunità per un maggiore coinvolgimento politico, infatti, i rischi legati alla disinformazione, alla manipolazione dell'opinione pubblica e all'erosione della fiducia sono ancora troppo lampanti e minacciosi per la salvaguardia dell'integrità dei processi democratici.

1.1 Uso strategico dei *deepfake* nella campagna elettorale

Come è stato analizzato nei capitoli precedenti, l'avvento di tecnologie basate sull'intelligenza artificiale ha introdotto nuove dinamiche nel panorama politico globale, con un impatto significativo sui processi elettorali e sulla stabilità democratica. Tra queste tecnologie, i *deepfake* in particolare emergono come uno strumento potente e potenzialmente insidioso, capace di influenzare l'opinione pubblica e alterare gli esiti delle competizioni politiche. La definizione stessa di "*deepfake* politico", precedentemente citata, sottolinea la natura intenzionale e manipolatoria di tali contenuti, il cui scopo primario è quello di esercitare un'influenza cognitiva sul pubblico di riferimento.[323]

L'impiego strategico dei *deepfake* nelle campagne elettorali si manifesta attraverso diverse tipologie, ognuna con specifiche finalità e modalità di impatto. Tra le tipologie di *deepfake* politici identificate dalla ricerca scientifica, troviamo quelli utilizzati per dipingere ironicamente o screditare figure politiche: si tratta di contenuti che possono variare da una semplice satira a vere e proprie campagne diffamatorie volte a danneggiare

[322] E. A. VINOGRADOVA, *op. cit.*, pp. 47-48.
[323] *Ivi*, pp. 44-45, 51.

la credibilità di un candidato o di un partito. Un'altra categoria pericolosa è rappresentata dai *deepfake* di prove, creati per presentare false evidenze alle autorità giudiziarie, con potenziali conseguenze significative in termini legali e politici. Inoltre, esistono i *deepfake* di minaccia o violenti, impiegati per costringere individui a compiere determinate azioni o per arrecare danno. La crescente sofisticazione di queste tecnologie permette anche la creazione di *deepfake* misti, che combinano elementi delle diverse tipologie, rendendo più difficile la loro identificazione e regolamentazione.[324]

Bisogna, comunque, tenere sempre a mente che l'obiettivo principale di un *deepfake* politico è, nella gran parte dei casi, legato alla volontà di influenzare il comportamento degli elettori, e di conseguenza, l'esito delle elezioni. Questo scopo può essere raggiunto agevolmente attraverso la disseminazione di informazioni false e la creazione di narrazioni fittizie, causando un consequenziale indebolimento dell'opposizione politica.[325] Il più grave *output* che ne consegue è la progressiva erosione della fiducia del pubblico verso i media ufficiali e le istituzioni democratiche: nello specifico, numerosi studi sociologici hanno dimostrato come l'uso di notizie false politiche durante le elezioni generi preoccupazione e sfiducia nei confronti dei media ufficiali tra il pubblico di riferimento.[326] Di fatto, un sondaggio condotto nel Regno Unito da *Luminate*[327] nel 2023 ha rivelato che oltre il 70% dei cittadini si dichiarasse preoccupato per l'impatto dei *deepfake* sulle imminenti elezioni nel Paese.[328]

Oggigiorno le campagne elettorali moderne ricorrono sempre più spesso a tecniche di *marketing* per la distribuzione di notizie false politiche, in particolare attraverso la pubblicità mirata; questo è un approccio strategico che può radicalmente alterare la situazione politica, permettendo di indirizzare messaggi manipolatori a segmenti specifici dell'elettorato. Di fatto, l'utilizzo di tecniche di *microtargeting*[329] può amplificare

---

[324] *Ivi*, p. 47.
[325] M. B. E. ISLAM, M. HASEEB, H. BATOOL, N. AHTASHAM, Z. MUHAMMAD, *op cit.*, pp. 458-460, 466.
[326] E. A. VINOGRADOVA, *op. cit.*, pp. 45-46.
[327] Fondazione creata nel 2018 con il fine di invogliare le persone a partecipare attivamente alla vita politica e a farsi carico del proprio dovere civico, affinché a tutti possa essere garantita la stessa capacità di accesso all'informazione.
[328] E. A. VINOGRADOVA, *op. cit.*, pp. 45-46.
[329] Il *microtargeting* è una strategia di *marketing* che consente di raggiungere segmenti di pubblico molto specifici con messaggi personalizzati, utilizzando dati digitali per profilare gli utenti. In sostanza, si tratta

l'effetto dei *deepfake*, adattando i contenuti alla suscettibilità del pubblico di riferimento e rendendoli una forma di disinformazione efficace e realistica.[330]

Tuttavia, è da tenere in considerazione che l'impiego strategico dei *deepfake* non si limita alla creazione di contenuti malevoli. Più recentemente sta, infatti, emergendo una nuova categoria di applicazioni politiche definite "*softfake*": si tratta di *deepfake* politicamente motivati ma non malevoli, impiegati da politici e dal *team* che si occupa di campagna elettorale durante le elezioni per raggiungere nuovi gruppi *target*, diffondere idee politiche, costruire rappresentazioni di sé favorevoli o persino implicare *endorsement* da personalità note (a volte anche defunte)[331], similmente a quanto accaduto nel caso di Trump e Taylor Swift.

In sostanza, i *softfake* permettono a partiti e candidati di raggiungere comunità marginalizzate, interagire più direttamente con i cittadini e personalizzare i messaggi elettorali. Tuttavia, nonostante la loro apparente innocuità, anche questi comportano rischi significativi, essendo in ogni caso in grado di ingannare gli spettatori sulle reali capacità, *background* o posizioni dei candidati. La mancata etichettatura chiara di tali contenuti aggrava ulteriormente le preoccupazioni, benché anche contenuti *softfake* etichettati possano veicolare informazioni false o fuorvianti. Un esempio emblematico è rappresentato dalle cosiddette *"robocall"* [332] utilizzate durante le elezioni indiane del 2024, che inventavano posizioni politiche attribuendole falsamente ai candidati.[333]

Negli Stati Uniti, la preoccupazione che siano utilizzati *deepfake* durante le elezioni è elevata, a causa del crescente timore che possano essere impiegati per screditare candidati o diffondere informazioni false sulle procedure di voto. Nonostante finora non vi siano state prove lampanti che un singolo *deepfake* sia stato in grado di influenzare decisamente l'esito di un'elezione democratica su scala globale, la loro proliferazione ha comunque

---

di indirizzare la pubblicità e i contenuti a gruppi di persone con caratteristiche, interessi e comportamenti simili.

[330] E. A. VINOGRADOVA, *op. cit.*, pp. 48-49.

[331] M. PAWELEC, *op. cit.*, pp. 13-15.

[332] Una *robocall* è una telefonata automatizzata, ovvero una chiamata effettuata da un sistema computerizzato che riproduce messaggi preregistrati a un gran numero di persone contemporaneamente. Sono spesso utilizzate per *telemarketing* o truffe, ma anche per scopi legittimi come notifiche o promemoria.

[333] M. PAWELEC, *op. cit.*, p. 15.

contribuito ad erodere ulteriormente la fiducia del pubblico nelle istituzioni e ad aumentare la polarizzazione sociale. Tuttavia, è importante notare che negli USA sono stati compiuti sforzi per regolamentare questo fenomeno, con 17 Stati che, al 2024, avevano adottato leggi riguardanti i *deepfake* politici, seppur frammentate.[334]

Tale crescente preoccupazione non riguarda solo gli Stati Uniti, in quanto a livello globale nessun Paese si è dimostrato estraneo a queste tecniche di manipolazione elettorale. Un esempio lampante è la campagna presidenziale argentina del 2023, durante la quale entrambi i candidati e i loro *team* hanno ampiamente diffuso *deepfake*. Similmente, nelle elezioni parlamentari slovacche del 2023, un *deepfake* audio diffamatorio nei confronti del candidato progressista Michal Šimečka è circolato pochi giorni prima del voto, rendendo difficile una correzione tempestiva. In Turchia, durante le elezioni presidenziali del 2023 Muharrem İnce, uno dei tre candidati, si è addirittura ritirato dalla corsa dopo la diffusione non consensuale di un *deepnude*.[335]

Paesi come India, Cina, Singapore, Gran Bretagna, Corea del Sud, Australia e Giappone, oltre all'Unione Europea, stanno attivamente contrastando la diffusione di contenuti falsi dannosi, anche a livello legislativo.[336] In Corea del Sud, la Commissione Nazionale Elettorale[337] ha rivelato 129 casi di utilizzo di *deepfake* correlati alle imminenti elezioni tra la fine del 2023 e l'inizio del 2024, tutti in violazione della legge sull'elezione dei funzionari pubblici.[338] In particolare, in tale occasione alcuni candidati hanno persino utilizzato *avatar* generati al computer per animare il dibattito politico, riscuotendo un enorme successo tra il pubblico.

Nondimeno, anche in contesti non strettamente elettorali, i *deepfake* politici vengono impiegati strategicamente per destabilizzare situazioni politiche internazionali: un esempio emblematico riguarda alcuni attori filorussi che hanno condotto una campagna di disinformazione mirata contro la Francia, il presidente francese Emmanuel Macron e il Comitato Olimpico Internazionale in vista e durante le Olimpiadi di Parigi 2024. Questa

[334] *Ivi*, pp. 7-10.
[335] *Ivi*, p. 7.
[336] E. A. VINOGRADOVA, *op. cit.*, pp. 45-46.
[337] Istituzione costituzionale indipendente della Corea del Sud, istituita per gestire elezioni libere ed eque, *referendum* nazionali e altre questioni amministrative riguardanti i partiti politici e i fondi.
[338] E. A. VINOGRADOVA, *op. cit.*, pp. 52-53.

campagna includeva un "documentario" denigratorio sul Comitato Olimpico Internazionale narrato da un clone vocale sintetico dell'attore Tom Cruise. Un altro esempio di impiego strategico, sebbene non direttamente legato a competizioni elettorali tradizionali, riguarda il caso dell'*ex leader* pakistano Imran Khan e il suo utilizzo di un *deepfake* per rivolgersi a un comizio elettorale *online* nel dicembre 2023, sorpassando le barriere del confinamento dato dalla sua permanenza in carcere.[339]

In sostanza, l'impiego strategico dei *deepfake* sembra essere mirato anche a incrementare la discordia sociale e a diminuire la fiducia del pubblico di riferimento nei media.[340] La capacità di creare contenuti manipolatori su misura per specifici segmenti dell'elettorato può esacerbare le divisioni politiche preesistenti e rafforzare le convinzioni partitiche, rendendo più difficile un dibattito pubblico basato su fatti condivisi.[341] Non è infatti da escludere la possibilità che la diffusione di *deepfake* di *leader* politici possa rapidamente influenzare la coscienza collettiva, generare sfiducia nel governo e potenzialmente innescare disordini o, *in extremis*, colpi di Stato. Essi possono deteriorare l'ambiente informativo, aumentando la probabilità che informazioni false si diffondano durante le campagne; la creazione di *deepfake* via via più convincenti rende altrettanto difficile per i cittadini distinguere la realtà dalla finzione, erodendo inevitabilmente la loro fiducia nelle fonti di informazione tradizionali.[342]

In definitiva, strumenti come le campagne di disinformazione mirate alla creazione di *softfake* apparentemente innocui nonché la diffusione di video falsi con l'intento di minare la reputazione e il successo di un candidato dello schieramento opposto, sono impiegati in modi sempre più sofisticati e, simultaneamente al crescente (ab)uso di *deepfake*, l'azione concertata a livello legislativo, tecnologico ed educativo sembra essere l'unica soluzione per mitigare i rischi e salvaguardare la fiducia dei cittadini nel sistema democratico.

[339] M. PAWELEC, *op. cit.*, p. 14.
[340] E. A. VINOGRADOVA, *op. cit.*, pp. 48-49.
[341] M. PAWELEC, *op. cit.*, pp. 8-9.
[342] E. A. VINOGRADOVA, *op. cit.*, pp. 47-49.

1.2 Il prodotto di un *deepfake* virale: tra disinformazione e innovazione politica

Assodata la capacità di un *deepfake* politico di guadagnare viralità così velocemente da rendere difficile, se non impossibile, arginare la diffusione e le conseguenze di tali contenuti, appare lecito domandarsi quanto questa tecnologia a nostra disposizione possa essere considerata un'innovazione politica e quanto invece il suo impiego sia troppo rischioso in termini di diffusione esponenziale di disinformazione.

Se da un lato le tecnologie basate sull'intelligenza artificiale possono offrire nuovi strumenti per la comunicazione politica, come ad esempio *avatar* generati al computer per interagire con l'elettorato[343], o la personalizzazione estrema dei messaggi elettorali attraverso i *softfake*, l'impiego di *deepfake* manipolativi e virali si discosta radicalmente dalla natura di innovazione costruttiva.

Il *deepfake* virale a scopo manipolatorio trascende totalmente la sfera della semplice innovazione comunicativa per configurarsi primariamente come una pericolosa forma di disinformazione e un'arma di manipolazione psicologica di massa. La sua efficacia risiede nella capacità di ingannare la percezione del pubblico, sfruttando la credibilità intrinseca che il formato audiovisivo ancora possiede nell'era digitale: un *deepfake* ben realizzato può apparire indistinguibile da un contenuto autentico, soprattutto se alla portata di un pubblico tecnologicamente analfabeta.[344]

I danni che possono scaturire dall'impiego di un *deepfake* in ambito elettorale non sono assolutamente da sottovalutare: la possibilità di estorsione, diffamazione, intimidazione, bullismo e minamento della fiducia implicano un danno psicologico non trascurabile. Un *deepfake* virale può facilmente trasformarsi in uno strumento di diffamazione su larga scala, danneggiando irreparabilmente sia la reputazione che la carriera di un individuo.

Similmente, è possibile innescare un grave danno anche a livello finanziario: se un *deepfake* virale ritrae un politico o un *leader* aziendale in una situazione finanziariamente compromettente, le conseguenze economiche possono essere significative e immediate.

---

[343] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *Artificial intelligence, disinformation and media literacy proposals around deepfakes, Observatorio (OBS*) Journal*, 2024, pp. 188-189.
[344] *Ibidem.*

In ultima istanza, la conseguenza più palese riguarda l'aspetto sociale: la manipolazione delle preferenze elettorali e la diffusione di informazioni false circa le procedure di voto possono minare la fiducia non solo nei confronti di un singolo candidato e del suo partito o schieramento di riferimento, ma anche verso il processo democratico nel suo complesso.[345]

Oltretutto, la capacità di sfruttare piattaforme di *social media* e sistemi di messaggistica istantanea per una diffusione capillare e spesso anonima aggrava ulteriormente il problema.[346] Numerosi studi affermano che, entro il 2026, una percentuale elevatissima di contenuti *online* potrebbe essere generata sinteticamente, rendendo l'uso di *deepfake* una potenziale fonte comune di *cybercrime* e interferenza elettorale.[347]

Se la diffusione crescente di *deepfake* politici si individua, dunque, come una minaccia alla corretta informazione, piuttosto che come una mera innovazione, allora è necessario munirsi delle giuste precauzioni prima che il danno diventi irreversibile. Oltre all'imprescindibile regolamentazione legislativa, anche lo sviluppo di tecnologie di rilevamento è notevolmente valido per agire in maniera più capillare e tempestiva: la ricerca si concentra fondamentalmente sulla creazione di strumenti basati sull'intelligenza artificiale capaci di identificare *deepfake* attraverso l'analisi di anomalie nei dati audiovisivi.[348] L'uso della *blockchain* (libro mastro condiviso e immutabile che facilita il processo di registrazione delle transazioni e di monitoraggio degli *asset* in una rete aziendale) è anche esplorato come *framework* per la verifica dell'autenticità dei contenuti digitali.[349]

Allo stesso modo, non è trascurabile l'alfabetizzazione mediatica: iniziative di educazione pubblica che mirino a sensibilizzare i cittadini sui rischi dei *deepfake* potrebbero fornire loro le competenze necessarie per riconoscerli e valutarne criticamente

[345] E. A. VINOGRADOVA, *op. cit.*, pp. 48-49.
[346] M. R. SHOAIB, Z. WANG, M. T. AHVANOOEY, J. ZHAO, *Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models, International Conference on Computer and Applications (ICCA), IEEE*, 2023, pp. 1-7.
[347] E. A. VINOGRADOVA, *op. cit.*, p. 57.
[348] M. N. A. LATIF AL WAROI, *False Reality: Deepfakes in Terrorist Propaganda and Recruitment, Security Intelligence Terrorism Journal (SITJ), Vol. 01 No. 01*, 2024, pp. 44.
[349] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *op. cit.*, p. 189.

l'autenticità.[350] Corsi e materiali informativi vengono sviluppati per contrastare la disinformazione e promuovere un consumo consapevole dei media; le organizzazioni di *fact-checking* nascono proprio con l'intento di verificare i contenuti e di smentire la veridicità dei *deepfake* virali.[351]

Per rispondere, dunque, al quesito iniziale, il prodotto di un *deepfake* virale in campo elettorale non può essere considerato una legittima innovazione politica: pur sfruttando tecnologie avanzate in piena legalità, la sua natura intrinsecamente manipolatoria e il suo potenziale di causare danni significativi alla sfera pubblica e democratica lo collocano inequivocabilmente nella categoria della pericolosa disinformazione. La sua capacità di ingannare su vasta scala, erodere la fiducia, polarizzare la società e influenzare gli esiti elettorali rappresenta una seria minaccia all'integrità dei processi democratici a livello globale che non può essere elusa mascherandola come una mera satira. La viralità, in questo contesto, forse non è un indicatore di successo o innovazione positiva, ma piuttosto un moltiplicatore della pericolosità della diffusione di informazioni false e volutamente dannose.

1.3 Risposte pubbliche e private: il ruolo delle piattaforme digitali

Gli Stati Uniti registrano un'attività legislativa crescente a livello statale volta a proibire la manipolazione di registrazioni audio e materiali visivi con l'intento di influenzare gli esiti elettorali. Alcune proposte normative considerano l'obbligo per le aziende produttrici di *deepfake* di auto-identificare i propri contenuti attraverso filigrane digitali, al fine di accrescere la consapevolezza del pubblico sulla potenziale non autenticità di tali materiali.[352] Parallelamente, si sta valutando l'adeguatezza del quadro giuridico esistente: la stessa Casa Bianca, attraverso l'Ufficio di Politica Scientifica e Tecnologica, ha iniziato a pubblicare rapporti preliminari sulla preparazione al futuro dell'IA, segnalando un riconoscimento governativo circa la significatività della questione.[353]

---

[350] M. R. SHOAIB, Z. WANG, M. T. AHVANOOEY, J. ZHAO, *op. cit.*, pp. 1-7.
[351] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *op. cit.*, pp. 176-181.
[352] M. CHAWKI, *Navigating legal challenges of deepfakes in the American context: a call to action, Cogent Engineering*, *11 (1),* 2024, p. 1.
[353] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *op. cit.*, p. 178.

Tuttavia, la distinzione tra un uso espressivo protetto, come la satira o il commento sociale, e una manipolazione dannosa con intenti ingannevoli non è sempre nitida e il Primo Emendamento della Costituzione statunitense rende la differenza tra le due ancora più sottile. In questo scenario, sono state proposte modifiche alle leggi sul diritto alla propria immagine (*Right of Publicity Acts*) per includere disposizioni specifiche riguardanti le "repliche digitali" manipolate, cercando di proibire usi non consensuali che potrebbero essere percepiti come autentici da una persona ragionevole.[354]

La Sezione 230 del *Communications Decency Act* (CDA230), ha tradizionalmente conferito alle piattaforme *online* una significativa protezione dalla responsabilità per i contenuti pubblicati dagli utenti; tuttavia, recenti sviluppi giuridici, come le pronunce della Corte Suprema nel 2023, hanno messo in discussione l'estensione di questa immunità, stabilendo che spetta al legislatore, e non alla magistratura, determinarne i confini. Ciò apre la strada a possibili futuri interventi legislativi che potrebbero ridefinire la responsabilità delle piattaforme nella diffusione di contenuti dannosi, inclusi i *deepfake*. Il dibattito, ancora una volta, si concentra sulla necessità di bilanciare la protezione della libertà di espressione *online* con l'imperativo di contrastare la diffusione di disinformazione e manipolazione.[355]

In questo contesto, è possibile che l'assenza di una regolamentazione federale onnicomprensiva e la protezione costituzionale accordata a determinate forme di espressione potrebbero implicitamente consentire la proliferazione di alcune categorie di *deepfake*, in particolare quelle con intenti satirici o di commento sociale che però non superano la soglia della diffamazione o dell'incitamento alla violenza. La difficoltà intrinseca nel definire con precisione i confini tra espressione protetta e disinformazione dannosa rende ardua una proibizione totale e potrebbe condurre a una tolleranza implicita di forme di manipolazione digitale che rientrano nella sfera del discorso politico protetto.[356]

---

[354] A. PREMINGER, M. B. KUGLER, *The right of publicity can save actors from deepfake armageddon*, *Berkeley Technology Law Journal, Forthcoming Northwestern Public Law Research Paper No. 23-52*, 2024, pp. 148, 153.

[355] G. GOSZTONYI, F. G. LENDVAI, *Online platforms and legal responsibility: A contemporary perspective in view of the recent U.S. developments, Masaryk University Journal of Law and Technology*, 2024, pp. 125-129, 136-137.

[356] A. R. HUBER, Z. WARD, *op. cit.*, pp. 14-18.

Rispetto all'atteggiamento poco deciso e chiaro dell'apparato statale statunitense, le piattaforme digitali si trovano senza dubbio in una posizione molto delicata. Da un lato, sono sempre più consapevoli del loro ruolo nella diffusione di disinformazione, inclusi i *deepfake*, e per questo alcune di esse hanno implementato politiche volte a contrastare la circolazione di contenuti manipolati, specialmente in contesti elettorali;[357] tuttavia, la moderazione dei contenuti su piattaforme con un'utenza globale rappresenta una sfida di proporzioni immense, e persistono preoccupazioni riguardo al livello di controllo esercitato dalle grandi aziende tecnologiche sull'informazione politica e al potenziale rischio di *bias* nelle loro politiche di moderazione.[358]

Mentre il CDA230 ha storicamente offerto una protezione significativa, la sua futura interpretazione e possibili modifiche legislative potrebbero, però, portare a un ripensamento del ruolo delle piattaforme, non più considerate come semplici mezzi di informazione.[359]

Un'ulteriore difficoltà per le piattaforme risiede nella distinzione tra opinioni e fatti, un aspetto cruciale nella valutazione della disinformazione. I *deepfake*, per loro natura, mirano a presentare come veritieri degli eventi o delle dichiarazioni che nella realtà non lo sono, rendendo la linea di demarcazione particolarmente labile.[360] Inoltre, la rapidità con cui i *deepfake* possono essere creati e diffusi, spesso sfruttando tecniche di *microtargeting* per raggiungere segmenti specifici di pubblico, unita alla difficoltà di distinguerli dai contenuti autentici, rende estremamente complesso per le piattaforme intervenire in modo tempestivo ed efficace.[361]

Con riferimento specifico ai *deepfake* politici, è inutile nascondere che oramai i *social media* sono il luogo entro cui si forma e si diffonde la comunicazione politica, il che apre la strada all'utilizzo dell'IA per influenzare le opinioni politiche e le elezioni. Le piattaforme si trovano quindi a dover bilanciare il loro ruolo come spazi di libera

[357] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *op. cit.*, p. 189.
[358] E. A. VINOGRADOVA, *op. cit.*, pp. 47-48.
[359] G. GOSZTONYI, F. G. LENDVAI, *op. cit.*, pp. 125-129, 134-137.
[360] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *op. cit.*, p. 190-191.
[361] E. A. VINOGRADOVA, *op. cit.*, pp. 45, 52-56.

espressione con la necessità di proteggere gli utenti dalla manipolazione e dalla disinformazione veicolata dai *deepfake* politici.[362]

Se la posizione statunitense riguardo il campo di azione delle piattaforme sembra essere alquanto vaga, sul versante europeo, invece, tutt'altro approccio è stato adottato con il *Digital Services Act*, il quale ha già iniziato a mettere in discussione il ruolo delle piattaforme come potenziali "editori" capaci di influenzare la portata dei contenuti, introducendo obblighi più espliciti in termini di responsabilità e moderazione.[363]

L'Unione Europea, mediante l'applicazione del DSA, tiene molta considerazione delle responsabilità affibbiate alle piattaforme circa la diffusione di contenuti generati tramite intelligenza artificiale, e non si tira indietro nel sanzionare eventuali atteggiamenti non conformi al regolamento. È il caso di quanto successo recentemente con X, la piattaforma acquistata da Elon Musk nell'ottobre del 2022, precedentemente conosciuta come *Twitter*: l'Unione Europea ha infatti avviato a dicembre 2023 un'indagine approfondita su X ai sensi del *Digital Services Act*, con l'accusa che la piattaforma non stia adempiendo ai suoi obblighi nella lotta contro la diffusione di contenuti illegali e la manipolazione dell'informazione.[364]

In particolare, l'UE punta ad esaminare il funzionamento degli algoritmi di raccomandazione dei contenuti di X, sospettando che possano essere stati manipolati per dare maggiore visibilità a *post* e politici di estrema destra.[365] La Commissione Europea ha infatti richiesto a X l'accesso a documenti interni sui suoi algoritmi e sulle sue pratiche di moderazione dei contenuti, sottolineando la necessità di fare luce sulla conformità della piattaforma con gli obblighi del DSA. L'UE teme che la manipolazione algoritmica possa influenzare indebitamente le elezioni europee, amplificando alcune narrazioni politiche a scapito di altre e mettendo a rischio la correttezza del processo democratico. Alcuni

[362] A. RUDNIEVA, A. O. РУДНЄВА*, op. cit.*, pp. 174-175.
[363] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *op. cit.*, p. 177.
[364] M. STRAUSS, P. BLENKINSOP, *EU steps up probe into Musk's X, days ahead of Trump inauguration, Reuters*, 2025.
[365] L. O' CARROLL, *EU asks X for internal documents about algorithms as it steps up investigation, The Guardian*, 2025.

politici europei hanno effettivamente accusato Musk di interferire nelle elezioni, ad esempio attraverso le sue dirette *streaming* con figure politiche controverse.[366]

Elon Musk, dal suo canto, antepone il Primo Emendamento e la libertà di espressione a qualsiasi altro diritto, andando quindi a colpire e prevaricare altri diritti tutelati dal DSA. La Commissione ha in ogni caso chiarito che Musk è libero di esprimere le proprie opinioni, ma sta valutando se gli algoritmi di X amplifichino in modo distorto alcune narrazioni, potenzialmente danneggiando la parità di condizioni nelle competizioni elettorali.[367]

Questa azione legale dimostra un approccio europeo più proattivo e regolamentare nei confronti della responsabilità delle piattaforme rispetto a quanto tradizionalmente osservato negli Stati Uniti con il CDA230: mentre gli Stati Uniti si concentrano maggiormente sulla protezione della libertà di espressione, coerentemente con l'ideologia di Musk, l'UE sembra adottare un approccio più interventista, mirando a garantire un ambiente informativo *online* più sicuro e trasparente, soprattutto in vista di importanti appuntamenti elettorali.[368]

2. Censura e controllo nell'era digitale cinese: il caso Xi Jinping

Contrariamente a quanto avviene negli Stati Uniti, La Cina ha stabilito un rapporto profondo e strategico con l'intelligenza artificiale, considerandola uno strumento cruciale non solo per lo sviluppo tecnologico ed economico, ma anche per le sue strategie di sicurezza nazionale e per il mantenimento del controllo sociale e politico.[369]

Più recentemente, sotto la guida di Xi Jinping, la situazione della censura in Cina ha subito significative trasformazioni, segnando un allontanamento dalle metodologie precedenti e introducendo un sistema più centralizzato, tecnologicamente avanzato e pervasivo.[370]

---

[366] M. STRAUSS, P. BLENKINSOP, *EU steps up probe into Musk's X, days ahead of Trump inauguration, Reuters*, 2025.
[367] *Ibidem.*
[368] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *op. cit.*, p. 186.
[369] L. Y. HUNTER, C. D. ALBERT, J. RUTLAND, K. TOPPING, C. HENNIGAN, *Artificial intelligence and information warfare in major power states: how the US, China, and Russia are using artificial intelligence in their information warfare and influence operations, Defense & Security Analysis*, 2024, pp. 3-4.
[370] B. HILLMAN, K. CHIEN-WEN, *Political and Social Control in China: The Consolidation of Single-Party Rule, ANU Press*, 2024, p.1.

Inizialmente, la moderazione dei contenuti *online* era gestita in gran parte da censori impiegati dal governo e moderatori delle piattaforme per monitorare e rimuovere materiale politicamente sensibile. Tuttavia, l'esplosione dei contenuti generati dagli utenti sui *social media*, sulle piattaforme di notizie e sui siti di condivisione di video ha reso la censura manuale insufficiente a gestire l'enorme volume di informazioni: questo ha portato alla transizione verso sistemi di censura basati sull'intelligenza artificiale, che ora costituiscono la spina dorsale del quadro normativo di internet in Cina.[371]

L'evoluzione della censura basata sull'IA è stata ampiamente guidata dai progressi tecnologici nell'elaborazione del linguaggio naturale, nel *deep learning* e nell'analisi del *sentiment*, parallelamente a severe normative governative che impongono una sorveglianza *online* completa. Piattaforme accreditate come *WeChat*, *Weibo* e *Douyin* hanno integrato strumenti di censura basati sull'IA, consentendo la scansione automatizzata dei contenuti, il riconoscimento di *pattern* e il filtraggio predittivo. I sistemi IA sono estremamente funzionali in questo ambito poiché in grado di segnalare, sfocare o rimuovere contenuti istantaneamente, riducendo significativamente il tempo necessario per applicare le politiche di censura; questa transizione ha, infatti, segnato un punto di svolta nella gestione della rete in Cina: a differenza dei moderatori umani, i sistemi IA sono in grado di elaborare milioni di *post* al minuto e di monitorare efficientemente le narrazioni *online*, capacità imprescindibile specialmente durante quei periodi politicamente sensibili.[372]

Questa rinnovata determinazione e i nuovi apparati per domare i *social media* e controllare il discorso *online* sembrano ormai essere al centro delle politiche di Xi Jinping: la censura è infatti diventata più internalizzata, sistematica e di vasta portata sotto la sua *leadership*. Questa concettualizzazione della censura da parte di Xi detta la continuità e i cambiamenti nelle politiche del partito sul controllo dei media: mentre la censura di Stato è ulteriormente confermata a livelli strutturali e legislativi, essa è ora realizzata da una rete di agenti censori, con il partito e il presidente stesso all'apice del potere. L'autorità e il controllo del Partito Comunista Cinese (PCC) su agenzie e individui

[371] Y. CHEN, *The Accuracy and Biases of AI-Based Internet Censorship in China, Journal of Research in Social Science and Humanities*, 2025, pp. 27-29.
[372] *Ibidem*.

avvengono in gran parte attraverso una "auto-gestione" foucaultiana[373], in cui le piattaforme internet e persino i *netizen*[374] sono spinti a diventare i censori di loro stessi.[375]

Lo scopo, come definito da Xi Jinping, è quello di ottenere un'ecologia *online* chiara e incontaminata piena di "energia positiva", un termine popolare manipolato per rappresentare la superiorità ideologica del PCC e per reprimere la libera espressione *online* con il pretesto che questa incarni qualcosa di meramente negativo per l'agenda nazionale. Il presidente cinese ha da subito percepito la necessità urgente di rafforzare il controllo su internet, che considera come "la più grande variante sul campo di battaglia del discorso pubblico", oltre che "una spina nel fianco piantata dall'Occidente per abbattere la Cina": così facendo, Xi ha avvertito i quadri superiori e i propagandisti che si incorrerebbe in errori storici irreversibili se il partito perdesse la sua presa ferma sul potere di guidare e gestire. Ciò ha dato origine al cosiddetto "*dao*"[376] di Xi sulla nuova censura, incarnato nei suoi discorsi sul controllo dei media da parte dello Stato-Partito, tradotto nel discorso ufficiale attraverso leggi e politiche, praticato dai suoi propagandisti a tutti i livelli e materializzato incrementalmente nella realtà digitale presente oggi in Cina. Tale prospettiva si discosta dalle linee tradizionali sulla censura autoritaria repressiva e fondata su un approccio *top-down*, enfatizzando la censura internalizzata attraverso una rete di istituzioni e agenti censori auto-gestiti.[377]

Xi Jinping concettualizza un sistema di governo in cui i fornitori di servizi, così come i *netizen* comuni, sono spinti ad assumere un ruolo proattivo nel mantenere e sorvegliare uno spazio digitale incontaminato; fallire nell'esecuzione di una censura corretta e tempestiva comporta infatti questioni giurisdizionali con la filiale locale del CAC (*Central Cyberspace Affairs Commission*) o, peggio ancora, sanzioni dirette. Attraverso le sue pesanti azioni punitive, lo Stato-Partito promuove un ambiente intimidatorio in cui

[373] L'approccio foucaultiano si basa sull'idea che il potere non sia solo un'entità repressiva, ma che permei la società e influenzi il modo in cui pensiamo, agiamo e ci relazioniamo.
[374] Un *netizen*, o più raramente un *cybercitizen*, è una persona che partecipa attivamente alla vita di internet, spesso con l'obiettivo di migliorarla. Si tratta di un termine che definisce gli utenti di internet che si impegnano nella vita *online*, frequentando comunità, partecipando a conversazioni e contribuendo alla discussione.
[375] B. HILLMAN, K. CHIEN-WEN, *op. cit.*, pp. 3, 10-11.
[376] In cinese, "*dao*" (道) significa "via", "sentiero" o "principio". In filosofia, il *dao* è un concetto centrale del pensiero cinese, soprattutto nel taoismo, e rappresenta la via dell'universo, il modo in cui tutto funziona.
[377] B. HILLMAN, K. CHIEN-WEN, *op. cit.*, pp. 3, 10, 14-15.

le piattaforme internet, in qualità di censori delegati, rispondono con eccessiva cautela ed entusiasmo, censurando qualsiasi informazione che potrebbe attirare l'attenzione delle autorità superiori e a sua volta innescare una sanzione: ciò consente al governo di essere meno visibile nel processo di censura effettivo pur rimanendo presente nel suo controllo attraverso la delega della censura. Il problema maggiore è che la maggior parte della censura delegata è attuata come prevenzione piuttosto che come mitigazione, poiché i fornitori di servizi sono spinti a praticare una stretta censura pre-pubblicazione e una rapida censura post-pubblicazione, piuttosto che aspettare che le informazioni sensibili si sviluppino in potenziali crisi di opinione pubblica. Oltre alla pre-censura attiva, la censura manuale attraverso la segnalazione, in cui i *netizen* comuni sono stati trasformati in complici del fenomeno, contribuisce anche all'identificazione e al silenziamento di informazioni sensibili o "dannose".[378]

Questo sistema di censura risulta complesso e stratificato, combinando filtraggio di parole chiave, analisi del *sentiment*, modelli di *machine learning* e supervisione umana: infatti, a differenza dei metodi di censura tradizionali, i sistemi basati sull'IA possono elaborare enormi quantità di contenuti *online* in tempo reale, consentendo alle autorità di rilevare e sopprimere discussioni sensibili prima che si diffondano. L'efficienza di questo sistema è stata evidente soprattutto durante l'epidemia iniziale di COVID-19, quando la censura basata sull'IA ha giocato un ruolo cruciale nella soppressione delle segnalazioni dei *whistleblower* e del giornalismo indipendente, oltre ad essere stata impiegata per sopprimere il dissenso *online*, le proteste e le narrative dell'opposizione.[379]

In sostanza, il governo cinese ha integrato l'IA in una vasta gamma di settori e regioni, e cerca attivamente di utilizzarla per migliorare il controllo sulla popolazione, compresa la profilazione e il controllo delle minoranze etniche; questa integrazione dell'IA in quasi ogni aspetto della tecnologia conferisce alla Cina vantaggi specifici nel controllo sociale e nella gestione delle informazioni, un aspetto ulteriormente rafforzato dalla *Cyber Security Law* del 2017[380]. Il governo ha persino creato la *Strategic Support Force* (SSF)

[378] *Ivi*, pp. 10-11, 16-17.

[379] L. Y. Hunter, C. D. Albert, J. Rutland, K. Topping, C. Hennigan, *op. cit.*, p. 17.

[380] La legge cinese sulla *Cybersecurity* (网络安全法) è entrata in vigore nel giugno del 2017. Gli obiettivi della legge sono descritti nell'articolo 1 come segue: garantire la sicurezza della rete; proteggere la sovranità del *cyber*spazio, la sicurezza nazionale e gli interessi pubblici.

nel 2015 per generare vantaggi strategici nel ciberspazio e integrare le capacità di guerra cibernetica, elettronica e psicologica. [381]

L'obiettivo dichiarato del governo sarebbe quello di garantire l'autenticità dei contenuti, la stabilità sociale e prevenire la manipolazione politica basata sull'IA, assicurando che la tecnologia *deepfake*, nello specifico, non mini la fiducia pubblica e la sicurezza nazionale[382], ma nella pratica il governo potrebbe altresì utilizzare la tutela degli "interessi nazionali" come scusa per censurare *deepfake* che promuovono narrazioni alternative o che mettono in discussione la versione ufficiale degli eventi, specialmente in periodi di tensione sociale o politica. La paura di incorrere nella censura a causa della vaghezza di questa clausola di tutela degli "interessi nazionali" potrebbe anche generare uno spaventoso effetto *chilling*[383], per cui i cittadini si auto-censurano per evitare possibili future ripercussioni.[384] Ne consegue che questa mancanza di trasparenza e di definizioni chiare in merito a cosa rientri in questa categoria solleva serie preoccupazioni riguardo la libertà di espressione e la possibilità che contenuti legittimi vengano soppressi.

Alla luce di quanto detto, non si può quindi escludere l'eventualità che la tecnologia sia effettivamente impiegata, in maniera più o meno occulta, per diffondere narrazioni favorevoli al regime o per screditare figure percepite come avversarie, soprattutto considerando il pervasivo controllo esercitato dallo Stato cinese sui media e sui flussi informativi. Ad esempio, un rapporto del Dipartimento di Stato degli Stati Uniti spiega come la Cina cerchi di contrastare la manipolazione delle informazioni nell'arena internazionale "inondando" le conversazioni *online* per soffocare i messaggi che percepisce come sfavorevoli ai suoi interessi sui motori di ricerca e sui *feed* dei *social media*: le ricerche mostrano come i media statali cinesi appaiano in modo prominente nei risultati di ricerca per termini chiave su piattaforme come *Google News*, *Bing News* e *YouTube*, in parte proprio perché l'IA alimenta i motori di ricerca e ne decide ogni risultato.[385]

---

[381] L. Y. Hunter, C. D. Albert, J. Rutland, K. Topping, C. Hennigan, *op. cit.*, pp. 15-16.
[382] S. Li, *The Social Harms of AI-Generated Fake News: Addressing Deepfake and AI Political Manipulation, Digital Society & Virtual Governance, Volume 1, Issue 1*, 2025, pp. 72-88.
[383] L'effetto *chilling* è un effetto inibitorio che consiste nella rinuncia a esercitare un proprio diritto per il timore di sanzioni legali.
[384] G. Gosztonyi, F. G. Lendvai, *op. cit.*, p. 132.
[385] L. Y. Hunter, C. D. Albert, J. Rutland, K. Topping, C. Hennigan, *op. cit.*, p. 18.

L'ascesa dell'IA generativa (GenAI) come *ChatGPT*, poi, apre nuove frontiere digitali per la propaganda e la disinformazione, essendo questi strumenti in grado di offrire immensa flessibilità e adattabilità a scopo di censura. Le autorità cinesi hanno ben presto realizzato sia la potenziale minaccia che la promessa dell'IA generativa: se, da un lato, la Cina è il primo Stato al mondo a implementare regolamentazioni che disciplinano l'uso domestico di tali strumenti e ad applicare la censura, d'altro canto, le agenzie di stampa e attori più clandestini ne stanno esplorando l'uso nei media per rafforzare l'immagine internazionale della Cina.[386] Pertanto, sotto il governo di Xi Jinping, la situazione della censura sembra essersi evoluta in un sistema sofisticato che utilizza l'IA non solo per il controllo e la soppressione interna, ma che considera anche attivamente il potenziale dell'IA generativa e dei *deepfake* come strumenti per plasmare le narrazioni sia a livello nazionale che internazionale, bilanciando la necessità di controllare la tecnologia internamente con il suo potenziale uso per scopi di influenza esterna.[387]

È, tuttavia, importante considerare che alcune analisi in lingua inglese potrebbero presentare una visione distorta o parziale, focalizzandosi in modo sproporzionato sugli aspetti negativi dello sviluppo e dell'utilizzo delle tecnologie di intelligenza artificiale e *cybersecurity* in Cina, così come accade in Russia, potenzialmente dipingendole come Paesi *leader* in questo campo, ma in un'ottica prevalentemente dispregiativa. Pertanto, è auspicabile approcciarsi alla letteratura scientifica e alle analisi in lingua inglese con un occhio critico, tenendo presente la possibilità di pregiudizi o focalizzazioni selettive che potrebbero influenzare la rappresentazione del ruolo di specifiche nazioni nello sviluppo e nell'utilizzo delle tecnologie di intelligenza artificiale.[388]

2.1 Manipolazione di video per il controllo sociale e politico

Nell'ambito delle strategie di informazione e di guerra informativa adottate dalle principali potenze, la Cina impiega senza dubbio una vasta gamma di tecnologie dell'informazione in diversi settori, con un *focus* particolare sulla cosiddetta "guerra di informatizzazione" o, in cinese, "*xinxi-hua*", che riguarda l'applicazione della tecnologia

[386] K. DRINHAUSEN, M. OHLBERG, I. KARÁSKOVÁ, G. STEC, *Image control: how China struggles for discourse power, Merics Report*, 2023, p. 11.
[387] B. HILLMAN, K. CHIEN-WEN, *op. cit.*, pp. 3, 10, 14-15.
[388] E. A. VINOGRADOVA, *op. cit.*, p. 58.

dell'informazione a tutti gli aspetti delle operazioni militari. Essenzialmente questa strategia include l'uso di *big data* e intelligenza artificiale per rafforzare la *leadership* del Partito Comunista Cinese e comprendere al meglio i cittadini.[389]

Non è un caso che strutture di propaganda sostenute dallo Stato siano equipaggiate con tecnologia *deepfake* potenziata dall'IA, sotterfugio che consente la creazione di narrazioni false più realistiche, generando immagini false e persino *videoclip* che coinvolgono figure chiave, allo scopo di supportare qualsiasi narrazione che i cosiddetti "*troll*" stiano cercando di promuovere. L'uso della tecnologia connessa a internet mira fondamentalmente a indebolire, manipolare e fuorviare le informazioni consumate dalle persone, ritenendo che ciò possa favorire gli obiettivi politici e militari dello Stato: questo si manifesta perlopiù nell'iniezione di disinformazione o misinformazione (in parte ottenute tramite attacchi informatici) e di notizie false che una maggioranza degli esposti crede siano vere al momento. Nella pratica vengono coordinate campagne di *post* inautentici per creare l'illusione di un ampio supporto popolare per una politica, un individuo o un punto di vista, anche quando tale supporto non esiste.[390]

Il PCC utilizza sia la "propaganda dura", rigida e didattica, sia la "propaganda morbida", più coinvolgente, che include film, serie TV, programmi di intrattenimento e contenuti sui *social media*. Queste forme di propaganda sono impiegate dal governo autoritario del PCC per monopolizzare il discorso pubblico, imporre il controllo governativo sulla società e influenzare i valori, le emozioni e gli atteggiamenti dei cittadini verso il governo e le politiche pubbliche; il controllo estensivo sui media conferisce, infatti, ai *leader* cinesi una maggiore facilità nel plasmare e promuovere la propria immagine rispetto alle *élite* democratiche. Ad esempio, sono stati analizzati i comunicati stampa ufficiali che documentano le attività pubbliche dei Segretari del Partito Provinciale (PPS) dal 2016 al 2022: questi rappresentano il canale principale per la diffusione di informazioni sulle attività dei *leader* e, sebbene aderiscano a determinate norme politiche, consentono flessibilità. I PPS hanno, però, sufficiente potere per plasmare i temi e le narrazioni dei

[389] L. Y. Hunter, C. D. Albert, J. Rutland, K. Topping, C. Hennigan, *op. cit.*, p. 16.
[390] *Ibidem.*

loro comunicati stampa in modo da adattarli ai loro obiettivi di costruzione dell'immagine.[391]

Sotto la *leadership* di Xi Jinping, il PCC ha rafforzato in modo più intenso il controllo sul sistema di propaganda, con politiche di censura più severe che limitano la libertà dei *reportage* dei media commerciali: ciò significa che i funzionari locali, specialmente i *senior leader*, non possono facilmente concedere interviste ai media commerciali. Mentre i *leader* locali, che in precedenza avevano un ampio seguito su piattaforme di *social media* come *Weibo*, hanno dovuto disattivare i loro *account* per evitare reazioni negative a commenti inappropriati. Di conseguenza, per le *élite* locali nell'era Xi Jinping, i comunicati stampa ufficiali sono diventati quasi l'unico canale attraverso cui mettere in mostra la loro immagine. Questo ha portato gli alti funzionari locali a prestare maggiore attenzione alle loro parole e azioni negli eventi pubblici, il che li ha spinti a esercitare un maggiore controllo sui dipartimenti di propaganda locali, garantendo che i *report* delle loro attività pubbliche fossero gestiti attentamente per soddisfare le aspettative dell'autorità centrale.[392]

Tuttavia, i comunicati stampa delle attività dei PPS mostrano *focus* e narrazioni molto distinti tra loro, dando vita a diverse preferenze di costruzione dell'immagine: queste preferenze includono le immagini di "competenza", "benevolenza", "fedeltà al partito" e "versatilità". L'immagine di "fedeltà al partito" è ben distinta dalle nozioni tradizionali di "benevolenza" o "competenza" poiché sottolinea la capacità dei funzionari di dimostrare obbedienza e lealtà politica all'autorità centrale del PCC e rafforzare la *leadership* generale del partito nelle loro rispettive regioni. Ad esempio, *reportage* dei media ufficiali mostrano i PPS in scenari diversi per proiettare immagini specifiche, come pragmaticità, vicinanza alla gente, lealtà al partito, o accessibilità. In regioni sensibili legate alla sovranità nazionale e alla sicurezza (come Tibet e Xinjiang), i funzionari aumentano la loro visibilità negli affari del partito, specialmente in termini di sicurezza politica e di presentazione di un fronte unito, riflettendo l'intento di creare uno stile di *leadership* forte per ridurre l'insoddisfazione sociale o le proteste pubbliche.[393]

---

[391] Y. YAN, Z. YANG, *Portraying Competence, Benevolence or Party Loyalty? Political Propaganda and the Image-Building of Political Elites in China, Comparative Politics*, 2025, pp. 1-6.
[392] *Ivi*, p. 5.
[393] *Ivi*, pp. 2, 10-11, 19.

Il PCC mobilita e amplifica il sostegno alle proprie posizioni, utilizzando cittadini, diaspora ed *élite* straniere; e al fine di gestire l'immagine proiettata di sé all'estero, il PCC esclude le voci cinesi, i ragionamenti e le visioni che non si allineano con la linea ufficiale di partito.[394] In regioni come Xinjiang, Tibet e Mongolia Interna, il PCC è addirittura ricorso ad *influencer* di *social media* popolari, donne e appartenenti a minoranze, per diffondere la narrativa del PCC secondo cui le condizioni politiche, economiche e sociali siano ideali così da respingere o ignorare le preoccupazioni sui diritti umani. I ricercatori suggeriscono che gli *influencer* siano stati probabilmente manipolati attraverso quelli che vengono definiti come "contenuti generati da utenti professionali", prodotti con l'aiuto di agenzie di gestione degli *influencer* note come "*multi-channel networks*". Queste agenzie sono direttamente controllate e finanziate dal PCC e sono progettate per propagare la narrativa del PCC; i *post* delle stesse vengono, di fatto, spesso prioritizzati dagli algoritmi dei motori di ricerca IA grazie all'elevato volume e frequenza di pubblicazione, riducendo la visibilità dei contenuti non affiliati al PCC che invece si occupano di sollevare preoccupazioni sui diritti umani.[395]

In realtà, dei ricercatori in Cina hanno affermato di aver sviluppato un'intelligenza artificiale "lettore del pensiero", in grado di misurare la lealtà dei cittadini al Partito Comunista Cinese: questo verrebbe fatto analizzando le espressioni facciali e le onde cerebrali dei cittadini in risposta a informazioni politiche. Sebbene la pubblicazione di questa ricerca sia stata rapidamente cancellata in seguito a diffuse critiche, le informazioni su tale sviluppo hanno contribuito alla decisione del Dipartimento del Commercio degli Stati Uniti di aggiungere l'*Academy of Military Medical Sciences* (AMMS) cinese e undici dei suoi istituti di ricerca alla propria *Entity List*[396]. La scelta statunitense è dipesa dall'utilizzo, da parte di questi istituti, dei processi biotecnologici a supporto di usi finali militari cinesi, come la presunta "arma di controllo cerebrale".[397]

[394] K. DRINHAUSEN, M. OHLBERG, I. KARÁSKOVÁ, G. STEC, *op. cit.*, pp. 7-8, 20.
[395] L. Y. HUNTER, C. D. ALBERT, J. RUTLAND, K. TOPPING, C. HENNIGAN, *op. cit.*, pp. 18-19.
[396] L'*Entity List* è un elenco di restrizioni commerciali pubblicato dal *Bureau of Industry and Security* del Dipartimento del Commercio degli Stati Uniti, composto da determinate persone, entità o governi stranieri. È pubblicato come Supplemento 4 della Parte 744 del Codice dei regolamenti federali.
[397] L. Y. HUNTER, C. D. ALBERT, J. RUTLAND, K. TOPPING, C. HENNIGAN, *op. cit.*, p. 15.

In definitiva, la manipolazione di video e contenuti digitali in Cina è una componente della più ampia strategia di informatizzazione e controllo del PCC, che utilizza IA e, più nello specifico, *deepfake* per creare narrazioni false realistiche, influenzare l'opinione pubblica, monopolizzare il discorso pubblico e plasmare positivamente l'immagine dei *leader*. Questa forte indole propagandistica è supportata da un estensivo controllo sui media e sui *social media*, inclusa la prioritizzazione algoritmica di contenuti affiliati allo Stato.[398]

Se è vero che i cittadini sono tutelati dalla disciplina delle *Deep Synthesis Provisions*, che impongono obblighi in capo ai fornitori di servizi, come la verifica, e l'etichettatura per contenuti potenzialmente fuorvianti, è altresì vero che queste normative, pur mirando a proteggere dalle attività illecite, non pongono un'enfasi significativa sull'alfabetizzazione informativa del pubblico come misura di difesa fondamentale, a differenza di quanto discusso, ad esempio, nel contesto europeo. Nel caso cinese, la "protezione" da queste tecnologie sembra inscindibile dal mantenimento del controllo statale.[399]

2.2 Reazioni della comunità internazionale e impatti geopolitici

Dinanzi alle emergenti minacce poste dall'uso dell'intelligenza artificiale e, in particolare, della tecnologia *deepfake*, per scopi di disinformazione e manipolazione, così come abbiamo visto accadere in Cina, la comunità internazionale mostra una crescente preoccupazione per l'uso dell'IA e dei *deepfake* nell'ambito elettorale e, più in generale, per l'integrità delle informazioni. La mancanza di *standard* riconosciuti a livello internazionale sulla regolamentazione di queste tecnologie può portare ad abusi o, al contrario, a una certa accondiscendenza a livello nazionale.[400]

Sussiste, in ogni caso, un generale appello alla cooperazione internazionale per affrontare le implicazioni dei *deepfake* e per migliorare l'alfabetizzazione mediatica del pubblico, essendo questo un elemento essenziale per contrastare l'attività illegale dei *deepfake* e

[398] K. DRINHAUSEN, M. OHLBERG, I. KARÁSKOVÁ, G. STEC, *op. cit.*, pp. 14-15.
[399] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 17-19.
[400] E. A. VINOGRADOVA, *op. cit.*, pp. 48, 53.

mitigarne l'impatto, un aspetto che gli approcci regolamentari attuali tendono a trascurare concentrandosi piuttosto su produttori, fornitori e piattaforme.[401]

La possibilità che la Cina esporti le sue tecnologie di controllo basate sull'IA a regimi autoritari è una preoccupazione reale, dal momento che alcuni Stati potrebbero pensare di adottare l'approccio cinese, anziché reagirvi negativamente. Questo aspetto potrebbe rivelarsi molto pericoloso dal momento in cui la Cina integra in modo significativo l'IA nelle sue strategie e tattiche di *Information Warfare and Influence Operations* (IWIO)[402], considerando queste attività nel contesto di un conflitto continuo con l'occidente. L'identità autoritaria del regime cinese consente infatti una raccolta dati più aggressiva sia a livello nazionale che internazionale, dati che vengono poi utilizzati in algoritmi di IA per le operazioni IWIO, diffondendo disinformazione, propagando le narrazioni del PCC e cercando di minare la fiducia nei governi democratici.[403]

La Cina utilizza l'IA per migliorare il controllo della popolazione e per profilare e controllare le minoranze etniche, com'è lampante nella regione dello Xinjiang, dove le autorità cinesi tentano attivamente di manipolare e dominare il discorso globale sul trattamento degli Uiguri[404] *online* e sui *social media*. La strategia adottata è quella di inondare le conversazioni per soffocare i messaggi percepiti come sfavorevoli e garantendo che i media statali cinesi figurino in modo prominente nei risultati dei motori di ricerca, con l'IA che alimenta quasi ogni parte di tali motori. [405]

Per quanto riguarda invece l'uso di IA generativa, come gli strumenti simili a *ChatGPT*, questo è considerato in Cina come un'opportunità per migliorare le operazioni di influenza, riducendo i costi e consentendo la creazione di messaggi più personalizzati e autentici, potenzialmente facilitando interazioni "organiche" con *bot* di propaganda. La Cina ha, di fatto, riconosciuto sia il potenziale che la minaccia posti da tali strumenti: da

[401] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 17-19.
[402] Attività volte ad influenzare l'opinione pubblica o ad alterare il comportamento di un gruppo *target* attraverso l'uso di informazioni, spesso con l'obiettivo di raggiungere vantaggi politici o militari. Si tratta di una tattica di guerra spesso definita come "guerra psicologica" o "operazioni psicologiche".
[403] L. Y. HUNTER, C. D. ALBERT, J. RUTLAND, K. TOPPING, C. HENNIGAN, *op. cit.*, pp. 5-7.
[404] Gli Uiguri sono un'etnia turcofona di religione islamica che vive nel nord-ovest della Cina, soprattutto nella regione autonoma dello Xinjiang, insieme ai cinesi Han. Gli Uiguri costituiscono la maggioranza relativa della popolazione della regione.
[405] L. Y. HUNTER, C. D. ALBERT, J. RUTLAND, K. TOPPING, C. HENNIGAN, *op. cit.*, pp. 8-9.

un lato, agenzie di stampa e attori più clandestini in Cina stanno esplorando l'utilizzo di questi strumenti nei media per rafforzare l'immagine internazionale del Paese; dall'altro, le autorità cinesi hanno adottato un approccio regolamentare proattivo, essendo la Cina la prima regione al mondo a implementare normative che disciplinano l'uso domestico di tali strumenti e ad applicarne la censura tramite normative come le *Provisions on the Administration of Deep Synthesis Internet Information Services*.[406]

Come già analizzato in precedenza, l'approccio regolamentare cinese sui *deepfake* è incentrato sul soggetto, ponendo l'accento sulle organizzazioni e gli individui che offrono servizi utilizzando la tecnologia di *deep synthesis*. Questa regolamentazione richiede ai fornitori di servizi di rispettare le leggi, le norme sociali ed etiche, e adottare una corretta direzione politica, assumendosi la responsabilità per i contenuti generati, inclusi il divieto di produrre o diffondere *deepfake* con contenuti illegali o informazioni. Risulta chiaro come questo tipo di approccio miri principalmente a salvaguardare la stabilità sociale e garantire che i contenuti si allineino con la morale sociale, sebbene ciò possa potenzialmente limitare la libertà di espressione.[407] Ancora una volta, dunque, in contrasto con gli Stati Uniti, che fanno invece della libertà di espressione il caposaldo della loro legislazione.

In realtà, anche gli Stati Uniti utilizzano l'IA nelle loro operazioni IWIO, ma con un'enfasi diversa, principalmente concentrata sulla rilevazione e il contrasto delle operazioni di influenza avversarie. In particolare, governi autoritari e centralizzati come quelli di Cina e Russia hanno una maggiore capacità di raccogliere dati sia sulle loro popolazioni interne che esternamente, rispetto alle democrazie come gli Stati Uniti. Questo avviene tramite metodi più aggressivi, spesso senza alcuna preoccupazione per la *privacy* o le libertà civili, imprescindibili negli Stati Uniti dove le istituzioni pongono maggiori restrizioni alla capacità di raccolta e utilizzo di dati su larga scala: tali limitazioni derivano dai principi democratici, dalla protezione dei diritti alla *privacy* e dalla supervisione politica e burocratica che disciplina la raccolta e l'uso dei dati di sorveglianza interna.[408] È bene tenere a mente che la regolamentazione statunitense sui

[406] K. DRINHAUSEN, M. OHLBERG, I. KARÁSKOVÁ, G. STEC, *op. cit.*, p. 11.
[407] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 11-14.
[408] L. Y. HUNTER, C. D. ALBERT, J. RUTLAND, K. TOPPING, C. HENNIGAN, *op. cit.*, p. 9.

deepfake segue un paradigma basato sull'applicazione, concentrandosi principalmente sull'uso dei *deepfake* in contesti specifici e pragmatici come le elezioni e la pornografia, con un approccio considerato più clemente e guidato dal mercato rispetto alla Cina.[409]

Bisogna altresì osservare che la mancanza di un approccio centralizzato alla raccolta dati negli Stati Uniti rappresenta uno svantaggio rispetto a Stati come la Cina. Pertanto, gli Stati Uniti tendono a porre maggiore enfasi sullo sviluppo di programmi IA per rilevare e contrastare le minacce di IWIO, utilizzando l'intelligenza artificiale in modo difensivo piuttosto che offensivo, servendosi dei dati per identificare possibili discussioni o immagini di disinformazione; questo è il caso di strumenti come i *deepfake detector*s. Dunque, gli Stati Uniti generalmente separano le attività in tempo di pace e quelle in tempo di guerra nel dominio informativo, limitando le loro capacità di IWIO quando non sono in conflitto, a differenza di Cina e Russia che si impegnano costantemente in attività offensive come questione di strategia e politica. In effetti, attraverso strategie che includono il cosiddetto "*surveillance capitalism*" [410], questi Stati possono identificare e micro-targettizzare individui e gruppi specifici in base alle loro abitudini sociali e punti di vista politici, una capacità di cui chiaramente possono servirsi in misura maggiore rispetto alle democrazie per via delle loro strutture di governo autoritarie.[411]

La Cina, ad esempio, utilizza metodi di sorveglianza per raccogliere enormi quantità di dati che potenziano algoritmi d'intelligenza artificiale per scopi come il controllo della popolazione, la targhettizzazione di minoranze e, per la politica estera, la conduzione di IWIO volte a dividere e polarizzare le società democratiche, incluso gli Stati Uniti. Questi dati aiutano a personalizzare gli algoritmi IA per raggiungere gli obiettivi di *information warfare* della Cina, tra cui la diffusione internazionale delle narrative politiche del PCC e l'aumento della divisione nelle democrazie occidentali per minare la fiducia nei loro governi. Agenti cinesi, spesso supportati da IA, operano sui *social media* per propagare narrative favorevoli e indebolire la percezione dei *leader* e dei cittadini nemici.

---

[409] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 13-14.

[410] Il "capitalismo della sorveglianza" è un concetto introdotto da Shoshana Zuboff, che descrive un modello economico in cui le aziende raccolgono, analizzano e vendono dati personali per trarre profitto, spesso sfruttando tecniche di sorveglianza avanzate per influenzare il comportamento dei consumatori.

[411] L. Y. HUNTER, C. D. ALBERT, J. RUTLAND, K. TOPPING, C. HENNIGAN, *op. cit.*, pp. 1-2.

L'integrazione dell'IA in quasi ogni aspetto della tecnologia cinese consente vantaggi specifici nel controllo sociale e nella gestione delle informazioni: si ritiene infatti che la Cina stia usando tecnologie IA per migliorare il controllo della popolazione e che esporterà queste capacità a governi autoritari, oltre che pianificare di utilizzare l'intelligenza artificiale per monitorare e controllare lo spazio informativo relativo ai cittadini cinesi. Similmente, la Russia utilizza le informazioni raccolte tramite sorveglianza per sperimentare l'analisi guidata dall'IA al fine di identificare potenziali dissidenti politici e impiegando algoritmi per aumentare la polarizzazione politica negli Stati Uniti, ad esempio targhettizzando veterani o alimentando proteste e contro-proteste.[412]

Alla luce di quanto detto, la domanda cruciale da porsi è se queste misure adottate dalla Cina costituiscano principalmente una tutela degli interessi nazionali o una vera e propria censura: mentre le disposizioni cinesi mirano ufficialmente a salvaguardare la stabilità sociale impedendo che i *deepfake* violino gli *standard* legali o la morale sociale, la definizione di "morale sociale" è concettualmente vaga e ampia. Questo approccio basato sul soggetto e l'enfasi sull'allineamento con la morale sociale rischiano, infatti, di rendere i fornitori riluttanti a offrire servizi che potrebbero produrre contenuti controversi, potenzialmente riducendo la libertà di espressione e aprendo la strada ad una possibile censura di Stato.[413]

L'approccio cinese è, infatti, spesso accusato di modellare la repressione, dove un vasto apparato di propaganda, tipico elemento distintivo dei Paesi autoritari come la Cina, è propedeutico a monopolizzare il discorso pubblico e influenzare valori e atteggiamenti verso il governo.[414] L'esportazione di questo modello di censura basato sull'IA ad altri regimi autoritari, fornendo strumenti di moderazione dei contenuti *AI-based* a Paesi come Iran, Venezuela ed Etiopia, suggerisce che l'obiettivo va oltre la semplice tutela degli interessi nazionali, inclinando verso un controllo centralizzato delle informazioni e la soppressione del dissenso, il che risulterebbe più coerente con la nozione di censura.[415] Pertanto, sebbene le normative cinesi siano presentate come misure per la stabilità e la

[412] *Ivi*, p. 23.
[413] G. ZHENG, J. SHU, K. LI, *op. cit.*, pp. 12-15.
[414] Y. YAN, Z. YANG, *op. cit.*, 2025, pp. 3, 5.
[415] L. Y. HUNTER, C. D. ALBERT, J. RUTLAND, K. TOPPING, C. HENNIGAN, *op. cit.*, p. 26.

sicurezza nazionale, le preoccupazioni per la censura e la limitazione della libertà di espressione sono, senza dubbio, fondate.[416]

In definitiva, mentre la Cina persegue attivamente un uso offensivo e di controllo dell'IA e dei *deepfake* per scopi di IWIO e di stabilità interna, gli Stati Uniti si concentrano maggiormente sulla difesa e il contrasto con un approccio regolamentare mirato, in contrapposizione con l'Unione Europea che invece auspica ad una regolamentazione più ampia e preventiva, in un contesto internazionale dove la consapevolezza dei rischi sta crescendo ma mancano ancora *standard* globali e strategie di risposta coordinate per affrontare efficacemente le sfide poste dall'uso di queste tecnologie.

3. *Deep porn* e tutela delle celebrità: il caso Taylor Swift

L'avvento delle nuove tecnologie ha profondamente trasformato l'industria dell'intrattenimento, presentando al contempo sia nuove opportunità sia sfide significative, specialmente per le celebrità: nell'industria dell'intrattenimento e nel *marketing*, infatti, il potenziale dei *deepfake* è stato percepito come rivoluzionario ma anche minaccioso. Un *deepfake* di un attore può a tutti gli effetti apparire come l'attore stesso in un modo che la *computer grafica* (CGI) o una controfigura non potrebbero mai eguagliare, rappresentando potenzialmente una competizione diretta o ingannando il pubblico. È diventato possibile far recitare attori defunti in nuove produzioni, rivedere scelte di *casting* mesi dopo le riprese, e simulare comparse elettronicamente.[417]

Ma l'IA è stata utilizzata anche per innovare un linguaggio pubblicitario ormai obsoleto e per far rifiorire interesse in quei settori da cui le persone si sentono sempre meno attratte: un esempio a riguardo è stata la trovata di far "tornare in vita" l'artista Salvador Dalì nel museo a lui dedicato a Petersburg, riuscendo ad attirare una grande quantità di persone.[418]

L'IA può anche essere impiegata per generare contenuti per i *fan* o per sostituire attori in caso di necessità[419], oltre ad assistere i giornalisti in compiti come trascrizione, traduzione, classificazione di contenuti e riconoscimento di immagini, applicazione che

[416] M. PAWELEC, *op. cit.*, pp. 16-17.
[417] A. PREMINGER, M. B. KUGLER, *op. cit.*, pp. 102-103.
[418] M. CHAWKI, *op. cit.*, p. 6.
[419] A. PREMINGER, M. B. KUGLER, *op. cit.*, pp. 102-103.

potrebbe risultare adeguata anche alla gestione dell'immagine pubblica o la produzione di contenuti per le celebrità.[420]

Tuttavia, l'impatto dell'IA e dei *deepfake* sulle celebrità presenta anche un lato oscuro significativo: questa tecnologia, così potente, comporta rischi notevoli se non gestita correttamente, minacciando la sicurezza e la *privacy* degli individui.[421] Bisogna tenere a mente che i *deepfake* possono essere utilizzati in modi palesemente nefasti, tra cui la rappresentazione di celebrità impegnate in atti sessuali tabù, la fabbricazione di dichiarazioni false o la simulazione di azioni che non hanno mai compiuto, circostanze che minano anche la fiducia nei media e nelle comunicazioni autentiche. I *deepfake* possono infatti causare danni alla reputazione ed economici molto maggiori rispetto alle tradizionali forme di appropriazione dell'immagine: possono fungere da sostituti degli individui in contesti commerciali, permettendo di utilizzare l'immagine di una persona per vendere beni o servizi o recitare in qualsiasi scena. Inoltre, dato il loro realismo, i *deepfake* possono minacciare l'autonomia di un individuo dirottando funzionalmente il suo registro di comportamento; un *deepfake* sufficientemente potente può causare danni alla reputazione che anche il più sofisticato *team* di pubbliche relazioni non può annullare. Non da meno è il potenziale danno economico, basti pensare ad attori ben pagati, che possono guadagnare milioni per ruoli in film o serie TV, i quali affronterebbero potenzialmente una sbalorditiva perdita economica se le loro repliche digitali fossero utilizzate al loro posto. Gli attori temono che le loro "copie digitali" vengano create e utilizzate perpetuamente, a volte per un compenso minimo o nullo (come nel caso della proposta di scansionare comparse per una giornata di paga e usarle per sempre).[422]

Questi potenziali danni chiaramente sono fonte di preoccupazione nell'industria dell'intrattenimento, come percepito correttamente dalla *Screen Actors Guild* (SAG): i membri di SAG-AFTRA (*Screen Actors Guild-American Federation of Television and Radio Artists*) si sono uniti ai loro colleghi della *Writers Guild of America* in uno sciopero nel luglio 2023, chiedendo protezioni più rigorose contro le minacce poste dall'IA. Dopo

[420] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *op. cit.*, p. 189.
[421] J. KAUR, K. SHARMA, M. P. SINGH, *Exploring the Depth: Ethical Considerations, Privacy Concerns, and Security Measures in the Era of Deepfakes*, in *Navigating the World of Deepfake Technology, IGI Global*, 2024, p. 149.
[422] A. PREMINGER, M. B. KUGLER, *op. cit.*, p. 157.

ben 118 giorni di sciopero, gli attori hanno ottenuto una concessione significativa che richiede agli studi di ottenere il consenso degli attori prima di utilizzare immagini generate dall'IA e di compensarli a un tasso commisurato alla loro *performance* dal vivo.

Le leggi sul diritto di pubblicità, che tradizionalmente regolano l'uso commerciale dell'immagine di una persona, faticano però a controllare questa nuova forma di sfruttamento: le licenze tradizionali, spesso scritte in modo molto ampio per coprire l'uso dell'immagine "in tutti i media" e "in tutte le forme", non potevano prevedere l'uso di *deepfake*. Tuttavia, c'è una differenza fondamentale tra i *deepfake* e altre forme di immagine, come foto, CGI meno realistiche, o impersonatori, essendo i *deepfake* in grado di sostituire perfettamente il lavoro della persona e danneggiarne la reputazione in modi senza precedenti. Per affrontare ciò, sarebbe auspicabile considerare i *deepfake* come un tipo speciale di utilizzo dell'immagine; ciò comporterebbe la richiesta di licenze specifiche ed esplicite per la creazione e l'uso di *deepfake* commerciali, negoziate separatamente dalle autorizzazioni generali per l'uso dell'immagine, il che obbligherebbe le parti contrattuali a considerare esplicitamente la creazione di *deepfake ex ante*. Questa necessità di specificità è già vista come coerente con le disposizioni dell'accordo sindacale tra SAG-AFTRA del 2023, che richiedeva un nuovo consenso per riutilizzare l'immagine di una persona in un nuovo filmato, a meno che non fosse stata fornita una descrizione specifica dell'uso previsto nel contratto iniziale.[423]

Alcuni attori, come Keanu Reeves, hanno persino inserito clausole nei loro contratti per vietare le modifiche digitali alla loro recitazione. Reeves, nello specifico, ha definito i *deepfake* "spaventosi", decidendo di proteggersi legalmente attraverso accordi individuali. Anche se in altri casi i *deepfake* sono stati creati con licenza e approvazione, come nel caso di Bruce Willis che ha concesso in licenza il suo "gemello digitale" per una pubblicità, o Val Kilmer che ha utilizzato l'IA per ricreare la sua voce, tuttavia, l'assenza di un sistema standardizzato per garantire che i filmati non vengano abusati lascia le parti nel caos per cercare di proteggere i propri interessi economici e reputazionali.[424]

[423] *Ivi*, 118, 142-143, 146.
[424] *Ivi*, 103, 105-106.

Le celebrità sono senza dubbio tra le principali vittime dei *deepfake* per diverse ragioni: essendo figure pubbliche, le loro immagini, voci e video sono ampiamente disponibili su internet e sui *social media*, fornendo un vasto bacino di dati per l'addestramento degli algoritmi di *deepfake*; inoltre, la loro notorietà rende i video che li coinvolgono più virali e d'impatto, poiché vedere una figura conosciuta in una situazione inattesa o compromettente tende a generare maggiore attenzione e diffusione.[425]

Storicamente, i primi *deepfake* apparsi pubblicamente nel 2017 coinvolgevano lo scambio di volti di celebrità femminili in video pornografici non consensuali. Questo ha stabilito un precedente e ha reso le celebrità, in particolare le donne, un obiettivo frequente per i creatori di *deepfake* malintenzionati. [426]

Il fenomeno dei *deepfake* pornografici, o *Non-Consensual Intimate Deepfakes* (NCID), è, infatti, particolarmente dannoso e colpisce in modo sproporzionato le donne, incluse le celebrità femminili: questo tipo di contenuto implica l'uso di tecniche di IA per creare materiale pornografico iperrealistico senza il consenso della persona raffigurata, trovando un impiego particolarmente inquietante nella produzione di materiale non consensuale.[427] I *deepfake* che raffigurano donne in contesti pornografici spesso non solo cercano di modificare l'immagine pubblica di queste donne, ma anche di invalidarle a livello sociale e ridurle al silenzio nella sfera pubblica.[428] Un esempio noto di questa pratica ha visto protagonista una celebrità del calibro di Scarlett Johansson, i cui video *deepfake* pornografici sono stati ampiamente diffusi senza il suo consenso.[429] Questa forma di *deepfake* rappresenta una significativa minaccia disinformativa per la società, una forma di violenza sessuale di genere, e un esempio di violenza infrastrutturale della piattaforma, che si manifesta quando le piattaforme non proteggono sufficientemente le vittime.[430]

---

[425] M. CHAWKI, *op. cit.*, p. 6.
[426] M. PAWELEC, *op. cit.*, p. 6.
[427] R. RANJAN*, Deep fake porn: Duplicity and Dystopia, Law in the Age of Disruption. Understanding the Impact of Technology, Lex Assisto Media and Publications*, 2023, pp. 110-116.
[428] M. GARRIGA, R. RUIZ INCERTIS, R. MAGALLÓN ROSA, *op. cit.*, pp. 186-190.
[429] M. CHAWKI, *op. cit.*, p. 6.
[430] M. J. RIEDL, A. NEWELL, *Reporting Image-Based Sexual Violence: Deepfakes, #ProtectTaylorSwift, and Platform Responsibility, Proceedings of the TPRC2024 The Research Conference on Communications, Information and Internet Policy*, 2024, pp. 14-16.

Attualmente la vicenda che più di tutte ha acceso i riflettori sui pericoli di *deepfake*, e nello specifico di *deep porn*, per le celebrità a livello globale è stato il caso di Taylor Swift. Abbiamo già precedentemente trattato dei video e delle immagini *deepfake* che la ritraevano falsamente mentre sosteneva Donald Trump, con un *post* che ha raggiunto oltre 10,3 milioni di visualizzazioni.[431] Nel gennaio 2024, immagini *deepfake* pornografiche non consensuali generate dall'IA raffiguranti la *popstar* sono state ampiamente diffuse sulla piattaforma X: queste immagini, create con strumenti di AI generativa come *Microsoft Designer*, hanno rapidamente accumulato milioni di visualizzazioni e condivisioni, diventando virali prima che le autorità della piattaforma riuscissero ad intervenire;[432] una di queste immagini ha attirato oltre 45 milioni di visualizzazioni, 24.000 ripubblicazioni e centinaia di migliaia di *like* e *bookmark* in sole 17 ore.

La reazione dei *fan* di Taylor Swift, noti come "*Swifties*", è stata rapida e organizzata: hanno formato un collettivo, anche sotto l'*hashtag #ProtectTaylorSwift,* per segnalare attivamente il materiale *deepfake* pornografico e chiederne la rimozione, nonché per limitarne un'ulteriore diffusione futura; hanno inoltre inondato la piattaforma e gli *hashtag* corrispondenti con contenuti legittimi per rendere i *deepfake* più difficili da trovare. In risposta alla proliferazione dei contenuti, X ha temporaneamente bloccato tutte le ricerche che contenessero la dicitura "Taylor Swift" sulla piattaforma; [433] tuttavia, la risposta ritardata di X alle NCID ampiamente disseminate che hanno preso di mira Taylor Swift evidenzia le limitazioni delle piattaforme nel gestire i danni *online*[434]: sebbene le piattaforme forniscano strumenti di segnalazione, il lavoro effettivo di segnalazione ricade sugli utenti. Bisogna anche tenere a mente che nel caso di figure potenti e famose come Taylor Swift, le piattaforme rispondono più rapidamente bloccando i termini di ricerca in maniera più tempestiva rispetto ai casi di vittime meno conosciute, ma, nonostante ciò, il meccanismo di segnalazione non è abbastanza efficiente per limitare e fermare la proliferazione di tali contenuti. Questo caso, infatti, avvia anche una riflessione

[431] K. C. DUC TRUONG, *Reputation (Not Taylor's Version): Regulating Artificial Intelligence Hallucinated Deepfakes of Public Figures, Journal of Law, Technology & Policy*, 2024, p. 452.
[432] B. KIRA, *Deepfakes, the Weaponisation of AI Against Women and Possible Solutions, VerfBlog*, 2024, pp. 1-6.
[433] M. J. RIEDL, A. NEWELL, *op. cit.*, p. 1.
[434] B. KIRA, *Deepfakes, the Weaponisation of AI Against Women and Possible Solutions, VerfBlog*, 2024, pp. 1-6.

su come la fama e la potenza di una celebrità possano influenzare la rapidità della risposta di una piattaforma rispetto a vittime meno potenti, facendoci allo stesso tempo ragionare sul fatto che però nessuno è esente da questi attacchi tecnologici.[435]

Taylor Swift, come precedentemente osservato, negli ultimi anni non è risultata del tutto estranea alla circolazione di contenuti *deepfake* che la ritraessero come protagonista, come nel caso della controversia con Donald Trump. Nel 2023, sulla piattaforma *TikTok*, erano inoltre emersi dei *deepfake audio* che, riproducendo la sua voce, venivano utilizzati per diffondere brani musicali presentati ingannevolmente come nuovi singoli tratti da un album in imminente uscita, ma che in realtà erano stati generati mediante intelligenza artificiale. Analogamente all'atteggiamento adottato in questi precedenti episodi, anche in merito alla diffusione di suoi NCID, Taylor Swift ha deciso di non affrontare pubblicamente la controversia, attuando probabilmente una mossa strategica per evitare di aumentare l'esposizione del contenuto dannoso e l'*engagement*: evitando di interagire con questi *deepfake*, ha infatti ridotto il rischio di potenzialmente promuovere o dare maggiore visibilità ai video, proteggendo così la sua reputazione.

Da un punto di vista meramente legale, la protezione contro i *deepfake*, in particolare per le celebrità, è un campo in evoluzione e presenta diverse lacune: sostanzialmente, il diritto non ha ancora raggiunto la rivoluzione tecnologica portata dai *deepfake*.[436] Le leggi esistenti sulla *privacy* e sulla diffamazione sono spesso inadeguate per affrontare l'abuso di IA generativa e la vasta diffusione di disinformazione tramite *deepfake*.[437] Il diritto di pubblicità (*right of publicity*), che protegge l'uso non autorizzato dell'immagine di un individuo, lotta a contenere questa nuova forma di sfruttamento dell'identità; tuttavia, le protezioni tradizionali per gli usi espressivi, che consentono la rappresentazione di figure del mondo reale in *biopic* o drammi storici, sono troppo ampie se applicate a repliche digitali come i *deepfake*. Vi è chiaramente una necessità di cambiamenti nel modo in cui il diritto di pubblicità tratta gli usi espressivi: la creazione e l'uso di *deepfake* dovrebbero costituire un tipo distinto di uso dell'immagine nel contesto del diritto di pubblicità, richiedendo permessi più espliciti, anche per usi espressivi.[438] Tuttavia, in casi come

[435] M. J. Riedl, A. Newell, *op. cit.*, pp. 13-14.
[436] A. Preminger, M. B. Kugler, *op. cit.*, p. 102.
[437] J. Kaur, K. Sharma, M. P. Singh, *op. cit.*, pp. 147, 149.
[438] A. Preminger, M. B. Kugler, *op. cit.*, pp. 108, 148.

quello di Taylor Swift, dove i *deepfake* non sono creati o distribuiti per scopi commerciali, un reclamo basato sul diritto di pubblicità non sarebbe in ogni caso abbastanza.[439]

Le azioni legali per diffamazione o "*false light*" (falsa rappresentazione) sono altre vie potenziali per potersi difendere, ma anch'esse presentano sfide, specialmente quando si tratta di contenuti generati dall'IA e della difficoltà nell'identificare i responsabili originali su internet.[440] In quest'ottica, la legge sulla protezione dei dati personali è rilevante in quanto i *deepfake* spesso utilizzano dati privati (immagini, audio) senza consenso[441]; abbiamo visto come esistano anche risposte nel diritto penale di alcuni Paesi, come il caso dell'Italia, contro i *deepfake* pornografici, ma la loro efficacia nel proteggere le vittime resta oggetto di dibattito. Alcuni Stati negli USA e il governo federale hanno invece iniziato a creare schemi statutari e regolatori per combattere i *deepfake* usati nel discorso politico.[442]

In definitiva, il caso di Taylor Swift risulta particolarmente emblematico poiché in grado di evidenziare, oltre le lacune nelle leggi sulla diffamazione e sui contenuti sessualmente espliciti diffusi senza consenso, anche come i *deepfake* vengano generati e diffusi semplicemente per sensazionalismo e disinformazione; oltretutto, questa vicenda ha posto l'accento anche sulla responsabilità delle piattaforme e sull'eventualità che anch'esse vengano considerate responsabili per i contenuti *deepfake* diffusi.[443] Il caso di Taylor Swift potrebbe servire da catalizzatore per il Congresso degli Stati Uniti per creare opzioni legali per perseguire creatori e distributori e risarcire le vittime, evidenziando la papabile questione della responsabilità della piattaforma.[444]

3.1 Diffusione di contenuti non consensuali e danni reputazionali

La tecnologia *deepfake*, è bene rammentare, è emersa pubblicamente per la prima volta nel 2017 proprio con la comparsa *online* sulla piattaforma *Reddit* di video sessualizzati non consensuali che sostituivano i volti di attori pornografici con quelli di celebrità

[439] K. C. Duc Truong, *op. cit.*, p. 452.
[440] *Ivi*, p. 475.
[441] R. Ranjan, *op. cit.*, pp. 110-116.
[442] A. Preminger, M. B. Kugler, *op. cit.*, p. 139.
[443] B. Kira, *Deepfakes, the Weaponisation of AI Against Women and Possible Solutions, VerfBlog*, 2024, pp. 1-6.
[444] M. J. Riedl, A. Newell, *op. cit.*, p. 1.

femminili.[445] Sebbene, dunque, siano molteplici i campi di utilizzo dei *deepfake*, in origine questi sono stati pensati e diffusi proprio come contenuti espliciti non consensuali a danno delle celebrità. Dati empirici rivelano che fino al 98% di tutti i *deepfake online* sono pornografici o immagini sessualizzate non consensuali. Uno studio del 2019 ha inoltre evidenziato che circa il 96% dei *deepfake* contenesse materiale pornografico non consensuale; non è difficile intuire che il 99% di questi contenuti prenda di mira le donne.[446] Inizialmente, le vittime più frequenti erano celebrità femminili, inclusi nomi noti come per l'appunto Taylor Swift, ma anche le attrici Gal Gadot, Emma Watson, e Kristen Bell.[447]

Tuttavia, l'accessibilità crescente della tecnologia e la ridotta quantità di dati di addestramento necessari rendono ora possibile creare immagini sessualizzate non consensuali di qualsiasi donna che abbia condiviso foto sui *social media*, estendendo il rischio ben oltre le figure pubbliche.[448] Nel caso delle celebrità, la diffusione di *deepfake* non consensuali, in particolare di natura pornografica (spesso definiti anche "*image-based abuse*" oltre che "*deep porn*"), ha conseguenze devastanti per le celebrità prese di mira, impattando significativamente sia la loro reputazione sia la loro sfera economica.

Sul piano reputazionale, i *deepfake* possono compromettere irreparabilmente l'immagine pubblica di un individuo; questi contenuti sono realizzati per sembrare autentici, utilizzando filmati reali e suoni realistici, mirano a rendere sempre più difficile per il pubblico distinguere la finzione dalla realtà. Queste tecnologie hanno il potenziale di dirottare l'intera persona di un individuo e alterare fondamentalmente la percezione che il pubblico ha di questo, ritraendolo in comportamenti criminali, socialmente tabù o comunque in disaccordo con i suoi reali valori. Mentre i *biopic* tradizionali separano chiaramente le scene interpretate dagli attori dal materiale d'archivio reale, i *deepfake* possono fondere queste distinzioni, facendo apparire le scene ricreate dagli attori autentiche quanto i filmati d'archivio veri; questo offusca il confine tra storia reale e interpretata, con il potenziale di riscrivere la narrazione sociale riguardo al comportamento, alle inclinazioni e alle associazioni di una figura pubblica. La

[445] M. PAWELEC, *op. cit.*, p. 6.
[446] K. MANIA, *op. cit.*, p. 118
[447] M. J. RIEDL, A. NEWELL, *op. cit.*, p. 2.
[448] M. PAWELEC, *op. cit.*, p. 10.

conseguenza più diretta è che la reputazione della persona possa essere permanentemente macchiata, poiché il pubblico non ha altra garanzia se non la parola della vittima che gli atti in questione non siano mai accaduti.[449]

Per quanto riguarda la fattispecie di *deep porn*, nello specifico, le vittime si ritrovano a subire frequente umiliazione, perdita di controllo e intimidazione; trattasi di una situazione che si amplifica notevolmente quando sono coinvolte celebrità conosciute a livello globale. Tutto ciò può avere un grave impatto sulla salute mentale di chi ne è vittima, in quanto si potrebbero sviluppare un profondo disagio psicologico, ansia, paura e, nei casi più estremi, pensieri suicidi. Queste esperienze dannose derivano in gran parte dalla violazione della *privacy* sessuale della persona raffigurata: non si può ignorare che la diffusione di *deepfake* sessualmente espliciti costituisca una grave invasione della *privacy,* imprescindibile per la garanzia alle donne di partecipare alla vita pubblica. I contenuti diffusi possono esporre le vittime a un vasto pubblico *online* che le vede come meri oggetti sessuali da sfruttare ed esporre; ciò significa che, nonostante i creatori possano affermare che si tratti di semplice intrattenimento o divertimento innocuo, le conseguenze sulle vittime sono profonde, impattanti e concrete, indipendentemente dal realismo della rappresentazione.[450]

Oltre al danno diretto alla reputazione e al benessere psicologico, la vasta diffusione *online* dei *deepfake* non consensuali facilita l'*harassment* e lo *stalking*: l'ampio pubblico può infatti partecipare a danneggiare la vittima attraverso commenti, visualizzazioni, interazioni e condivisioni, oltre ad esporle a *stalking online*. La *Cyber Civil Rights Initiative*[451]*,* un gruppo per i diritti civili che lotta contro il *revenge porn* e i *deepfake* sessualmente espliciti, promuovendo la sensibilizzazione per le vittime, ha condotto un sondaggio con 1.606 risposte, di cui 361 si sono auto-dichiarate come vittime di *revenge porn*, mentre altre 67 si sono dichiarate vittime di foto pornografiche alterate con il famoso *software* di fotoritocco *Photoshop*. In questo studio, il 49% delle vittime ha

[449] A. PREMINGER, M. B. KUGLER, *op. cit.*, pp. 103, 119.
[450] M. PAWELEC, *op. cit.*, p. 10.
[451] La *Cyber Civil Rights Initiative* (CCRI) è un'organizzazione *no profit* che fornisce risorse e supporto a vittime di abusi *online*, in particolare quelli che coinvolgono immagini intime non consensuali, come il *revenge porn* e la *sextortion*. La CCRI lavora per proteggere i diritti civili *online* e promuovere la sicurezza degli utenti.

riferito di essere stata molestata o perseguitata *online* a causa del materiale diffuso su internet, mentre il 30% ha riferito di essere stata molestata o perseguitata *offline* dagli utenti che avevano visto questi contenuti *online*.

Sul fronte economico, i *deepfake* rappresentano una minaccia sostanziale per i guadagni e le opportunità di carriera delle celebrità, in particolare nel settore dell'intrattenimento: un *deepfake* di un artista può apparire identico all'artista in carne ed ossa, fungendo da concorrenza diretta o ingannando il pubblico. La tecnologia *deepfake* consente di raffigurare celebrità mentre utilizzano prodotti che non hanno mai provato o ricoprono ruoli che non hanno mai interpretato e questa capacità sta diventando fonte di preoccupazione e allarmismo ad Hollywood. Considerando che gli attori possono guadagnare decine di milioni di dollari per un singolo ruolo cinematografico o oltre un milione per episodio televisivo, la potenziale perdita economica per questi professionisti è enorme e non trascurabile dall'industria.[452]

D'altronde, le celebrità sono rese particolarmente vulnerabili a questa forma di sfruttamento anche a causa delle clausole contrattuali ampiamente utilizzate nel settore dell'intrattenimento: termini generici come il diritto di utilizzare la somiglianza "in tutti i media, sia ora noti che in futuro ideati" e "in tutte le forme, inclusi, senza limitazione, immagini o video digitalizzati, in tutto l'universo in perpetuo" erano in passato limitati dalle possibilità tecnologiche; tuttavia, con l'avvento dei *deepfake*, queste clausole eccessivamente ampie possono essere interpretate per giustificare usi che le parti contraenti non avrebbero potuto immaginare, lasciando gli artisti esposti a sfruttamento non remunerato. Storicamente, il diritto di pubblicità è emerso per proteggere l'uso commerciale dell'identità di un individuo, concettualizzato sia come un aspetto dei diritti della *privacy* (il diritto di essere lasciato solo) sia, più frequentemente, come un interesse di proprietà volto a proteggere le opportunità di profitto derivanti dal valore commerciale della propria identità. Le celebrità, riconoscendo questo valore, hanno a lungo concesso in licenza il loro volto, la loro voce e altri indicatori di identità a studi cinematografici e aziende pubblicitarie; tali accordi, inclusi nelle clausole sul "*use of likeness*" nei contratti di *performance* e intrattenimento, sono stati a lungo oggetto di intense negoziazioni.[453]

---

[452] A. PREMINGER, M. B. KUGLER, *op. cit.*, pp. 118-119, 147.
[453] *Ivi*, 143-145.

Tuttavia, la natura di tali clausole è spesso estremamente ampia e genericamente formulata, ad esempio vi sono licenze che concedono diritti irrevocabili per "adattare, riprodurre, distribuire e visualizzare" fotografie, video e audio, o il diritto di "copiare, riprodurre, fotografare, distribuire, trasmettere, mandare in onda, esibire, trascrivere, digitalizzare, mostrare, tutelare con diritto d'autore, concedere in licenza, trasferire, riprodurre, tradurre, modificare o altrimenti usare perpetuamente in tutto il mondo in tutti i media ora esistenti e in seguito ideati". In passato, i problemi creati da questi termini di licenza molto ampi erano limitati dalle restrizioni tecnologiche; si poteva fare solo un certo lavoro nel ricostruire l'identità di qualcuno utilizzando fotografie statiche, filmati video o registrazioni vocali tradizionali. Di fatto, presumibilmente gli artisti intendevano per "*likeness*" l'uso tradizionale di fotografie, video, voce o, forse, un'imitazione o una ricreazione non letterale, ma di certo non avrebbero potuto nemmeno immaginare un risultato simile a quello prodotto dall'intelligenza artificiale.[454]

L'avvento della tecnologia *deepfake* ha radicalmente alterato questo equilibrio, e le limitazioni tecnologiche che in precedenza mitigavano la portata delle clausole ampie oggi non esistono più. La prevalenza di queste disposizioni ampie, facilmente incorporabili nei contratti di intrattenimento tramite risorse *open-source* e contratti standardizzati, crea una forza normativa che erode il potere contrattuale degli artisti, i quali si trovano così a concedere diritti molto più preziosi di quanto apparentemente pattuito, poiché i termini contrattuali, concepiti in un'epoca pre-*deepfake*, possono essere interpretati per consentire la creazione e l'uso di repliche digitali realistiche senza consenso aggiuntivo. La capacità di possedere le *performance* degli attori e tutti i loro derivati digitali potrebbe permettere agli studi di aggiornare film e serie televisive senza sostituire i membri del *cast*, o addirittura sostituire un singolo membro del *cast* in caso di scandalo o disputa. Questo non richiederebbe più l'uso di tute con sensori o tecnologie fisiche simili; il *software* da solo potrebbe far passare una controfigura per l'attore reale e a costi praticamente nulli. Considerando, poi, i considerevoli stipendi offerti agli attori per riprendere i ruoli in *reboot* e *spin-off*, la perdita di tali opportunità a favore dei *deepfake* rappresenta massive perdite economiche per questa categoria. La conseguenza diretta di tutto ciò è che gli artisti siano di fatto indotti a concedere implicitamente il

[454] *Ivi*, pp. 140-143.

diritto di creare e utilizzare i loro *deepfake*, diritti che hanno un valore economico significativo nel mercato attuale e futuro, senza che ciò sia stato esplicitamente negoziato in precedenza o adeguatamente compensato.[455]

3.2 Tutela della *privacy* e protezione dei personaggi pubblici

La protezione della *privacy* e dei diritti dei personaggi pubblici e delle celebrità nell'era digitale, in particolare nel contesto emergente dei *deepfake*, rappresenta una sfida complessa che si articola attraverso diversi livelli di tutela giuridica. I *deepfake*, pur potendo costituire un uso della somiglianza di una persona, sono considerati in ogni caso "speciali" a causa delle loro capacità di manipolazione realistica, sollevando preoccupazioni significative.[456] Ancora una volta, la tutela della *privacy* dei personaggi pubblici è organicamente disciplinata e garantita soltanto in Europa mediante i Regolamenti già trattati in precedenza come il GDPR, il regolamento UE 2018/1725 (sul trattamento dei dati personali da parte delle istituzioni UE), e le direttive 2002/58/CE (ePrivacy) e UE 2016/680 (sul trattamento dei dati personali a fini di contrasto).[457]

Per il GDPR il trattamento dei dati personali è lecito solo in presenza di una base giuridica; ciò significa che, nel contesto dei *deepfake*, soprattutto se contenenti immagini realistiche di persone, il trattamento potrebbe riguardare categorie particolari di dati personali, come i dati biometrici, se trattati con un dispositivo tecnico specifico che consente l'identificazione univoca o l'autenticazione. Il trattamento di tali dati è generalmente vietato, salvo specifiche eccezioni basate sul consenso esplicito dell'interessato o su importanti motivi di interesse pubblico.[458]

In questo contesto, gli interessati, compresi i personaggi pubblici, godono di diversi diritti per controllare i propri dati personali: tra i più rilevanti vi sono il diritto di accesso, il diritto di rettifica e il diritto all'oblio. Quest'ultimo, nello specifico, permette di chiedere la cancellazione di dati non più necessari, trattati illecitamente, o per i quali il consenso è stato revocato. Il diritto all'oblio è particolarmente rilevante per i contenuti *online* e si

[455] *Ivi*, pp. 143-146.
[456] R. RANJAN*, op. cit.*, pp. 110-116.
[457] AI Act, art. 1, paragrafo 7.
[458] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 (GDPR).

applica anche se l'interessato aveva prestato il consenso quando era minore, clausola molto importante nel mondo dell'intrattenimento. Tuttavia, il diritto all'oblio presenta delle limitazioni: ad esempio, se il trattamento è necessario per l'esercizio del diritto alla libertà di espressione e di informazione, per adempiere a un obbligo legale, per l'esecuzione di un compito di interesse pubblico, per finalità di archiviazione nel pubblico interesse, ricerca scientifica o storica, o per l'accertamento, l'esercizio o la difesa di un diritto in sede giudiziaria; in tutti questi casi gli interessati godono comunque del diritto di opposizione al trattamento e diritti relativi alle decisioni basate esclusivamente sul trattamento automatizzato (inclusa la profilazione), avendo diritto a un intervento umano e a una spiegazione della decisione.[459]

La creazione e diffusione di *deepfake* che coinvolgono personaggi pubblici può rientrare in trattamenti di dati che presentano un rischio elevato per i diritti e le libertà delle persone fisiche. In questi casi, il titolare del trattamento è spesso tenuto a effettuare una valutazione d'impatto sulla protezione dei dati (DPIA) prima di procedere: una DPIA è richiesta in particolare per la valutazione sistematica e globale di aspetti personali basata su trattamenti automatizzati, inclusa la profilazione, per trattamenti su larga scala di categorie particolari di dati o dati relativi a condanne penali, o per la sorveglianza sistematica su larga scala di zone accessibili al pubblico. Per i trattamenti effettuati da un'autorità pubblica o per trattamenti su larga scala di dati sensibili, è inoltre spesso richiesta la designazione di un responsabile della protezione dei dati (DPO), che ha il compito di informare e fornire consulenza, sorvegliare l'osservanza del regolamento e cooperare con l'autorità di controllo.[460]

Per i personaggi pubblici, la protezione della *privacy* inevitabilmente si scontra e deve essere bilanciata con il diritto alla libertà d'espressione e di informazione, sancito dall'articolo 11 della Carta dei Diritti Fondamentali dell'UE. La legge interna degli Stati membri dell'UE può prevedere esenzioni o deroghe a molte disposizioni del GDPR (come principi, diritti dell'interessato, obblighi del titolare e responsabile, trasferimenti, autorità di controllo, cooperazione) quando ciò è necessario per conciliare questi diritti fondamentali, specialmente per trattamenti a scopi giornalistici o di espressione

[459] *Ibidem.*
[460] *Ibidem.*

accademica, artistica o letteraria.[461] La giurisprudenza costituzionale italiana e quella europea (CEDU, Corte di Giustizia UE) hanno sottolineato più volte la necessità di bilanciare la libertà di manifestazione del pensiero, che include il diritto di informare/essere informati e la libertà di opinione/critica/satira, con altri diritti fondamentali, tra cui la reputazione e i diritti altrui. I limiti alla libertà di espressione, come definiti dall'articolo 10, paragrafo 2, della CEDU e recepiti nel diritto dell'UE, devono essere previsti dalla legge, e devono perseguire uno scopo legittimo (es. protezione della reputazione o dei diritti altrui, o per sicurezza pubblica) ed essere necessari e proporzionati in una società democratica.[462]

L'applicazione di tali limiti nel contesto *online*, chiaramente, può essere più restrittiva: nel caso specifico dei *deepfake*, la natura intrinsecamente falsificata del contenuto rende particolarmente acuto il conflitto, richiedendo un intervento normativo che si interroga sull'opportunità di regolamentare il fenomeno senza operare un divieto generalizzato, che altrimenti costituirebbe una compressione della libertà di espressione. L'approccio europeo, a differenza di quello statunitense fortemente influenzato dal Primo Emendamento che protegge anche dichiarazioni false a meno che non causino danni legalmente riconoscibili, e di quello cinese caratterizzato da un controllo statale più stringente, mira comunque a trovare sempre un punto di equilibrio.[463]

La disciplina prevista dall'AI Act e la proposta di legge nazionale italiana si collocano in questo solco, ponendo un forte accento sulla trasparenza, che si coniuga nell'obbligo principale per i contenuti *deepfake* di etichettatura univoca al fine di rendere palese la natura artificiale del contenuto e negare l'elemento dell'apparente verità, che è fondamentale per la diffamazione. Questa trasparenza non deve però impedire la libertà di espressione artistica, satirica o di finzione, un aspetto cruciale per i personaggi pubblici che sono spesso oggetto di parodia; tuttavia, anche in questi casi, devono essere rispettate adeguate tutele per i diritti e le libertà di terzi. È bene ricordare che l'obbligo di etichettatura non si applica se il contenuto è stato sottoposto a revisione umana o controllo editoriale con assunzione di responsabilità editoriale.[464]

---

[461] *Ibidem.*
[462] *Ibidem.*
[463] A. ORLANDO, *op. cit.*, p. 321.
[464] S. TROZZI, *op. cit.*, pp. 239-241.

In definitiva, l'attuale quadro normativo e tecnico riflette, il difficile, ma necessario, bilanciamento tra la protezione dell'individuo, la libertà di espressione (inclusa la critica e la satira sui personaggi pubblici) e la responsabilità degli intermediari *online*, con un'attenzione crescente verso la trasparenza come strumento chiave per consentire agli utenti di distinguere il vero dal falso e salvaguardare così i diritti fondamentali nell'era digitale. È appurato che non esista un divieto assoluto ai *deepfake*, ma piuttosto un tentativo di gestire i rischi associati a queste tecnologie tramite obblighi specifici e la possibilità di intervento delle autorità competenti, comprese quelle per la protezione dei dati.[465]

Ne consegue che, sebbene la struttura normativa europea sia ancora in fase di implementazione, così come le persone comuni, anche le celebrità ed i personaggi pubblici godano di una maggiore tutela della *privacy* in territorio europeo rispetto a quello statunitense, fortemente e irrimediabilmente ancorato alla supremazia della libertà di espressione e del Primo Emendamento.

4. Riflessioni finali dai casi studio: implicazioni giuridiche e sociali

I casi studio analizzati hanno messo in evidenza la duplice natura dei *deepfake*: da un lato, uno strumento capace di stimolare innovazione e creatività, dall'altro, una minaccia concreta nel panorama digitale contemporaneo.

Il caso delle elezioni presidenziali statunitensi del 2024 ha offerto un esempio emblematico dell'utilizzo strategico dei *deepfake* nel contesto elettorale: questi strumenti, infatti, sono stati utilizzati per diffondere disinformazione e manipolare l'opinione pubblica, dimostrando come i contenuti falsati possano alterare la percezione degli eventi, compromettere gli avversari politici e, in ultima analisi, erodere la fiducia nei processi democratici. La loro diffusione rapida e spesso incontrollata, favorita dalle moderne tecnologie digitali, amplifica il rischio di danni su larga scala; tuttavia, le risposte fornite da attori pubblici e privati, incluse le iniziative intraprese dalle piattaforme *online*, sebbene ancora caratterizzate da efficacia disomogenea, evidenziano un'accresciuta

[465] A. ORLANDO, *op. cit.*, p. 324-327.

consapevolezza del problema e sottolineano l'importanza di responsabilizzare questi intermediari nella gestione dei contenuti manipolati.

Il caso della Cina ha offerto una prospettiva diversa, concentrandosi sull'uso dei *deepfake* nel quadro del controllo sociale e politico. Attraverso normative rigide e un approccio centralizzato orientato al mantenimento degli interessi nazionali e dei valori socialisti fondamentali, il governo cinese ha cercato di regolamentare lo sviluppo e l'utilizzo di queste tecnologie. Da un lato, tali misure impongono ai fornitori di servizi obblighi significativi, come la verifica dell'identità degli utenti e l'etichettatura dei contenuti manipolati. Dall'altro, la loro portata ampia e la vaghezza di definizioni quali "interessi nazionali" sollevano interrogativi critici riguardo alla censura e al possibile effetto deterrente sulla libertà di espressione. Inoltre, l'uso potenziale dei *deepfake* come strumento per promuovere narrazioni favorevoli al regime o screditare gli oppositori dimostra come il controllo dell'informazione prevalga sulla tutela dei diritti individuali, creando un netto contrasto con gli approcci adottati dalle democrazie occidentali. Le reazioni internazionali a queste pratiche e le conseguenze geopolitiche che ne derivano rimarcano la natura globale della problematica e la necessità di una cooperazione internazionale che sappia però confrontarsi con profonde divergenze normative e politiche tra i vari Paesi.

Infine, il caso di Taylor Swift ha messo in luce le gravi ricadute personali e reputazionali associate all'uso improprio dei *deepfake* non consensuali; la diffusione di *deep porn* senza il consenso della vittima ha evidenziato lacune significative nelle protezioni esistenti e la vulnerabilità anche delle figure pubbliche più influenti. Questo episodio ha dimostrato quanto sia semplice creare contenuti manipolati sfruttando la grande quantità di dati facilmente reperibili *online* (immagini, video e audio), oltre alle enormi difficoltà nel limitarne la circolazione una volta caricati in rete. Le conseguenze vanno ben oltre il danno all'immagine pubblica, toccando tematiche cruciali come la *privacy*, la sicurezza personale e persino implicazioni economiche e giuridiche legate alla perdita del controllo sulla propria identità digitale. Parallelamente, il caso ha fatto emergere le sfide affrontate dalle piattaforme digitali nel gestire tali contenuti e dalle vittime nel cercare giustizia o individuare i responsabili. Diverse soluzioni sono state proposte, tra cui l'applicazione di normative esistenti in ambito di protezione dei dati personali o del diritto d'autore, e

l'introduzione di nuove leggi che affrontino esplicitamente la manipolazione artificiale a scopi sessuali, come accaduto in Italia; tuttavia, l'efficacia concreta di queste misure rimane ad oggi oggetto di dibattito e ulteriore riflessione.

Dal punto di vista giuridico, l'analisi comparativa rivela come i casi studio confermino l'esistenza di paradigmi normativi distinti a livello globale. L'approccio europeo, rappresentato dall'AI Act e dal GDPR, si configura come un modello "basato sul ciclo di vita", orientato a regolamentare le diverse fasi della creazione e diffusione dei *deepfake* attraverso un sistema basato sul rischio, con particolare enfasi sulla trasparenza e la protezione dei diritti fondamentali. L'AI Act impone obblighi specifici di etichettatura per i contenuti generati o manipolati artificialmente, seppur con alcune eccezioni, mentre il GDPR offre strumenti per la protezione dei dati personali e diritti per gli interessati. Tuttavia, l'applicabilità di tali diritti nel contesto dei *deepfake* incontra significative difficoltà pratiche. Il DSA, pur non focalizzato esclusivamente sui *deepfake*, interviene nel contrasto alla diffusione di contenuti illegali e stabilisce obblighi di trasparenza per le piattaforme digitali. In generale, l'Unione Europea mira a bilanciare innovazione e tutela attraverso un approccio più interventista rispetto agli Stati Uniti.

Negli Stati Uniti, infatti, emerge un approccio "basato sull'applicazione", frammentario e fortemente influenzato dalla priorità data al Primo Emendamento: la regolamentazione tende a concentrarsi su utilizzi specifici e considerati dannosi, come i *deepfake* a scopo politico in periodi elettorali o la pornografia non consensuale. La dottrina del *marketplace of ideas* e le protezioni garantite dalla Sezione 230 del CDA limitano la possibilità di restrizioni generalizzate sui contenuti, anche se recenti sviluppi legali e proposte legislative suggeriscono un crescente riconoscimento della necessità di interventi. L'approccio statunitense è prevalentemente reattivo (*ex post*), focalizzato sul risarcimento delle vittime e sulle sanzioni per usi lesivi, in netto contrasto con l'approccio cinese più proattivo.

Per quanto riguarda invece le implicazioni sociali dei *deepfake* messe in luce dai casi studio analizzati, queste sono ampie e complesse. La capacità di generare contenuti falsi che appaiano, però, autentici mina la fiducia pubblica nelle fonti di informazione e nelle istituzioni; la disinformazione, amplificata dalla facilità di diffusione *online*, rappresenta una seria minaccia ai processi democratici, accentuando le divisioni politiche e

potenzialmente conducendo a disordini sociali. A livello individuale, il danno reputazionale e alla dignità delle persone coinvolte è significativo, come evidenziato dal caso di Taylor Swift. L'uso dei *deepfake* può inoltre minare l'autodeterminazione informativa degli individui, ossia la possibilità di controllare l'uso dei propri dati personali. Un aspetto fondamentale per ridurre questi rischi risiede nell'alfabetizzazione mediatica, ovvero nell'educazione dei cittadini a riconoscere e valutare criticamente i contenuti digitali; tuttavia, le normative si concentrano ancora in larga misura su produttori, fornitori e piattaforme, trascurando talvolta l'importanza di responsabilizzare *in primis* gli utenti.

In conclusione, i casi studio esaminati nel presente lavoro confermano che i *deepfake* costituiscono una sfida articolata che necessita di soluzioni integrate e collaborazioni internazionali. Sebbene approcci normativi differenti (ciclo di vita in UE, applicazione negli USA, orientamento al soggetto in Cina) riflettano diversità culturali e giuridiche, vi è una crescente consapevolezza dei rischi connessi e della necessità di trovare un equilibrio tra progresso tecnologico e protezione dei diritti fondamentali e dell'interesse pubblico. Strumenti essenziali per affrontare questa sfida includono la trasparenza garantita dall'etichettatura, la responsabilità delle piattaforme, lo sviluppo di tecnologie di rilevamento e soprattutto un maggiore investimento nell'alfabetizzazione mediatica. La continua evoluzione tecnologica rende indispensabile un aggiornamento costante dei quadri normativi e delle strategie difensive, favorendo una cooperazione internazionale per assicurare che l'intelligenza artificiale si sviluppi in modo etico e responsabile, preservando la fiducia pubblica e la stabilità delle istituzioni democratiche.

CONCLUSIONI

Il presente lavoro di tesi si è concentrato sull'analisi dettagliata della natura, dello sviluppo, delle applicazioni, dei rischi e del quadro giuridico relativo ai *deepfake*, indagandone anche l'impatto concreto attraverso lo studio di specifici casi emblematici. Da questa analisi emergono la complessità e la rapida evoluzione di una tecnologia che si colloca al crocevia tra progresso digitale e implicazioni socioculturali. La dualità dei *deepfake* si manifesta nell'essere sia strumenti d'innovazione, con effetti positivi in ambiti quali l'intrattenimento e la creatività, sia veicoli di minacce rilevanti per la diffusione di disinformazione, la manipolazione dell'opinione pubblica e la violazione della dignità umana e della *privacy*.

Nel primo capitolo, è stata affrontata una disamina delle principali tipologie di *deepfake* corredata da un'analisi delle tecniche di produzione e dei rischi connessi, con particolare attenzione alle implicazioni sulla proliferazione di *fake news* e disinformazione. È stato evidenziato come l'elevato grado di sofisticazione tecnologica renda sempre più arduo distinguere il vero dal falso, sollevando questioni giuridiche che necessitano di regolamentazioni adeguate.

Nel secondo capitolo è stato analizzato il confronto fra i principali modelli regolatori internazionali: quello europeo, statunitense e cinese. L'Unione europea, in particolare, è emersa come pioniera nella disciplina normativa attraverso strumenti come il Regolamento Generale sulla Protezione dei Dati (GDPR) e l'AI Act, che adottano un approccio basato sul ciclo di vita delle tecnologie e sulla valutazione del rischio. L'AI Act, nello specifico, prevede obblighi stringenti in tema di trasparenza per i contenuti generati o manipolati artificialmente, tra cui l'etichettatura obbligatoria, seppur con alcune eccezioni. Diversamente, il contesto normativo degli Stati Uniti si presenta più frammentato e decentralizzato, fortemente influenzato dalla salvaguardia – quasi assoluta – della libertà di espressione sancita dal Primo Emendamento e dalla dottrina del "*marketplace of ideas*". L'approccio statunitense appare orientato ad un intervento circoscritto, con regolamentazioni mirate a settori come le elezioni o la pornografia non consensuale, adottando prevalentemente misure tempestive, ma limitate. Inoltre, la Sezione 230 del *Communications Decency Act* offre un significativo grado di immunità alle piattaforme digitali rispetto ai contenuti generati dagli utenti, sebbene tale protezione

sia oggetto di crescenti discussioni. La Cina si distingue per un approccio centralizzato e rigoroso alla regolamentazione dei *deepfake* tramite le *Deep Synthesis Provisions*. Questo modello si focalizza sulla responsabilizzazione diretta dei fornitori di servizi tecnologici e impone il divieto di una vasta gamma di contenuti considerati contrari agli interessi nazionali. Sebbene tale approccio cerchi di preservare la stabilità sociale e il perseguimento di valori nazionali, esso solleva interrogativi in merito alla censura e alle limitazioni imposte alla libertà espressiva.

Il terzo capitolo ha, poi, illustrato casi studio concreti per esplorare l'impatto pratico dei *deepfake*. L'analisi del loro utilizzo durante le elezioni presidenziali degli Stati Uniti nel 2024 ha mostrato come questa tecnologia possa essere impiegata strategicamente per influenzare l'opinione pubblica, configurandosi come un rischio per la trasparenza dei processi democratici. Il caso legato a Taylor Swift ha invece evidenziato le lacune normative nella protezione delle figure pubbliche contro *deepfake* intimi non consensuali, mettendo in discussione il ruolo delle piattaforme digitali nella prevenzione di tali abusi. Inoltre, è stato esaminato come il Partito Comunista Cinese si serva dell'intelligenza artificiale e dei *deepfake* per finalità di censura e controllo sociale, evidenziando la manipolazione di contenuti audiovisivi a scopo propagandistico e repressivo, nonché il ruolo dell'intelligenza artificiale nella sorveglianza dei cittadini. Infine, si sono esaminate le reazioni della comunità internazionale, sottolineando le implicazioni geopolitiche e la necessità di una risposta normativa globale.

In definitiva, i *deepfake* costituiscono una problematica complessa e attuale, sollevando questioni in ambito tecnologico, giuridico, politico e sociale. I vari paradigmi normativi internazionali rispecchiano approcci differenti nel bilanciare l'innovazione tecnologica, la tutela dei diritti fondamentali (e, in particolare, la libertà di espressione) con l'interesse pubblico. Strumenti cruciali per affrontare tale sfida includono la trasparenza mediante l'etichettatura, la responsabilizzazione delle piattaforme, lo sviluppo di tecnologie di rilevamento avanzate e, in particolare, l'alfabetizzazione mediatica dei cittadini.

L'inarrestabile progresso della tecnologia *deepfake* richiede un costante aggiornamento dei quadri normativi e delle strategie difensive. È desiderabile un maggiore coordinamento e cooperazione a livello internazionale tra le potenze mondiali per elaborare un approccio congiunto ai rischi associati all'intelligenza artificiale, affinché il

suo sviluppo avvenga in maniera etica e responsabile, salvaguardando al contempo la fiducia pubblica e la stabilità delle istituzioni democratiche nell'era digitale.

BIBLIOGRAFIA

ABID A., ROY S.K., LEES-MARSHMEN J. ET AL., *Political social media marketing: a systematic literature review and agenda for future research, Electron Commer Res 25*, 2025.

ADEBAYO A., *Campaigning in the Age of AI: Ethical Dilemmas and Practical Solutions for The UK and US. International Journal of Social Science and Human Research. 07*, 2024.

AMORE N., *La tutela penale della riservatezza sessuale nella società digitale. Contesto e contenuto del nuovo cybercrime disciplinato dall'art. 612-ter c.p., Rivista di Diritto Penale Contemporaneo*, 2020.

ANDREW J. *Elite in China face austerity under Xi's rule, New York Times*, 27 March 2013.

AYSON M. E., *Visual Propaganda in the Time of COVID-19: China's Image Repair in State Media Political Cartoons*, 2022.

Arriverà quest'anno l'agenzia spagnola per l'intelligenza artificiale, Notizie.AI, 2023.

ARRUZZOLI F., *Deepfake – Significato, Storia, evoluzione, ICT Security Magazine*, 2022.

ARTERO MUÑOZ A., RUIZ DE TOLEDO RODRÍGUEZ C. F., MAIRAL MEDINA P., *Agencia Española de Supervisión de la Inteligencia Artificial, la clave para un desarrollo tecnológico ético, justo y sostenible, Revista Española de Control Externo, vol. XXV, n. º 74-75*, 2023.

BALMAS M., *Tell Me Who is Your Leader, and I Will Tell You Who You Are: Foreign Leaders' Perceived Personality and Public Attitudes toward Their Countries and Citizenry, American Journal of Political Science*, 62, 2018.

BAMMAN D., O'CONNOR B., SMITH N.A., *Censorship and deletion practices in Chinese social media, First Monday, 17*, 2012.

BARBERA D., *Il sistema Intel per riconoscere i deepfake con una precisione del 96%. FakeCatcher è una tecnologia sviluppata con la State University di New York che riconosce i falsi dal flusso sanguigno del viso, Wired*, 2022.

BASSINI M., LIGUORI L., POLLICINO O., *Sistemi di Intelligenza Artificiale, responsabilità e accountability. Verso nuovi paradigmi?, Intelligenza artificiale, protezione dei dati personali e regolazione, Giappichelli editore*, 2018.

BEDDINGFIELD S., *China's Strategic Use of Digital Media, Liberty University Journal of Statesmanship & Public Policy: Vol. 5: Iss. 1, Article 6*, 2024.

BEHUN R. J., OWENS E., *Youth and Internet Pornography: The Impact and Influence on Adolescent Development, Routledge*, 2019.

BOS L., VAN DER BRUG W., DE VREESE C., *How the Media Shape Perceptions of Right-Wing Populist Leaders, Political Communication, 28*, 2011.

BRADY A., *Marketing Dictatorship: Propaganda and Thought Work in Contemporary China, Lanham, Rowman & Littlefield Publishers*, 2009.

BREGLER C., COVELL M., SLANEY M., *Video Rewrite: Visual Speech Synthesis from Video, ISCA Speech*, 1997.

CASSANO G., TASSONE B., GALLI C., FRANCESCHELLI V., *Diritto industriale e diritto d'autore nell'era digitale, Giuffrè*, 2022.

CASADEI M., *Intelligenza artificiale, moda al test dell'AI Act: aziende in ritardo. Servono policy e formazione, Il Sole 24 Ore*, febbraio 2025.

СВИРИДОВА Е.А., *Rules for the Use of Deepfake Technologies in the Law of the USA and the People's Republic of China: Adaptation of Foreign Experience in Legal Regulation, Современное право*, 2024.

CHAWKI M., *Navigating legal challenges of deepfakes in the American context: a call to action, Cogent Engineering, 11 (1)*, 2024.

CHEN X., CAO R., HASHIM N. B., KAMARUDIN S. B., YE HE C. R., *Navigating New Media: The Impact of Short Video Platforms on Political News Consumption Among Chinese University Students. Environment and Social Psychology, 10(1)*, 2025.

CHEN Y., *The Accuracy and Biases of AI-Based Internet Censorship in China, Journal of Research in Social Science and Humanities*, 2025.

CHEN Z., *Perception of Crisis Responsibility: Examining Ai-Generated Deepfake Content and Public Response to Taylor Swift, Theses – ALL. 896,* 2024.

CHEUNG T. M., *The rise of China as a cybersecurity industrial power: Balancing national security, geopolitical and development priorities, Journal of Cyber Policy 3(3)*, 2018.

CHIN S. J., *Institutional origins of the media censorship in China: The making of the socialist media censorship system in 1950s Shanghai, Journal of Contemporary China 27*, 2018.

CIVITARESE MATTEUCCI S.*, La dignità umana come principio "autonomo" per giustificare la tutela dei diritti sociali, Diritto pubblico, 1*, 2022.

COMANDÈ G., *Intelligenza artificiale e responsabilità tra liability e accountability. Il carattere trasformativo dell'IA e il problema della responsabilità, Analisi giuridica dell'economia, 1*, 2019.

DAILEY H., *"I Hate Taylor Swift": Everything Donald Trump Has Ever Said About the Pop Star, Billboard*, 2025.

DAWKINS R., *The Selfish Gene, Oxford University Press*, 2006.

DE GREGORIO G., *The market place of ideas nell'era della post-verità: quali responsabilità per gli attori pubblici e privati online*?, *Medialaws*, 2019.

DOLHANSKY B., HOWES R., PFLAUM B., BARAM N., CANTON FERRER C., *The Deepfake Detection Challenge (DFDC) Preview Dataset, ArXiv*, 2019.

DONG X., BAO J., CHEN D., ZHANG T., ZHANG W., YU N., CHEN D., WEN F., GUO, B., *Protecting Celebrities with Identity Consistency Transformer, ArXiv*, 2022.

DRINHAUSEN K., OHLBERG M., KARÁSKOVÁ I., STEC G., *Image control: how China struggles for discourse power, Merics Report*, 2023.

DUC TRUONG K. C., *Reputation (Not Taylor's Version): Regulating Artificial Intelligence Hallucinated Deepfakes of Public Figures, Journal of Law, Technology & Policy*, 2024.

EASTTOM W., *Deepfake Technology: Emerging Threats and Security Implications, International Conference on Cyber Warfare and Security, 20,* 2025.

FALLETTA P., *Lezioni di diritto pubblico del digitale, Cedam*, 2024.

FATIMA S., *Legal and Ethical Implications of Deepfake Technology: Exploring the Intersection of Free Speech, Privacy, and Disinformation, Illinois Institute of Technology*, 2025.

FLORIDI L., *L'etica dell'intelligenza artificiale, Raffaello Cortina Editore*, 2022.

GALVANO F., BADIALI L., *Analisi comportamentale applicata al Deepfake, Behaviour Analysis Team*, 2025.

GARRIGA M., RUIZ INCERTIS R., MAGALLÓN ROSA R., *Artificial intelligence, disinformation and media literacy proposals around deepfakes, Observatorio (OBS*) Journal*, 2024.

GOODFELLOW I., BENGIO Y., COURVILLE A., *Deep learning, The MIT Press*, 2016.

GOSZTONYI G., LENDVAI F. G., *Online platforms and legal responsibility: A contemporary perspective in view of the recent U.S. developments, Masaryk University Journal of Law and Technology*, 2024.

GRAVES F., GABRIEL G., *Right to Not Be Forgotten (Sometimes): Celebrity Privacy Rights in a Data-Driven World, Landslide® Magazine, Volume 13, Number 2*, 2020.

GUADAMUZ A., *Impact of Artificial Intelligence on IP Policy, World Intellectual Property Organization*, 2017.

HENRY N., POWELL A., *Embodied Harms: Gender, Shame, and Technology-Facilitated Sexual Violence, Violence Against Women, 21(6)*, 2015.

HILLMAN B., CHIEN-WEN K., *Political and Social Control in China: The Consolidation of Single-Party Rule, ANU Press*, 2024.

HOOVER A., *"If Taylor Swift Can't Defeat Deepfake Porn, No One Can", Wired*, 2024.

HUBER A. R., WARD Z., *Non-consensual intimate image distribution: Nature, removal, and implications for the Online Safety Act, European Journal of Criminology, 22 (1)*, 2025.

HUNTER L. Y., ALBERT C. D., RUTLAND J., TOPPING K., HENNIGAN C., *Artificial intelligence and information warfare in major power states: how the US, China, and Russia are using artificial intelligence in their information warfare and influence operations, Defense & Security Analysis*, 2024.

ISLAM M. B. E., HASEEB M., BATOOL H., AHTASHAM N., MUHAMMAD Z., *AI Threats to Politics, Elections, and Democracy: A Blockchain-Based Deepfake Authenticity Verification Framework*, Blockchains 2, 2024.

IWI P. S., YULIANINGSIH W., *Restorative Justice for Children of Cyber-Porn Offenders, Awang Long Law Review, Vol. 7, No. 1*, 2024.

KALWANI H., *The price of fame: celebrity privacy rights, Indian Journal of Integrated Research in Law, Volume 3, Issue 1*, 2023.

KANSTEINER W., *Censorship and Memory: Thinking Outside the Box with Facebook, Goebbels, and Xi Jinping, Journal of Perpetrator Research, 4*, 2021.

KAUR J., SHARMA K., SINGH M. P., *Exploring the Depth: Ethical Considerations, Privacy Concerns, and Security Measures in the Era of Deepfakes, in Navigating the World of Deepfake Technology, IGI Global*, 2024.

KESHAV A., The Vanishing Act: Celebrities' Right to be Forgotten, 6 (5), IJLSI, 2024.

KIRA B., *Deepfakes, the Weaponisation of AI Against Women and Possible Solutions, VerfBlog*, 2024.

KIRA B., *When non-consensual intimate deepfakes go viral: The insufficiency of the UK Online Safety Act, Computer Law & Security Review 54*, 2024.

KRAUSE D., *The EU AI Act and the Future of AI Governance: Implications for U.S. Firms and Policymakers*, 2025.

LATHAM & WATKINS, *UK Online Safety Act 2023: A primer on the new law for relevant service providers*, 2024.

LATIF AL WAROI M. N. A., *False Reality: Deepfakes in Terrorist Propaganda and Recruitment, Security Intelligence Terrorism Journal (SITJ), Vol. 01 No. 01*, 2024.

LEWIS D., LASEK-MARKEY M., GOLPAYEGANI D., PANDIT H. J., *Mapping the Regulatory Learning Space for the EU AI Act, ArXiv*, 2025.

LI S., *The Social Harms of AI-Generated Fake News: Addressing Deepfake and AI Political Manipulation, Digital Society & Virtual Governance, Volume 1, Issue 1*, 2025.

LIU J., *Internet Censorship in China: Looking Through the Lens of Categorisation, Journal of Current Chinese Affairs*, 2024.

LOHSSE S., SCHULZE R., STAUDENMAYER D., *Liability for Artificial Intelligence and the Internet of Things, Münster Colloquia on EU Law and the Digital Economy IV, Nomos*, 2019.

LONGO A., *Il Parlamento europeo approva l'AI Act, cosa cambierà per le nostre aziende?, Il Sole 24 Ore*, 2023.

MANIA K., *Legal Protection of Revenge and Deepfake Porn Victims in the European Union: Findings From a Comparative Legal Study, Trauma Violence & Abuse 25(1)*, 2022.

MANIA K., *The Legal Implications and Remedies Concerning Revenge Porn and Fake Porn: A Common Law Perspective, Sexuality & Culture, 24*, 2020.

MBIOH W., *Beyond possessive agency: TikTok, YouTube, and the inadequacies of GDPR, OSA, DSA, and AIA, Oñati International Institute for the Sociology of Law*, 2025.

METSELAAR L., *Framing Deepfake Technology in European Union Governance: Discursive strategies and regulatory responses to deepfake technology, Management Society and Technology Universiteit Twente*, 2025.

MIOTTI A., WASIL A., *Combatting deepfakes: Policies to address national security threats and rights violations, ArXiv*, 2024.

MURPHY G., CHING D., TWOMEY J., LINEHAN C., *Face/Off: Changing the face of movies with deepfakes. PLoS ONE 18(7)*, 2023.

NG Y., *An error management approach to perceived fakeness of deepfakes: The moderating role of perceived deepfake targeted politicians' personality characteristics, Current Psychology. 42*, 2022.

O' CARROLL L., *EU asks X for internal documents about algorithms as it steps up investigation, The Guardian*, 2025.

ORLANDO A., *La regolamentazione del deepfake in Europa, Stati Uniti e Cina, Medialaws, 2024*.

PAGALLO U., *Intelligenza Artificiale e diritto. Linee guida per un oculato intervento normativo, in Sistemi intelligenti, Rivista quadrimestrale di scienze cognitive e di intelligenza artificiale, 3*, 2017.

PALMERINI E., STRADELLA E., *Law and technology: the challenge of regulating technological development, Pisa University Press*, 2013.

PAPA A., *Intelligenza artificiale e decisioni pubbliche tra tecnica, politica e tutela dei diritti, Federalismi.it, 22*, 2022.

PASCUAL M. G., *El Gobierno aprueba la norma para el buen uso de la IA, que obliga a etiquetar contenidos creados con esta tecnología, El País*, 2025.

PAWELEC M., *Deepfakes: Manipulation on Demand? Evolution of Deepfake Technology, Societal Impact, and the Path Forward, Heinrich Böll Foundation Tel Aviv, Israel Public Policy Institute*, 2025.

PIETY T. R., *Market Failure in the Marketplace of Ideas: Commercial Speech and the Problem that Won't Go Away, 41 Loyola of Los Angeles Law Review*, 2007.

PREMINGER A., KUGLER M. B., *The right of publicity can save actors from deepfake armageddon, Berkeley Technology Law Journal, Forthcoming Northwestern Public Law Research Paper No. 23-52*, 2024.

QUINTAIS J. P., *Generative AI, Copyright and the AI Act, Computer Law & Security Review 56*, 2025.

RAMLUCKAN T., *Deepfakes: The Legal Implications, Proceedings of the 19th International Conference on Cyber Warfare and Security*, 2024.

RAMOS F., *Deepfake: Análisis de sus implicancias tecnológicas y jurídicas en la era de la Inteligencia Artificial. Derecho Global. Estudios sobre Derecho y Justicia, IX*, 2024.

RANJAN R., *Deep fake porn: Duplicity and Dystopia, Law in the Age of Disruption. Understanding the Impact of Technology, Lex Assisto Media and Publications, 2023*.

RIEDL M. J., NEWELL A., *Reporting Image-Based Sexual Violence: Deepfakes, #ProtectTaylorSwift, and Platform Responsibility, Proceedings of the TPRC2024 The Research Conference on Communications, Information and Internet Policy*, 2024.

RUDNIEVA A., РУДНЄВА A. O., *Innovative information technologies in election political communications, Epistemological studies in Philosophy, Social and Political Sciences, 2024*.

RUFFO A., *Il disordine informativo e l'Intelligenza Artificiale; tra insidie e possibili strumenti di contrasto, Medialaws, 2024*.

SCOTT D., *Deepfake Porn Nearly Ruined My Life, Elle*, 2020.

SHARMA A., SHARMA H., *Right To Be Forgotten - An Analysis, Indian Journal of Law and Legal Research*, 2022.

SHOAIB M. R., WANG Z., AHVANOOEY M. T., ZHAO J., *Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models, International Conference on Computer and Applications (ICCA), IEEE*, 2023.

SIAROHIN A., LATHUILIÈRE S., TULYAKOV S., RICCI E., SEBE N., *First Order Motion Model for Image Animation, Advances in neural information processing systems 32*, 2019.

SINGH P., DHIMAN B., *Exploding AI-Generated Deep fakes and Misinformation: A Threat to Global Concern in the 21st Century, J Robot Auto Res, 5(1)*, 2024.

SONG R., *Faking It: A Proposed Solution to Counter Nonconsensual Pornographic Deepfakes, 31 Wash. & Lee J. Civ. Rts. & Soc. Just. 157*, 2025.

STURINO F., *Deepfake Technology and Individual Rights, Social Theory and Practice*, 2023.

STRAUSS M., BLENKINSOP P., *EU steps up probe into Musk's X, days ahead of Trump inauguration, Reuters*, 2025.

TAMPUBOLON M., *Digital Face Forgery and the Role of Digital Forensics. International Journal for the Semiotics of Law, Revue internationale de Sémiotique juridique*, 2023.

The number of deepfake videos online is spiking. Most are porn, CNN Business, 2019.

TREMOLADA L., *Autoregolamentazione, trasparenza e sorveglianza: i nodi da sciogliere dell'AI Act, Il Sole 24 Ore*, 2023.

TROZZI S., *La dimensione costituzionale dell'intelligenza artificiale generativa. La tutela della dignità umana nell'era del deepfake, in Diritto Pubblico Europeo Rassegna online, 1*, 2024.

TUNG H. WU W., *What Can Comparative Authoritarianism Tell Us About China Under Xi Jinping (and Vice Versa)?, Issues & Studies, 57*, 2021.

VALENTI F. V., *Il deep fake: la nuova sfida dell'intelligenza artificiale generativa, Derecom 37*, 2024.

VINOGRADOVA E. A., *Potential threats of unauthorized use of political deepfakes during political elections: international experience, Мировая политика*, 2024.

What is Machine Learning?, *IBM Cloud Education*, 2020.

WILSON J., *Deepfake: Post the Bruce Willis Controversy What Disruption To Entertainment Could Be Caused, Forbes*, 2022.

YAN Y., YANG Z., *Portraying Competence, Benevolence or Party Loyalty? Political Propaganda and the Image-Building of Political Elites in China, Comparative Politics*, 2025.

YAVUZ C., *Criminalisation of the Dissemination of Non-consensual Sexual Deepfakes in the European Union. A Comparative Legal Analysis, Researching the boundaries of sexual integrity, gender violence and image-based abuse, Vol 95, Issue 2, Revue Internationale de Droit Pénal*, 2024

ZENGIN A., *Deepfake: Implications and Solutions in the EU, L'Europe Unie, no. 21*, 2024.

ZHANG L., *Internet control in China, Journal of Contemporary China 15(49)*, 2006.

ZHENG G., SHU J., LI K., *Regulating deepfakes between Lex Lata and Lex ferenda - a comparative analysis of regulatory approaches in the U.S., the EU and China, Crime, Law and Social Change*, 2024.

RIFERIMENTI NORMATIVI

Administrative Regulations on Online Audio and Video Information Services, 18 novembre 2019. Cfr. *China issues regulation for online audio, video services, in english.gov.cn,* 30 novembre 2019.

California Assembly Bill 602 (AB 602).

Camera dei deputati, Introduzione dell'articolo 612-quater del Codice penale, in materia di manipolazione artificiale di immagini di persone reali allo scopo di ottenerne rappresentazioni nude, A.C. 2986, XVIII legislatura.

Communications Decency Act del 1996.

Criminal Justice and Courts Act 2015.

Disegno di legge S. 1146 - 19ª Legislatura (Disegno di legge sull'intelligenza artificiale).

Direttiva 96/9/CE del Parlamento europeo e del Consiglio, dell'11 marzo 1996, relativa alla tutela giuridica delle banche di dati.

Direttiva 2001/29/CE del Parlamento europeo e del Consiglio, del 22 maggio 2001, sull'armonizzazione di taluni aspetti del diritto d'autore e dei diritti connessi nella società dell'informazione.

Direttiva 2002/58/CE del Parlamento europeo e del Consiglio, del 12 luglio 2002, relativa al trattamento dei dati personali e alla tutela della vita privata nel settore delle comunicazioni elettroniche (direttiva ePrivacy).

Direttiva (UE) 2016/680 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativa alla protezione delle persone fisiche con riguardo al trattamento dei dati personali da parte delle autorità competenti a fini di prevenzione, indagine, accertamento e perseguimento di reati o esecuzione di sanzioni penali, nonché alla libera circolazione di tali dati e che abroga la decisione quadro 2008/977/GAI.

Direttiva (UE) 2019/790 del Parlamento europeo e del Consiglio, del 17 aprile 2019, sul diritto d'autore e sui diritti connessi nel mercato unico digitale e che modifica le direttive 96/9/CE e 2001/29/CE.

Direttiva (UE) 2024/1385 del Parlamento europeo e del Consiglio, del 14 maggio 2024, sulla lotta alla violenza contro le donne e alla violenza domestica.

Florida, S.B. 1798, 24 giugno 2022.

Garante per la Protezione dei Dati Personali, *Vademecum*, dicembre 2020.

H.R.5586 - *DEEPFAKES Accountability Act*.

Indiana, H.B. 1133, 12 marzo 2024.

Louisiana, S.B. 1 (Act), 2 giugno 2023.

Mississippi, S.B. 2577, 30 aprile 2024.

Nex Mexico, H.B. 182, 5 marzo 2024.

Oregon, S.B. 1571, 27 marzo 2024.

Provisions on the Administration of Deep Synthesis Internet Information Services, 25 novembre 2022.

Regolamento (CE) n. 45/2001 del Parlamento europeo e del Consiglio, del 18 dicembre 2000, concernente la tutela delle persone fisiche in relazione al trattamento dei dati personali da parte delle istituzioni e degli organismi comunitari.

Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (Regolamento generale sulla protezione dei dati personali - GDPR).

Regolamento (UE) 2018/1725 del Parlamento europeo e del Consiglio, del 23 ottobre 2018, sulla protezione delle persone fisiche in relazione al trattamento dei dati personali da parte delle istituzioni, degli organi e degli organismi dell'Unione e sulla libera

circolazione di tali dati, e che abroga il regolamento (CE) n. 45/2001 e la decisione n. 1247/2002/CE

Regolamento (UE) 2022/2065 del Parlamento Europeo e del Consiglio del 19 ottobre 2022 relativo a un mercato unico dei servizi digitali e che modifica la direttiva 2000/31/CE (Regolamento sui servizi digitali - DSA).

Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale (AI Act).

S.3696 - *DEFIANCE Act of* 2024.

S.4875 - *NO FAKES Act of* 2024.

South Dakota, S.B. 9, 13 febbraio 2022.

Tennessee, H.B. 2091, 26 marzo 2024 (ELVIS Act).

Texas, S.B. 751, 2019.

UK Online Safety Act 2023.

Virginia, H.B. 2678, 2019

Washington, S.B. 1999, 6 giugno 2024.