MODERARE I *DEEPFAKE*: POTERI E LIMITI DEL *DIGITAL SERVICES ACT**

di Michela Parente**

Sommario. 1. Il paradigma del caso singolo. – 2. Il rischio della proliferazione dei contenuti sintetici: la destrutturazione della sfera pubblica. – 3. La regolamentazione dei *deepfake* come contenuti illegali. – 3.1. La regolamentazione dei *deepfake* come contenuti dannosi. – 4. Criticità dell'attuale *framework* normativo – 5. Modelli alternativi di marcatura dei contenuti digitali.

1. Il paradigma del caso singolo. La l.n. 132/2025 *Disposizioni e deleghe al Governo in materia di intelligenza artificiale* ha introdotto all'interno del Codice penale l'articolo 612-*quater* che disciplina il nuovo reato di «Illecita diffusione di contenuti generati o alterati con sistemi di intelligenza artificiale». In base alla norma, «Chiunque cagiona un danno ingiusto ad una persona, cedendo, pubblicando o altrimenti diffondendo, senza il suo consenso, immagini, video o voci falsificati o alterati mediante l'impiego di sistemi di intelligenza artificiale e idonei a indurre in inganno sulla loro genuinità, è punito con la reclusione da uno a cinque anni.»[1].
La fattispecie è stata ribattezzata dalla stampa «reato di *deepfake*»[2]; una definizione tuttavia impropria, poiché la norma non incrimina la mera produzione o diffusione di contenuti falsi in quanto tali, ma sono necessari due ulteriori elementi costitutivi: l'idoneità del contenuto a indurre in inganno sulla sua autenticità e la produzione di un danno ingiusto in capo alla persona offesa. La disposizione appare pertanto destinata a trovare applicazione prevalentemente in concorso con fattispecie penali già esistenti, quali la diffamazione aggravata, la sostituzione di persona, la truffa, la frode informatica, l'estorsione, o la diffusione illecita di immagini o video sessualmente espliciti.
La nuova incriminazione si inserisce dunque in una tendenza già consolidata, che riconosce rilevanza giuridica ai singoli *deepfake* soltanto nella misura in cui essi veicolino contenuti già illeciti secondo l'ordinamento.
Non si distingue, in altre parole, tra contenuti umani e sintetici: ciò che rileva è l'effetto prodotto, indipendentemente dal modo in cui in cui il contenuto è nato.

---

* *Sottoposto a referaggio.*
** Dottoranda in Copyright Law and Cultural Economy – Accademia di Belle Arti di Napoli.
[1] Prosegue «Il delitto è punibile a querela della persona offesa. Si procede tuttavia d'ufficio se il fatto è connesso con altro delitto per il quale si deve procedere d'ufficio ovvero se è commesso nei confronti di persona incapace, per età o per infermità, o di una pubblica autorità a causa delle funzioni esercitate».
[2] Un *deepfake è definito* all'art. 3 co. 1, 60) dell'*Artificial Intelligence Act* 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»; una definizione volutamente più ampia rispetto alla prima formulata in ambito europeo, risalente al 2021 in cui erano *i deepfake* erano definiti come *«manipulated or synthetic audio or visual media that seem authentic, and which feature people that appear to say or do something they have never said or done, produced using artificial intelligence techniques, including machine learning and deep learning.»* Cfr. Cfr. M. van Huijstee *et al.*, *Tackling deepfakes in European policy*, in *EPRS/European Parliamentary Research Service*, 2021, 1.

La soglia di intervento, pertanto, dovrebbe essere identica per entrambi i casi[3].
Ad esempio, una registrazione audio contraffatta di un leader politico è lesiva del dibattito democratico sia che sia prodotta con strumenti analogici, sia che sia generata da un algoritmo[4]. Il *revenge porn* rappresenta un problema preesistente all'avvento dei sistemi di intelligenza artificiale capaci di *declothing*.
Soprattutto, il dibattito relativo ai *deepfake* è stato spesso inserito nel solco delle riflessioni sulla disinformazione; tema endemico alle democrazie liberali che ha però, con l'espansione dovuta ai social media, generato una divergenza negli approcci regolatori: l'adozione di interventi normativi per cercarne di limitarne gli effetti dannosi oppure una tendenza al *laissez-faire* basata sulla fiducia nelle capacità di autoregolamentazione del mercato delle idee[5].
Nelle democrazie occidentali, ad un'analisi dello *ius conditum*, l'approccio settoriale – che si focalizza sui singoli *deepfake* – pare essere quello prevalente: le differenze nel trattamento giuridico dei contenuti sintetici non derivano da una diversa valutazione della fonte ma dai diversi limiti imposti alla libertà di espressione.
Negli Stati Uniti, dove la sentenza Alvarez[6] garantisce al falso protezione costituzionale ai sensi del Primo Emendamento a meno che non rappresenti un pericolo legale concreto, i *deepfake* sono regolamentati solo quando si intersecano con *compelling interest* considerati di innegabile rilevanza dall'ordinamento federale, come il pudore sessuale e la dignità umana[7].
Il *Take It Down Act*[8], firmato dal Presidente Trump 19 maggio 2025, criminalizza la pubblicazione non consensuale di contenuti intimi, inclusi quelli generati tramite intelligenza artificiale, e impone alle piattaforme digitali l'obbligo di rimuovere tali immagini entro 48 ore dalla segnalazione della vittima.
Al di là di quest'unico atto approvato dal Congresso, si registrano a livello statale sporadici interventi a tutela di altri interessi: si tratta perlopiù di limiti ai *deepfake* politici nei periodi elettorali[9], norme a tutela dei consumatori e della proprietà intellettuale[10].

[3] S. A Fisher, J. W. Howard, B. Kira, *Moderating Synthetic Content: The Challenge of Generative AI*, in *Philosophy & Technology*, 37, 2024, 9.
[4] Lungo il corso della storia, ci sono state molte fotografie false, sin da subito anche a scopo propagandistico. D'altro canto, Così, ad esempio, Stalin le utilizzava per cancellare i nemici, uno dei deepfake più famosi è la foto di Mussolini a cavallo, e nel 1990 Newsweek scrisse per la prima volta «con le foto finte, il regime cinese negherà atrocità realmente accadute».
[5] Cfr. O. Pollicino, *La prospettiva costituzionale sulla libertà di espressione nell'era di Internet*, in *MediaLaws – Rivista di diritto dei media,* 1, 2018, 7.
[6] Corte Suprema USA, *United States v. Alvarez*, 567 U.S. 709 (2012).
[7] F. Abbondante, *La tirannia degli algoritmi e la libertà di manifestazione del pensiero*, in *i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale*, 12-1/3, 2019, 13. Parzialmente di diverso parere M.Manetti secondo cui la differenza tra il regime europeo e quello statunitense starebbe invece tra i «diritti altrui» con cui la libertà di espressione viene bilanciata, «poiché i Paesi europei includono tra i "diritti altrui" l'onore e la dignità dei gruppi razziali, religiosi, nazionali, spesso identificati con i sentimenti di identità di questi ultimi, mentre negli Stati Uniti la tutela di tali sentimenti non è ammessa» Cfr. M. Manetti, *Alla ricostruzione della sfera pubblica*, in *Rivista AIC,* 3, 2025, 308.
[8] Acronimo che sta per «*Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks*».
[9] Tralasciando i 45 Stati che hanno aggiunto obblighi per i *deepfake* in ambito pornografico, 28 Stati hanno emanato leggi che regolano i *deepfake* politici. In 18 di questi stati, la legge si applica ai materiali politici distribuiti entro un certo numero di giorni prima o dopo un'elezione. Cfr. Ballotpedia, *Ballotpedia's State of Deepfake Legislation 2025 Annual Report*, 2024, online a: https://ballotpedia.org/Ballotpedia%27s_State_of_Deepfake_Legislation_2024_Annual_Report.
[10] Cfr. A. Orlando, *La regolamentazione del deepfake in Europa, Stati Uniti e Cina*, in *MediaLaws – Rivista di diritto dei media*, 1, 2024, 317-321.

Non diversa è la situazione nell'Unione Europea: nonostante le innegabili differenze tra il modello *market-driven* americano e quello *right-driven* europeo[11], nell'ambito dei contenuti sintetici, ciò che distingue i due sistemi è la diversa calibrazione della libertà di espressione nel Vecchio Continente causata dalle nuove tecnologie.

Sin dal *Code of conduct on countering illegal hate speech online* del 2016, l'Unione Europea ha progressivamente ristretto le maglie del discorso lecito. Concentrandosi su ciò che via via diventava rilevante – dal terrorismo[12], al *microtargeting* politico[13], alla propaganda anti-vax – è stata inaugurata una strategia di regolazione indiretta del discorso pubblico, fondata sulla responsabilizzazione delle piattaforme digitali e spesso sulla rimozione (o effetti assimilabili alla rimozione, come il *delisting*) di contenuti considerati falsi o dannosi.

La tendenza si è progressivamente consolidata sino all'adozione del Reg. UE n. 2022/2065 (*Digital Services Act*), che trasforma i precedenti «Impegni» contenuti nei Codici di *soft law* in doveri giuridicamente vincolanti, legittimando quindi i poteri di intervento da parte delle piattaforme sulle comunicazioni online. La giurisprudenza della Corte di giustizia e della Corte europea dei diritti dell'uomo ha accompagnato questo processo, con una tutela della libertà di espressione «spesso "cedevole" rispetto alla tutela di altri interessi concorrenti»[14].

Ed è in questo contesto che si sono andati a collocare i *deepfake*: in virtù della loro diffusione attraverso la rete, rientrano nell'ambito di applicazione del quadro normativo che ha come finalità creare un «ambiente online sicuro, prevedibile e affidabile»[15].

Ma il DSA è una normativa basata sul caso concreto: asimmetrica, che assegna agli *hosting provider*, in ragione del loro diritto di proprietà sulle arene digitali, il compito di distinguere caso per caso i contenuti con diritto d'asilo nel dibattito pubblico da quelli che non lo hanno[16].

Permane quindi il criterio focalizzato sui *deepfake* illeciti o dannosi, oggetto di intervento attraverso le norme esistenti.

Concentrarsi su singoli contenuti, i cui effetti sono analizzabili nel breve periodo, ha però due rischi complementari: uno per eccesso, per la discrezionalità lasciata alle piattaforme nella gestione dei contenuti, strumentalizzabile ai fini censori; uno per difetto, perché si trascura l'impatto sistemico e le conseguenze a lungo termine derivanti dall'ingresso massivo di contenuti artificiali nel dibattito pubblico.

Finché l'analisi resta ancorata ai singoli contenuti e ai loro effetti immediatamente lesivi, il fenomeno dei *deepfake* appare governabile attraverso le categorie tradizionali dell'illecito e la neonata categoria del «dannoso». Ciò si rivela però insufficiente nel momento in cui il problema non è più il contenuto in sé, ma l'impatto cumulativo e strutturale derivante dalla loro proliferazione massiva nel dibattito pubblico.

È questa la seconda prospettiva che merita di essere esplorata: una prospettiva che considera *deepfake* non solo per gli effetti immediatamente lesivi di singoli contenuti, ma anche *ex se*.

---

[11] A. Bradford, *Digital Empires-The Global Battle to Regulate Technology*, Oxford, 2023, 33 ss.

[12] C. Caruso, *L'hate speech a Strasburgo: il pluralismo militante del sistema convenzionale,* in *Quaderni Costituzionali*, 4, 2017, 966-976.

[13] E. Shattock, *Fake News in Strasbourg: Electoral Disinformation and Freedom of Expression,* in *European Journal of Law and Technology*, 1, 2022, 3 ss.

[14] O. Pollicino, *La prospettiva costituzionale sulla libertà di espressione nell'era di Internet*, cit., 13.

[15] *Digital Services Act,* Considerando 3. I *deepfake* sono anche oggetto dell'*Artificial Intelligence Act*. L'argomento sarà trattato in modo più dettagliato nel paragrafo 3.

[16] G. De Minico, *Nuova tecnica per nuove diseguaglianze. Case law: Disciplina Telecomunicazioni, Digital Services Act e Neurodiritti*, in *Federalismi.it*, 6, 2024, 17.

2. Il rischio della proliferazione dei contenuti sintetici: la destrutturazione della sfera pubblica. I contenuti manipolati tramite intelligenza artificiale generativa possono essere considerati un *quid novis*, non riconducibile a problematiche sociali già affrontate in passato, a causa dell'evoluzione qualitativa e quantitativa della disinformazione.

Sul piano qualitativo, i contenuti sintetici diventano molto più credibili e pericolosi per tre motivi principali. Innanzitutto, sono specificamente progettati per mantenere la coerenza temporale: i video non fanno salti logici e non riproducono strani sfarfallii ma riescono a replicare perfettamente l'aspetto di una persona. A questo si aggiunge che la clonazione vocale (bastano pochi secondi di audio per replicare il modo di parlare di una persona) ha superato la «soglia dell'indistinguibile»[17].

La vera novità, però, è che questi strumenti non sono più per pochi esperti. Grazie a piattaforme come Sora di OpenAI o Veo di Google, chiunque può generare video professionali partendo da una semplice descrizione testuale.

Qui il salto quantitativo: la democratizzazione dell'IA generativa ha causato l'imprevedibile e inedita crescita dei contenuti sintetici: un aumento da 500,000 del 2023 a una stima di 8 miliardi nel 2025; secondo l'Europol entro il 2026 il 90% dei contenuti online potrebbe essere generato in modo sintetico[18].

In questo contesto, la *Dead Internet Theory*[19], un tempo relegata alle teorie cospirazioniste, viene promossa a fenomeno meritevole di attenzione[20].

Non è un colpo di scena ma l'evoluzione logica delle *echo chamber*[21]: i social media hanno progressivamente trasformato lo spazio dell'interazione umana digitale a luogo di dominio di algoritmi ottimizzati per massimizzare l'*engagement*, che premiano la redditività economica dei contenuti rispetto all'esperienza dell'utente[22]. La sostituzione dei contenuti artificiali a quelli umani è funzionale a questo scopo: bot e sistemi di intelligenza artificiale simulano conversazioni, imitano tendenze e producono automaticamente messaggi studiati per incentivare una permanenza sempre più prolungata sulle piattaforme. Ciò accresce l'esposizione degli utenti alla pubblicità, con un conseguente incremento dei ricavi per le società di social media[23].

La trasformazione dei social media in ambienti strutturati e finalizzati al consumo impatta in profondità la dimensione individuale e culturale dell'esperienza online.

La sostituzione delle interazioni significative con scambi superficiali e serializzati contribuisce, specialmente per le giovani generazioni, all'isolamento e alla dilagante incapacità di creare connessioni.

---

[17] S. Lyu, *2026 will be the year you get fooled by a deepfake, researcher says. Voice cloning has crossed the indistinguishable threshold,* in *fortune.com*, 27/12/2025, online a: https://fortune.com/2025/12/27/2026-deepfakes-outlook-forecast/.

[18] M. Negreiro, *Children and deepfakes*, in *EPRS | European Parliamentary Research Service,* 2025, 2-3, online a: https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/775855/EPRS_BRI(2025)775855_EN.pdf.

[19] La *Dead Internet Theory* sostiene che la maggior parte dei contenuti, delle interazioni e degli utenti che incontriamo online non sia umana, ma generata da bot, algoritmi e intelligenza artificiale.

[20] P. Muzumdar et al., *The Dead Internet Theory: A Survey on Artificial Interactions and the Future of Social Media*, in *Asian Journal of Research in Computer Science*, 1, 2025, 68.

[21] Per *echo chamber* si intende l'ambiente digitale in cui un individuo è esposto esclusivamente a informazioni, idee e opinioni che confermano e rinforzano il proprio sistema di credenze preesistente. In tale contesto, l'architettura degli algoritmi di raccomandazione, ottimizzata per massimizzare l'interazione, isola l'utente da prospettive divergenti, alimentando fenomeni di polarizzazione sociale e frammentazione del discorso pubblico. Cfr. E. Mostacci, *Critica della ragione algoritmica: Internet, partecipazione politica e diritti fondamentali*, in *Costituzionalismo.it,* 2, 2019, 96.

[22] Y. Walter, *Artificial influencers and the dead internet theory*, in *AI & Society,* 40, 2025, 239.

[23] P. Muzumdar *et al., The Dead Internet Theory,* cit., 68.

Inoltre, gli algoritmi orientati a mantenere l'attenzione tendono a privilegiare format replicabili e ad alta resa, penalizzando la creatività. Ne deriva un ecosistema informativo non pluralista, ripetitivo e privo di stimoli, nel quale i contenuti originariamente prodotti da esseri umani vengono progressivamente oscurati da materiali generati artificialmente[24].

Ma, soprattutto, il proliferare di contenuti sintetici e la sensazione di artificiosità e alienazione che caratterizza sempre più l'esperienza in Rete, rende, da un lato, più agevole negare la realtà; dall'altro induce le persone a presumere in maniera generalizzata la falsità di ciò che leggono. Alla diffusione della menzogna – oggetto specifico della disinformazione – si aggiunge quindi la delegittimazione del vero: in un ecosistema informativo saturo di verosimile, gli individui tendono a sviluppare un atteggiamento di scetticismo generalizzato, una progressiva sfiducia verso qualsiasi contenuto, indipendentemente dalla sua effettiva veridicità.

Si tratta del fenomeno del «*Liar's Dividend*». Teorizzato da Chesney e Citron, il «dividendo del bugiardo» indica l'effetto per cui diventa più semplice negare e mentire su eventi reali, sostenendo che sono *deepfake*[25].

«Dividendo» perché è quanto i bugiardi guadagnano dall'alfabetizzazione della società in queste materie: più le persone sono informate sui pericoli dei *deepfake*, più diventeranno sospettose e tenderanno a credere a chi gli dice che un contenuto, seppur vero, è falso[26].

La diffusione di contenuti falsi presentati come veritieri, unitamente alla diffusione di contenuti veritieri percepiti come falsi, produce un effetto preciso: l'abbattimento del principio di realtà aristotelico, ma soprattutto, dopo l'effetto *filter bubble,* lo sgretolamento definitivo della «enciclopedia comune», la base conoscitiva condivisa, necessaria al funzionamento delle democrazie liberali[27].

In definitiva, la teoria della sfera pubblica habermasiana della sfera pubblica[28], l'interpretazione funzionale della libertà di espressione da parte della Consulta[29], la giurisprudenza della Corte Europea dei Diritti dell'Uomo in materia di libertà di informazione[30] affermano tutte un identico principio fondamentale: ogni dibattito

---

[24] Ivi, 70-71.

[25] B. Chesney, D. Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, in *California Law Review*, 6, 2019, 1785.

[26] Un caso emblematico è quello del ministro degli affari economici della Malesia Azmin Ali, oggetto di un video a sfondo sessuale di cui è ancora discussa la veridicità. Il «dividendo del bugiardo», nonché più in generale i *deepfake*, sono anche però ampiamente sfruttati da Donald Trump, la cui retorica politica si fonda sulla delegittimazione sistematica delle informazioni scomode, etichettando ogni informazione meno favorevole come «*fake news*» sul suo social *Parler*.

[27] E. Mostacci, *Critica della ragione algoritmica*, cit., 60.

[28] Per Habermas, la sfera pubblica rappresenta uno spazio intermedio tra lo Stato e la società civile; un luogo ideale e materiale in cui i privati cittadini si riuniscono per interagire liberamente e discutere razionalmente di questioni di interesse comune. Funge da organo di controllo critico nei confronti del potere politico e la sua legittimità poggia sul presupposto che il dibattito sia guidato dall'accesso paritario a informazioni veritiere. Sulla genesi e sulla funzione della sfera pubblica come fondamento della democrazia deliberativa: J. Habermas, *Storia e critica dell'opinione pubblica,* Roma-Bari, 1974.

[29] La teoria funzionale valorizza la dimensione partecipativa del diritto di manifestazione del pensiero: essa parte dal presupposto che i sistemi democratici si basano sulla libera volontà e sull'autodeterminazione degli individui, senza le quali le elezioni non avrebbero senso. Da ciò discende che l'articolo 21 ha un duplice nucleo e accanto alla libertà di esprimersi si configura l'attività volta ad informarsi, per cui è ammissibile – o addirittura necessario – porre limiti al profilo attivo in modo che quello passivo sia realizzato nella maniera più completa possibile. E. Cheli, *La giurisprudenza della Corte costituzionale italiana in tema di media*, in *Relazione alla Corte costituzionale*, 2010, 3. La Consulta ha aderito a questa interpretazione in molteplici sentenze. Cfr. Corte cost., sentt. nn. 48/1964, 112/1993, 155/2002.

[30] Per tutte: Corte EDU, 8 novembre 2016, *Magyar Helsinki Bizottság c. Ungheria*, ric. n. 18030/11, § 106, ove si riconduce la libertà di informazione alla sua funzione sistemica di presidio di un dibattito pubblico informato.

sull'indirizzo politico da imprimere al Paese e su come plasmare la realtà presuppone un accordo preliminare su quale sia la realtà.
Il pericolo per questo principio che deriva dai *deepfake* è già stato delineato dal primo atto europeo dedicato al fenomeno, il report *Tackling deepfakes in European policy*, redatto nel 2021 per il *Panel for the Future of Science and Technology* del Parlamento europeo. Il documento individua nei *deepfake* un potenziale di disgregazione dell'ecosistema informativo, per i loro «effetti a cascata»: perdita di fiducia nelle istituzioni, indebolimento della coesione sociale, compromissione della percezione condivisa della realtà[31].
Anche l'UNESCO, nel rapporto *Synthetic Content and its Implications for AI Policy*, sottolinea come i contenuti sintetici contribuiscano a erodere la fiducia tanto negli individui quanto nelle istituzioni, incidendo negativamente sulla formazione dell'opinione pubblica e sul funzionamento stesso delle società democratiche[32].

Del declino della verità si avvantaggia l'autoritarismo: «se il pubblico non crede più a ciò che vede e sente e la verità diventa una questione di opinione, il potere fluisce verso coloro che sono più in vista, finendo per rafforzare chi è già autorità»[33].
Dinanzi a questo scenario, l'interrogativo si sposta sull'adeguatezza degli strumenti vigenti e sull'idoneità del *Digital Services Act* ad un essere un meccanismo di governo dell'ecosistema informativo nel suo complesso.
Il DSA affronta il suo compito individuando due macro-aree di intervento[34]: i contenuti illegali e una seconda area di contenuti dannosi/rischiosi/*awful-but-lawful*[35] da individuare in base agli «eventuali rischi sistemici nell'Unione derivanti dalla progettazione o dal funzionamento del loro servizio e dei suoi relativi sistemi, compresi i sistemi algoritmici, o dall'uso dei loro servizi»[36].
A questa dicotomia corrispondono due sistemi di risposta. Il primo è basato sulla rimozione e riguarda la generalità dei prestatori di servizi. Il secondo, invece, è un modello di «gestione» del rischio, che il legislatore europeo ha scelto di limitare esclusivamente alle piattaforme e ai motori di ricerca di dimensioni molto grandi (VLOP e VLOSE).
La domanda che si impone, a questo punto, è se i *deepfake*, come fenomeno unitario, possano essere sussunti in uno dei due regimi del *Digital Services Act*, e se tale operazione sia in grado di offrire una risposta adeguata alla crisi della sfera pubblica.

3. La regolamentazione dei deepfake come contenuti illegali. I contenuti illegali sono oggetto di un intricato sistema procedimentale che impone alle piattaforme di rimuoverli, pur limitando i loro poteri attraverso garanzie procedurali[37].

---

[31] Cfr. M. van Huijstee *et al., Tackling deepfakes,* cit., 4-6.
[32] M. Squicciarini et al., *Synthetic Content and its Implications for AI Policy – A Primer*, in *unesco.org*, 30, online a : https://unesdoc.unesco.org/ark:/48223/pf0000392181.
[33] B. Chesney, D. Citron, *Deep Fakes: A Looming Challenge*, cit., 1786.
[34] Al Considerando 2 si introducono «obblighi di diligenza per i prestatori di servizi intermediari per quanto riguarda il modo in cui dovrebbero contrastare i contenuti illegali» ma anche «la disinformazione online e altri rischi per la società»; ugualmente, al Considerando 5 le «attività illegali» sono integrate da quelle «comunque dannose».
[35] L'espressione è stata coniata da D. Keller per indicare quei contenuti che non sono di per sé illeciti ma risultano sgraditi nell'ambito del discorso pubblico. Cfr. D. Keller, *Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users,* in *University of Chicago Law Review*, 07/06/2022.
[36] *Digital Services Act,* art. 34, co. 1. Al secondo comma si individuano quattro aree di rischio: diffusione di contenuti illegali; effetti negativi, attuali o prevedibili, per l'esercizio dei diritti fondamentali; eventuali effetti negativi, attuali o prevedibili, sul dibattito civico e sui processi elettorali, nonché sulla sicurezza pubblica; qualsiasi effetto negativo, attuale o prevedibile, in relazione alla violenza di genere, alla protezione della salute pubblica e dei minori e alle gravi conseguenze negative per il benessere fisico e mentale della persona.
[37] G. De Gregorio, *The Normative Power of Artificial Intelligence*, in *Indiana Journal of Global Legal Studies*, 2, 2023, 75.

La presenza di un contenuto illecito apre per i singoli utenti, per i segnalatori attendibili e per le pubbliche autorità la possibilità di notificarlo alla piattaforma; notifica cui consegue l'obbligo di rimozione, pena la responsabilità dell'*hosting provider* per il contenuto stesso[38].
La decisione di rimozione deve essere accompagnata da una motivazione chiara e dalla possibilità di appello[39].
Ciò non toglie, tuttavia, che si è comunque in presenza di privati cui spetta la decisione sul diritto di parola degli utenti; decisione che sarà esaminata da un giudice in base a parametri legali solo in seconda – o finanche in terza[40] – battuta.
Questo approccio così restrittivo previsto per i contenuti illeciti – di per sé criticato dalla dottrina che ritiene che la procedimentalizzazione si traduca nella legittimazione delle *Big Tech* come arbitri del dibattito e nella delega in bianco di poteri pubblici alle prime[41] – non può quindi essere esteso indiscriminatamente a tutti i contenuti sintetici.

Assoggettarli al regime dei contenuti illegali previsto dal *Digital Services Act* risulterebbe manifestamente incompatibile con l'impianto costituzionale europeo, scontrandosi frontalmente con i principi fondamentali dell'ordinamento giuridico, in *primis* con la libertà di espressione sancita all'articolo 10 della Convenzione Europea dei Diritti dell'Uomo e all'articolo 11 della Carta dei Diritti Fondamentali dell'Unione Europea e l'interpretazione che ne è stata data dalle Corti.
Nonostante la menzionata «ostilità» delle Corti europee per la libertà di espressione nel mondo digitale[42], ogni ipotesi di limitazione dei contenuti sintetici deve necessariamente confrontarsi con il rigido scrutinio previsto dall'articolo 10, paragrafo 2, della CEDU. Tale norma impone che qualsiasi restrizione sia «prevista dalla legge», persegua uno degli obiettivi tassativamente elencati nella norma e, soprattutto, «necessaria in una società democratica», ovvero rispettosa del canone di proporzionalità.
Sebbene sotto il profilo della legalità, la categorizzazione *ex ante* di tutti i contenuti sintetici possa apparire formalmente più agevole rispetto all'arduo compito di codificare *per tabulas* la verità, e pur ammettendo che ragioni di sicurezza nazionale possano fungere da base giustificativa per l'intervento, un divieto generalizzato soccomberebbe inesorabilmente al vaglio di proporzionalità.
I *deepfake*, al di là dei rischi concreti e mediaticamente visibili, costituiscono uno strumento espressivo utilizzato in ambiti artistici, creativi e culturali[43]. Hanno contribuito alla realizzazione di opere di valore sociale e documentaristico, alla valorizzazione dell'arte e alla sensibilizzazione pubblica, come dimostra il loro utilizzo in campagne di sensibilizzazione pubblica, in documentari per proteggere l'identità di soggetti vulnerabili e nella ricostruzione

[38] *Digital Services Act*, art. 16 per i meccanismi di segnalazione e azione; art. 9 per gli ordini di contrastare i contenuti illegali da parte delle Autorità; art. 22 per i segnalatori attendibili; art. 6, c. 1, lett. b) per la responsabilità condizionata delle piattaforme. A questi meccanismi va aggiunta la clausola del Buon Samaritano: la possibilità che la piattaforma svolga «indagini volontarie di propria iniziativa o di adottare altre misure volte a individuare, identificare e rimuovere contenuti illegali» (art.7).
[39] Ivi, art. 17 per l'obbligo di motivazione; art. 20 per l'obbligo di gestione interna dei reclami.
[40] Oltre al sistema di gestione interna dei reclami, è predisposto anche un meccanismo di risoluzione extragiudiziale delle controversie (art. 21). In tal modo, pur non essendo mai precluso il ricorso giudiziale, gli utenti potrebbero scegliere di rivolgersi alla giurisdizione ordinaria anche in ultima istanza, esperite prima le possibilità interne al DSA.
[41] G. De Minico, *Le fonti del diritto: un argine all'intelligenza artificiale?*, in *Rivista AIC*, 3, 2025, 96; M. Zalnieriute, *Against procedural fetichism: a call for a new digital constitution*, in *Indiana Journal of Global Legal Studies*, 2, 2023, 253.
[42] O. Pollicino, *La prospettiva costituzionale sulla libertà di espressione nell'era di Internet*, cit., 21.
[43] Cfr. A. Orlando, *La regolamentazione del deepfake in Europa, Stati Uniti e Cina*, cit., 308.

virtuale di interventi storici[44]. Solo qualora i costi sociali derivanti dall'impiego dell'IA generativa risultassero eccessivi, potrebbe rendersi necessario sacrificare i benefici che essa apporta pur di prevenirne gli effetti deleteri[45].
A ciò si aggiunga una criticità di ordine tecnico: l'imponente volume di dati immessi quotidianamente in rete renderebbe la moderazione umana impraticabile, rendendo inevitabile il ricorso a sistemi di rilevamento automatizzato che presentano margini di errore tuttora insostenibili[46].
Per la Corte di giustizia, internet è uno «degli strumenti principali con cui gli individui esercitano il loro diritto alla libertà di espressione e informazione» e «sebbene l'articolo 10 della CEDU non vieti in quanto tale qualsiasi restrizione preliminare ad un mezzo di diffusione, siffatte restrizioni presentano tuttavia pericoli così grandi per il rispetto di tale diritto fondamentale da dover iscriversi in un contesto giuridico particolarmente rigoroso»[47].
Pertanto, l'impiego di filtri algoritmici caratterizzati da un elevato numero di falsi positivi, unito alla loro natura di controllo preventivo sulle comunicazioni, finirebbe per configurare una forma di sorveglianza generalizzata, incompatibile con lo Stato di diritto[48].
Lo stesso *AI Act*, in una sua versione precedente, classificava i *detection system* come sistemi ad alto rischio, a dimostrazione della consapevolezza istituzionale dei pericoli connessi al loro utilizzo[49].
Alla luce di tutto ciò, estendere quanto previsto per i contenuti illeciti ai *deepfake*, con conseguente divieto generalizzato e monitoraggio da parte delle piattaforme, non solo non sarebbe auspicabile sul piano politico e culturale, ma risulterebbe anche giuridicamente insostenibile. Si porrebbe in palese violazione del canone di proporzionalità e non supererebbe il controllo di legittimità richiesto dallo *strict scrutiny* della Corte di Giustizia.
Ogni intervento normativo dovrà pertanto rispettare il principio secondo cui la libertà di espressione, anche digitale, è la regola, e la censura l'eccezione.
La soluzione non risiede, pertanto, nel paternalismo pubblico, che finirebbe per comprimere il pluralismo delle opinioni, bensì nel recupero di una rigorosa distinzione tra fatti accertabili e mere opinioni.

3.1. La regolamentazione dei deepfake come contenuti dannosi. Esclusa la possibilità di ricondurre i *deepfake* in quanto tali al regime dei contenuti illegali o estendere la normativa per questi prevista senza violare i principi fondamentali della libertà di espressione, l'analisi deve necessariamente spostarsi verso il secondo binario: quello dei contenuti che comportano «rischi per la società».
Ci si addentra così nell'area grigia dei contenuti cosiddetti *awful but lawful*: espressioni che, pur rimanendo formalmente entro i confini della legalità definiti dagli ordinamenti nazionali, si rivelano potenzialmente nocive per la tenuta dei processi democratici o per la tutela dei diritti fondamentali.

---

[44] S. Chandler, *Why Deepfakes are a Net Positive for Humanity*, in *Forbes*, 09/03/2020 online a: https://www.forbes.com/sites/simonchandler/2020/03/09/why-deepfakes-are-a-net-positive-for-humanity/.
[45] In questa prospettiva, l'IA generativa potrebbe manifestarsi con una tale mole di contenuti dannosi da rendere preferibile un'esclusione totale. S.A. Fischer, ritiene comunque che tale approccio pessimistico sia prematuro. Cfr. S. A Fisher, J. W. Howard, B. Kira, *Moderating Synthetic Content*, cit., 6.
[46] Cfr. F. R. Moreno, *Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content*, in *International Review of Law, Computers & Technology*, 3, 2024, 318.
[47] Corte di Giustizia dell'Unione Europea, *Poland v Council and European Parliament* C-401/19, §46, 26/04/2022.
[48] Cfr. F. R. Moreno, *Generative AI and Deepfakes*, cit., 315.
[49] M. Cazzaniga, *Una nuova tecnica (anche) per veicolare disinformazione: le risposte europee ai deepfakes*, in *MediaLaws – Rivista di diritto dei media*, 1, 2023, 182,

In questo ambito, il legislatore europeo è consapevole di muoversi su un crinale estremamente scivoloso. Se il principio cardine del DSA è «rendere illegale online ciò che è già illegale offline», l'estensione della vigilanza a ciò che è «solo» sgradito rischia di scivolare verso forme di censura o eccessiva moderazione.
La scelta di occuparsi dei contenuti dannosi ha, infatti, sollevato critiche: mentre nel mondo fisico la semplice falsità di un'affermazione non costituisce reato e l'incitamento all'odio richiede spesso un nesso concreto con un'azione dannosa, nello spazio digitale si tende ad applicare restrizioni anche in assenza di un danno effettivo e immediato[50].
In tale cornice, i *deepfake* rappresentano il paradigma del contenuto «terribile ma lecito». Possono circolare liberamente finché non integrano fattispecie criminose specifiche (come la diffamazione o la *non-consensual deepfake pornography*), poiché l'Unione Europea ha consapevolmente evitato di istituire un diritto alla verità o di equiparare il falso all'illecito. Coerentemente, il DSA non obbliga le piattaforme a rimuovere i *deepfake* in quanto tali o ad occuparsene *tout court*, bensì richiede la mitigazione dei rischi associati a quelli potenzialmente dannosi.

Non potendo intervenire sulla censura del contenuto, ci si sposta sul piano della neutralizzazione degli effetti negativi.
La strategia normativa europea proposta, a partire dai primi Codici di condotta fino all'*Artificial Intelligence Act* e al *Digital Services Act*, mira a introdurre un principio di trasparenza che funga da «mappa»: negli atti europei è sempre più centrale la scelta di ricorrere all'etichettatura come strumento regolatorio. Lasciare il contenuto artificiale visibile ma riconoscibile come tale.
L'etichettatura dei contenuti sintetici – strumento principe per consentire di riconoscere il falso – concilia due esigenze: da un lato, la salvaguardia della libertà di espressione, evitando interventi censori; dall'altro, la promozione dell'educazione mediatica o, come definita dal Garante Privacy, «quantità industriale di spirito critico»[51].
Imperniare la regolamentazione dei *deepfake* sulla trasparenza significa anche poter intervenire in maniera più capillare, su più livelli del loro ciclo di vita: dalla fase tecnologica e creativa, a quella della circolazione e della ricezione[52].
Le prime due dimensioni, quelle relative alla creazione del contenuto, sono oggetto di regolamentazione da parte dell'*AI Act*, mentre le successive – diffusione e ricezione – rientrano nell'ambito del DSA.
In particolare, l'articolo 50 dell'*AI Act*, corredato dai Considerando 132-137 e destinato a entrare in vigore nel marzo 2026, si rivolge a due categorie principali di soggetti, disciplinando in maniera più dettagliata le eccezioni rispetto agli obblighi previsti.
Da un lato, l'articolo 50.4 impone ai «fornitori di sistemi di IA, inclusi i sistemi di AI per finalità generali, che generano contenuti audio, immagine, video o testuali sintetici» di contrassegnare «in un formato leggibile meccanicamente e rilevabili come generati o manipolati artificialmente» i contenuti. Il Considerando 133 esemplifica tali contrassegni con «filigrane, identificazioni di metadati, metodi crittografici per dimostrare la provenienza e l'autenticità dei contenuti, metodi di registrazione, impronte digitali o altre tecniche, a seconda dei casi».

[50] G. De Minico, *Nuova tecnica per nuove disuguaglianze,* cit., 18.
[51] M. Cazzaniga, *Una nuova tecnica (anche) per veicolare disinformazione*, cit., 186.
[52] Queste quattro delle possibili dimensioni regolatorie riconosciute nel report *Tackling deepfakes,* cui deve aggiungersi la dimensione dell'audience. Cfr. M. Van Uistee *et al.*, *Tackling deepfakes in European policy*, cit., 5, online a: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.

Ciò però solo se «tecnicamente possibile, tenendo conto delle specificità e dei limiti dei vari tipi di contenuti, dei costi di attuazione e dello stato dell'arte generalmente riconosciuto, come eventualmente indicato nelle pertinenti norme tecniche» che, a norma del comma 7 del medesimo articolo, saranno contenuti nei codici di buone pratiche incoraggiati e agevolati dall'Ufficio IA[53]. Inoltre, l'obbligo non sorge se i sistemi AI svolgono «una funzione di assistenza per l'editing standard o non modificano in modo sostanziale i dati di input forniti dal *deployer*» o se «autorizzati dalla legge ad accertare, prevenire, indagare o perseguire reati.».
Dall'altro lato, anche i cosiddetti *deployer* – cioè coloro che integrano tali sistemi o gli utenti finali – sono tenuti a dichiarare che il contenuto è stato generato o modificato artificialmente. Anche in questo caso, però, si tratta di un obbligo ampiamente derogabile: non vale se l'utilizzo rientra in opere manifestamente artistiche, creative, satiriche o fittizie, per le quali si cerca un bilanciamento con la fruizione dell'opera stessa. L'obbligo decade anche se il contenuto è diffuso per finalità di informazione pubblica, a condizione che vi sia stata una revisione umana o un controllo editoriale e che esista una responsabilità editoriale chiaramente individuata[54].

Il risultato è una disciplina che, ancor prima di entrare in vigore, sembra dichiarare la propria inefficacia: priva di reali strumenti operativi, estremamente vaga nelle eccezioni, che si appella alla correttezza morale dei creatori.
L'ampiezza e la flessibilità degli obblighi previsti dall'AIA trovano la loro ragion d'essere nel coordinamento con il *Digital Services Act*, che si pone come il vero fulcro regolatorio per la gestione dei *deepfake* nella fase della loro circolazione.
Ciò emerge dal Considerando 136 dell'AIA, il quale afferma che «Gli obblighi imposti dal presente regolamento ai fornitori e ai *deployer* di taluni sistemi di IA, volti a consentire il rilevamento e la divulgazione del fatto che gli output di tali sistemi siano generati o manipolati artificialmente, sono molto importanti per contribuire all'efficace attuazione del regolamento UE 2022/2065».
La normativa europea continua quindi a delegare l'effettiva moderazione dei contenuti alle grandi piattaforme online e ai motori di ricerca più influenti, i cosiddetti VLOP e VLOSE[55], «data l'importanza che le piattaforme online di dimensioni molto grandi […] rivestono nel facilitare il dibattito pubblico, le operazioni economiche e la diffusione al pubblico di informazioni, opinioni e idee e nell'influenzare il modo in cui i destinatari ottengono e comunicano informazioni online»[56].
Sebbene, come anticipato, il *Digital Services Act* non disciplini specificamente i *deepfake* come categoria, la potenziale distorsione del dibattito civico che questi potenzialmente causano rientra tra i «rischi sistemici» che le piattaforme devono identificare e mitigare, ai sensi degli articoli 34 e 35.

[53] Questi codici rientrano però a pieno titolo nel modello co-regolativo, con rischio di una «inevitabile torsione rapace, egoistica delle regole armonizzate, ideate a vantaggio esclusivo dell'impresa.» Cfr. G. De Minico, *Le fonti del diritto*, cit., 89-90.
[54] Ciò sconfessando le indicazioni contenute nel report *Tackling Deepfake* che ammonivano contro l'utilizzo di eccezioni troppo ampie per questo obbligo. Cfr. A. Orlando, *La regolamentazione del deepfake in Europa, Stati Uniti e Cina*, cit., 313.
[55] «*The assumption seems to be that platforms will be playing cat-and-mouse with anonymous bad actors who will distribute deepfakes and must be detected and labelled*». Cfr. E. Hine, L. Floridi, *New Deepfake Regulations in China Are a Tool for Social Stability, but at What Cost?*, in *Nature Machine Intelligence*, 4, 2022, 609.
[56] *Digital Services Act,* Considerando 75.

Tra le misure proposte per la mitigazione dei rischi di cui all'articolo 35 c'è infatti anche «il ricorso a un contrassegno ben visibile per fare in modo che un elemento di un'informazione, […] che a una persona appaia falsamente autentico o veritiero, sia distinguibile»[57].
Il quadro si completa con lo *Strenghtened Code of Information*, acquisito nel DSA a partire dal 13 febbraio 2025[58]. In esso, con gli Impegni 14 e 15, i firmatari promettono di contrastare la disinformazione e i deepfake malevoli, rafforzando le politiche contro le manipolazioni e di «prendere in considerazione» gli obblighi di trasparenza previsti dalla proposta di legge sull'IA.
Dalla lettura congiunta dell'*AI Act* e del *Digital Services Act* emerge un quadro normativo che, almeno in apparenza, persegue un obiettivo condiviso: distinguere ciò che è artificiale da ciò che è reale.

Inquadrando quindi la totalità dei *deepfake* come un fattore di rischio sistemico, il *framework* europeo dovrebbe far scattare l'obbligo, in uno qualsiasi dei momenti del ciclo di vita del contenuto sintetico – durante la sua creazione, nella fase della condivisione o al momento della fruizione da parte dell'utente – dell'apposizione di un'etichetta che renda manifesta la natura artificiale del contenuto. Una marcatura che, in teoria, dovrebbe impedire all'utente di esserne tratto in inganno.

4. Criticità dell'attuale framework normativo. Se il principio di trasparenza evocato dall'*AI Act* e dal *Digital Services Act* fosse sostenuto da una disciplina coerente, stringente e realmente operativa, esso potrebbe costituire uno strumento efficace per contrastare i rischi sistemici derivanti dalla proliferazione dei *deepfake*. L'etichettatura, infatti, impedirebbe che l'artificiale continui a presentarsi come reale, contribuendo alla progressiva delegittimazione del vero e alla destrutturazione della sfera pubblica.
Rendere riconoscibile la natura artificiale di un contenuto significherebbe restituire agli utenti le condizioni minime per orientarsi nello spazio informativo, ricostruendo quella base conoscitiva condivisa senza la quale il dibattito democratico perde significato. Un simile intervento, agendo sulla trasparenza anziché sulla rimozione, consentirebbe inoltre di ridimensionare il ruolo delle piattaforme come arbitri sostanziali della libertà di espressione, poiché la valutazione dei contenuti verrebbe spostata dal potere discrezionale degli intermediari alla capacità critica degli individui.
Tale ragionamento sarebbe valido in presenza di una normativa di dettaglio, un obbligo che imponga realmente ai soggetti l'etichettatura dei contenuti sintetici.
Come, tuttavia, si evince dalla proliferazione di *deepfake* che ingannano anche gli occhi più attenti, l'obbligo di marcatura, seppur formalmente contenuto in molteplici atti, è attualmente assente.
Relativamente all'*AI Act,* oltre alla sospesa entrata in vigore, alla genericità e alle ampie riserve non è previsto per i *provider* il criterio del «contrassegno ben visibile», ma unicamente che «i contenuti siano marcati in un formato leggibile meccanicamente e rilevabili come generati o manipolati artificialmente». La previsione appare finalizzata a impedire che i bot *crawler* che si occupano di reperire dati per il *training* acquisiscano contenuti di scarsa qualità, quelli sintetici, piuttosto che a tutelare gli utenti finali[59].

[57] La gestione dei rischi sistemici è anche presente nell'*AI Act*, anche questa come obbligo futuro che dovrà essere contenuto nei Codici di buone pratiche (Art. 56, co. 2, lett. *b)*.
[58] Si sostiene che però questo Codice non sia stato molto influente perché le piattaforme non hanno cambiato le loro *policy* di moderazione dopo la firma. Cfr. M. Cazzaniga, *Una nuova tecnica (anche) per veicolare disinformazione*, cit., 180.
[59] Secondo A. Knott le aziende hanno interesse che i contenuti sintetici siano rilevanti per evitare il «collasso del modello» che deriva dall'allenare un *foundation model* sul proprio output. Cfr. A. Knott, D. Pedreschi, R.

Nel contesto del *Digital Services Act* la situazione non è più rassicurante. L'adozione di un «contrassegno ben visibile» viene infatti citata meramente a titolo esemplificativo come possibile misura da adottare per mitigare i rischi sistemici. La norma non impone, però, alcun obbligo effettivo alle piattaforme: si limita a suggerire, lasciando piena libertà di scelta sia nell'individuazione del rischio sia nella risposta. Si tratta quindi di un potere, non di un dovere.

In quanto non-obbligo, l'etichettatura può essere tanto ignorata quanto applicata solo in modo sporadico e selettivo, finendo per colpire unicamente quei *deepfake* che già intersecano altre forme di illecito o di dannosità. Il risultato è che la logica del «caso singolo» sopravvive anche nel regime dei rischi sistemici, lasciando irrisolto il problema di fondo: come consentire agli utenti di orientarsi in un ecosistema informativo strutturalmente saturo di contenuti sintetici.

Questo potere può inoltre essere esercitato in forme anche più radicali, come la rimozione, a scapito della trasparenza. Tra le misure prospettate dall'articolo 35 figura, infatti, anche «l'adeguamento dei sistemi algoritmici, compresi i sistemi di raccomandazione». In assenza di vincoli stringenti, nulla impedisce alle piattaforme di optare per strategie di *shadowbanning* per contrastare la viralità dei *deepfake*. Agire sulla limitazione della visibilità anziché sulla consapevolezza dell'utente finisce per configurare una forma di censura occulta.

Un esempio concreto aiuta a comprendere come tutto questo si traduca nella realtà. Nel *DSA Risk Assessment Report* (2024) presentato da TikTok, la piattaforma sostiene di applicare una politica mista di rilevamento automatico, segnalazione da parte degli utenti e rimozione dei *deepfake*. Si afferma che quasi sei milioni di video sono stati etichettati come *AI-generated* nell'ultimo anno[60].

Il report non specifica, tuttavia, quanto queste etichette siano effettivamente visibili, né quanto spesso raggiungano effettivamente gli utenti. L'impossibilità di fare affidamento sull'autodisciplina dei prestatori in assenza di obblighi cogenti è stata confermata da un'inchiesta di *Wired* che ha rivelato come il sistema di etichettatura di TikTok risulti totalmente inattivo nel Sud globale, lasciando le popolazioni di queste aree, spesso le più esposte a campagne di disinformazione, prive di qualsiasi forma di protezione informativa[61].

L'assenza di un obbligo di etichettatura chiaro e uniforme, unita all'eccessiva discrezionalità lasciata alle *Big Tech*, rischia di trasformare uno strumento di protezione in un meccanismo che deforma ulteriormente la qualità del dibattito pubblico.

Il diritto costituzionale pare essere entrato in un vicolo cieco: continuare a occuparsi dei singoli contenuti artificiali dannosi non attenua il loro potenziale rischio sistemico, ma il tentativo di regolamentare i *deepfake* come categoria giuridica autonoma stringe il dibattito pubblico in una morsa tra due mali costituzionali.

Da un lato, assoggettarli alla categoria dei contenuti illeciti risulta manifestamente sproporzionato e in contrasto con la libertà di espressione: come illustrato *supra*, l'estensione del regime dei contenuti illegali a tutti i contenuti sintetici non supererebbe il vaglio di proporzionalità della Corte di Giustizia. Dall'altro, la via alternativa percorsa dal legislatore europeo – quella della trasparenza attuata dall'etichettatura – non è immune da vizi.

Sul versante dell'*AI Act*, gli obblighi di marcatura dei contenuti sintetici si rivelano estremamente parziali.

Chatila, et al., *Generative AI models should include detection mechanisms as a condition for public release*, in *Ethics and Information Technology*, 25, 2024, 4.

[60] TikTok, *DSA Risk Assessment Report (2024)*, in www.tiktok.com/transparency/en/dsa-transparency/.

[61] V. Elliott, *TikTok Lite Leaves up to 1 Billion Users With Fewer Protections*, in *wired.com*, 23/07/2024, online a: https://www.wired.com/story/tiktok-lite-global-south/.

La previsione di cui all'articolo 50.4 pare riferibile alla leggibilità meccanica dei metadati, invisibili per loro natura a chi fruisce del contenuto e pertanto non attenua in alcun modo il fenomeno di «erosione della verità» che mette in crisi la fiducia nelle istituzioni[62].

L'impianto del DSA, invece, si presenta come astrattamente in contrasto con la *rule of law* intesa in senso sostanziale[63]: nel dare alle piattaforme digitali di dimensione molto grandi – pur con la possibile collaborazione del giudice[64] – il potere di identificazione e successiva mitigazione dei rischi sistemici, inclusi quelli derivanti dai *deepfake* (artt. 34 e 35 DSA, *supra*), il legislatore ha di fatto abdicato alla responsabilità pubblica di definire con norma sostanziale i confini del discorso lecito, trasferendola ad algoritmi opachi e meccanismi di *governance* privata. Questo fenomeno di privatizzazione della censura si scontra con il regime di tutela dei diritti fondamentali sanciti dall'articolo 10 della CEDU e dall'articolo 11 della Carta di Nizza.

Come anticipato, infatti, ciascuna ingerenza alla libertà di espressione deve essere, per la Corte EDU, «prevista dalla legge», perseguire uno degli scopi legittimi indicati nel secondo paragrafo dell'articolo 10 e «necessaria in una società democratica».

Il sistema delineato dal DSA non soddisfa compiutamente nessuna di queste tre condizioni nel dominio specifico dei contenuti sintetici.

Non soddisfa il requisito della «previsione legale» in senso sostanziale, perché il legislatore ha preferito procedimentalizzare i meccanismi di moderazione[65] anziché dettare regole sostanziali che definiscano con sufficiente precisione le condotte vietate e i criteri di intervento. Non soddisfa il requisito di necessità, perché la discrezionalità lasciata alle piattaforme nella valutazione dei rischi sistemici e nella scelta delle misure di mitigazione eccede quanto necessario per il perseguimento degli scopi legittimi invocati. Non soddisfa, infine, il requisito democratico, perché le limitazioni alla libertà di espressione non discendono da un atto del potere pubblico responsabile e sindacabile, ma l'autoregolazione degli operatori economici prevale – per una combinazione di necessità tecnologica, rispetto del diritto all'iniziativa economica privata e rapporti di forza – sull'esigenza di autoregolazione per la tutela dei diritti[66]. Infatti, in concreto, le piattaforme filtrano,

---

[62] È la cd. «*truth decay*» ovvero la progressiva perdita di rilevanza dei fatti e dell'analisi oggettiva nel dibattito pubblico, a favore di opinioni, emozioni ed esperienze personali. Il concetto è stato formalizzato e studiato approfonditamente dalla RAND Corporation, un istituto di ricerca statunitense, per analizzare la disinformazione dilagante e la polarizzazione politica. Le sue caratteristiche sono: confusione su fatti e interpretazioni degli stessi, confini labili tra opinione e dato, diminuzione della fiducia nei confronti delle istituzioni, dei media tradizionali e delle fonti autorevoli e proliferazione dei canali di informazione. Cfr. J. Kavanagh, M. Rich, *Truth Decay,* Stati Uniti, 2018.

[63] Alcuni studiosi – come Fuller o Raz – adottano una definizione «formale» della *rule of law*; altri, tra cui Bingham e Cass, preferiscono definirla in termini sostanziali o «pesanti». Nel primo caso, l'attenzione si concentra sulle fonti e sulle forme della legge. D'altro canto, chi interpreta la *rule of law* in maniera sostanziale include anche requisiti sul contenuto della legge: è una guida e un limite per gli attori pubblici quando interferiscono con diritti e libertà. Tom Bingham, nel suo libro *The Rule of Law*, con riferimento alla definizione formale, sostiene «La respingerei categoricamente a favore di una definizione «spessa», che includa la protezione dei diritti umani nel suo ambito. A mio parere non si può ritenere che uno Stato che reprime o perseguita ferocemente una parte del suo popolo rispetti la *rule of law*». Cfr. J. Waldron (voce), *The Rule of Law*, in E.N. Zalta, U. Nodelman (a cura di), *The Stanford Encyclopedia of Philosophy*, Stanford, 2023.

[64] Il DSA prevede infatti che la segnalazione o l'ordine di rimozione dei contenuti illeciti possa arrivare dall'autorità giudiziaria (art. 5, co. 2 e art. 9, co. 1).

[65] *Supra*, nota 39.

[66] N. Pica, *La tutela della libertà di informazione nel Digital Services Act tra contrasto alle manipolazioni algoritmiche e limiti alla content moderation,* in *MediaLaws – Rivista di diritto dei media,* spec. 1, 2024, 53 ss.

declassano o rimuovono i contenuti basandosi unilateralmente sui propri «Termini di Servizio» anziché sui limiti previsti dalla legge[67].
Il risultato è che ciò che è consentito alle piattaforme «fare» – moderare, etichettare, oscurare – prevale su ciò che è consentito agli utenti «dire», in contrasto con l'orientamento europeo secondo cui le libertà economiche fondamentali andrebbero interpretate «quanto più possibile nel senso di non consentire alcun provvedimento che ecceda i limiti di un intervento lecito sui diritti fondamentali in questione e di non permettere quindi alcuna misura che non si concili con i diritti fondamentali»[68].
Né risulta dirimente la previsione che i meccanismi di risoluzione extragiudiziale delle controversie facciano «salvo il diritto del destinatario del servizio di avviare, in qualsiasi fase, procedimenti per contestare tali decisioni dinanzi a un organo giurisdizionale» (art. 21, co. 1 DSA): l'accesso alla giustizia ordinaria, successivo e residuale rispetto alla moderazione privata, non compensa la violazione degli obblighi positivi dello Stato (riconosciuti dalla CEDU) di predisporre un quadro legislativo e amministrativo idoneo a tutelare e assicurare un pluralismo effettivo dell'informazione[69], né soddisfa le esigenze di *Drittwirkung* dei diritti fondamentali nel rapporto tra privati[70].
I rischi si proiettano dal piano individuale sui sistemi di democrazia rappresentativa e sui presupposti del dibattito pubblico deliberativo.
Se la legittimità delle decisioni politiche e la formazione dell'indirizzo politico presuppongono un elettorato capace di orientarsi autonomamente tra opinioni diverse partendo da una base fattuale condivisa, l'incoerenza del *framework* normativo attuale scardina questa premessa.
Il quadro che emerge è, infatti, quello di un ordinamento che ha scelto la strada della trasparenza senza dotarla degli strumenti per renderla effettiva. L'assenza di un obbligo di marcatura cogente negli atti europei amplifica il fenomeno della polarizzazione algoritmica: quando i *deepfake* e i contenuti sintetici non vengono intercettati ed etichettati, l'utente-elettore si trova immerso in bolle informative in cui il «voto libero» viene progressivamente svuotato di significato. Il votante diventa consumatore: le sue preferenze vengono tracciate e i suoi bisogni anticipati[71]; ciò si traduce in contenuti sintetici «cuciti su misura» per colpire le sue vulnerabilità cognitive e i suoi pregiudizi. L'atto del voto rischia di degradare da espressione di una scelta politica ponderata a mera reazione emotiva a stimoli artificialmente manipolati. A questo si aggiunge una discrezionalità delle piattaforme nella «mitigazione del rischio» derivante dai *deepfake* così ampia da poter sfociare in forme di rimozione eccessiva che soffocano il discorso legittimo, ovvero essere orientata a massimizzare logiche commerciali a scapito del pluralismo.
Finché il legislatore europeo non tradurrà il principio di trasparenza in obblighi sostanziali, uniformi e giustiziabili, il principio di sovranità popolare resterà esposto al rischio di essere

[67] O. Grandinetti, *Le piattaforme digitali come poteri privati e la censura online*, in *Rivista Italiana di Informatica e Diritto*, 1, 2022, 179.
[68] R. Torino, *Libertà d'impresa e diritto alla protezione dei dati personali*, in *MediaLaws – Rivista di diritto dei media*, 1, 2023, 68. L'autore richiama le conclusione della dell'Avvocato generale Stix Hackl nella causa C-36/02, §56, e pur ritenendo non contestabile «l'impostazione di fondo dei sistemi giuridici europei secondo cui la dignità umana è valore primario e insuperabile, […] da cui si fa discendere la primazia dei diritti della personalità sulle libertà economiche», sostiene la necessità di valorizzare maggiormente la libertà d'impresa e il legittimo interesse economico, invitando a non ostacolare l'innovazione tecnologica in nome di una super-protezione aprioristica dei dati personali.
[69] O. Grandinetti, *Le piattaforme digitali come poteri privati e la censura online*, cit., 181.
[70] Ivi, 179.
[71] L. Califano, *Brevi riflessioni su privacy e costituzionalismo al tempo dei big data,* in *Federalismi.it,* 9, 2017, 5.

svuotato dall'interno, con cittadini chiusi in realtà fittizie che impediscono il confronto tra le diverse idee e opinioni presenti nella società[72].

5. Modelli alternativi di marcatura dei contenuti digitali. L'inefficacia degli attuali strumenti basati sull'etichettatura discrezionale impone di esplorare modelli alternativi di governo dei contenuti sintetici.

Una prima linea d'intervento, che resterebbe compatibile con l'assetto attuale dell'*AI Act*, potrebbe concretizzarsi all'interno dei Codici di condotta previsti dall'articolo 56, i quali, se validati dalla Commissione, possono assumere efficacia generale. Una proposta avanzata da A. Knott *et al.* suggerisce un obbligo mirato per i *foundation model*: rendere meccanismi di rilevamento dei contenuti generati come un prerequisito, una condizione di esistenza per la loro messa a disposizione sul mercato[73]. Un vincolo di questo tipo rappresenterebbe un primo passo verso una maggiore tracciabilità dei contenuti sintetici, agendo a monte del ciclo di diffusione.

L'etichettatura di tutte le creazioni artificiali continua, tuttavia, a presentare delle criticità.

La prima complicazione riguarda le limitazioni tecniche del *labeling*: difficoltà nell'inserimento di marcatori, assenza di standardizzazione e potenziali pregiudizi nei confronti dei non madrelingua. I sistemi di rilevamento, inoltre, spesso generano falsi positivi, classificando erroneamente testo umano come prodotto da intelligenza artificiale[74]. I *watermark*, infine, sono vulnerabili a manipolazioni, rimozioni o attacchi di *spoofing*, con il rischio di comprometterne l'affidabilità e di danneggiare la reputazione degli sviluppatori[75].

In secondo luogo, non è da escludere che un obbligo generalizzato di filigrana sui contenuti sintetici venga percepito come una forma indiretta di restrizione della libertà di espressione. In un quadro come quello europeo, in cui l'articolo 10 della CEDU – nella lettura costante della Corte EDU e della Corte di Giustizia – tutela non solo il contenuto, ma anche i mezzi di espressione, anche la semplice marcatura sistematica potrebbe apparire come un ostacolo all'esercizio di questa libertà non rispettoso del principio di proporzionalità. L'idea che ogni contenuto sintetico debba portare un marchio potrebbe essere interpretata come una forma di censura preventiva.

Infine, è necessario essere realistici: in un ambiente informativo dove i contenuti sintetici sono già la maggioranza – e destinati a crescere – ogni tentativo di regolare singolarmente ciò che è artificiale rischia di risultare un compito di Sisifo.

Da questa constatazione potrebbe derivare una proposta controintuitiva, ma potenzialmente più pragmatica: anziché continuare a inseguire l'identificazione dell'artificiale, si potrebbe concentrarsi sul segnalare ciò che è reale. Non si tratterebbe di sostituire l'etichettatura dei contenuti sintetici, ma di integrarla, con un sistema speculare che si fondi sulla verifica e certificazione dell'autenticità.

Si potrebbe sfruttare il «Dividendo del bugiardo» e la presunzione di falso fino a prova contraria, concentrandosi invece sulla prova contraria: decidere di segnalare, al contrario, ciò che è vero.

---

[72] C. Casonato, *Costituzione e intelligenza artificiale: un'agenda per il prossimo futuro*, in *BioLaw Journal. Rivista di BioDiritto*, 2, 2019, 715.

[73] *«We propose that any organization that develops a LLM intended for public use should be required by law to demonstrate a reliable detection tool for the content the model generates, as a condition for its release to the public. After release, the detection tool should be freely available to the public.»* Cfr. A. Knott, D. Pedreschi, R. Chatila, *et al.*, *Generative AI models should include detection mechanisms as a condition for public release*, cit., 3.

[74] Per una panoramica tutti i limiti dell'etichettatura dei contenuti cfr. T. Madiega, *Generative AI and watermarking*, in *EPRS | European Parliamentary Research Service*, 2023, 3 ss.

[75] E. Hine, L. Floridi, *New Deepfake Regulations in China Are a Tool for Social Stability, but at What Cost?,* cit., 608.

L'idea è quella di affiancare ai meccanismi di contrasto del falso un modello speculare e complementare, volto a valorizzare ciò che è verificato e attendibile, attraverso la creazione di uno spazio informativo qualificato. Mediante il *fact-checking*, il contenuto (testuale, visivo, audiovisivo o sonoro) potrebbe essere accompagnato da un'indicazione visibile di autenticità, idonea a segnalare all'utente che l'informazione è stata sottoposta a un processo di verifica secondo criteri pubblici e trasparenti.

Un simile strumento non configurerebbe un meccanismo di verità imposta dall'alto, né l'istituzione di una verità di Stato, ma un modello procedurale fondato sulla tracciabilità e sulla verifica, nel rispetto della libertà di espressione e del pluralismo informativo.

Il dibattito francese sulla *labellisation* dimostra tuttavia quanto sia delicato questo terreno. La proposta avanzata dal Presidente Macron di introdurre un'etichettatura dei media ritenuti *fiables* – ovvero affidabili – ispirata ai criteri della *Journalism Trust Initiative* promossa da *Reporters Without Borders*[76], è stata immediatamente criticata come il tentativo di istituire un «Ministero della verità» e paragonata alla *Pravda* sovietica[77].

A livello europeo, non sarebbe necessario avviare un nuovo processo regolatorio o introdurre un'inedita autorità di certificazione dell'informazione. Un sistema di *fact-checking* orientato alla valorizzazione dell'autenticità potrebbe innestarsi sull'architettura già prevista dal *Digital Services Act*, che già delinea un modello di governance distribuita dell'informazione, fondato su una pluralità di attori e di garanzie procedurali.

I meccanismi di segnalazione da parte degli utenti, il riconoscimento di segnalatori attendibili, obblighi di trasparenza e sistemi di audit indipendenti e la responsabilizzazione graduata delle piattaforme previsti dal DSA e oggi orientati principalmente alla rimozione dei contenuti illeciti e dannosi, potrebbero essere ricalibrati per assolvere a una funzione costruttiva: individuare e rendere visibile l'informazione verificata e affidabile.

L'obiettivo è ripristinare un ruolo sociale degli «intermediari della cultura»: un'intermediazione plurale e collettiva in cui, anziché a decidere cosa debba essere escluso dallo spazio pubblico, si è chiamati a ricostruire l'«enciclopedia comune» imprescindibile per il funzionamento del dibattito pubblico[78].

Ne deriverebbe un ecosistema informativo articolato su tre livelli: una prima area di contenuti contrassegnati come artificiali; una zona intermedia di contenuti non etichettati; e una terza area, fondata su procedure condivise di verifica, in cui i contenuti siano dichiaratamente autentici. Questo meccanismo consentirebbe condizioni procedurali minime per distinguere i fatti dalle opinioni, nel rispetto del paradigma deliberativo.

Si potrebbe così ripristinare un nucleo minimo di realtà condivisa, non imposto dall'alto ma risultante da processi collettivi, verificabili e trasparenti: un ancoraggio fragile ma necessario per la tenuta della sfera pubblica e, con essa, delle democrazie.

Abstract. La regolazione giuridica dei deepfake è oggi fondata su un approccio centrato sui singoli contenuti, che interviene solo quando essi integrano fattispecie già illecite o comunque dannose.

[76] Una certificazione, lanciata nel 2021 dall'organizzazione di monitoraggio dei media Reporter Senza Frontiere, che si concentra su come vengono prodotti i contenuti giornalistici e sull'etica che circonda questo processo, piuttosto che giudicare i singoli contenuti. I criteri di certificazione sono stati sviluppati da un comitato di 130 esperti, tra cui giornalisti, varie istituzioni, enti regolatori, editori e membri dell'industria tecnologica.

[77] E. Nilsson-Julien, *Macron's plan to 'label' trustworthy media manipulated and taken out of context*, in *euronews.com*, 05/12/2025, online a: https://www.euronews.com/my-europe/2025/12/05/macrons-plan-to-label-trustworthy-media-manipulated-and-taken-out-of-context.

[78] J. M. Balkin, *How to Regulate (and Not Regulate) Social Media (November 8, 2019),* in *Journal of Free Speech Law,* 71, 2021, 78.

Questo modello, recepito anche dal *Digital Services Act*, è tuttavia insufficiente di fronte all'impatto sistemico della proliferazione dei contenuti sintetici, che destabilizza la sfera pubblica e compromette le condizioni stesse del dibattito democratico. Poiché l'estensione del regime dei contenuti illegali risulterebbe incompatibile con la libertà di espressione, i *deepfake* vengono collocati nell'area dei contenuti *awful but lawful* e gestiti attraverso obblighi di mitigazione del rischio e trasparenza.
Il *framework* normativo proposto dal DSA e dall'*Artificial Intelligence* risulta però poco efficace. Si propone, pertanto di integrare la trasparenza sull'artificiale con sistemi di certificazione e *fact-checking* volti a valorizzare l'autenticità dei contenuti, al fine di ricostruire l'«enciclopedia comune» necessaria alla tenuta delle democrazie liberali.

Abstract. The legal regulation of deepfakes is currently based on a case-by-case approach, which intervenes only when such material falls within pre-existing categories. This model, also adopted by the Digital Services Act, doesn't appear adequate to tackle the systemic impact of the mass proliferation of synthetic content, which destabilises the public sphere and undermines the very conditions of democratic debate.
Extending the regime applicable to illegal content doesn't appear compatible with freedom of expression. Deepfakes should be overall placed within the category of awful-but-lawful content and addressed through risk-mitigation and transparency obligations.
However, the framework for such content established by the DSA and the Artificial Intelligence Act proves largely ineffective. The article therefore proposes complementing transparency about the artificial nature of content with systems of certification and fact-checking aimed at enhancing the visibility of authentic information, in order to rebuild the «shared epistemic base» necessary for the functioning of liberal democracies.

Parole chiave. Deepfake – Digital Services Act – Contenuti awful-but-lawful – Sfera pubblica – Disinformazione.

Key words. *Deepfake – Digital Services Act – Awful but lawful content – Public sphere – Disinformation.*