EDICIONES COMPLUTENSE

Derecom
e-ISSN: 1988-2629

ARTÍCULOS

Il deep fake: la nuova sfida dell'intelligenza artificiale generativa

Fenice Valentina Valenti
Università degli Studi Magna Graecia di Catanzaro (Italia) ✉

https://dx.doi.org/10.5209/dere.98114 Recibido: 15/08/2024 • Evaluado: 26/08/2024 • Aceptado: 02/09/2024

IT **Abstract:** L'intelligenza artificiale ha inaugurato quello che i filosofi definiscono la "quarta rivoluzione", ponendo l'uomo davanti a molteplici sfide. Sotto il profilo informatico e giuridico, infatti, uno dei maggiori problemi che si profila nel vasto orizzonte delle nuove tecnologie concerne il *Deep Fake* — il c.d. il falso profondo — e la relativa difficoltà di accertamento nell'ambito della *digital forensics.* Recenti studi hanno confermato, inoltre, che l'impatto di tali strumenti sul meccanismo cognitivo porti l'essere umano a fidarsi di un dato falso e questo anche in misura maggiore rispetto a un dato "vero". Per cercare di ovviare a tale pericolo, l'UE ha inserito le tecniche manipolative nel catalogo delle pratiche vietate previsto dall'AI Act, sebbene il continuo progresso tecnologico imponga una riflessione sul rapporto diritto e tecnologie, nonché sull'idoneità dei moderni strumenti giuridici a far fronte alla sfida dell'intelligenza artificiale.
Lo scopo del presente lavoro è indagare il funzionamento dell'IA generativa di un dato falso, soffermandosi sulle criticità nell'accertamento dei *Deep Fake,* nonché sulle risposte legislative adottate negli altri ordinamenti giuridici.
Parole chiave: intelligenza artificiale; deep fake; digital forensics.

ENG The deep fake: the new challenge of generative artificial intelligence

Abstract: Artificial intelligence has ushered in what philosophers call the 'fourth revolution', presenting mankind with multiple challenges. In fact, from a computer and legal point of view, one of the biggest problems looming on the vast horizon of new technologies concerns the *Deep* Fake —and the relative difficulty of ascertaining it in the field of *digital forensics.* Recent studies have confirmed that such tools' impact on the cognitive mechanism leads human beings to trust false data to a greater extent than 'true' data. In an attempt to remedy this danger, the EU has included manipulative techniques in the catalogue of prohibited practices provided for in the AI Act, although continued technological progress calls for reflection on the relationship between law and technology, as well as on the suitability of modern legal instruments to meet the challenge of artificial intelligence.
This paper aims to investigate the functioning of the generative AI of fake data, focusing on the critical issues of *Deep Fake,* as well as the legislative responses adopted in other legal systems.
Keywords: artificial intelligence; deep fake; digital forensics.

Cómo citar: Valentina Valenti, F. (2024). Il deep fake: la nuova sfida dell'intelligenza artificiale generativa. *Derecom* 37, 9-18. https://dx.doi.org/10.5209/dere.98114

1. Introduzione

«*La riconosco quando la vedo*». È la celebre frase pronunciata nel 1964 da Potter Stewart, giudice della Suprema Corte degli Stati Uniti d'America, quando fu chiamato a decidere cosa potesse essere ritenuto osceno e cosa no[1]. Successivamente il prof. Floridi[2], nei suoi scritti, ebbe a rilevare come «*l'amicizia, l'IA e molte altre cose nella vita sono come la pornografia: non sono definibili...[] ma le riconosciamo quando le vediamo*». Sono passati poco meno di due anni e questa definizione, per alcuni versi, è stata messa in discussione.

L'intera comunità scientifica si interroga sulla portata dell'intelligenza artificiale, cercando di coglierne le potenzialità e prevenirne i rischi. Lo sviluppo tecnologico, tuttavia, corre a ritmi serrati e presenta al giurista sfide nuove ancor prima che, quest'ultimo, abbia avuto modo di assimilare le precedenti. È noto in ogni latitudine, infatti, come le nuove tecnologie abbiano mutano rapidamente i termini del dibattito: dal dato digitale, alla *privacy online*; dal metaverso agli algoritmi, fino a giungere all'intelligenza artificiale generativa.

Il *deep-fake* — ossia il c.d. falso profondo — è il nuovo risvolto dell'IA e concerne i sistemi informatici in grado di manipolare i dati, i video e le immagini con un grado di realismo tale da non riuscire più a distinguere il vero dal falso. Sebbene possa avere anche dei vantaggi[3], è principalmente conosciuto in una accezione negativa, in quanto i rischi legati all'utilizzo delle tecniche manipolative sono molteplici e coinvolgono diversi profili della sfera soggettiva dell'individuo che, non di rado, sfociano in condotte penalmente rilevanti[4].

Si pensi al diritto dell'identità digitale e ai reati della sfera sessuale: proliferano, infatti, i casi in cui l'intelligenza artificiale generativa viene impiegata per la diffusione di immagini pornografiche e pedopornografiche in cui si utilizza il volto di persone realmente esistenti[5].

È notizia recente la sentenza emessa dal Tribunale per i minorenni di Badajoz, in Spagna, con la quale lo scorso luglio soni stati condannati quindici ragazzi, rei di aver prodotto e diffuso immagini di nudo non consensuale di almeno 20 coetanee utilizzando *software* d'intelligenza artificiale[6]; si possono ricordare, ancora, le indagini condotte dalla *Guardia Civil* di Siviglia in merito alla diffusione di materiale pedopornografico[7] creato con l'intelligenza artificiale; sebbene si tratti di una indagine ancora in fase embrionale è destinata ad alimentare il dibattito pubblico, anche alla luce del fatto che tali strumenti vengono utilizzati in misura prevalente contro il genere femminile[8].

Si assiste, poi, ad un aumento dei reati informatici, in special modo le truffe e le frodi informatiche: spopolano in *internet* le campagne pubblicitarie per la compravendita di prodotti per la cura del corpo in cui ci si avvale di immagini di volti noti del cinema, della tv e, in generale, dello *star-system*.

In alcuni casi, il profitto del reato conseguito mediante strumenti di IA può raggiungere cifre vertiginose: è di dominio mondiale, infatti, la truffa consumata ai danni di un dipendente di una multinazionale cinese che, grazie ad una videochiamata, è stato convinto ad effettuare un pagamento per l'equivalente di circa venticinque milioni di dollari. Dalle informazioni rilasciate dalla polizia di Hong Kong[9] determinante, nella realizzazione dell'intento criminoso, è stata la conferenza da remoto in cui hanno partecipato *manager* della società-madre, avente sede legale in Inghilterra, nonché alcuni colleghi di lavoro del dipendente, ignaro del fatto, però, che gli individui fossero stati clonati da un sistema di intelligenza artificiale.

Ulteriori problemi, non secondari, si riscontrano nel campo della (dis)informazione e della diffamazione e, come opportunamente rilevato in dottrina[10], i *deep-fake* sollevano problemi di *privacy* e *copyright*, poiché le rappresentazioni visive delle persone nei video non sono le copie esatte di un materiale già esistente quanto, piuttosto, nuove rappresentazioni generate dall'IA. La manipolazione dei dati digitali, inoltre, può compromettere il corso delle elezioni politiche, ridurre la fiducia nelle istituzioni ed essere utilizzati come armi per creare crisi nazionali e internazionali[11].

Per comprendere la reale portata del fenomeno, può essere utile riportare qualche dato: il 16 luglio 2024 *Bitdefender* — la nota società di *software*— ha

---

1 La citazione è di FLORIDI L.,(2022), *L'etica dell'intelligenza artificiale,* Raffale Cortina Editore, Milano, p. 41, in cui l'A. spiega come tale espressione venne utilizzata dal giudice nella causa *Jacobellis v. Ohio*, 378 U.S. 184 (1964). Per maggiori approfondimenti è possibile leggere la sentenza al seguente link https://supreme.justia.com/cases/federal/us/378/184/ (consultato il 20 luglio 2024).

2 FLORIDI L.,(2022), *L'etica dell'intelligenza artificiale,* Raffale Cortina Editore, Milano, p. 41.

3 Un utilizzo positivo tali tecnologie lo si riscontra nel documentario *Welcome to Chechnya*, in cui si affronta la persecuzione omossessuale — e in generale della comunità Lgbtq — in Cecenia; il volto (reale) dei testimoni è stato sostituito con quello di altri soggetti americani che hanno prestato il loro consenso. L'intelligenza artificiale ha, dunque, creato un volto nuovo consentendo di tenere segreta la reale identità degli intervistati, salvaguardando la loro integrità fisica e, al contempo, le loro testimonianze.

4 Il tema del furto dell'identità digitale e i deep-fake è stata oggetto di specifico approfondimento da

5 Sull'argomento si v. ÁNGELES JAREÑO LEAL, *El derecho a la imagen íntima y el Código penal La calificación de los casos de elaboración y difusión del deepfake sexual*, in *Revista Electrónica de Ciencia Penal y Criminología*, RECPC 26-09 (2024);

6 *Cfr.* https://www.diarioconstitucional.cl/2024/07/18/un-ano-de-libertad-vigilada-para-15-menores-por-manipular-imagenes-de-ninas-que-las-mostraban-desnudas/ (consultato il 10 agosto 2024)

7 Si v. https://www.interior.gob.es/opencms/en/detail-pages/article/Investigados-cinco-jovenes-por-crear-y-difundir-imagenes-de-20-menores-desnudas-por-inteligencia-artificial/ (consultato il 10 agosto 2024).

8 Sull' argomento si v. SIMO SOLER, E., (2023), *Retos jurídicos derivados de la Inteligencia Artificial Generativa Deepfakes y violencia contra las mujeres como supuesto de hecho*, in *InDret,* 2, pp. 493-515.

9 Si v. https://www.ansa.it/osservatorio_intelligenza_artificiale/notizie/societa/2024/02/08/a-hong-kong-il-deepfake-ti-chiama-in-videocall-per-truffarti_c4d6b374-484c-4e6e-825d-345995a24ed8.html (consultato il 10 agosto)

10 GARCÍA-ULL, F. J. (2021). «*Deepfakes: el próximo reto en la detección de noticias falsas*». Anàlisi: Quaderns de Comunicació i Cultura, 64, 103-120. DOI: https://doi.org/10.5565/rev/analisi.3378

11 ARSLAN, F., (2023) *Deepfake Technology: A Criminological Literature Review*, in *The Sakarya Journal of Law*, 2023, 22, 1, pp. 709 e ss.

reso pubblici i risultati della ricerca condotta sull'impatto dei *deep-fake* nel trimestre marzo-maggio 2024[12]: sebbene l'indagine fosse circoscritta al solo settore sanitario — nello specifico, la vendita di prodotti "miracolosi" in grado di curare il cancro — è emerso come il raggio d'azione di tali campagne fraudolente abbia coinvolto l'Europa, l'Asia, il Nord America, il Medio Oriente e l'Australia e come si siano diffuse grazie alle piattaforme *social* di *Meta*, *Instagram* e *Messenger.*

Ciò dimostra quanto sia difficile, oggi, riconoscere, l'intelligenza artificiale poiché è sempre più arduo fare ragionevole affidamento sulla percezione umana.

A fronte di un fenomeno che ha assunto, già da tempo, confini internazionali, bisogna chiedersi, anzitutto, se la regolamentazione comunitaria e internazionale possa rappresentare un utile ausilio ovvero se sia auspicabile ricercare altrove — e nello specifico negli strumenti informatici — la chiave di lettura per prevenire le manipolazioni dell'IA generativa.

2. Il funzionamento dell'intelligenza artificiale generativa. Brevi cenni

Prima di interrogarci sulle insidie nel rilevamento dei *deep-fake* è necessario soffermarsi, seppur brevemente, sul funzionamento dei sistemi di intelligenza artificiale di tipo generativo e, per far ciò, è necessario individuare alcune premesse concettuali di base.

Anzitutto, l'espressione "intelligenza artificiale" è una locuzione generale che ricomprende diversi algoritmi computazionali in grado di svolgere compiti che tipicamente richiedono l'intelligenza umana, come la comprensione del linguaggio naturale, il processo decisionale, la scelta e l'apprendimento dalle precedenti esperienze[13].

In secondo luogo, è doveroso chiarire che il *machine learning* (c.d. apprendimento automatico) e il *deep learning* (apprendimento profondo), sebbene vengano utilizzati come sinonimi, individuano modelli computazionali diversi.

Il primo è un sottocampo dell'IA che si occupa dello sviluppo di algoritmi in grado di risolvere autonomamente dei compiti senza essere esplicitamente programmati per far ciò; il secondo rappresenta, invece, una forma più evoluta del *machine learning,* che sfrutta le reti neurali artificiali per modellare rappresentazioni complesse nei grandi insiemi di dati, rilevandone automaticamente le correlazioni e i modelli. Le reti neurali sono dei modelli computazionali ispirati alla struttura e al funzionamento del cervello umano, costituiti da strati interconnessi di neuroni artificiali e sono in grado di comprendere gli strati più nascosti dell'architettura, ai grazie ai quali possono apprendere le rappresentazioni gerarchiche delle caratteristiche dei dati, portando a un miglioramento delle prestazioni. L'apprendimento profondo, in tal modo, è in grado di elaborare dati ad alta dimensionalità in vari domini, da quelli monodimensionali — come segnali e testi — a quelli multidimensionali come le immagini, i video o gli audio[14].

Gli sviluppi registrati nel campo dell'apprendimento profondo hanno aperto la strada all'intelligenza artificiale generativa, che rappresenta un nuovo modello di IA per generare contenuti nuovi sulla base di dati reali. Tali modelli ? grazie alla comprensione della distribuzione complessa dei dati ? possono produrre dei risultati che assomigliano molto ai dati del mondo reale. Mutuando le regole statistiche, infatti, l'IA generativa apprende le distribuzioni di probabilità ad alta dimensione, partendo da un "insieme finito" di dati di addestramento, per creare nuovi campioni simili che assomiglino a un'approssimazione della classe di dati di addestramento sottostante[15]. I modelli più innovativi, ossia quelli che sfruttano le reti neuronali, generano contenuti di qualità significativamente più elevata in quanto il sistema algoritmo si concentra sulla generazione probabilistica di nuovi dati invece di determinare i confini decisionali dei dati esistenti[16]; tale aspetto rappresenta, inoltre, il principale fattore di differenziazione tra un modello di *machine learning* e uno di *deep-learning.*

Sebbene l'attenzione globale all'IA generativa sia relativamente recente, essa non rappresenta un *novum* tra gli addetti ai lavori che, in realtà, la studiano da oltre un decennio. I sistemi *de quibus* hanno un indubbio potenziale, anche alla luce del fatto che i campi di applicazioni sono eterogenei; tuttavia, essi portano con sé delle sfide che non possono essere sottaciute, primo fra tutti il pregiudizio strettamente correlato alla qualità dei dati utilizzato nella fase di addestramento ovvero di inferenza (verifica) che può portare all'allucinazione dell'intelligenza artificiale; in secondo luogo, l'alta imprevedibilità comporta un problema di trasparenza, sebbene su tale profilo gli scienziati si stiano interrogando sulla possibilità di sviluppare un sistema di intelligenza artificiale "spiegabile", ossia che l'essere umano sia in grado di capire l'*iter* decisionale elaborato dal modello di IA[17].

Da ultimo, la maggiore sfida è rappresentata dall'uso improprio dell'intelligenza artificiale e rappresenta un problema prioritario. Tali strumenti sono facilmente accessibili — anche ad un costo irrisorio o quasi nullo — e possono produrre effetti dirompenti sul tessuto sociale come dimostrano, ad esempio, le notizie d'oltreoceano ove l'IA viene (anche) impiegata per far veicolare teorie sulla legittimità delle elezioni e orientare il voto degli elettori.

Rischi, questi, perfettamente noti sia tra le Istituzioni che tra i ricercatori i quali, ognuno con i propri mezzi, cerca di studiare delle contromisure.

3. Le prime risposte normative: l'AI ACT

Gli ordinamenti giuridici, di fronte all'incessante sviluppo dell'intelligenza artificiale, hanno avvertito la necessità di una regolamentazione al fine di individuare principi, limiti e divieti nell'utilizzo di tali strumenti.

---

12 Si v. https://www.bitdefender.com/blog/labs/deep-dive-on-supplement-scams-how-ai-drives-miracle-cures-and-sponsored-health-related-scams-on-social-media/?srsltid=AfmBOooO59aXF9Zau2seg6_9mMpThASmh2tn5iPtsbGXc0wXhcBoVsR%2F (consultato il 25 luglio 2024)

13 BANH, L., STROBEL, G., (2023), *Generative artificial intelligence. Electron Markets,* 33, p. 63.

14 GOODFELLOW, I., BENGIO, Y., & COURVILLE, A. (2016). *Deep learning*. The MIT Press.

15 RUTHOTTO, L., & HABER, E. (2021). *An introduction to deep generative modeling.* GAMM-Mitteilungen, 44(2); BANH, L., STROBEL, *op. cit.*, p. 63.

16 BANH, L., STROBEL, *op. cit.*, p. 63.

17 *Ibidem*.

L'Unione Europea, con l'emanazione del Regolamento Europeo (UE) 2024/1689 del 13 giugno 2024 — d'ora in poi *AI Act* o regolamento —, ha raggiunto l'ambizioso obiettivo di disciplinare, per la prima volta, l'intelligenza artificiale, seppur con qualche inevitabile criticità. La prima non può che risiedere nel fattore "tempo", in quanto fisiologicamente il diritto regolamenta i fenomeni sociali quando questi sono già abbiano già prodotto effetti; in secondo luogo, il rapido sviluppo delle tecnologie genera nuovi quesiti e delinea i contorni di nuove sfide che non possono essere soddisfatte sufficientemente dal regolamento unionale. Tra questi, a parere di chi scrive, vi è anche il fenomeno del *deep-fake*.

I rischi derivanti dalla manipolazione cognitiva artificiale — nonché la capacità di incidere sul processo volitivo dell'individuo — non hanno lasciato indifferente il legislatore europeo, sebbene abbiano avuto un ruolo marginale in sede di legiferazione.

Il Regolamento tiene in debito conto l'abuso dell'intelligenza artificiale il cui utilizzo, lungi dal contribuire al funzionamento del mercato interno, persegua pratiche manipolatorie, di sfruttamento o di controllo sociale. Come facilmente intuibile, si tratta di pratiche dannose e contrarie ai valori sposati dall'Unione Europea, quali il rispetto della dignità umana, della libertà, dell'uguaglianza, della democrazia, dello Stato di diritto e dei diritti fondamentali sanciti dalla Carta, ivi compresi il diritto alla non discriminazione, alla protezione dei dati e alla vita privata e i diritti dei minori. I sistemi *de quibus,* inoltre, possono sfruttare la vulnerabilità degli individui in ragione delle condizioni economiche e/o sociali e di disabilità psico-fisica.

Per quanto le perplessità in commento siano ineccepibili, in punto di regolamentazione *l'AI Act* non prende una posizione ben definitiva sul *deep-fake*, con ciò ingenerando alcune difficoltà di inquadramento giuridico.

Sotto il profilo definizionistico, l'*AI Act* all'art. 3 comma 60 descrive il *deep-fake* come «*un'immagine o un contenuto audio o video generato o manipolato dall'IA che assomiglia a persone, oggetti, luoghi, entità o eventi esistenti e che apparirebbe falsamente autentico o veritiero a una persona*».

Orbene, con espresso riferimento alle tecniche manipolative, il regolamento prevede diversi livello di rischio[18] — alto[19] e sistemico — e vieta «*l'immissione sul mercato, la messa in servizio o l'uso di un sistema di IA che utilizza tecniche subliminali che agiscono senza che una persona ne sia consapevole o tecniche volutamente manipolative o ingannevoli aventi lo scopo o l'effetto di distorcere materialmente il comportamento di una persona o di un gruppo di persone, pregiudicando in modo considerevole la loro capacità di prendere una decisione informata, inducendole pertanto a prendere una decisione che non avrebbero altrimenti preso, in un modo che provochi o possa ragionevolmente provocare a tale persona, a un'altra persona o a un gruppo di persone un danno significativo*»[20].

In linea teorica, dunque, il *deep-fake* potrebbe rientrare nelle pratiche vietate alla luce della sua elevata capacità disinformativa e manipolativa sulla società digitale.

Tuttavia, come già accennato, il regolamento europeo calibra l'intelligenza artificiale su più livelli di rischio ed ha individuato due diverse categorie: i "*sistemi di IA con finalità generali a rischio sistemico*" e i "*determinati sistemi di IA*".

Per quanto l'*AI Act* non prenda posizione sul punto, non può escludersi che i *deepfake* possano entrare in frizione con i diritti fondamentali o con la sicurezza dell'UE. In tal caso, ci si troverebbe dinanzi a sistemi di intelligenza artificiale a rischio sistemico. Ai sensi dell'art. 3 comma 65, un sistema è tale quando «*rappresenta un rischio specifico per le capacità di impatto elevato dei modelli di IA per finalità generali, avente un impatto significativo sul mercato dell'Unione a causa della sua portata o di effetti negativi effettivi o ragionevolmente prevedibili sulla salute pubblica, la sicurezza, i diritti fondamentali o la società nel suo complesso, che può propagarsi su larga scala lungo l'intera catena del valore*».

In ossequio ad esigenze di completezza si fa presente che per "finalità generali" si intende la capacità di un modello algoritmo — anche laddove sia addestrato con grandi quantità di dati utilizzando l'autosupervisione su larga scala — che sia caratterizzato una generalità significativa e sia in grado di svolgere con competenza una vasta gamma di compiti, indipendentemente dalle modalità con cui il modello sia stato immesso sul mercato, che possa essere integrato in una varietà di sistemi o applicazioni a valle, ad eccezione dei modelli di IA utilizzati per attività di ricerca, sviluppo o di prototipazione prima dell'immissione sul mercato[21].

Ancora, è doveroso ricordare che il rischio sistemico, ai sensi dell'art. 51, per essere tale deve soddisfare una delle condizioni di seguito illustrate: deve presentare una capacità di impatto elevato valutata sulla base di strumenti tecnici e metodologie adeguati; deve presentare una capacità o un impatto elevato valutato *ex officio* dalla Commissione o a seguito di una segnalazione qualificata del gruppo di

---

18 Come noto l'UE, sin dalla prima proposta di regolamentazione, aveva sposato un sistema di classificazione basato sulla diversità intensità di rischio: minimo, limitato, alto e inaccettabile. La versione definitiva approvata dal Parlamento, invece, contiene copiosa normativa sui sistemi ad alto rischio e su quelli a rischio. Sul punto si v. https://digital-strategy.ec.europa.eu/it/policies/regulatory-framework-ai (consultato il 6 agosto 2024)

19 I sistema di classificazione ad alto rischio — di non immediata intellegibilità — prevede tre diverse ipotesi: in primo luogo, un sistema può essere considerato tale se soddisfa due condizioni, ossia il prodotto sistema di IA deve essere utilizzato come componente di sicurezza di un prodotto, o il sistema di IA è esso stesso un prodotto, disciplinato dalla normativa di armonizzazione dell'Unione elencata nell'allegato I e il prodotto — il cui componente di sicurezza a norma della lettera a) — è il sistema di IA, o il sistema di IA stesso in quanto prodotto, è soggetto a una valutazione della conformità da parte di terzi ai fini dell'immissione sul mercato o della messa in servizio di tale prodotto ai sensi della normativa di armonizzazione dell'Unione elencata nell'allegato I; in secondo luogo, un sistema può essere ad altro rischio quando si tratta di sistemi di intelligenza artificiali in settori ricompresi nell'allegato III del regolamento, salvo il caso esso non rappresenti un rischio significativo di danno per la salute, la sicurezza o i diritti fondamentali delle persone fisiche, anche nel senso di non influenzare materialmente il risultato del processo decisionale.; infine, un sistema è sempre ad alto rischio quando effettua la profilazione delle persone.

20 *Cfr.* art. 5, comma 1, lettera a) del regolamento.

21 *Cfr.* art. 3, comma 63 del regolamento.

esperti scientifici, tenendo conto dei criteri sulla trasparenza indicati nell'allegato XII[22]. Secondo *l'AI ACT,* inoltre, si presume che un modello di IA per finalità generali abbia capacità di impatto elevato quando la quantità cumulativa di calcolo utilizzata per il suo addestramento sia superiore a 10^{25}.

Per tali sistemi algoritmi il legislatore europeo ha previsto degli obblighi per i fornitori, la cui disciplina è contenuta negli artt. 53, 54 e 55. A tal fine, essi devono redigere e mantenere aggiornata la documentazione del modello, dalla fase di addestramento fino ai risultati della valutazione; l'aggiornamento *de quo* include anche i dati nel caso in cui il fornitore intendesse integrare il modello, nonché la sintesi dei contenuti utilizzati per l'addestramento del modello di IA (art. 53); tutta la documentazione in commento deve essere messa a disposizione dell'Agenzia per IA per il tramite del rappresentante del fornitore del modello di IA, il quale deve cooperare con gli uffici dell'Unione e fornire tutte le informazioni necessarie per dimostrare la conformità del modello di IA alle prescrizioni del Regolamento europeo. Infine, trattandosi di modelli con rischio sistemico, i fornitori devono predisporre una valutazione dei modelli alla luce dei protocolli, nonché attenuare i possibili rischi sistemici nell'UE, garantendo un livello adeguato della sicurezza informatica e dell'infrastruttura fisica del modello[23].

Un espresso riferimento al *deep-fake*, invece, lo si ritrova negli obblighi di trasparenza per i fornitori e i *deployers* di "*determinati sistemi di IA*", di cui all'art. 50 del Regolamento.

L'obiettivo principale è colmare il *gap* informativo degli individui affinché quest'ultimi siano consapevoli di interagire con un sistema di intelligenza artificiale.

A tal fine è previsto l'obbligo per i fornitori dei sistemi di IA, compresi i sistemi di IA per finalità generali, che generino contenuti audio, immagine, video o testuali sintetici, di garantire che gli *output* del sistema di IA siano marcati in un formato leggibile meccanicamente e siano rilevabili come generati o manipolati artificialmente. 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 che il testo in questione sia stato generato o manipolato artificialmente (artt. 50, commi 2 e 4 ultimo periodo).

Con espresso riferimento al *deep-fake*, il comma 4 stabilisce che «*i deployer di un sistema di IA che genera o manipola immagini o contenuti audio o video che costituiscono un "deep fake" rendono noto che il contenuto 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. Qualora il contenuto faccia parte di un'analoga opera o di un programma 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*».

Si evince, dunque, che il regolamento europeo non proibisce "in assoluto" l'impiego, l'utilizzo o la circolazione dei *deep-fake*; infatti, quando essi siano frutto della creatività artistica e della libertà di espressione dell'individuo si impone al creatore del contenuto di rilevare il fatto che esso sia frutto di una manipolazione. In tale modo si è cercato di bilanciare diritti contrapposti ma egualmente tutelati: la manifestazione del pensiero da un lato, il diritto ad una corretta informazione dall'altro.

Un'ulteriore eccezione, infine, è prevista nel caso in cui il contenuto manipolato venga impiegato per la prevenzione dei reati.

A parere di chi scrive, la normativa europea in tema sul tema dei *deep-fake* presta il fianco ad alcune critiche.

Anzitutto, la tecnica legislativa impiegata non è immediatamente intellegibile, il che rende più difficoltosa la ricostruzione normativa dei sistemi di rischio con riferimento a specifici modello di IA.

In secondo luogo, non si ritiene che imporre l'obbligo di dichiarare che il contenuto sia stato manipolato sia idoneo ad assolvere una funzione deterrente; del resto, le sanzioni previste dagli artt. 99-100-101 sono espressamente previste per le aziende che immettano nel mercato prodotti — o servizi — di IA non conformi ai requisiti previsti dal regolamento.

Tuttavia, i sistemi di intelligenza artificiale — ancorché sviluppati da aziende specializzate — possono essere (e sono) utilizzati dai singoli individui, il quali possono addestrare il sistema nella creazione di contenuti falsi e/o manipolati. Del resto, i sistemi di apprendimento profondo — proprio perché tali — sono dei sistemi aperti; ciò significa che ogni singolo *output* generato entra a far parte del set di base che, a sua volta, verrà utilizzato per migliorare le risposte di previsione.

Il legislatore europeo ha prediletto un sistema di responsabilità calibrato sulle aziende produttrici del prodotto ma, a parere di chi scrive, non sembra sufficiente a limitare, in qualche modo, l'utilizzo di pratiche manipolatorie.

Al di là delle aziende, ci si dovrebbe interrogare sulla possibilità di rintracciare un eventuale uso distorto dell'intelligenza artificiale affinché le sanzioni previste possano effettivamente svolgere una funzione deterrente anche in capo ai singoli fruitori.

Inoltre, il tema del *deep-fake* all'interno dell'IA ACT è affrontato solo marginalmente all'interno degli obblighi informativi dei determinati modelli di IA. Per tale ragione si ritiene che l'efficacia di tale Regolamento resti alquanto dubbia: infatti, l'assenza di una espressa classificazione del rischio per

---

[22] L'allegato XII richiede una serie di informazioni necessarie per assolvere l'obbligo di trasparenza, quali: una descrizione generale del modello di IA per finalità generali, comprendente i compiti che il modello è destinato a eseguire e il tipo e la natura dei sistemi di IA in cui può essere integrato; le politiche di utilizzo accettabili applicabili; la data di pubblicazione e i metodi di distribuzione; il modo in cui il modello interagisce o può essere utilizzato per interagire con hardware o software che non fanno parte del modello stesso; le versioni del software pertinente relative all'uso del modello di IA per finalità generali; l'architettura e il numero di parametri; la modalità (ad esempio testo, immagine) e il formato degli input e degli output; la licenza per il modello. È richiesta, altresì, una descrizione degli elementi del modello e del processo relativo al suo sviluppo, compresi i mezzi tecnici necessari per integrare il modello di IA per finalità generali nei sistemi di IA; la modalità (ad esempio testo, immagine, ecc.) e il formato degli input e degli output e la loro dimensione massima; informazioni sui dati utilizzati per l'addestramento, la prova e la convalida, se del caso, compresi il tipo e la provenienza dei dati e le metodologie di organizzazione.

[23] *Cfr.* art. 55.

i sistemi di IA che generano *deep-fake*, nonché le eccezioni agli obblighi di trasparenza consentono di sfuggire agilmente agli obblighi previsti in tema di manipolazione artificiale.

3.1. (segue). La legislazione cinese e le proposte federali USA

La regolamentazione dell'intelligenza artificiale non è una prerogativa solo dell'Unione Europea, ma si tratta di una esigenza condivisa a livello internazionale che vede la partecipazione attiva di altri due attori di indubbia valenza geopolitica: gli Stati Uniti d'America e la Cina sebbene, al momento, lo sviluppo delle rispettive discipline di settore siano ancora in una fase embrionale[24].

Diversamente da quel che accede per l'intelligenza artificiale, alla luce delle informazioni sinora disponibili, l'unico ordinamento giuridico ad avere promulgato una legge specifica sul *deep-fake* sembra essere, al momento, la Repubblica Popolare Cinese.

La legge in commento, entrata in vigore il 10 gennaio 2023, è la prima del suo genere ad essere direttamente indirizzata al fenomeno dei *deep-fake*, poiché specifica le linee guida cui i *providers* devono attenersi e introduce la protezione dei diritti di *privacy* degli utenti[25]. Quest'ultima, infatti, è al centro della regolamentazione, poiché la legge riconosce il diritto degli utenti di contestare l'utilizzo della propria immagine per la creazione di video manipolati da condividere in *internet*. I fornitori di IA devono registrare i propri servizi e sottoporre, con cadenza periodica, i propri codici e i dati ad una revisione statale. I gestori delle applicazioni (cc.dd. "*app*") di *deep-fake* devono, invece, indicare espressamente la presenza di un video alterato classificandolo come "modificato" e non originale. Sono vietati in modo assoluto tutti i contenuti che non siano stati approvati dagli enti governativi, nonché quelli considerati illegittimi per diffusione di false informazioni[26].

Il governo cinese ha previsto, inoltre, l'istituzione di una agenzia *ad hoc* per il controllo dei contenuti di *deep-fake*: la *Cyberspace Administration of China ha* il compito di vigilare sul rispetto degli obblighi imposti ai *providers*.

Più nello specifico, il governo cinese ha previsto l'obbligo, per i fornitori, di implementare i sistemi algoritmi per garantire la sicurezza delle informazioni; sviluppare e migliorare i sistemi di gestione per la verifica degli utenti e la valutazione algoritmica; valutare l'etica della tecnologia, delle informazioni, la sicurezza dei dati e la protezione delle informazioni personali; prevenire le frodi nelle reti di telecomunicazioni e fornire una risposta agli incidenti di sicurezza, nonché implementare misure di salvaguardia tecnica[27].

Al fine di contrastare in maniera efficace le tecniche di *deep-fake* la Cina ha previsto, altresì, l'obbligo in capo ai fornitori di IA di verificare le reali identità degli utenti tramite i loro numeri di cellulare, i numeri dei documenti di d'identità, i codici di credito sociale unificati o tramite il servizio pubblico nazionale di verifica dell'identità online ed essi non possono fornire servizi di diffusione di informazioni agli utenti che non abbiano verificato le proprie informazioni.

Inoltre, se il fornitore del servizio accerta che i sistemi di IA vengono utilizzati per produrre, copiare, pubblicare o diffondere informazioni false, devono prontamente adottare misure per smentire le voci, conservare i relativi registri e inviare un rapporto al dipartimento informazioni e ad altri dipartimenti pertinenti[28].

Onde evitare di ingenerare confusione tra gli utenti, i contenuti divulgati in *internet* e manipolati con gli strumenti di intelligenza artificiale devono avere un contrassegno ben visibile che informi l'utente che il contenuto visionato sia stato modificato o generato dal sistema, in special modo quando si tratta di: servizi che simulano persone fisiche per generare e/o modificare testi, dialoghi e scritture; servizi che generano voci sintetiche o simulano voci, oppure alterano in modo significativo le caratteristiche dell'identità personale; servizi che creano immagini o video di persone, come la generazione di volti, lo scambio di volti, la manipolazione dei volti e la manipolazione dei gesti; nonché la modifica degli attributi personali e quelli che generano o modificano scene di simulazione immersiva.

È previsto, altresì, per i fornitori l'obbligo di archiviazione dei contenuti, che devono essere contrassegnati da un numero di archiviazione e devono contenere il relativo collegamento al registro pubblico di archiviazione sui siti *web* e sulle applicazioni.

Ebbene, l'approccio cinese al *deep-fake* è contraddistinto da una gestione delle responsabilità distribuita tra i fornitori dei modelli di IA e i singoli

---

[24] Per quanto riguarda gli USA, infatti, la legge sull'intelligenza artificiale è ancora in fase di evoluzione: sono molte le proposte federali e statale sul tema e nessuna di esse è stata formalmente approvata. Allo stato, l'unica legge federale è il *National Artificial Intelligence Initiative Act* avente ad oggetto a ricerca e lo sviluppo nel campo dell'AI e il *National Artificial Intelligence Initiative Office*, responsabile della supervisione e dell'implementazione della strategia nazionale statunitense sull'AI. Sul punto si v. https://www.agendadigitale.eu/cultura-digitale/leggi-sullintelligenza-artificiale-ecco-la-complessa-roadmap-usa/ (consultato il 3 luglio 2024). Sebbene anche la Cina non abbia ancora approvato una legge ad hoc, una bozza del testo legislativo — ancorché non definitiva — è stata condivisa dal Governo cinese nell'ottobre 2023. Il *Basic security requirements for generative artificial intelligence service,* questo il titolo, si prefigge il compito di offrire una regolamentazione di dettaglio dallo spiccato taglio tecnico (dalla progettazione alla messa in commercio), al fine di avere maggiori possibilità di efficacia. L'ultima versione (2023) è disponibile in lingua inglese al seguente indirizzo https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai/ (consultato l'8 agosto 2024). Per approfondimenti sulla regolamentazione tra Cina e UE, sia consentito il rinvio a Roberts, H., Cowls, J., Hine, E., Morley, J., Wang, V., Taddeo, M., & Floridi, L. (2022). Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes. *The Information Society*, *39*(2), 79-97. https://doi.org/10.1080/01972243.2022.2124565

[25] Il testo della legge 互联网信息服务深度合成管理规定 (Regulations on deep synthesis management of Internet information services) è consultabile al seguente link https://perma.cc/JE3W-PF26 (consultato il 7 agosto 2024).

[26] Per approfondimenti si v. L. GOLDIN, *La Cina regolamenta l'uso deep fake*, in *Istituto Confucio,* consultabile al seguente link https://www.istitutoconfucio.unimi.it/2023/01/la-cina-regolamenta-luso-deep-fake/#:~:text=Il%20governo%20cinese%20ha%20emesso,politici%20utilizzati%20in%20chiave%20satirica. (consultato l' 8 agosto 2024)

[27] *Cfr.* art. 7 della legge cinese.

[28] *Cfr.* art. 11.

utilizzatori, posto che senza la condivisione dei reali dati personali l'accesso non è consentito. I dati identificativi dell'utente, inoltre, consentono di risalire alla persona specifica nel caso in cui i contenuti divulgati non siano in linea con gli standard legislativi, il cui contenuto è controllo, gestito e conservato mediante un processo di tracciamento che grava sul fornitore del servizio.

La filiera del controllo digitale è forse, in linea teorica, quella maggiormente efficace per contrastare le pratiche di *deep-fake*.

Tuttavia, la legge in commento si spinge oltre, mediante un controllo capillare che travalica le reali necessità di informazione, *privacy* e tutela dell'immagine, abbracciando un controllo politico per il quale la sicurezza e la reputazione nazionale assumono la priorità assoluta.

I termini del discorso mutano sensibilmente, invece, se si volge lo sguardo agli Stati Uniti d'America. Per quanto il fenomeno dei *deep-fake* sia particolarmente sentito — anche a livello politico — non vi è attualmente una legge che disciplini il fenomeno.

Tra le varie proposte federali attualmente in cantiere meritano attenzione il *Copied Act*[29] e il "*No Fakes Act*" [30]. Il primo si prefigge l'obiettivo di contrastare l'aumento dei *deep-fake* dannosi mediante 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'intelligenza artificiale e ritenere i trasgressori responsabili degli abusi. A tal fine, è necessario standardizzare le informazioni sulla provenienza dei contenuti, la filigrana e il rilevamento dei contenuti sintetici. Così facendo si vorrebbero incrementare gli *standard* di trasparenza per identificare quali contenuti siano stati generati o manipolati dall'IA, nonché accertare la provenienza dei contenuti di IA.

Il "*No Fakes Act*", invece, intende tutelare l'immagine e la voci delle persone clonate, sinora, in maniera tendenzialmente illegittima dagli algoritmi, riconoscendo un diritto federale ad ogni individuo e prevedendo l'obbligo per le piattaforme di procedere alla rimozione dei contenuti non autorizzati dal legittimo proprietario.

Il disegno di legge in commento, diversamente dal precedente, ha un campo di applicazione più settorializzato e si prefigge il compito di tutelare il diritto d'autore, special modo di attori e cantanti.

Tali misure — in special modo il *Copied Act* — ove dovessero essere approvate, consentirebbero di fare un passo in avanti nella lotta ai contenuti manipolati e porre un freno alla disinformazione, anche in considerazione del fatto che negli USA, attualmente, non esiste una legge che proibisca la rimozione, la disattivazione o la manomissione delle informazioni sulla provenienza dei contenuti. Il disegno di legge in commento, invece, proibirebbe a chiunque, comprese le piattaforme *Internet*, i motori di ricerca e le società di *social media*, di interferire con le informazioni sulla provenienza dei contenuti digitali.

4. Le difficoltà delle scienze forensi nell'accertamento dei *deep-fake*

Alla luce degli effetti giuridici prodotti dai *deep-fake,* la necessità di vagliare la genuinità delle informazioni che circolano in rete ha portato allo sviluppo di una vasta gamma di strumenti e tecniche nell'ambito della *digital forensic,* una branca delle scienze forensi che si occupa dell'analisi del materiale rinvenuto nei dispositivi digitali. Si tratta di analisi tecniche che accertano l'originalità e la genuinità delle voci registrate, delle immagini, dei suoni, dei video, analizzandone le proprietà tecniche. L'accertamento della veridicità dei *file* digitali nell'era dei *deep-fake* varia a seconda del tipo di contenuto da analizzare e le difficoltà che si riscontrano sono differenti a seconda che si tratti di audio, di immagini o di video.

Più nel dettaglio, in tema di analisi vocale i recenti studi condotti nel campo dell'informatica[31] dimostrano che gli individui faticano a riconoscere una voce naturale da una prodotta da un computer, poiché i sistemi di *deep-learning* riescono a far apparire il discorso verbale estremamente naturale inducendo in errore l'interlocutore.

Per individuare, allora, le possibili manipolazioni degli audio gli studiosi hanno sperimentati diversi metodi, quali: l'analisi delle informazioni contenute nel dispositivo[32] e sul rumore di fondo dell'audio[33]; l'accertamento sul dominio di frequenza del segnale[34]; l'apprendimento auto-supervisionato per il rilevamento e l'addestramento delle configurazioni acustiche[35]; l'utilizzo delle caratteristiche ad alta frequenza per identificare e rilevare il parlato e, infine, la costruzione di un sistema neuronale per l'analisi della genuinità delle caratteristiche[36].

Tuttavia, sulla base degli ultimi risultati prodotti sulle riviste di settore, il metodo più efficace per smascherare le voci e i dialoghi generati con l'intelligenza artificiale sarebbe il sistema di classificazione

---

29 Bill Draft (RIL24710 NP8) "To require transparency with respect to content and content provenance information, to protect artistic content, and for other purposes". Il testo, proposto dai senatori Ms. CANTWELL, Mrs. BLACKBURN, e Mr. HEINRICH è consultabile al seguente link: https://www.commerce.senate.gov/services/files/3012CB20-193B-4FC6-8476-DDE421F3DB7A (consultato il 2 agosto 2024).

30 Bill Draft (EHF24841 1M0) "*To protect intellectual property rights in the voice and visual likeness of individuals, and for other purposes*" Si v. il testo in https://www.coons.senate.gov/imo/media/doc/no_fakes_act_bill_text.pdf (consultato l'8 agosto 2024). Sul tema si v. anche Geng, Yinuo. (2023). *Comparing "deepfake" regulatory regimes in the United States, the European Union, and China. G*eorgetown Law Technology Review, 7(1), 157-178.

31 Si veda la ricerca condotta da AL-Shakarchy, N.D., Abdullah, Z.N., Alameen, Z.M. *et al., Audio verification in forensic investigation using light deep neural network*. *Int. j. inf. tecnol.* **16**, 2813-2821 (2024). https://doi.org/10.1007/s41870-024-01812-2

32 Muh MAQHAAHMG (2021) *Digital audio forensics: microphone and environment classification using deep learning*. IEEE Access 9:62719-62733

33 CHEN J, XIANG S, HUANG H, LIU W (2016) *Detecting and locating digital audio forgeries based on singularity analysis with wavelet packet. Multimed Tools* Appl 75:2303-2325;

34 XIAODAN LIN XK, (2017) *Exposing speech tampering via spectral phase analysis. Digit Signal Process* 60:63-74;

35 LEI Z, YANG Y, LIU C, YE J (2020) *Siamese convolutional neural network using gaussian probability feature for spoofing speech detection school of computer and information engineering*. Jiangxi Normal University, Nanchang, pp 1116-1120;

36 SHIM HJ, HEO HS, JUNG JW, YU HJ (2020), *Self-supervised pretraining with acoustic configurations for replay spoofing detection*. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. Vol. 2020, pp. 1091-1095;

basato su reti neurali profonde, i cc.dd. *light deep neural network*[37].

Si tratta di un sistema basato su una rete neurale convoluzionale unidimensionale, con un nuovo modello di architettura per la verifica dell'audio che si fonda sulla predizione della relazione tra un modello vocale, la determinazione delle mappe di caratteristiche e l'approssimazione della funzione interna sconosciuta a un insieme di dati di ingresso e di uscita corrispondenti. Il funzionamento del sistema in commento può essere scisso in due diversi momenti: la fase di pre-elaborazione e quella di verifica. Nella prima avviene l'addestramento del modello, mediante un *data-set* contenente sia voci reali che artefatte, al fine di migliore le prestazioni del modello nel riconoscere gli audio veri da quelli creati artificialmente; nella seconda fase, invece, si estraggono e si analizzano le caratteristiche degli audio per accertare se effettivamente si tratti di un audio reale oppure no.

Tuttavia, come dichiarato dagli sviluppatori, il modello *de quo* è particolarmente utile nel caso in cui si disponga di una quantità limitata di dati per l'addestramento[38], ma non è attualmente in grado di svolgere l'accertamento in tempo reale, sicché potrebbe presentare qualche profilo di criticità nei casi in cui le esigenze forensi presentino tempistiche particolarmente stringenti.

Profili di maggiore criticità si registrano nel campo delle immagini, in quanto l'identificazione e l'autenticazione dei media digitali rappresentano una sfida non secondaria nelle indagini forensi. Gli studiosi negli ultimi sei anni hanno sviluppato diversi sistemi per cercare di "smascherare" foto, video e audio artefatti.

Ad esempio, alcuni ricercatori[39] hanno proposto un modello basato su due diversi sistemi ricorrenti addestrabili *end-to-end*: uno che estrae le caratteristiche più critiche dei dati e un sistema che procede all'analisi sequenziale; alcuni studiosi dell'Università di New York hanno proposto un modello basato su 4 distinte reti neuronali addestrato utilizzando volti reali e artefatti[40]; gli scienziati dell'Università di Albany (USA)[41] hanno sperimentato un metodo basato sul rilevamento del battito degli occhi; alcuni ricercatori hanno proposto una rete a due flussi: la prima rete di classificazione dei volti cattura le prove degli artefatti di manomissione e la seconda la rete di triplette, basata sulle caratteristiche della steganalisi, cattura le prove residue del rumore locale[42]; altri[43] hanno utilizzato maschere generate al computer per ritagliare e regolare le aree del volto per rilevare la manipolazione; il sistema *ForensicTransfer* (FT), invece, è in grado di differenziare efficacemente le immagini originali da quelle contraffatte[44]; alcuni hanno sviluppato un *software multi-tasking*[45] per rilevare e segmentare simultaneamente immagini e video manipolati; alcuni ricercatori hanno sviluppato un modello per rilevare video di ritratti sintetici sfruttando i segnali biologici[46], mentre in vengono utilizzati i riflessi mancanti e i dettagli mancanti nelle aree degli occhi e dei denti[47].

Ebbene, ognuno dei modelli sinora proposti si scontra con alcune evidenti criticità perché i sistemi di *deep-fake* creano video e immagini realistici grazie a sistemi di intelligenza artificiale generativa in grado di imitare la gestualità, le voci e il volto di un soggetto. Tra i sistemi maggiormente utilizzati vi sono le reti generative avversarie (GAN) e *autoencoder*[48], che richiedono l'addestramento di due reti neurali: un generatore e un discriminatore. Il primo produce immagini e video manipolati, mentre il secondo viene addestrato per determinare se un'immagine o un video è autentico o falso. Inoltre, per comprimere e ricostruire i dati viene utilizzata una classe di reti neurali chiamata *autoencoder*. Quest'ultimi vengono addestrati a riconoscere gli aspetti più cruciali di un'immagine – quali, ad esempio, le caratteristiche del viso – per comprimerla in una quantità minore di dati che vengono utilizzati per ricreare l'immagine originale[49]. Trattandosi di sistemi di auto-apprendimento, i modelli sono in grado di generare *output* con un grado di precisione e realismo sempre maggiore, con

---

37 HIREN MEWADA QN, AL-ASAD JF, ALMALKI FA, KHAN AH, ALMUJALLY NA, EL-NAKLA S (2023) *Gaussian-filtered high-frequency-feature trained optimized BiLSTM network for spoofed-speech classification.* Sensors (Basel) 23(14):1-24

38 AL-SHAKARCHY, N.D., ABDULLAH, Z.N., ALAMEEN, Z.M. *et al., Audio verification in forensic investigation using light deep neural network*. *Int. j. inf. tecnol.* 16, 2813-2821 (2024). https://doi.org/10.1007/s41870-024-01812-2

39 D. GÜERA and E. J. DELP, "*Deepfake Video Detection Using Recurrent Neural Networks*", *2018 15 th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* , pp. 1-6, 2018.

40 Y. LI and S. LYU, "*Exposing DeepFake Videos by Detecting Face Warping Artifacts*", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 46-52.

41 Y. LI, M. CHANG and S. LYU, "*In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking*", *2018 IEEE International Workshop on Information Forensics and Security (WIFS) Hong Kong Hong Kong*, pp. 1-7, 2018

42 P. ZHOU, X. HAN, V. I. MORARIU and L. S. DAVIS, "Two-Stream Neural Networks for Tampered Face Detection", *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1831-1839, 2017.

43 E. SABIR, J. CHENG, A. JAISWAL, W. ABDALMAGEED, I. MASI and P. NATARAJAN, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos", *Workshop on Applications of Computer Vision and Pattern Recognition to Media Forensics with CVPR*, pp. 80-87.

44 D. COZZOLINO, J. THIES, A. RÖSSLER, C. RIESS, M. NIESSNER and L. VERDOLIVA, "*ForensicTransfer: Weakly-supervised domain adaptation for forgery detection*", *arXiv*, 2018.

45 H. H. NGUYEN, F. FANG, J. YAMAGISHI and I. ECHIZEN, "*Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos*", *arXiv*, 2019.

46 U. A. CIFTCI, I. DEMIR and L. YIN, "*FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals*", *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

47 F. MATERN, C. RIESS and M. STAMMINGER, "Exploiting visual artifacts to expose deepfakes and face manipulations", *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 83-92, 2019.

48 Gli *autoencoder* (auto-codificatori) sono una classe di reti neurali orientati all'elaborazione di immagini con una caratteristica centrale: la rete neurale viene considerata *non come una scatola nera ma bensì due*: l'encoder responsabile per elaborare l'input e convertirlo in un formato intermedio in uno spazio vettoriale (chiamato *latent space*); il decoder prende in input la rappresentazione nel *latent space* di un'immagine e la converte in un'immagine di output. Per approfondimenti si veda A. CISTERNINO, *L'autoencoder svelato: ecco come l'AI crea le immagini*, in *Agenda Digitale*, consultabile in https://www.agendadigitale.eu/cultura-digitale/lautoencoder-svelato-ecco-come-lai-crea-le-immagini/ (consultato il 4 agosto 2024).

49 TAMPUBOLON, M., *Digital Face Forgery and the Role of Digital Forensics. Int J Semiot Law* 37, 753-767 (2024). https://doi.org/10.1007/s11196-023-10030-1

conseguente difficoltà nell' individuare i contenuti frutto di creazione.

Un sistema non dissimile avviene per la creazione dei video, in quanto le reti neuronali vengono utilizzare per la creazione di video in formato 3D. Infatti, il rilevamento comporta il riconoscimento delle principali caratteristiche facciali dell'individuo e le proietta su un video o un'immagine esistente. In questo modo, è possibile produrre un *deep-fake* che imita con precisione le espressioni facciali e il linguaggio del corpo dell'individuo[50].

In letteratura, inoltre, non vi è una uniformità di vedute sulla tipologia di modello da utilizzare: secondo alcuni ricercatori[51], infatti, i sistemi di *deep learning* sebbene consentano di ottenere risultati molto accurati nel rilevamento dei *deep-fake* comportano degli svantaggi che non possono essere sottaciuti, quali: la comprensibilità, l'interpretabilità, la complessità dei dati e l'elevato costo computazionale[52]; diversamente, i modelli di *machine learning* sembrerebbero avere una maggiore intellegibilità e, sulla base degli ultimi studi condotti in via sperimentale, avrebbero imparato a riconoscere i *deep-fake* con una precisione pari al 99%.

Tuttavia, sono due le criticità con le quali bisogna confrontarsi: i modelli di *deep learning* basati sulle reti neuronali generano un problema di interpretabilità dei dati, perché le correlazioni individuate ed elaborate dal sistema non sono sempre lineari e ciò incide in maniera significativa sulla loro comprensione. Inoltre, i ricercatori hanno evidenziato come l'addestramento di tali modelli richieda una enorme quantità di dati tali da incidere sui tempi di addestramento; i modelli basati sul *machine learning,* però, basandosi su un sistema di dati molto più contenuto comportano un sacrificio in termini di accuratezza delle correlazioni.

Il fatto che nella comunità scientifica vi sia una diversità di vedute sugli strumenti tecnici da utilizzare per superare i pericoli del *deep-fake* — ognuno dei quali ha dei pro e contro — aiuta a comprendere come il problema sia molto più complesso di quanto, forse, si possa pensare.

La mancanza di una visione comune si traduce anche nell'assenza di protocolli e linee di intervento condivise da impiegare nell'ambito delle scienze forensi, le quali si trovano ad affrontare tali sfide con risorse non sempre adeguate al caso concreto. A ciò bisogna aggiungere che i sistemi di intelligenza artificiale per la creazione dei *deep-fake* si sviluppano in tempi ristretti e gli esperti forensi difficilmente riescono a tenere il ritmo.

Inoltre, vi è un altro aspetto sul quale soffermarsi: come rilevato in dottrina[53], la maggior parte dei falsi digitali viene creata sulla base di video o immagini realmente esistenti, il cui contenuto è accessibile sulle piattaforme *social*, sicché l'accertamento effettuato mediante un confronto con il video (o la foto) reale rappresenta un ottimo ausilio poiché offre indici significativi; i problemi, tuttavia, aumentano, nel caso in cui non si disponga del contenuto digitale reale ma sì tratti di un contenuto creato *ex novo* dall'intelligenza artificiale. In tali casi la possibilità dei sistemi di rilevare e smascherare un *deep fake* richiede più tempo e maggiori risorse che non sempre è possibile ottenere.

5. Conclusioni

Alla luce delle riflessioni suesposte è possibile trarre qualche conclusione. Anzitutto, i fatti di cronaca citati nel corso dell'esposizione — che pure rappresentano una minima parte — dimostrano inequivocabilmente quanto l'utente sia quasi inerme dinanzi alla tecnologia. Se in alcuni casi una maggiore accortezza avrebbe consentito di scrivere un epilogo diverso — ad esempio nel caso delle sistematiche truffe *online* — in altri si è totalmente privi di difesa ed un plastico esempio è la truffa milionaria di Hong Kong. Se da un lato ciò dimostra l'altro grado di sofisticatezza che le nuove tecnologie possono raggiungere, dall'altro dimostra pure che la sicurezza informatica dovrà necessariamente adeguarsi al cambiamento onde offrire una tutela efficace e ciò avrà inevitabilmente ripercussioni sul tessuto sociale.

Un mutamento, questo, per alcuni versi già in atto. L'ordinamento giuridico italiano, per esempio, ha recentemente introdotto nuove fattispecie di reato nel codice penale mediante le quali si incriminano le diffusioni di contenuti generati con l'intelligenza artificiale, sebbene tali previsioni non siano esenti da critiche.

Rimanendo in tema di legiferazione, lo stesso regolamento europeo presenta criticità dovute al fatto che le previsioni in esso contenute non appaiono idonee a far fronte alle inside dei *deep-fake*, in primo luogo perché il fenomeno non è stato oggetto di una specifica trattazione; in secondo luogo, è regolamentato in maniera disorganica nelle poche disposizioni presenti nell'IA ACT. Inoltre, e questo vale per l'intero regolamento, il legislatore europeo ha prediletto una regolamentazione eccessivamente burocratizza che, a parere di scrive, mal si concilia con le esigenze di mercato e con i tempi di sviluppo delle tecnologie.

Ciò, invero, non stupisce: il rapporto tra il diritto e la tecnologia è storicamente "difficile", poiché la *lex digitalis* influenza il comportamento umano in quanto consente l'interazione con gli individui. Come rileva Sartor, «*la tecnologia condiziona ogni altra forma di regolazione, compresa quella giuridica, e l'uomo, per mezzo dello strumento del diritto, deve essere capace di governarla, raggiungendo un difficile equilibrio, idoneo a non limitare l'evoluzione tecnica e, allo stesso tempo, capace di non determinare la prevalenza della tecnologia sulla regolazione giuridica*».

Le tradizionali regole di legiferazione si sono mostrate, spesso, inappaganti: è avvenuto con l'*e-commerce*, con gli *smart contract* e in generale con tutto ciò che ha una dimensione immateriale[54]; il *deep-fake* non è da meno.

---

50 *Ibidem.*

51 M. S. RANA, B. MURALI and A. H. SUNG, "*Deepfake Detection Using Machine Learning Algorithms*," 2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI), Niigata, Japan, 2021, pp. 458-463, doi: 10.1109/IIAI-AAI53430.2021.00079.

52 *Ibidem.*

53 TAMPUBOLON, M., (2024), *Digital Face Forgery and the Role of Digital Forensics. Int J Semiot Law* **37**, 753-767; Paul Joseph, D., and Norman, and Jasmine(2019), An analysis of digital forensics in cyber security. In *First International Conference on Artificial Intelligence and Cognitive Computing: AICC 2018*, 701-708. Singapore: Springer.

54 O. POLLICINO, G. CAMERA, *La legge è uguale anche sul web. Dietro le quinte del caso Google-Vividown*, EGEA, 2010, pp. 2 e ss.

Se la legge volesse offrire un contributo effettivo nella lotta ai falsi artificiali sarebbe necessario tracciare il contenuto dei dati (così come previsto nella legge cinese), sebbene questi comporti una compromissione delle libertà individuali.

Né possono sottacersi le difficoltà che la stessa informatica riscontra nell'individuare dei falsi profondi. Al di là delle sperimentazioni tra gli addetti ai lavori che, proprio per l'alto grado di tecnicismo non possono essere comprese dall'utente medio di *internet*, è necessario individuare alcuni indici che siano in grado di far comprendere all'individuo la possibilità di trovarsi dinanzi a contenuti non originali e disinformativi. A tal fine, il 1° agosto 2024 Google ha annunciato un nuovo aggiornamento al suo algoritmo per contrastare i *deep-fake*, con l'obiettivo di salvaguardare le aziende e le persone dalle frodi e dal danno reputazionale. L'intento è quello di ridurre la visibilità dei contenuti manipolati sul motore di ricerca e rendere maggiormente individuabili i contenuti veritieri. Per tale ragione è stata implementata la funzione della ricerca di *Google Lens* che fornisce informazioni sulle immagini, quali l'indicizzazione e la divulgazione sui siti. Ciò consente di identificare rapidamente i contenuti manipolati e a prendere decisioni informate sulla loro veridicità.

Ciò dimostra, ancora una volta, che la tecnologia detta i tempi e le regole, al di là della presenza o meno di una prescrizione normativa.

Al netto delle difficoltà di accertamento di cui si è dato conto nel corso della presente trattazione, a parere di chi scrive sembrerebbe che solo l'intelligenza artificiale potrebbe essere in grado di smascherare l'intelligenza artificiale. Il mezzo è lo stesso, ma il fine è diverso e la differenza non è di secondaria importanza. Probabilmente è la massima rappresentazione del fatto che, se ben utilizzata, l'IA può contribuire nel mantenimento dei principi della società moderna. Richiede, però, ingenti sforzi — anche economici — perché l'IA muta velocemente e le "contromisure" devono svilupparsi in parallelo. Ad oggi, le sperimentazioni portate avanti dagli scienziati, per come si è visto, non sono esenti da criticità ma ciò non significa che nel prossimo futuro non verranno sviluppati nuovi modelli di IA in grado di individuate i contenuti generati artificialmente.

Nel mentre, però, sarebbe quantomeno necessario che si aumentare la conoscenza informatica della società civile, fornendo gli strumenti di base per sviluppare una maggiore consapevolezza sui rischi delle nuove tecnologie ed evitare che l'uomo rimanga inerme dinanzi al governo della tecnologia.

Fonti e documenti consultati

Regolamento Europeo (UE) 2024/1689 del 13 giugno 2024;

Bill Draft (RIL24710 NP8) "*To require transparency with respect to content and content provenance information, to protect artistic content, and for other purposes*";

Bill Draft (EHF24841 1M0) "*To protect intellectual property rights in the voice and visual likeness of individuals, and for other purposes*";

Il testo della legge cinese 互联网信息服务深度合成管理规定 (Regulations on deep synthesis management of Internet information services) accessibile al seguente link https://perma.cc/JE3W-PF26