# Una nuova tecnica (anche) per veicolare disinformazione: le risposte europee ai *deepfakes**

Martina Cazzaniga


## Abstract

Il contributo mira ad esaminare il "*deepfake*", un'innovativa tecnica basata sull'Intelligenza Artificiale che può essere utilizzata anche per veicolare disinformazione. Dopo alcune considerazioni preliminari di carattere costituzionale, si procede con un'analisi della posizione assunta dall'Unione europea in materia di *deepfake*, approfondendo quanto previsto all'interno dello *Strengthened Code of Practice on Disinformation* e della proposta di Regolamento sull'Intelligenza Artificiale. Infine, si cerca di riflettere sulle implicazioni sul piano costituzionale di una regolamentazione del *deepfake*.

The paper aims to examine the so-called "deepfake", an innovative technique based on Artificial Intelligence that can be used to convey disinformation. After some preliminary constitutional considerations, the paper focuses on the European Union approach to deepfakes by analyzing the Strengthened Code of Practice on Disinformation and the proposed Regulation on Artificial Intelligence. Finally, it presents a reflection about the constitutional implications of a deepfake regulation.


## Sommario




## Keywords

*deepfake* – disinformazione – falso - Intelligenza Artificiale - Unione europea


* L'articolo è stato sottoposto, in conformità al regolamento della Rivista, a referaggio "a doppio cieco".

## 1. Cosa sono i *deepfakes*: la manipolazione alla portata di tutti, anche a scopo disinformativo

Nel marzo 2023, Donald Trump ed Emmanuel Macron sono i protagonisti di alcune immagini condivise sul *web* che in pochissimo tempo sono state visualizzate da milioni di utenti[1]. I fatti che riportano destano evidente stupore: nella prima serie di scatti fotografici, l'ex Presidente degli Stati Uniti appare in balìa di alcuni poliziotti nell'atto di arrestarlo, mentre nella seconda il Presidente francese si trova in strada, tra manifestanti e forze dell'ordine, durante le proteste organizzate contro la riforma del sistema pensionistico[2]. Nonostante il realismo che le caratterizza, queste immagini sono false. Divulgare informazioni fuorvianti o menzognere non è mai stato così facile, dato che è possibile veicolare disinformazione ricorrendo a molteplici tecniche: è proprio su una di queste, il "*deepfake*", utilizzata per creare le immagini false di Trump e Macron[3], che si intende concentrare l'attenzione.

La divulgazione di tali contenuti manipolati ha iniziato a "smuovere le coscienze", aprendo una discussione circa l'utilizzo del *deepfake* con finalità disinformative. Tuttavia, prima di affrontare questo tema è necessario fare "un passo indietro" per comprendere almeno a grandi linee questa nuova tecnologia.

La parola inglese "*deepfake*" nasce nel 2017 e deriva dal *nickname* di un utente di Reddit conosciuto su questo sito *web* per la sua "abilità" di produrre video pornografici del tutto "originali": questi filmati venivano creati "ritagliando" i volti degli attori che veramente avevano recitato per la realizzazione del video e sostituendoli con i volti di vari personaggi – soprattutto donne – di fama mondiale. Il termine *deepfake*, il quale fino a qualche anno fa indicava semplicemente il *nickname* scelto da questo utente, è oggi utilizzato per riferirsi a questa particolare tecnica di intelligenza artificiale che permette di creare contenuti falsi e manipolati.

Chiunque provasse a digitare questa parola nella barra di ricerca di un qualsiasi *browser*, si accorgerebbe facilmente che sono sempre più numerose le notizie, provenienti da tutto il mondo, che riguardano il *deepfake*. Infatti, anche se in molti probabilmente non sono al corrente dell'esistenza di questa particolare tecnica, la sua diffusione è sempre più ampia: secondo i dati riportati da *Sensity*[4] nel dicembre 2020 i *deepfakes* presenti *online* erano già più di ottantacinquemila[5] e nell'ultimo biennio si è registrata una crescita esponenziale.

Ma che cosa è un *deepfake*? Nel dicembre 2020, il Garante per la Protezione dei Dati Personali ha fornito la seguente definizione della tecnologia del *deepfake*: «I *deepfake* sono foto, video e audio creati grazie a *software* di intelligenza artificiale (IA) che, partendo da contenuti reali (immagini e audio), riescono a modificare o ricreare, in modo

[1] E. Franceschini, *Trump in manette e Macron tra i manifestanti: quelle foto false dei leader viste da milioni di utenti*, in *La Repubblica*, 23 marzo 2023.

[2] *Ibidem*.

[3] *L'arresto di Trump simulato dall'intelligenza artificiale: è bufera sulla foto fake*, in *La Stampa*, 23 marzo 2023,

[4] Si tratta di una società con sede ad Amsterdam che si occupa anche di rilevare e tracciare i *deepfakes online*. Cfr. Europol Innovation Lab, *Facing reality? Law enforcement and the challenge of deepfakes*, 2022, 11.

[5] *How to Detect a Deepfake Online: Image Forensics and Analysis of Deepfake Videos*, in *Sensity*, 2021.

estremamente realistico, le caratteristiche e i movimenti di un volto o di un corpo e a imitare fedelmente una determinata voce»[6].

La parola è composta da due diversi concetti: difatti, si tratta di un neologismo[7] formato dall'unione del termine *fake*, che nella lingua inglese significa comunemente "falso", e *deep*, il quale fa invece riferimento ad una particolare tecnologia di intelligenza artificiale[8]. Infatti, con queste tecniche[9] ultra-innovative è possibile manipolare audio, immagini e video autentici[10], al fine di creare dei contenuti multimediali falsi e distorti[11]: bastano alcune immagini e una connessione Internet per "incollare" il viso di una persona al corpo di un'altra e farle dire e fare ciò che, nella realtà, non ha mai detto o fatto.

Il diffondersi di tecnologie in grado di imitare in modo così fedele la realtà non deve stupire: secondo quanto riportato dalla Commissione per le libertà civili, la giustizia e gli affari interni presso il Parlamento europeo, in futuro sarà possibile creare falsificazioni estremamente realistiche grazie a tecnologie sempre più sofisticate, delle quali ne

---

[6] Garante per la protezione dei dati personali, *Deepfake. Il falso che ti «ruba» la faccia e la privacy*, 28 dicembre 2020.

[7] *Ibidem*.

[8] Il termine "*deep*" a sua volta rimanda al concetto di *Deep Learning*. Il *Deep Learning*, la cui traduzione letterale è "apprendimento profondo", costituisce una sottocategoria del *Machine Learning* (letteralmente "apprendimento automatico") e indica quella branca dell'intelligenza artificiale che fa riferimento agli algoritmi ispirati alla struttura e alla funzione del cervello, chiamati "reti neurali artificiali". Cfr. N. Boldrini, *Deep Learning, cos'è l'apprendimento profondo, come funziona e quali sono i casi di applicazione*, in *Ai4business.it*, 26 settembre 2022.

[9] Le tecnologie di *deep learning* principalmente utilizzate per la creazione di *deepfakes* sono due. La prima è quella a cui i creatori di *deepfakes* ricorrono più spesso: essa prende il nome di *Generative Adversarial Networks* (meglio conosciuta con l'acronimo GAN), cioè il sistema con il quale è possibile realizzare le sostituzioni facciali. Il funzionamento della tecnologia GAN può essere riassunto nel modo che segue: un primo algoritmo è in grado di individuare i frammenti in cui i due soggetti (quello che prende il posto dell'originale e quello che viene sostituito con il volto altrui) hanno espressioni simili; a questo punto interviene un secondo algoritmo che svolge il successivo passaggio di posizionamento facciale, ovvero sovrappone concretamente i due volti in questione. Sostanzialmente, questi algoritmi di apprendimento automatico analizzano il materiale multimediale a disposizione e sono in grado di crearne uno altrettanto di qualità paragonabile. La seconda tecnologia prende invece il nome di *Autoencoders*: si tratta di un tipo di rete neurale che è in grado di estrarre informazioni sulle caratteristiche facciali apprese da immagini e utilizzarle per crearne delle altre con espressioni diverse. Cfr. F. Bertoni, *Deepfake, ovvero Manipola et impera. Un'analisi sulle cause, gli effetti e gli strumenti per la sicurezza nazionale, nell'ambito dell'utilizzo malevolo dell'intelligenza artificiale ai fini di disinformazione e propaganda*, in *Ciberspazio e diritto*, 20, 62, 2019, 15-16; M. van Huijstee et al., *Tackling deepfakes in European policy*, European Parliamentary Research Service, 2021, 7-8.

[10] Oltre ai filmati, i quali costituiscono sicuramente la forma più nota di *deepfakes*, possono essere diffusi anche *deepfakes* realizzati con tecnologie di c.d. *voice cloning*, le quali permettono di ricreare la voce umana (in questo caso il *deepfake* consiste in una *clip* audio); in alcuni casi invece non viene realizzato un filmato ma una immagine statica, ovvero le c.d. *deepfake images* (ad esempio per creare immagini false da usare come foto del profilo dei *social network*); infine, sono anche da menzionare le tecnologie che permettono di realizzare delle sintesi testuali dato che può essere utile creare anche dei testi falsi in grado di imitare il modo pressoché unico che il *target* ha di parlare e scrivere, considerando anche il suo vocabolario. Su questo tema cfr. T. C. Helmus, *Artificial intelligence, deepfakes, and disinformation. A primer*, in *Rand.org*, 2022, 4-6; M. van Huijstee et al., *Tackling deepfakes in European policy,* cit., 7-15.

[11] Commissione europea, *Technology and democracy. Understanding the influence of online technologies on political behaviour and decision-making*, JRC Science for policy report, agosto 2020, 111.

è un esempio anche la realtà aumentata[12].

Il *deepfake* trova le sue origini proprio nel contesto della pornografia non consensuale[13]: spesso questi materiali sono utilizzati per compiere atti di *revenge porn* ai danni di soggetti fragili – principalmente di sesso femminile – o anche come mezzo di *sextortion*. Ancora oggi, oltre il 95% dei *deepfakes* presenti *online* è costituito proprio da video non consensuali a sfondo pornografico[14].

Sulla base di questi dati, non sorprende il fatto che i *deepfakes* vengano spesso presentati innanzitutto – o esclusivamente – come una nuova "minaccia" da limitare il più possibile. Ad alimentare i timori legati al *deepfake*, si aggiunge il fatto che questi contenuti multimediali manipolati possono essere altresì utilizzati per la commissione di reati (quali, ad esempio, estorsioni, diffamazioni, furti d'identità, frodi informatiche), e, inoltre, possono costituire una vera e propria "nuova frontiera" della disinformazione in grado di mettere a rischio l'assetto democratico[15].

Dunque, il tema del *deepfake* interseca anche quello della disinformazione *online*[16]: con l'avvento di Internet e dei *social network*, la proliferazione di notizie false è notevolmente aumentata, tanto che la disinformazione costituisce oggi un fenomeno dalle molteplici (e quasi infinite) sfaccettature, anche considerato che il suo impatto dipende da tecnologie interessate da celeri processi evolutivi[17], proprio come nel caso dei *deepfakes*. Infatti, sarebbe paradossale pensare che il modo di disinformare possa rimanere il medesimo per sempre: così come era diverso in passato[18] rispetto ad oggi, lo stesso varrà per il futuro.

Perché, dunque, concentrare l'attenzione sui *deepfakes* disinformativi? Quali effetti potrebbero conseguire dall'impiego di questa tecnologia come mezzo di disinformazione, ad esempio in ambito politico? I *deepfakes* raffiguranti Trump e Macron sono la dimostrazione di come questa tecnica possa essere impiegata per diffondere notizie

---

[12] J. Bayer et al., *Disinformazione e propaganda – impatto sul funzionamento dello Stato di diritto nell'Ue e nei suoi Stati membri*, Commissione LIBE presso il Parlamento europeo, marzo 2019, 6.

[13] F. Bertoni, *Deepfake, ovvero Manipola et impera. Un'analisi sulle cause, gli effetti e gli strumenti per la sicurezza nazionale, nell'ambito dell'utilizzo malevolo dell'intelligenza artificiale ai fini di disinformazione e propaganda*, cit., 13.

[14] Europol Innovation Lab, *Facing reality? Law enforcement and the challenge of deepfakes*, 11.

[15] I rischi associati ai *deepfakes* sono stati approfonditi anche nello studio dell'*European Parliamentary Research Service* dedicato a questa tecnologia. Cfr. M. van Huijstee et al., *Tackling deepfakes in European policy*, 29-35.

[16] Quello della disinformazione è un tema di estrema attualità. Il concetto di *fake news* raggiunge il suo apice di "popolarità" nel 2016, anno della Brexit ma anche delle elezioni presidenziali negli Stati Uniti. Da quel momento, il fenomeno è oggetto di una costante attenzione, tanto che alcuni studiosi hanno dedicato al tema interessanti approfondimenti. Cfr., *ex multis*, C. Hassan-C. Pinelli, *Disinformazione e democrazia. Populismo, rete e regolazione*, Venezia, 2022; S. Sassi, *Disinformazione contro costituzionalismo*, Napoli, 2021; A. Nicita, *Il mercato delle verità,* Bologna, 2021.

[17] P. Cesarini, *The Digital Services Act: a silver bullet to fight disinformation?*, in *MediaLaws.eu*, 2021. L'autore del contributo spiega che «[…] *disinformation is a multi-faceted phenomenon whose impact depends on fast-evolving technologies, service-specific vulnerabilities and constant shifts in manipulative tactics*».

[18] La disinformazione non è sicuramente un fenomeno nato a seguito della rivoluzione digitale, dato che episodi di strategica disinformazione risalgono ad epoche anche molto antiche. Per alcuni esempi cfr. C. Valditara, *Fake news: regolamentazione e rimedi*, in *Diritto dell'informazione e dell'informatica*, 2, 2021, 257 ss; C. Pinelli, *Disinformazione, comunità virtuali e democrazia: un inquadramento costituzionale*, in *Diritto pubblico*, 1, 2022, 173 ss.

non veritiere ma anche per finalità satiriche, specie nei casi, come quelli sopra descritti, ove il lettore medio comprende agevolmente che si tratta di un falso.
Ciò premesso, è evidente che *deepfakes* "politici", diffusi specie durante il periodo elettorale, possono costituire uno strumento di distorsione del dibattito pubblico e di manipolazione dei meccanismi di formazione del consenso[19].
La condivisione di un filmato compromettente (in realtà falso), probabilmente più convincente di una notizia solo letta[20], specie se divulgato a ridosso della fine della campagna elettorale, potrebbe incidere sull'esito delle elezioni, soprattutto qualora il soggetto leso non abbia, il tempo per smentirlo efficacemente[21].
Un ulteriore esempio concreto circa l'uso del *deepfake* come mezzo di disinformazione – anche se nel contesto bellico[22] – riguarda lo scontro militare che vede contrapposte la Federazione Russa e l'Ucraina a seguito dell'invasione di quest'ultima da parte del Cremlino avvenuta il 24 febbraio 2022. Il conflitto in corso è noto anche per essere a tutti gli effetti una guerra "mediatica", seppur non la prima nella storia[23], dato che i *social media* sono divenuti luogo privilegiato per la diffusione di informazioni riguardanti le ostilità. Inoltre, in questo caso la disinformazione deve essere considerata una componente essenziale dello scontro in atto: la diffusione di notizie false e fuorvianti è divenuta una vera e propria strategia per combattere il nemico, anche ricorrendo alla tecnica del *deepfake*.
Infatti, dopo circa un mese dall'inizio delle ostilità, divenne virale un video nel quale l'attuale Presidente ucraino Volodymyr Zelens'kyj dichiara a gran voce la resa del popolo ucraino e ordina ai propri soldati di deporre le armi[24]. Il video falso, realizzato appunto tramite la tecnica del *deepfake*, non venne creduto come vero[25], considerato anche il fatto che poco dopo Zelens'kyj smentì ufficialmente la notizia in prima persona[26]. E, in questo caso, i principali *social network*, i quali normalmente intervengono con

[19] M. Pawelec, *Deefakes and democracy (theory): how synthetic audio-visual media for disinformation and hate speech threaten core democratic functions*, in *Digital Society*, 1:19, 2022, 15.

[20] T. Dobber-N. Helberger-N. Metoui-D. Trilling-C. de Vreese, *Do (microtargeted) deepfakes have real effects on policital attitudes?*, in *International Journal of Press/Politics*, 26, 2021.

[21] B. Chesney-D. Citron, *Deep fakes: a looming challenge for privacy, democracy and national* security, in *California Law Review*, 107, 2019, 1178.

[22] Oltreché in ambito politico, la disinformazione può avere impatti non indifferenti sull'andamento degli scontri militari. Per un interessante approfondimento sulla disinformazione e sui *deepfakes* nei periodi bellici cfr. M. Mezzanotte, *Fake news, deepfake e sovranità digitale nei periodi bellici*, in *Federalismi.it*, 33, 2022.

[23] P. Suciu, *Is Russia's invasion of Ukraine the first social media war?*, in *Forbes*, 1 marzo 2022; K. Tiffany, *The myth of the first "Tik tok war"*, in *The Atlantic*, 10 marzo 2022.

[24] S. Burgess, *Ukraine war: deepfake video of Zelensky telling Ukrainians to "lay down arms" debunked*, in *SkyNews*, 17 marzo 2022.

[25] In realtà, soffermandosi più attentamente sul video, si può notare che lo stesso non fu realizzato con una precisione tale da renderlo particolarmente credibile: la testa di Zelens'kyj è troppo grande e sproporzionata rispetto al resto del corpo, quest'ultimo ha una risoluzione minore e appare più sgranato in rapporto al viso e, infine, la voce risulta essere più "profonda" rispetto all'originale. Cfr. J. Wakefield, *Deepfake presidents used in Russia-Ukraine war*, in *BBC News*, 18 marzo 2022.

[26] *Ibidem*. Secondo quanto riportato da BBC, il Presidente ucraino definì questa mossa una «*childish provocation*».

l'eliminazione del contenuto falso solo in casi circoscritti[27], procedettero prontamente alla rimozione del video[28].

Tra l'altro, poco tempo dopo non tardò ad arrivare il "contrattacco" ucraino: infatti, venne diffuso un video del Presidente russo Vladimir Putin mentre afferma di aver raggiunto un accordo di pace con lo Stato ucraino[29]. Anche in questo caso le principali piattaforme indicarono il contenuto come manipolato.

Tali esempi aiutano a comprendere le capacità di alterazione dell'opinione pubblica di questi video manipolati. Anche se finora i casi più eclatanti di *deepfakes* politici hanno coinvolto più che altro paesi stranieri[30], si tratta di una «forma di manipolazione dell'informazione [che] è destinata a verificarsi anche sul nostro territorio»[31] o che comunque non è escluso che possa verificarsi. Tra l'altro, Federica Bertoni sottolinea anche il fatto che quando in passato il legislatore italiano è intervenuto in ambito tecnologico, l'ha spesso fatto con normative d'urgenza e basate su una sorta di «onda emozionale»[32]: per questo motivo, ragionare in anticipo sul tema permette probabilmente di formulare delle valutazioni che sperano di essere più equilibrate.

## 2. Una riflessione circa la compatibilità costituzionale di un divieto assoluto di creazione e diffusione di *deepfakes*

In un futuro prossimo, la possibilità di creare *deepfakes* realistici potrebbe davvero divenire alla portata di tutti a causa del processo di «*democratization of deepfakes*»[33], in realtà già in atto da qualche anno: infatti, molti strumenti sono disponibili presso il pubblico per la realizzazione di *deepfakes*, tecnica ormai non più accessibile solo ad esperti informatici. Di conseguenza, non è irrealistico prospettare che chiunque potrebbe in tempi brevi divenire capace di creare questo tipo di contenuti, anche attraverso l'uso di

[27] Facendo riferimento ad esempio a Facebook, la piattaforma rimuove i contenuti falsi solo quando questi violano le proprie *policies*; in generale per le notizie false al posto della rimozione si prediligono misure più *soft* quali la riduzione della distribuzione del contenuto oppure l'informazione viene corredata di un contesto più completo per permettere agli utenti di decidere in modo autonomo quali contenuti ritenere affidabili. Tuttavia, uno dei casi che giustifica la rimozione è proprio quello dei contenuti multimediali manipolati nel caso in cui la manipolazione non sia evidente. Cfr. *facebook.com/combating-misinfo*; *transparency.fb.com/it-it/policies/community-standards/misinformation/*.

[28] *Meta e YouTube rimuovono video "deepfake" Zelensky*, in *Ansa.it*, 25 marzo 2022.

[29] *Il deepfake di Putin che dichiara la pace con l'Ucraina*, in *La Stampa*, 18 marzo 2022.

[30] Oltre ai *deepfakes* di Trump e Macron richiamati in apertura, ad oggi altri casi eclatanti hanno riguardo Stati come il Gabon, la Malesia e l'India. Cfr., rispettivamente, S. Cahlan, *How Misinformation Helped Spark an Attempted Coup in Gabon*, in *The Washington Post*, 13 febbraio 2020; H. Ajder-F. Cavalli-L. Cullen-G. Patrini, *The state of deepfakes: landscape, threats and impact*, Deeptrace lab, 2019, 10; C. Jee, *An Indian politiciain is using deepfake technology to win new voters*, in *MIT Technology Review*, 19 febbraio 2020.

[31] F. Bertoni, *Deepfake, ovvero Manipola et impera. Un'analisi sulle cause, gli effetti e gli strumenti per la sicurezza nazionale, nell'ambito dell'utilizzo malevolo dell'intelligenza artificiale ai fini di disinformazione e propaganda*, cit., 23.

[32] *Ibidem*. Federica Bertoni si riferisce soprattutto agli interventi volti a disciplinare fenomeni come il *cyberbullismo* ma anche l'utilizzo di strumenti come i captatori informatici.

[33] P. Jurcys-J. Kalpokiene-A. Liaudanskas-E. Meskys, *Regulating deep fakes: legal ad ethical considerations*, in *Journal of Intellectual Property Law and Practice*, 15(1), 2020, 24.

semplici applicazioni[34]: ad esempio, le immagini *deepfake* di Trump in arresto non sono state create da un abile informatico ma da un giornalista[35].

Non è un caso che alcuni ordinamenti stanno intervenendo per disciplinare l'utilizzo di questa nuova tecnica, muovendo così i primi passi nella regolamentazione del settore. Proprio in quest'ottica è interessante svolgere alcune riflessioni preliminari di carattere costituzionale.

Il primo tema che si pone è sicuramente quello del falso. Ogni *deepfake*, infatti, è oggettivamente un falso: ciascun contenuto creato attraverso questa nuova tecnologia risulta essere intrinsecamente simulatore, dato che si tratta di video (nella maggior parte dei casi) le cui immagini e i cui suoni sono rielaborati e adattati ad un contesto differente rispetto a quello originario e autentico.

Dunque, i *deepfakes* devono essere ritenuti in sé illeciti perché falsi? L'interrogativo rimanda, nel contesto italiano, alla ben nota questione circa la protezione costituzionale del discorso falso e, sul punto, ritengo di poter aderire a quell'indirizzo secondo cui il c.d. subiettivamente falso, pur non rientrando nella tutela costituzionale della libertà di parola[36], non debba essere considerato illecito in quanto tale[37] ma solo qualora determini la lesione di beni giuridici di rilievo costituzionale[38].

Una tutela ancora più ampia del discorso falso è riconosciuta dalla giurisprudenza della Corte Suprema degli Stati Uniti d'America, la quale ha ricondotto all'ambito di applicazione del Primo Emendamento anche le affermazioni false[39], persino quando

[34] Le più famose sono *Reface*, *FaceApp*, *ZaoApp* e anche *Fakeyou*. Nei confronti di quest'ultima il Garante per la Protezione dei Dati Personali ha aperto una istruttoria nell'ottobre 2022. Cfr. *Deepfake: Garante avvia istruttoria su app che falsifica le voci*, 12 ottobre 2022.

[35] E. Franceschini, *Trump in manette e Macron tra i manifestanti: quelle foto false dei leader viste da milioni di utenti*, cit. L'Autore dell'articolo spiega: «Talvolta c'è qualche dettaglio rivelatore che si tratta di una finzione. In una delle immagini su Trump arrestato dalla polizia newyorchese, l'ex-presidente sembra avere tre gambe. In maniera simile, una precedente immagine diventata virale sul web, quella di un poliziotto francese che abbracciava una manifestante, l'indizio era la mano guantata dell'agente, che aveva sei dita. Ma in altre foto tutto appare veritiero e reale».

[36] Secondo Carlo Esposito, dato che la Costituzione italiana tutela le espressioni del «proprio» pensiero, non si dovrebbero considerare riconducibili all'art. 21 le informazioni che l'autore sa essere false. Cfr. C. Esposito, *La libertà di manifestazione del pensiero nell'ordinamento italiano*, in *Rivista italiana per le scienze giuridiche*, IX, serie III anni 1957-1958, 84-85.

[37] Secondo la lezione di Paolo Barile, «neppure la diffusione di notizie false può essere considerata illecita in sé e per sé»: anche se si dovesse propendere per una non riconducibilità dell'affermazione subiettivamente falsa tra il novero delle manifestazioni del pensiero costituzionalmente tutelate, questo non comporterebbe automaticamente la sua illiceità. P. Barile, *Diritti dell'uomo e libertà fondamentali*, Bologna, 1984, 229; in questo senso anche A. Pace-M. Manetti, *Commentario della Costituzione. Rapporti civili. Art. 21. La libertà di manifestazione del proprio pensiero,* Bologna, 2006, 89; M. Bassini-G.E. Vigevani, *Primi appunti su fake news e dintorni*, in questa *Rivista*, 1, 2017, 21.

[38] Così, tra i molti, M. Bassini-G.E. Vigevani, *op. ult. cit.*, 15, dove si fa riferimento alla c.d. «teoria del bene giuridico costituzionalmente protetto». Sul tema, cfr. anche C. Melzi d'Eril, *Fake news e responsabilità: paradigmi classici e tendenze incriminatrici*, in questa *Rivista*, 1, 2017, 64.

[39] La sentenza cardine in tema di *false speech* nel contesto statunitense è sicuramente il caso, anche relativamente recente, *United States v. Alvarez*, nel quale la Corte si pronunciò sulla conformità del c.d. *Stolen Valor Act* rispetto al Primo Emendamento escludendone la legittimità costituzionale: l'atto non venne considerato conforme a Costituzione poiché puniva la dichiarazione falsa in quanto tale, senza dare nessuna importanza alla finalità del falso e all'eventuale perseguimento di un interesse o profitto. Cfr. *United States v. Alvarez*, 567 U.S. 709 (2012); O. Pollicino, *La prospettiva costituzionale sulla libertà di espressione nell'era di Internet*, in questa *Rivista*, 1, 2018, 75.

l'autore ne sia consapevole.
In questa prospettiva si può ritenere che i *deepfakes* non possano essere vietati a priori ma possano essere ritenuti illeciti solo sulla base di un giudizio *ex post*, in caso di lesione di altri interessi di rilevanza costituzionale.
Tuttavia, proprio perché i *deepfakes* possono causare dei danni sia alla persona (anche dal punto di vista psicologico) che alla società (soprattutto se utilizzati come mezzo di disinformazione) è comunque opportuno interrogarsi sull'introduzione di specifici limiti in grado di circoscrivere la possibilità del concretizzarsi di questi rischi o, comunque, di fornire un adeguato rimedio qualora questi si verifichino.
Bisogna però precisare che ogni intervento regolatorio in materia necessita di tenere in considerazione il fatto che la tecnica del *deepfake* può avere – e concretamente ha, anche se questo aspetto fa molto meno "rumore" – delle applicazioni positive[40], ovviamente a patto che la stessa venga utilizzata nel modo corretto. Richiamare questo aspetto è importante ai fini di un ragionamento giuridico, perché nel momento in cui si prospetta una normativa in materia bisognerà fare attenzione a non ostacolare i risvolti positivi del *deepfake*[41], i quali non solo sono leciti ma altresì incoraggiabili.
Per citare alcuni casi di "usi benefici" del *deepfake*, si può richiamare il possibile impiego nel settore cinematografico[42], pubblicitario e della moda[43], nonché in quello educativo[44] e anche per la realizzazione di opere d'arte[45]: se utilizzato in questo senso, il *deepfake* può essere considerato a tutti gli effetti uno strumento per esprimere la propria creatività[46]. Inoltre, questa tecnologia può anche essere utilizzata per fare satira, parodia o comunque per creare altri tipi di contenuti ricompresi nell'alveo di protezione della libertà di manifestazione del pensiero[47]. Un ulteriore ambito fondamentale in cui l'utilizzo del *deepfake* può davvero agevolare molto la ricerca è quello medico[48].
Proprio perché sono molti i settori in cui il *deepfake* può essere utilizzato con finalità meritevoli, prospettare un divieto assoluto e indiscriminato di ricorrere a questa tecnologia pare sproporzionato oltreché insostenibile sul piano costituzionale, anche perché di per sé la manipolazione informatica non può essere considerata un'attività illecita in quanto tale[49]. Una proibizione totale non solo limiterebbe eccessivamente la libera manifestazione del pensiero ma, più in generale, la sperimentazione in molti ambiti, da

[40] Europol Innovation Lab, *Facing reality? Law enforcement and the challenge of deepfakes*, 5.

[41] M. Feeney, *Deepfake laws risk creating more problems than they solve*, rilasciato dal *Regulatory transparency project of the Federalist Society*, 1° marzo 2021, 5.

[42] B. Chesney-D. Citron, *Deep fakes: a looming challenge for privacy, democracy and national* security, cit., 1770.

[43] D. Yadav-S. Salmani, *Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network*, *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India, 2019, 854.

[44] B. Chesney-D. Citron, *Deep fakes,* cit., 1769.

[45] P. Jurcys-J. Kalpokiene-A. Liaudanskas-E. Meskys, *Regulating deep fakes: legal ad ethical considerations*, cit., 29.

[46] *Ibidem*.

[47] M. Feeney, *Deepfake laws risk creating more problems than they solve,* cit., 5.

[48] M. van Huijstee et al., *Tackling deepfakes in European policy*, cit., 28-29.

[49] B. Chesney-D. Citron, *Deep fakes,* cit., 1788.

quello artistico fino a quello medico[50].
Da queste considerazioni ne discende che l'operazione di bilanciamento tra i vari diritti e interessi sarà delicata e, di conseguenza, da svolgere con cautela.

## 3. Lo stato dell'arte della regolazione dei *deepfakes* nell'Unione europea: i primi passi tra *Strengthened Code of Practice on Disinformation* e la proposta di Regolamento sull'Intelligenza Artificiale

La diffusione di questa tecnica sta sollevando quesiti tanto giuridici quanto di tipo etico e sociale in molte aree del mondo: ci si domanda quale approccio debba assumere un ordinamento democratico, se sia necessario un intervento normativo e, inoltre, quale sarebbe eventualmente la disciplina più adatta[51].
L'ordinamento italiano non prevede alcuna disciplina in materia di *deepfakes*[52] – e, tra l'altro, nemmeno in materia di disinformazione più in generale[53] – probabilmente anche perché ad oggi casi clamorosi di *deepfakes* concernenti personaggi pubblici italiani non ve ne sono stati.
Se negli Stati Uniti il culmine di attenzione circa questo tema risale soprattutto al periodo che va dal 2018 fino circa al 2021[54], quando alcuni Stati hanno adottato normative sul tema, l'Unione europea, invece, sta muovendo i primi passi per affrontare i *deepfakes* proprio in questo ultimo triennio.

[50] Ivi, 1789.

[51] P. Jurcys-J. Kalpokiene-A. Liaudanskas-E. Meskys, *Regulating deep fakes: legal ad ethical considerations*, cit., 24.

[52] Se il legislatore italiano ancora non si è avvicinato a questo tema, lo ha fatto però in alcune occasioni il Garante per la Protezione dei Dati personali. Il primo caso riguarda la pubblicazione, a dicembre 2020, di un *vademecum* intitolato «*Deepfake*, il falso che ti "ruba" la faccia e la *privacy*», nel quale viene illustrata la tecnica, i principali rischi e alcune raccomandazioni su come proteggersi dai *deepfakes*. Il Garante per la Protezione dei Dati Personali torna però ad occuparsi di *deepfakes* più recentemente, a quasi due anni di distanza dalla pubblicazione del *vademecum*: infatti, con un comunicato stampa rilasciato a Roma il 12 ottobre 2022, annunciò l'avvio di una istruttoria nei confronti della società *The Storyteller Company*, proprietaria dell'applicazione *Fakeyou*, la quale permette di fare pronunciare *file* di testo a voci false – anche se realistiche – di numerosi personaggi famosi, compresi politici italiani. Cfr. Garante per la protezione dei dati personali, *Deepfake. Il falso che ti «ruba» la faccia e la privacy*, 28 dicembre 2020; *Deepfake: Garante avvia istruttoria su app che falsifica le voci*, 12 ottobre 2022.

[53] In passato si sono verificati due tentativi di intervento in materia, entrambi falliti nonché molto discutibili. Ci si riferisce a due proposte di legge, ambedue depositate nell'ultimo anno di lavori della XVII legislatura: la prima data 7 febbraio 2017 (Atto Senato n. 2688 comunicato alla Presidenza il 7 febbraio 2017 – *Disposizioni per prevenire la manipolazione dell'informazione online, garantire la trasparenza sul web e incentivare l'alfabetizzazione mediatica* – disegno di legge Gambaro e altri), la seconda, invece, risale al dicembre dello stesso anno ed è stata presentata dai Senatori Zanda e Filippin (Atto Senato n. 3001 comunicato alla Presidenza il 14 dicembre 2017 – *Norme generali in materia di social network e per il contrasto della diffusione su internet di contenuti illeciti e delle fake news* – disegno di legge Zanda-Filippin).

[54] A livello statale, prevedono una legislazione positiva in materia la Virginia, il Texas, la California, lo Stato di Washington, quello di New York e, infine, il Massachusetts. A livello federale si nota un approccio più cauto: le poche disposizioni vigenti, infatti, delineano solo obblighi di studio e ricerca in materia. Ovviamente sono stati depositati anche disegni di legge più incisivi, i quali, tuttavia, hanno trovato un'accoglienza assai critica.

L'Unione europea non ha adottato una normativa specifica di contrasto ai *deepfakes*[55]. Tuttavia, questo non è sintomo di uno scarso interesse delle istituzioni europee rispetto al fenomeno: al contrario, soprattutto il Parlamento europeo da anni tiene alta l'attenzione rispetto ai potenziali effetti negativi dei *deepfakes*, ovviamente non trascurando i benefici che il ricorso a questa tecnica può comportare. Infatti, in numerose raccomandazioni e risoluzioni del Parlamento europeo è stato fatto riferimento a questa problematica anche in relazione al tema della propaganda politica e della disinformazione[56].

Alcune previsioni relative ai *deepfakes* sono contenute nello *Strengthened Code of Practice on Disinformation* del 2022[57]: questo codice costituisce un atto di *soft-law* che prevede una serie di impegni e misure specifiche che i firmatari[58] si impegnano a rispettare al fine di limitare la circolazione di notizie false *online*. Nella sezione denominata *Integrity of services*[59] i *deepfakes* vengono richiamati come caso di *manipulative behaviours* da contrastare.

Infatti, il *commitment* numero 14[60] stabilisce che, allo scopo di limitare i comportamenti non consentiti, i sottoscrittori debbano mettere a punto – oppure a rafforzare se le prevedevano già prima della firma del codice – *policies* di contrasto alla disinformazione veicolata attraverso l'utilizzo di queste tecnologie di recente sviluppo.

Il 9 febbraio 2023, ovvero sei mesi dopo la firma del "codice rafforzato", sono stati pubblicati dal neoistituito Centro per la trasparenza i primi *report*[61] presentati dai firmatari, i quali illustrano le misure adottate da ciascuno di loro per rispettare gli impegni assunti.

---

55 M. Liu-X. Zhang, *Deepfake technology and current legal status of it*, in *Proceedings of the 2022 3rd International Conference on Artificial Intelligence and Education (IC-ICAIE 2022)*, 2022, 1311.

56 L'Istituzione da sempre più interessata al tema è indubbiamente il Parlamento europeo, il quale ancora prima di dedicare a questa tecnologia un apposito studio del Servizio di Ricerca nel 2021 (più volte citato) ha richiamato i *deepfakes* in plurime raccomandazioni e risoluzioni. Cfr., *ex multis*, Raccomandazione del Parlamento europeo del 13 marzo 2019 al Consiglio e al vicepresidente della Commissione/alto rappresentante dell'Unione per gli affari esteri e la politica di sicurezza sul bilancio del seguito dato dal SEAE a due anni dalla relazione del PE sulla comunicazione strategica dell'UE per contrastare la propaganda nei suoi confronti da parte di terzi, P8_TA(2019)0187, 13 marzo 2019; Risoluzione del Parlamento europeo del 12 febbraio 2019 su una politica industriale europea globale in materia di robotica e intelligenza artificiale, P8_TA (2019)0081, 12 febbraio 2019, Nr. 178; Risoluzione del Parlamento europeo del 20 gennaio 2021 sull'intelligenza artificiale: questioni relative all'interpretazione e applicazione del diritto internazionale nella misura in cui l'UE è interessata relativamente agli impieghi civili e militari e all'autorità dello Stato al di fuori dell'ambito della giustizia penale, PE653.860v02-00, 4 gennaio 2021.

57 Il testo completo dello *Strengthened Code of Practice on Disinformation* del 2022 è reperibile all'indirizzo https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

58 Nella versione del 2018 firmatari erano principalmente piattaforme digitali e motori di ricerca con un elevato numero di internauti (come Facebook, Twitter, Tik Tok, Google e Mozilla) insieme ad alcuni attori del settore pubblicitario. Oggi, invece, il *Code of practice on disinformation* conta trentaquattro firmatari, tra cui anche organizzazioni di *fact-cheking* indipendenti, associazioni di categoria e persino membri della società civile.

59 *Strengthened Code of Practice on Disinformation*, 15.

60 Ivi, 15-16.

61 Commissione europea, *I firmatari del codice di buone pratiche sulla disinformazione presentano le loro prime relazioni di riferimento presso il Centro per la trasparenza*, 9 febbraio 2023.

Quello che si può constatare da questi *report* è che in realtà le piattaforme non hanno modificato le normative interne in tema di contenuti multimediali manipolati dopo l'introduzione del codice. Ad esempio, Meta[62] richiama le proprie *policies*[63] e, nonostante non siano previste per il prossimo futuro modifiche sostanziali, dichiara di aver intrapreso uno studio insieme all'Università dello Stato del Michigan al fine di predisporre dei sistemi di *reverse engineering* in grado di rilevare i contenuti manipolati, *deepfakes* compresi[64].

Molto più scarne le parti dedicate alle misure di contrasto ai *deepfakes* all'interno dei *report* di Twitter e Google.

Per quanto riguarda Twitter[65], la disciplina in materia di contenuti multimediali manipolati viene completamente riportata nella relazione[66] quasi con un "*copy and paste*" della propria pagina *web* dedicata[67], senza effettivamente aggiungere altre indicazioni. Alle altre domande, ovvero quelle che interrogano il firmatario circa le misure che intende intraprendere in futuro, Twitter non fornisce alcuna risposta[68].

Google[69], invece, richiama in maniera più specifica le sole regole previste per YouTube, anche se in modo molto schematico ed approssimativo, rispondendo negativamente all'interrogativo circa la volontà di implementare gli sforzi in questo settore per il futuro[70].

La conclusione che si può trarre, per lo meno per quanto riguarda il tema dei *deepfakes*, è che l'adozione del codice non pare essere stata particolarmente influente e in grado di orientare in modo decisivo le scelte delle piattaforme in questo settore.

Anche la proposta di Regolamento sull'Intelligenza Artificiale tratta il tema dei *deepfakes* delineando una disciplina minimale ma comunque rilevante. Bisogna ricordare che i *deepfakes* vengono realizzati attraverso tecniche di Intelligenza Artificiale e proprio per questa ragione la proposta concernente questa materia non ne tralascia un riferimento[71]. La disciplina dei *deepfakes* è presentata in due "punti" diversi della proposta, ovvero nell'art. 52 e nell'Allegato III.

Esaminando l'art. 52, par. 3,[72] si può subito notare che nella proposta non si opta

---

62 *Code of Practice on Disinformation – Meta Baseline Report*, disponibile all'url *disinfocode.eu*.

63 Ivi, 55. Sia rispetto a Facebook che a Instagram, il *report* spiega che «*we remove videos under this policy if specific criteria are met: (1) the video has been edited or synthesized, beyond adjustments for clarity or quality, in ways that are not apparent to an average person, and would likely mislead an average person to believe a subject of the video said words that they did not say; and (2) the video is the product of artificial intelligence or machine learning*».

64 Ivi, 66.

65 *Code of Pratice on Disinformation – Report of Twitter for the period H2 2022*, disponibile in *disinfocode.eu*.

66 Ivi, 24-27.

67 Cfr. Twitter, *Norme sui contenuti multimediali artificiosi e manipolati*.

68 *Code of Pratice on Disinformation – Report of Twitter for the period H2 2022*, 27.

69 *Code of Practice on Disinformation – Report of Google for the period 1 July 2022 – 30 September 2022*, sempre disponibile in *disinfocode.eu*.

70 Ivi, 79-80.

71 M. van Huijstee et al., *Tackling deepfakes in European policy*, cit., 37.

72 Art. 52, par. 3: «Gli utenti di un sistema di IA che genera o manipola immagini o contenuti audio o video che assomigliano notevolmente a persone, oggetti, luoghi o altre entità o eventi esistenti e che potrebbero apparire falsamente autentici o veritieri per una persona ("*deep fake*") sono tenuti a rendere

per un generale divieto di creazione e divulgazione di *deepfakes*[73]: infatti, ciò che viene imposto a coloro che utilizzano i sistemi di Intelligenza Artificiale che generano *deepfakes* è "solamente" un generale obbligo di trasparenza, cioè di rendere palese che il contenuto multimediale è stato generato o manipolato attraverso queste tecnologie. Si tratta, dunque, di un dovere di tipo informativo. Il fatto che siano previsti unicamente degli obblighi di segnalazione è una conseguenza della non inclusione dei *deepfakes* nei sistemi a rischio inaccettabile o alto ma, invece, della loro riconducibilità in quelli a rischio limitato[74].

Rispetto alla scelta di prevedere un obbligo di questo tipo sorgono alcune perplessità, espresse anche dalla dottrina statunitense in relazione al c.d. *Deepfakes Accountability Act*[75], il quale delineava un dovere simile.

Innanzitutto, bisogna considerare che se il creatore del *deepfake* è disposto a far sapere che il contenuto che sta diffondendo è falso, ne discende che questo non sarà quasi sicuramente il tipo di utente di cui "preoccuparsi", dato che con tutta probabilità non avrà delle intenzioni nocive, proprio perché disposto a "smascherare" la manomissione[76]. Effettivamente, se chi realizza questi materiali alterati ha l'intenzione di recare danni ad un soggetto preciso o alla società, conoscerà sicuramente anche delle tecniche per non svelare la propria identità[77], optando per divulgare il contenuto tramite un *account* falso o un *bot*. Di conseguenza, non sarebbe poi facilmente identificabile un soggetto in capo a cui ricondurre la responsabilità. In definitiva, una previsione di questo tipo non sembra essere idonea a costituire un deterrente tale da demotivare gli autori di *deepfakes* nocivi[78].

L'art. 52, par. 3, delinea anche delle eccezioni a questo dovere di "etichettamento" del *deepfake* come tale, le quali corrispondono ai casi di esercizio della libertà di manifestazione del pensiero ma non solo, dato che viene fatto riferimento anche alla libertà delle arti e delle scienze e all'uso autorizzato dalla legge per accertare, prevenire, indagare e perseguire reati. La previsione di deroghe al generale obbligo appena analizzato è opportuna proprio per evitare di gravare eccessivamente i creatori di *deepfakes* legittimi.

Alla posizione che l'Unione europea intenderebbe assumere in tema di *deepfake* non sono state risparmiate critiche da parte della dottrina, dato che alcuni commentatori

noto che il contenuto è stato generato o manipolato artificialmente. Tuttavia, il primo comma non si applica se l'uso è autorizzato dalla legge per accertare, prevenire, indagare e perseguire reati o se è necessario per l'esercizio del diritto alla libertà di espressione e del diritto alla libertà delle arti e delle scienze garantito dalla Carta dei diritti fondamentali dell'UE, e fatte salve le tutele adeguate per i diritti e le libertà dei terzi».

[73] A. Del Ninno, *La proposta di Regolamento UE sull'Intelligenza Artificiale: i profili operativi del nuovo quadro normativo europeo – Parte Quarta*, in *Diritto e Giustizia*, 2021, 12.

[74] Camera dei deputati Ufficio Rapporti con l'Unione europea, *Legge sull'intelligenza artificiale*, dossier n. 57, 12 novembre 2021, 10.

[75] Questo disegno di legge venne presentato dalla componente della *House of Representatives* degli Stati Uniti Yvette Clarke nel 2021 ma risulta ancora al solo stato di "*introduced*". U. S. Congress H.R. 2395, *Deepfakes Accountability Act*, 2021.

[76] D. Coldewey, *Deepfakes Accountability Act wuold impose unenforceable rules – but it's a start*, in *TechCrunch*, 13 giugno 2019.

[77] M. van Huijstee et al., *Tackling deepfakes in European policy*, cit., 46.

[78] M. Feeney, *Deepfake laws risk creating more problems than they solve*, cit., 9.

hanno palesato più di un dubbio circa queste previsioni[79]. In generale, soprattutto chi vede i *deepfakes* più come una minaccia che come un beneficio, lamenta l'eccessiva "timidezza" di questa normativa[80].

Tra l'altro, la proposta nella sua originaria formulazione lasciava del tutto irrisolte alcune questioni, come quelle relative a come dovesse avvenire la *disclosure* delle informazioni richieste. Effettivamente, il testo dell'art. 52, par. 3, così come proposto non presentava alcuna indicazione circa la messa in pratica di questo dovere di trasparenza, con la conseguenza che la sua applicazione avrebbe potuto risultare concretamente complicata: l'assenza di una disciplina sul punto non permetteva di comprendere se fosse sufficiente una indicazione generale oppure se ne fosse necessaria una di carattere più specifico[81].

Alcuni profili critici vengono sottolineati anche nello studio dedicato al tema dei *deepfakes* condotto dallo *European Parliamentary Research Service*, nel quale vengono forniti alcuni "suggerimenti" su come migliorare la prospettata disciplina in materia[82].

Oltre che nell'art. 52, par. 3, i *deepfakes* sono richiamati anche all'interno dell'Allegato III[83] così come originariamente proposto, il quale elenca i sistemi ad alto rischio: al punto n. 6 lettera c) sono previsti anche «i sistemi di IA destinati a essere utilizzati dalle autorità di contrasto per individuare i "*deep fake*" di cui all'articolo 52, paragrafo 3». La ragione di questa scelta iniziale risiede nel fatto che si riteneva che i *detection softwares* potessero costituire un rischio per i diritti e le libertà dei singoli individui, decidendo di permetterne l'utilizzo solo nel rispetto di rigorose garanzie[84].

Tuttavia, le previsioni circa i *deepfakes* hanno subìto modifiche con gli emendamenti presentati dal Consiglio[85]: come già osservato, sono due le situazioni nelle quali il *deepfake* viene richiamato nella proposta ed entrambe sono state oggetto di emendamenti. Questo fa riflettere circa la rilevanza del tema, soprattutto se si considera che probabilmente le tecniche per realizzare i *deepfakes* saranno sempre più affinate negli anni a venire e che si tratta di un settore molto controverso.

---

79 D. Messina, *La proposta di regolamento europeo in materia di Intelligenza Artificiale: verso una "discutibile" tutela individuale di tipo consumer-centric nella società dominata dal "pensiero artificiale"*, in questa *Rivista*, 2, 2022, 220. Ad esempio, l'Autore sottolinea che i sistemi con i quali si realizzano i *deepfakes* non sembrano da considerare a rischio elevato, dato che assenti nell'elenco dell'Allegato III. Tuttavia, il par. 4 dell'art. 52 precisa che i paragrafi precedenti «lasciano impregiudicati i requisiti e gli obblighi di cui al Titolo III del presente regolamento», il quale riguarda i sistemi ad alto rischio.

80 A. Del Ninno, *La proposta di Regolamento UE sull'Intelligenza Artificiale: i profili operativi del nuovo quadro normativo europeo – Parte Quarta*, cit., 12.

81 B. van der Sloot-Y. Wagensveld, *Deepfakes: regulatory challenges for synthetic society*, in *Computer Law and Security Review*, 46, 2022, 7.

82 M. van Huijstee et al., *Tackling deepfakes in European policy*, cit., 59. Tra i suggerimenti forniti si possono richiamare i seguenti: chiarire più precisamente quali pratiche sono da considerare proibite, proposta avanzata anche dall'Europol (cfr. Europol Innovation Lab, *Facing reality? Law enforcement and the challenge of deepfakes*, cit., 22); prevedere obblighi non solo in capo ai creatori di *deepfakes* ma anche ai fornitori di queste tecnologie; classificare la tecnologia nella categoria dell'alto rischio.

83 Allegati della Proposta di Regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'Unione, SEC(2021) 167 final - SWD(2021) 84 final - SWD(2021) 85 final, 21 aprile 2021.

84 Europol Innovation Lab, *Facing reality? Law enforcement and the challenge of deepfakes*, cit., 21.

85 Per visionare il nuovo testo così come proposto dal Consiglio cfr. *data.consilium.europa.eu*.

Da un lato, è stato variato l'elenco dei sistemi di Intelligenza Artificiale ad alto rischio previsto dall'Allegato III: infatti, i sistemi impiegati dalle autorità per individuare i *deepfakes* sono stati rimossi da questo novero, scomparendo così ogni riferimento a questa tecnica tra i sistemi *high risk*, dato che viene completamente eliminata la lettera c) dal punto n. 6.
D'altro lato, anche l'art. 52 della proposta di regolamento ha subìto cambiamenti[86]. La prima parte del par. 3 rimane immutata, mentre sulla seconda il Consiglio è intervenuto proponendo una modifica dei casi in cui il creatore di *deepfakes* è sollevato dall'obbligo di etichettamento: viene rimosso il riferimento ai *deepfakes* che rientrano nell'esercizio della libertà della scienza, mentre sussiste l'esonero nel caso di opere e programmi manifestatamente creativi, satirici, artistici o fittizi e quando l'uso è autorizzato dalla legge per accertare, prevenire, indagare e perseguire reati.
Oltre a intervenire sul novero delle eccezioni, viene anche aggiunto un par. 3-*bis* all'art. 52, il quale prevede che «le informazioni di cui ai paragrafi da 1 a 3 sono fornite alle persone fisiche in maniera chiara e distinguibile al più tardi al momento della prima interazione o esposizione».
La disposizione appena riportata aggiunge qualche informazione circa i tempi e le modalità con i quali dovrebbe avvenire la rivelazione della manipolazione. Tuttavia, le indicazioni fornite risultano eccessivamente vaghe, dato che è solo richiesto che le informazioni siano chiare, distinguibili e fornite al massimo entro il momento del primo "contatto" con il contenuto. Questa aggiunta non risolve il problema di come debba essere concretamente adempiuto l'obbligo informativo di cui al par. 3.
Questo è lo stato dell'arte della disciplina europea in materia di *deepfakes* nei primi mesi del 2023. Si tratta di un approccio prudente: da un lato, la scelta di evitare obblighi troppo stringenti o divieti è apprezzabile perché in questo modo si scongiura il rischio di un *chilling effect* nei confronti del ricorso a questa tecnologia anche a scopi benefici; dall'altro, tuttavia, l'incognita concerne il comprendere se un dovere di trasparenza così configurato sarà effettivamente rispettato oppure, al contrario, spesso disatteso.
Le riflessioni tratteggiate sono formulate pur sempre sulla base di un quadro giuridico ancora incerto e rispetto ad un settore ancora in divenire: proprio per questo motivo sarà utile monitorare il lavoro delle istituzioni europee nei prossimi mesi in modo da comprendere quale sarà la posizione definitiva sul punto, se e quando l'atto entrerà in vigore.

## 4. Quali prospettive per il *deepfake*?

Lo stato dell'arte, sotto il profilo sia tecnologico che giuridico, non consente certo di

[86] Proposta di nuovo testo per l'art. 52, par. 3: «Gli utenti di un sistema di IA che genera o manipola immagini o contenuti audio o video che assomigliano notevolmente a persone, oggetti, luoghi o altre entità o eventi esistenti e che potrebbero apparire falsamente autentici o veritieri per una persona ("*deep fake*") sono tenuti a rendere noto che il contenuto è stato generato o manipolato artificialmente. Tuttavia, il primo comma non si applica se l'uso è autorizzato dalla legge per accertare, prevenire, indagare e perseguire reati o se il contenuto fa parte di un'opera o di un programma manifestamente creativo, satirico, artistico o fittizio, fatte salve le tutele adeguate per i diritti e le libertà dei terzi».

prospettare soluzioni a questioni complesse e in divenire. Sia concessa, tuttavia, una riflessione finale sulle implicazioni sul piano costituzionale di una regolamentazione del fenomeno nonché sui confini che i principi in materia di libertà di espressione pongono al legislatore, sia esso europeo o nazionale.

Innanzitutto, come si è già cercato di dimostrare, la previsione di un divieto generale di diffusione dei *deepfakes* sembra porsi in palese contrasto con il diritto a manifestare il pensiero. Tuttavia, ciò non esclude l'ipotesi di interventi "chirurgici" che mirino a contrastare la distorsione del dibattito pubblico, senza comunque precludere lo sviluppo di questa tecnologia.

Un'altra via di intervento, che pone più di una perplessità, è quella che prevede la creazione di nuove fattispecie di reato per perseguire i creatori di *deepfakes* disinformativi[87]. Questa strada è stata percorsa, ad esempio, dallo Stato del Texas, la cui legge in materia criminalizza la creazione e la diffusione di *deepfakes videos* aventi l'intento di nuocere un candidato politico o di influenzare l'esito delle elezioni[88]. Questa scelta fu criticata da alcuni studiosi, tra cui Orin Samuel Kerr, che mostrarono un notevole scetticismo nei confronti della possibilità di perseguire queste condotte ricorrendo allo strumento penale[89].

In generale, la prospettiva di un intervento normativo limitativo dell'uso di questa tecnologia costituisce indubbiamente un terreno scivoloso, proprio come nel contesto della regolamentazione della disinformazione in generale. Rispetto alle altre modalità di divulgazione di notizie false, probabilmente i *deepfakes* possono giustificare un intervento maggiormente stringente, soprattutto se si accoglie la teoria per cui questi contenuti manipolati sono effettivamente maggiormente persuasivi e in grado di condizionare in modo più incisivo l'opinione pubblica[90]. Sono esempio di questa linea di approccio al *deepfake* le *policies* dei principali *social network*: con riguardo ai contenuti disinformativi in generale tendenzialmente le piattaforme preferiscono evitare di fare ricorso a misure "drastiche" e fortemente limitative come la rimozione del contenuto, mentre sembrano più inclini a ricorrere alla cancellazione in relazione ai contenuti manipolati con la tecnica in esame.

Come è stato accennato, i *deepfakes* disinformativi sono maggiormente percepiti come una potenziale minaccia durante il periodo delle elezioni. Proprio in quest'ottica, una opzione regolatoria che alcuni ordinamenti – come la California e il Texas nel contesto statunitense – prediligono è quella di vietare i *deepfakes* "politici" solo durante il periodo elettorale[91], proprio perché questo rappresenta uno dei momenti fondamentali della

[87] La valutazione di non opportunità di un ricorso allo strumento penale riguarda i *deepfakes* disinformativi. Invece, nel caso di quelli a contenuto sessualmente esplicito non consensuali oppure che permettono la commissione di illeciti, il ricorso allo strumento penale non va escluso a priori anche alla luce dei diversi beni giuridici coinvolti.

[88] Texas Senate Bill, *Relating to the creation of a criminal offense for fabricating a deceptive video with intente to influence the outcome of an election*, in *LegiScan.com*, 2019-2020.

[89] O. S. Kerr, *Should Congress pass a "deep fakes" law?*, in *The Volokh Conspiracy*, 31 gennaio 2019.

[90] Cfr. T. Dobber-N. Helberger-N. Metoui-D. Trilling-C. de Vreese, *Do (microtargeted) deepfakes have real effects on policital attitudes?*, cit.

[91] California Assembly Bill-730: *Elections: deceptive audio or visual media*, in *California Legislative Information*, 2019-2020; Texas Senate Bill, *Relating to the creation of a criminal offense for fabricating a deceptive video with*

"vita" democratica, durante il quale molto spesso si verifica una maggiore circolazione di notizie false. Questa scelta, ad esempio, è stata adottata anche in Francia in materia di disinformazione in generale: difatti, la legislazione circoscrive la durata degli obblighi più stringenti in capo alle piattaforme proprio al periodo elettorale[92].

Un'altra tematica centrale in relazione alle possibili misure di contrasto alla disinformazione – e anche a quella veicolata attraverso i *deepfakes* – è quella della trasparenza[93]. Effettivamente, quello contenuto nell'art. 52, par. 3, della proposta di Regolamento sull'Intelligenza Artificiale è un obbligo di trasparenza, poiché impone al creatore del *deepfake* di rivelare che non si tratta di un contenuto autentico. Come già sottolineato, il problema di questa prescrizione riguarda la sua concreta effettività, dato che è molto probabile che venga disattesa per le ragioni già esposte.

Il ragionamento si sposta allora sulla trasparenza della fonte, dato che è ragionevole pensare che i *deepfakes* siano parte di campagne sistematiche di disinformazione e siano diffusi da *account* appartenenti a utenti la cui vera identità rimane sconosciuta. Il tema dell'anonimato in rete è complesso: da un lato, assicurarlo permette a tutti di poter esprimere in modo libero pensieri e opinioni senza il timore spesso legato al doversi esporre personalmente[94]; tuttavia, l'anonimato contribuisce ad un vero processo di deresponsabilizzazione dell'autore dei contenuti[95], il quale, "nascosto" dietro un profilo adespoto, si sente legittimato a veicolare contenuti manipolati e distorsivi.

Per non giungere a negare completamente l'anonimato, in dottrina emergono alcune posizioni propense ad elaborare delle soluzioni di compromesso: ad esempio, una possibilità potrebbe essere quella di permettere agli utenti di partecipare al dibattito pubblico *online* tramite uno pseudonimo, facendo però conoscere la propria identità almeno al gestore della piattaforma e, dunque, fornendo i propri dati personali autentici in sede di iscrizione[96].

Tuttavia, al di là della discussione su come porsi nei confronti dell'anonimato, si potrebbe richiedere alle piattaforme di fornire più informazioni circa la provenienza dei

*intente to influence the outcome of an election*, in *LegiScan.com*, 2019-2020.

92 Legge organica n. 1201 del 22 dicembre 2018 relativa alla lotta contro la manipolazione dell'informazione; Legge ordinaria n. 1202 del 22 dicembre 2018 relativa alla lotta contro la manipolazione dell'informazione.

93 Su questo valore cardine si basano sia il *Digital Services Act* che anche il Codice rafforzato di buone pratiche sulla disinformazione.

94 M. Lamanuzzi, *Il problema della disinformazione in rete: i limiti del diritto penale e le potenzialità del nuovo Codice rafforzato di buone pratiche dell'UE*, in *medialaws.eu*, 2022.

95 In questo senso si esprime da tempo Giulio Enea Vigevani, il quale in un recente contributo afferma che «il principio di trasparenza sulla provenienza dei contenuti – ricavabile dall'interpretazione sistematica dell'art. 21 Cost., commi 3 e 5, che conduce a valorizzare i principi di responsabilità personale di chi diffonde informazioni o idee ed esclude le manifestazioni del pensiero in forma anonima dall'alveo della tutela costituzionale – consentirebbe al legislatore (nazionale o europeo) di imporre alle piattaforme di accertare e indicare se un contenuto proviene da una persona fisica, una persona giuridica, un organo governativo, un programma automatico, di segnalare da quale paese deriva, etc. Si tratterebbe di uno strumento in grado di frenare quell'inquinamento dello spazio dell'informazione, specie da parte di potenze straniere, che finisce per rendere l'esercizio di un "controllo sugli altri poteri", missione fondamentale dell'attività informativa, subalterno a logiche del tutto diverse». Cfr. G.E. Vigevani, *Piattaforme digitali private, potere pubblico e libertà di espressione*, in *Rivista di Diritto Costituzionale*, 1, 2023, 59.

96 M. Lamanuzzi, *Il problema della disinformazione in rete,* cit.

contenuti[97]: un'idea potrebbe essere quella di investire su algoritmi in grado di stabilire quando un contenuto viene, ad esempio, condiviso da un *bot* e non da un utente "umano". In questo caso, anche qualora il contenuto multimediale fosse un *deepfake*, non sarebbe di per sé vietata completamente la pubblicazione – e non sarebbe nemmeno vietato l'anonimato – ma una etichetta avviserebbe circa quale tipo di *account* l'ha condiviso. Ad esempio, se l'utente sapesse che il contenuto che sta visualizzando è stato condiviso da un *bot*[98], sarebbe probabilmente propenso ad una sua valutazione più critica e prudente.

Dal punto di vista delle soluzioni non propriamente giuridiche, è importante richiamare lo spirito critico che ogni utente dovrebbe "mettere in campo" quando si imbatte in contenuti pubblicati sul *web*: in questo senso appare particolarmente persuasivo Guido Scorza, quando, commentando il fenomeno dei *deepfakes*, sostiene che in vista del progressivo aumento della diffusione di queste tecnologie è necessario iniziare a «sviluppare quantità industriali di spirito critico»[99], che deve essere "allenato" già in età scolastica.

E in questa prospettiva anche educativa, sembrano nel complesso condivisibili le raccomandazioni formulate dagli studiosi della *Alliance of Democracies*[100], i quali si sono occupati proprio del rapporto tra *deepfake* e disinformazione[101].

Dopo una approfondita analisi di questa tematica, vengono appunto formulate alcune *recommendations* finali, le quali possono essere così riassunte[102]: intensificare la collaborazione tra Stati e piattaforme in modo da evitare che queste, soggetti privati, possano decidere in autonomia come bilanciare i diritti rilevanti in questi settori; considerare la trasparenza l'elemento "chiave" nella regolazione della disinformazione *online*; definire coerentemente i casi di rimozione dei contenuti manipolati, in modo da evitare il rischio di *over-removal*; infine, accompagnare la legislazione e le restrizioni da parte delle piattaforme con l'educazione mediatica.

In definitiva, il termine "*deepfake*" ha assunto una connotazione intrinsecamente negativa proprio perché i rischi tendono ad oscurare i benefici che possono derivare dal ricorso a questa tecnica[103]. Questo approccio ai *deepfakes*, conseguenza di una spesso diffusa inquietudine nei confronti di ciò che è nuovo o ignoto, rischia di intaccare il

---

97 In questo senso cfr. N. Diakopoulos-D. G. Johnson, *Anticipating and addressing the ethical implications of deepfakes in the context of elections*, 2020, disponibile in *ssrn.com*.

98 Ad esempio, dal 1° luglio 2019 nello Stato della California vige una legge che impone il dovere per i *social network* di etichettare i *bots* in modo da renderli riconoscibili agli utenti. Cfr. California State Senate, Bill no. 1001, chapter 892, 2018 (c.d. *Bolstering online transparency act*).

99 *La app Fakeyou, i falsificatori si innovano: tempi duri per la verità – Intervento di Guido Scorza*, 20 ottobre 2022.

100 La *Alliance of Democracies Foundation* è una organizzazione no profit fondata nel 2017 dall'ex Primo Ministro danese nonché ex Segretario Generale NATO Anders Fogh Rasmussen. L'organizzazione si dedica al progresso della democrazia e del libero mercato in tutto il mondo. Secondo quanto riportato dalla stessa organizzazione, la stessa mira a divenire «*the world's leading megaphone for the cause of democracy*».

101 C. Waldemarsson, *Disinformation, deepfakes and democracy. The european response to election interference in the digital age*, Alliance of Democracies, 2020.

102 Ivi, 21-22.

103 M. van Huijstee et al., *Tackling deepfakes in European policy*, cit., 49.

«colore di per sé neutro che [lo] strumento ha quando vede la luce»[104], creando così un pregiudizio nei confronti della tecnica stessa che potrebbe influenzare una eventuale regolamentazione. Proprio per questo motivo, si è cercato di tracciare le linee generali di un approccio giuridico ai *deepfakes* che sia rispettoso dei diritti fondamentali: una regolamentazione specifica in materia di *deepfake* può essere benvenuta, purché sia ragionata, frutto di un corretto bilanciamento tra diritti e in grado di cogliere la complessità del fenomeno.

Non resta ora che attendere gli sviluppi futuri in questo settore, i quali riguarderanno tanto l'evoluzione della tecnologia quanto quella della normativa.

[104] F. Bertoni, *Deepfake, ovvero Manipola et impera. Un'analisi sulle cause, gli effetti e gli strumenti per la sicurezza nazionale, nell'ambito dell'utilizzo malevolo dell'intelligenza artificiale ai fini di disinformazione e propaganda*, cit., 13.