LA TIRANNIA DEGLI ALGORITMI E LA LIBERTÀ DI MANIFESTAZIONE DEL PENSIERO

Lo stato dell'arte e le prospettive future

FULVIA ABBONDANTE

i-lex

i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale
Rivista quadrimestrale on-line: www.i-lex.it
Dicembre 2019
Fascicolo 12, 1-3
ISSN 1825-1927

i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale
Rivista quadrimestrale on-line: www.i-lex.it

LA TIRANNIA DEGLI ALGORITMI E LA LIBERTÀ DI MANIFESTAZIONE DEL PENSIERO
Lo stato dell'arte e le prospettive future

FULVIA ABBONDANTE*

Abstract: La letteratura distopica e fantascientifica si è spesso occupata del difficile rapporto fra verità e menzogna e dei suoi effetti sulla democrazia, sulla libertà degli individui e sulla manipolazione dell'informazione. Proprio l'avvento della rete sembra aver realizzato e materializzato le paure e pericoli che venivano mirabilmente descritti da numerosi autori di mondi futuri. Le questioni che oggi si pongono all'attenzione del giurista sono divenute particolarmente complesse sia per la difficoltà di 'comprendere' una realtà così articolata come il web sia per la rapidità con cui le innovazioni tecnologiche irrompono sulla scena. I tentativi, soprattutto negli ultimi anni, di disciplinare fenomeni come hate speech e fake news – soprattutto da parte dell'Unione Europea – si sono rivelati fallaci e per di più rischiosi poiché affidano il controllo sulla bontà e verità dei discorsi che circolano online alle piattaforme. Queste ultime al fine di evitare possibili sanzioni o l'introduzione di una 'hard regulation' utilizzano Intelligenza Artificiale che garantisce risultati solo quantitativi con un'evidente lesione della libertà di pensiero e di informazione. I recenti sviluppi della tecnologia hanno raggiunto peraltro livelli di complessità come dimostrato dalla messa a punto di forme di Intelligenza Artificiale in grado di produrre deep fake basate sull'uso di algoritmi grazie ai quali è possibile creare audio e video di persone reali a cui si fa dire o fare cose che non hanno mai fatto o detto determinando così, una contraffazione totale del dibattito elettorale, incidendo sul diritto di voto nella sua fondamentale esplicazione costituita dagli elementi della libertà e consapevolezza degli elettori. Quali le possibili soluzioni per impedire uno scenario così inquietante? Probabilmente ancora una volta i legislatori delegheranno compiti di controllo e di rimozione alle piattaforme non diversamente da quanto accade per le fake e gli hate speeches. Forse la soluzione non va cercata esclusivamente nella regolazione ma nel (ri)appropriarsi di un vocabolario che attinga dalla filosofia, dalla letteratura, dall'etica nuova linfa in grado di fornire chiavi di letture più ampie e più profonde dei fenomeni: prima umani e poi tecnici.

* Università di Napoli Federico II.

Parole chiave: libertà di manifestazione del pensiero; algoritmi; deep fake; libertà di voto.

> Eccolo lì, seduto in posa solenne alla sua grossa scrivania di quercia con la bandiera americana alle spalle. A Mosca, dove esisteva un duplicato di Megavac 6-v, c'era un altro Sim identico a quello, con dietro la bandiera sovietica. Per il resto ogni particolare (gli abiti, i capelli grigi, l'espressione rassicurante, paterna, matura e militaresca, il mento deciso) corrispondeva: i Sim erano stati costruiti ambedue in Germania, contemporaneamente, e i cavi erano stati collegati dai migliori tecnici disponibili fra gli uomini-Yance. E lì gli addetti alla manutenzione erano in costante attività: gli bastava socchiudere gli occhi ormai addestrati e coglievano ogni minimo segno di malfunzionamento, anche l'esitazione di una frazione di secondo. Qualsiasi cosa potesse abbassare il livello qualitativo richiesto, quello dell'autenticità più semplice e totale. Questo simulacro, fra i tanti di cui si occupavano gli uomini-Yance, esigeva la più grande somiglianza con la realtà che imitava.
> (P. K. Dick, *La penultima verità*, Fanucci editore, 2016)

1. Introduzione

L'evoluzione tecnologica ha determinato un mutamento significativo nel rapporto tra fruitore di notizie e ricerca e conseguimento delle stesse. Oggi non è il cittadino a cercare l'informazione ma è la notizia a cercare il cittadino/utente. Questo fenomeno è stato reso possibile proprio grazie alla combinazione di algoritmi in grado di profilare il cibernauta – attraverso lo sfruttamento di cd. 'big data' – e dunque di fornirgli le informazioni individuate sulla base dei propri gusti e preferenze[1].

Come è stato opportunamente osservato quello che oggi accade è "*un effetto di inscatolamento del nostro mondo informativo, di costruzione dei mondi di vita a nostra immagine e somiglianza offrendo a ognuno ciò che gli interessa*"[2].

[1] Sulla correlazione fra profilazione e informazione la letteratura è ampissima. Senza pretesa di esaustività si segnala il classico E. Pariser, *Il Filtro. Quello che Internet ci nasconde*, Milano, 2012; G. Pitruzzella, *La libertà di informazione nell'era di Internet*, in *MediaLaws*, 1, 2018, pp. 19 ss; M. Bianca, *La filter bubble e il problema dell'identità digitale*, in *MediaLaws*, 2, 2019, specificamente pp. 42-46.

[2] M. Calise, F. Musella, *Il Principe Digitale*, Laterza, 2019, p. 11.

Ma non basta. La diffusione dei 'social network' ha fatto sì che l'utilizzatore di Internet è egli stesso produttore di notizie e informazioni e, nel medesimo tempo, diffusore di 'news' e informazioni veicolate da terzi ma, come si diceva, tagliate su misura del cibernauta. Così che il 'web' da strumento di pluralismo informativo si è trasformato in un 'medium' che produce effetti distorsivi sulla circolazione delle idee trasformando il dibattito pubblico – che si alimenta e nutre dei diversi punti di vista – in asfittico e polarizzato[3]. Notizie distorte e discorsi incitanti all'odio, pur essendo manifestazioni diverse, producono un esito circolare nel senso che la disinformazione quasi sempre diviene il catalizzatore anche del discorso incitante all'odio e viceversa[4].

Si tratta di fenomeni peraltro antichi ma amplificati dall'avvento della 'Net'. Da un punto di vista strettamente giuridico le stesse definizioni di hate speech e di notizia falsa sono alquanto complesse e la stessa regolazione offline è variegata in ragione del modello 'culturale' prescelto per garantire la libertà di pensiero. Il diritto costituzionale interno delle democrazie liberali – ma anche numerosi documenti internazionali – considerano, infatti, tali manifestazioni del pensiero – salvo alcune e limitate eccezioni – lecite e incensurabili con una gradazione che appunto varia dalla ampissima tolleranza riconosciuta negli Stati Uniti, di cui si dirà meglio in seguito, a una visione che presenta numerose sfumature all'interno dei singoli paesi europei.

Negli ultimi anni la propagazione incontrollata di queste forme di pensiero hanno determinato un allarme sia negli utenti sia nei legislatori nazionali e non.

Tali fenomeni sono stati resi possibili dallo sviluppo e l'uso dell'Intelligenza Artificiale: a 'monte' nella fase di diffusione delle

[3] C.R. Sunstein, *Republic.com 2.0*, Princeton 2007 e di recente si v. Conference e-book by the European Centre for Press and Media Freedom (ECPMF), *Promoting dialogue between the European Court of Human Rights and the media freedom community. Freedom of expression and the role and case law of the European Court of Human Rights: developments and challenges*, 2017, *https://ecpmf.eu/files/ecpmfecthr_conference_e-book.pdf*, p. 5.

[4] F. Pizzetti, *Fake news e allarme sociale: responsabilità, non censura*, in M*ediaLaws*, 1, 2017, p. 50.

'fake news'[5] e 'hate speeches' – attraverso l'utilizzo dei cd. bot – e in quello a 'valle' della rimozione del contenuto odioso o falso.

2. Come l'algoritmo intercetta fake news e hate speech: la regolazione europea

L'utilizzo degli algoritmi nasce come risposta alla regolazione dei contenuti illegali veicolati sulle piattaforme 2.0 – affidata alla logica esclusivamente contrattuale e privatistica dei 'Terms of Contract' realizzati unilateralmente dai 'providers' e vincolanti per gli utilizzatori del servizio.

Gli intermediari, attraverso meccanismi di filtraggio, provvedono alla rimozione dei materiali caricati dal cibernauta contrari alla 'policy' dell'azienda. Sin dalle origini, peraltro, questi sistemi automatici da un punto di vista tecnico, presentano un'alta probabilità di produrrei cd. falsi positivi o negativi essendo l'identificazione del contenuto avulso da un processo valutativo in quanto limitato all'individuazione di una parola o di un'immagine e non al contesto in cui l'uno o l'altra sono immerse.

D'altro canto, trattandosi di una disciplina interna e connessa alla libertà di impresa ciascuna piattaforma definisce secondo propri standard l'espressione 'illegal content' che dunque variabile in base al significato che ciascun provider attribuisce a tale espressione creando, così, un'evidente disparità di trattamento fra utenti, oltre che un'incertezza complessiva su ciò che è lecito o illecito nell'ambiente digitale.

[5] Il termine fake news generalmente usato per indicare la diffusione di notizie false in realtà descrive una serie di fenomeni che hanno natura molto diversa tra loro: notizie satiriche, parodie, fabbricazione di notizie false, manipolazione, pubblicità e propaganda divenendo così un concetto ambiguo poiché accomuna ipotesi di manifestazione del pensiero legali da quelle invece che tendono ad alterale il dibattito pubblico. Recentemente l'High level Group on fake news and online disinformation, *A multi-dimensional approach to disinformation*, 2018, p. 10, ha proposto di utilizzare il termine disinformazione che indica appunto tutte quelle forme di notizie false, inaccurate o ingannevoli ideate, presentate e promosse per causare intenzionalmente pericolo pubblico o profitto. Condivide in dottrina tale impostazione O. Pollicino, E. Bietti, *Truth and deception across the atlantic: a roadmap of disinformation in the US and Europe*, in *Italian Journal of Public Law*, 2019, pp. 49-51.

Questa prima fase è stata condizionata dall'appartenenza 'Big Companies' all'ordinamento statunitense tendenzialmente contrario a qualsiasi forma di 'regulation' pubblica sia per ragioni connesse sia all'ampiezza del 'freedom of speech' sia per motivi di natura economica. Qualunque forma di intervento regolatorio avrebbe, infatti, finito per limitare l'innovazione tecnologica in fase nascente. Una posizione peraltro avallata nel tempo dalla giurisprudenza non solo della Corte Suprema ma anche dalle Corte Federali che in base ad una rilettura della Setino 230 (1) lettera a del 'Communication Decency Act' (d'ora innanzi CDA) ha interpretato il concetto di irresponsabilità del 'provider' in modo ampio e dunque favorendo la circolazione di qualsiasi contenuto nel 'web'[6].

Il modello di autoregolazione è peraltro implicito nella stessa 'Section 230' del CDA poiché tanto nella lett. b) quanto nella lett. c) viene sancita l'immunità per gli intermediari che, in buona fede, rimuovano contenuti leciti in eccesso o invece consentano la permanenza di materiale illecito (producendo quello che nella dottrina statunitense identifica con il nome di 'over-screening' e di 'underscreening').

Differente invece l'approccio europeo che non ha nel proprio DNA la concezione illuministica e soprattutto taumaturgica del 'marketplace of ideas' di matrice statunitense e anzi ha sempre manifestato un certo sospetto e diffidenza del mezzo Internet che traspare, soprattutto, dalla giurisprudenza della Corte Europea dei Diritti dell'Uomo[7].

La diversa sensibilità fra i due modelli diviene peraltro evidente dal 2015.

La crescita dei fenomeni connessi alla piramide d'odio e disinformazione (rispetto a tale ultimo problema, ha generato preoccupazione soprattutto per l'uso che ne è stato fatto durante alcune delle vicende politiche europee, e non solo,più significative: Brexit prima di tutto ma anche tornate elettorali dei paesi membri) ha determinato un'attenzione crescente del decisore politico sovranazionale che ha operato incidendo non sulla eventuale sanzionabilità di alcune forme di pensiero – che resta appannaggio delle singole realtà costituzionali

[6] Così K. Citron, B. Wittes, *The internet will not break: denying bad samaritans section 230 Immunity*, in *Fordham Law Review*, 86, 2, pp. 406-411.

[7] O. Pollicino, *La prospettiva costituzionale sulla libertà di espressione nell'era di Internet*, in G. Pitruzzella, O. Pollicino, S. Quintarelli, *Parole e potere: Libertà d'espressione, hate speech e fake news*, Egea, 2017.

dei paesi membri ma sulla capacità di diffusione – possibile grazie all'avvento della rete ritenuta il fattore da cui scaturisce la dimensione di pericolo.

Protagonisti della lotta a tali fenomeni sono divenuti i fornitori dei canali di comunicazione sui quali gli utenti operano.

Probabilmente non poteva essere diversamente se il rapporto con i mezzi di diffusione non è più bilaterale e statico come per gli altri media ma è trilaterale e dinamico nel senso che la multimedialità e l'interattività richiedono un soggetto terzo – le piattaforme appunto – che rendano possibile la comunicazione ad alto grado di complessità. Insomma, un maggiore coinvolgimento di quei soggetti che nell'architettura della rete rappresentano lo snodo fondamentale per la creazione e poi la propagazione di qualunque forma di pensiero.

Il primo, in ordine di tempo, è stato il 'Code of Conduct on countering illegal hate speech'[8] sottoscritto dalle più importanti più importanti piattaforme le quali si sono impegnate a rimuovere contenuti che incitano all'odio e al terrorismo entro le 24 dopo aver ricevuto una segnalazione dall'utente. Si tratta di una regolazione 'soft' che rientra in una cooperazione fra pubblico e privato che si traduce 'voluntary delegate enforcement'[9].

Di poco successiva è la 'Communication Tackling Illegal Content Online'[10] dove appunto è anche previsto, fra i contenuti illegali, le 'fake news'. I punti chiave di tale nuovo documento sono: 1) la predisposizione di misure proattive con lo scopo di prevenire la diffusione di 'illegal content' con l'uso di algoritmi; 2) procedure di 'stay down' per evitare il 're-upload' dei contenuti già rimossi; 3) il divieto di monitoraggio generalizzato

[8] Commission, *Code of Conduct on Countering Illegal Hate Speech Online*, 2016 in *http://ec.europa.eu/justice/fundamentalrights/files/hate_speech_code_of_conduct_en.pdf* .

[9] G.F. Frosio, *The Death of 'No Monitoring Obligations': A Story of Untameable Monsters*, in '*Journal of Intellectual Property, Information Technology and E-Commerce Law JIPITEC*', 8, 3, 2017, https://www.jipitec.eu/issues/jipitec-8-3-2017/4621/#N1030D.

[10] Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, *Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms*, 2017.

In questo caso per agevolare la rimozione rapida è evidente che le piattaforme utilizzino algoritmi con un'altissima probabilità di eliminare ciò che è lecito e lasciar circolare ciò che è illecito senza possibilità di poter contestare la decisione del fornitore di contenuti

Infine, la Comunicazione 'Tackling online disinformation: a European approach' e il conseguente il 'Code of Practice on Disinformation'[11] sono invece più specificamente dedicati alla disinformazione ed emerge una maggiore attenzione alla individuazione del termine di 'fake news'. Queste ultime sono definite infatti, esclusivamente le notizie verificabilmente false o fuorvianti create, presentate e diffuse per motivi economici e con l'obiettivo di disinformare.

Appare dunque evidente che nel tratteggiare il contenuto ritenuto illecito la Commissione abbia voluto far riferimento esclusivamente alla notizia mentre restano fuori dalla 'non binding regulation' la satira, gli errori giornalistici e la comunicazione politica. Tale ultima precisazione è apparsa a molti opportuna se non altro perché evita la possibilità concreta che soggetti privati operino in maniera non trasparente e su pressione dei Governi, attuando forme di censura privata del dissenso e del pensiero politico in generale[12].

La definizione peraltro non esaurisce la caratterizzazione delle 'fake news' oggetto di eventuale intervento da parte delle piattaforme perché deve arrecare minaccia alla politica democratica e a beni pubblici dell'UE quali la protezione della salute, l'ambiente o la sicurezza dei cittadini europei definizione alquanto ampia e generica che si presta ovviamente a una lettura estremamente discrezionale da parte dei providers. Gli strumenti indicati nel Codice attraverso un rinvio alla Comunicazione sono di natura molto diversa e tendono per così dire ad agire su più aspetti del problema: dalla necessità di migliorare i meccanismi di controllo nell'assegnazione della pubblicità, penalizzando quindi quei siti che propagano 'fake', a investimenti in tecnologie per la ricerca e l'indicizzazione di notizie affidabili l'incentivazione di meccanismi che favoriscano il pluralismo a scapito della polarizzazione al *favor* per le attività di 'fact checking'.

[11] Commission, *EU Code of Practice on Disinformation*, in https://ec.europa.eu/.

[12] M. Monti, *Il Code of Practice of Disinformation dell'UE: tentativi in fieri di contrasto alle fake notizie*, in *MediaLaws*,1, 2019, p. 321.

Appare evidente peraltro che come già in passato per la regolazione dei contenuti odiosi online tali forme di regulation avallano l'uso dell'Intelligenza Artificiale.

Tanto per fare un esempio, Facebook, per adempiere al codice di condotta della UE, ha introdotto l'impiego dell'algoritmo Rosetta – dal nome della famosa Stele Egiziana – la cui particolarità è quella dunque di analizzare non le singole parole quanto di operare sull'intero testo e sulle immagini e i video che l'accompagnano riducendo il margine di errore rispetto al contenuto eventualmente rimosso o bloccato segnalato dagli utenti.

I rischi per la libertà di pensiero risiedono nel fatto che i meccanismi automatici di filtraggio, per quanto sofisticati, si sono dimostrati quantitativamente efficienti ma qualitativamente inefficaci poiché esposti a margini di errore superiori rispetto alla selezione effettuata dall'essere umano. Il linguaggio antropico è fatto di molteplici sfumature che i congegni tecnici non possono prendere in considerazione: i pregiudizi soggettivi, qualitativi, etici o ragionevoli sono ignorati dalle macchine. La valutazione sull'odiosità del discorso e sulla disinformazione non è, infatti, un processo meccanico che si ferma all'esame delle parole in sé e per sé considerate ma presuppone le numerose nuance di cui il linguaggio si compone.

In secondo luogo, al di là della retorica sulla neutralità dell'Artificial Intelligence quest'ultima è sempre un prodotto dell'uomo e, dunque, nella fase di predisposizione del processo informatico, gli algoritmi sono inconsapevolmente condizionati dal background di chi lo crea. Non solo ma spesso il materiale da cui queste 'machine learning' imparano derivano dalla rete e poi dall'interazione umana dunque con il bagaglio culturale che gli uni e gli altri si portano dietro che inevitabilmente ripropongono stereotipi e pregiudizi che vengono veicolati in maniera inconsapevole quando si procede alla rimozione dei testi[13].

Vero è peraltro che il 'General Data Protection Regulation' impone la restrizione all'uso degli algoritmi e di meccanismi automatizzati che riguardino gli individui e imponendo, dunque, l'intervento umano

[13] M.C. Carrozza, C. Oddo, S. Orvieto, A. di Minin, G. Montemagni, *AI: profili tecnologici. Automazione e autonomia: dalla definizione alle possibili applicazioni dell'intelligenza artificiale*, in *Biolaw journal*, 3, 2019, pp. 240-241.

ma le restrizioni previste e quando debba esservi l'azione dell'uomo non è peraltro molto chiaro[14].

3. …. E quella di alcuni Stati membri

Anche alcuni paesi europei si sono dotati di una specifica disciplina in tal senso. In Francia, ad esempio, è stata promulgata la 'Loi organique et loi ordinaire relatives à la manipulation de l'information', del 22 dicembre 2018 con la quale si intende combattere la manipolazione delle informazioni nell'era digitale e ad arginare la diffusione di notizie false durante i periodi della campagna durante i tre mesi precedenti le elezioni nazionali. Tale legge presuppone l'intervento del giudice che esamina la domanda di provvedimenti provvisori deve valutando, entro 48 ore, se tali informazioni false siano diffuse 'in modo artificiale o automatizzato' e 'massiccio'. Nella sua decisione del 20 dicembre 2018, il Consiglio costituzionale ha specificato che il giudice poteva interrompere la diffusione delle informazioni solo se si manifestavano inesattezze o natura fuorviante delle informazioni e il rischio di alterare la sincerità delle informazioni. Inoltre, le piattaforme digitali (Facebook, Twitter, ecc.) sono soggette agli obblighi di trasparenza quando distribuiscono contenuti a pagamento. Quelli che superano un determinato volume di connessioni al giorno devono avere un rappresentante legale in Francia e rendere pubblici i loro algoritmi. Il Consiglio superiore dell'audiovisivo (CSA) può anche prevenire, sospendere o interrompere la diffusione di servizi televisivi controllati da uno stato straniero o sotto l'influenza di questo Stato e pregiudicare gli interessi fondamentali della nazione[15].

[14] O. Pollicino, E. Bietti, *Truth and deception across the Atlantic: A roadmap of Disinformation In The Us And Europe*, in *Italian Journal of PublicLaw*, 11, 1, 2019.

[15] Per un ampio commento sulla legge e sulla pronuncia del Conseil constitutionnel si v. C. Magnani, *Libertà d'informazione online e fake news: vera emergenza?Appunti sul contrasto alla disinformazione tra legislatori statali e politiche europee*, in *forumcostituzionale.it*.

In Italia vi sono stati tentativi di regolamentare la diffusione delle 'fake news' peraltro tutti falliti[16].

Infine, la Germania ha approvato una legge severa che introduce una pesante sanzione economica a carico dei social network che abbiano almeno due milioni di iscritti e non provvedano alla predisposi-

[16] Nel 2017 il disegno di legge 'Gambaro', introduceva la punibilità, con sanzioni fino a 5.000 euro, per la diffusione tramite piattaforme informatiche di notizie 'false, esagerate o tendenziose'. La disciplina non si applicava ai 'soggetti e prodotti editoriali sottoposti a registrazione' ed escludeva in modo arbitrario sia le piattaforme social sia le testate registrate. Pertanto erano sottoposti al regime punitivo solo i nuovi siti o blog che per dichiarare l'avvio della propria attività dovevano registrarsi tramite invio di pec presso il Tribunale, con l'intento di tracciare gli attori della rete e combattere l'anonimato. Era peraltro prevista una pena detentiva, con la reclusione fino a 12 mesi qualora le notizie pubblicate destassero 'allarme pubblico'. Il controllo sulla veridicità delle notizie diffuse,doveva essere effettuata dagli operatori della rete, che, dunque avrebbero dovuto provvedere al monitoraggio delle pubblicazioni quali e la rimozione dei contenuti inattendibili. Critica la dottrina per i numerosi profili di contrarietà sia all'art. 21 Cost. sia alla disciplina europea sulla responsabilità dei provider. Si v. M. Bassini, G.E. Vigevani, *Primi appunti su fake news e dintorni*, in *MediaLaws*, 1, 2017, p.15; M. Cuniberti, *Il contrasto alla disinformazione in rete tra logiche di mercato e (vecchie e nuove) velleità di controllo*, in *MediaLaws*, 1, 2017, pp. 26-40; C. Melzi D'Eril, *Fake news e responsabilità: paradigmi classici e tendenze incriminatrici*, MediaLaws, 1, 2017, pp. 60-67. Nel dibattito italiano c'è chi ha proposto la creazione di un'Authority o l'intervento giudici, si v. sul punto G. Pitruzzella, *La libertà di informazione*, cit., in particolare pp. 13-14, deputati alla verifica di ciò che vero o falso ex post 'sulla base di principi predefiniti, intervengano successivamente, su richiesta di parte e in tempi rapidi, per far rimuovere dalla rete quei contenuti che sono palesemente falsi o illegali o lesivi dei diritti fondamentali e della dignità umana'. Rispetto alla prima ipotesi quella di un'Autorità indipendente molte le voci critiche fra i tanti N. Zanon, *Fake news e diffusione dei social media: abbiamo bisogno di un'"Autorità Pubblica della Verità"?*, in *MediaLaws*, 1, 2018, pp. 12-17; mentre la previsione di un controllo da parte del giudice è , senza dubbio, in linea con la riserva di giurisdizione e le garanzie di bilanciamento fra i diversi diritti in gioco ma non depotenzia l'effetto diffusivo incontrollato e velocissimo che è appunto l'elemento di pericolo tipico della Rete. Un'ulteriore proposta annunciata ma mai depositata sanzionava le piattaforme che non avessero rimosso nei termini previsti dal testo i contenuti ritenuti lesivi della dignità della persona e "*contro la Repubblica*", da cinquecentomila fino a cinque milioni di euro con l'obbligo di redigere un resoconto dettagliato ogni sei mesi delle rimozioni effettuate, visibile sulla homepage dei gestori; nonché l'introduzione di un'aggravante per i delitti a sfondo razziale.

zione di una serie di misure finalizzate alla rimozione dei contenuti illegali (tra cui 'hate speeches' e 'fake news') nel termine di ventiquattro ore. A garanzia del principio di trasparenza, comporta la pubblicazione di un report alla metà di ogni anno nel quale devono essere indicati:

a) gli sforzi effettuati dal provider per impedire la diffusione di contenuto offensivo sulla sua piattaforma;
b) i criteri e le procedure utilizzate per cancellare o rimuovere;
c) numero di denunce intervenute e le ragioni delle segnalazioni;
d) informazioni sull'organizzazione;
e) numero delle contestazioni ricevute e trasmesse a un gruppo di esperti terzi per preparare la decisione;
f) il tempo occorso (entro le 24h/48h/1 settimana) fra la segnalazione e la rimozione o il blocco del contenuto illegale;
g) le misure utilizzate per comunicare al denunciante e all'utilizzatore la decisione circa l'esito delle notifiche oggetto di verifica[17].

4. Le nuove frontiere dell'Intelligenza Artificiale: la democrazia in pericolo?

Questo lo stato dell'arte.

Peraltro, i mutamenti tecnologici operano in modo rapido e imprevedibile rendendo difficile l'opera di adeguamento e di riflessione del giurista. Di recente la messa a punto di forme di Intelligenza Artificiale in grado di creare 'fake' attraverso i cd. 'bot' e le 'deep fake' rappresentano una non meno difficile (se non quasi impossibile) sfida.

Da un punto di vista tecnico i 'bot' sono dei software che, accedendo alla Rete sfruttano gli stessi canali degli utenti in grado di

[17] Per un commento sulla legge tedesca si vedano M.R. Allegri, *Ubi social, ibi ius. Fondamenti costituzionali dei social network e profili giuridici della responsabilità dei provider*, Franco Angeli, 2108, pp. 203 ss.; nonché G. De Gregorio, *The market pace of ideas nell'era della post-verità: quali responsabilità per gli attori pubblici e privati online?*, in *MediaLaws*, 1, 2017, pp. 91-104.

svolgere i compiti più vari in maniera completamente autonoma. Tale tecnologia rappresenta una delle molteplici applicazioni delle 'machine learning' definite in tal modo perché sono in grado di apprendere dai loro errori ma soprattutto dalle loro interazioni con persone reali. Ciò permette di migliorare le capacità di analisi del linguaggio umano e fornire, così, risposte sempre più puntuali ed esatte.

Quanto alla diffusione di 'hate speeches' e 'fake news' responsabili della diffusione possono essere considerati le cd. 'chatbot' (servizi di messaggistica istantanea per automatizzare un certo tipo di comunicazione diffusione di notizie, ad esempio, o assistenza clienti) ovvero i 'bot social', ovvero dei profili falsi utilizzati su vari social network per fare volume online.

Più inquietante è poi la nuova tecnologia in grado di produrre 'deep fake' anch'essa basata sull'uso di algoritmi grazie ai quali è possibile di creare audio e video di persone reali a cui si fa dire o fare cose che non hanno mai fatto o detto. La tecnica è basata sulla comunicazione tra due algoritmi cd. 'generator adversarial network' (più noto con l'acronimo GAN): un algoritmo madre neurale produce un video completo di audio che è trasmesso a un altro algoritmo che replica il contenuto sulla base dei dati immagazzinati in precedenza. Dallo scambio e il dialogo continuo fra i due algoritmi in cui ogni invio corrisponde la correzione degli errori e difetti produce un prodotto sempre più indistinguibile dal reale[18].

Lo scenario che ne viene fuori sembra uscito dalla più cupa letteratura distopica di cui Philip Dick è stato l'esponente fra i più significativi e, per certi versi profetico, se in una sua opera meno nota, richiamata in esergo intravedeva la possibilità di realizzare meccanismi in grado di falsificare il reale a tal punto da rendere la linea di discrimine fra realtà e menzogna inesistente. Un panorama preoccupante per la nostra democrazia, già peraltro fortemente compromessa da fattori economici, sociali e di valori.

Infatti, se le fake news secondo la definizione data dal Codice citato è una notizia verificabile falsa o fuorviante nei casi di deep fake la verifica diventa oltremodo difficile.

[18] R. Chesney, D. Citron, *21st Century-Style* Truth *Decay: Deep Fakes and Privacy, Free Expression, and National Security*, in *Maryland Law Review* 78, 4, 2019.

La dottrina statunitense per prima si è interrogata sulle possibili ricadute che queste particolari espressioni del pensiero possono produrre essendo le deep fake non solo diffuse ma oggetto di attenzione da parte delle agenzie governative, privati e accademia già da tempo.

Nella pronuncia 'Alvarez'[19] la Corte Suprema ha ritenuto protetta la bugia ai sensi del I Emendamento, pur individuando un primo e timidissimo spiraglio nell'indicare la possibilità di censura di quelle menzogne che possano arrecare un 'riconoscibile pericolo legale' e dunque solo sulla base di un compelling interest introdurre una legislazione punitiva del falso.

La diffusione delle deep fake non è espressione solo negativa dell'uso tecnologia ma anche un mezzo per la realizzazione nuove forme d'arte e in particolare della satira ma diviene molto problematica se consideriamo quel particolare segmento del discorso pubblico dove viene in gioco il momento più significativo della democrazia rappresentativa: quello delle campagne elettorali.

In un futuro non molto lontano si potrebbe realizzare una contraffazione totale del dibattito elettorale, determinando non solo la lesione della reputazione dei candidati ma del diritto di voto nella sua fondamentale esplicazione costituita dagli elementi della libertà e consapevolezza degli elettori. La confusione tra vero e falso, realtà e finzione determinerebbe una quasi impossibilità per il votante di scegliere consapevolmente il candidato, di distinguere fra cioè fra quello che egli ha realmente detto e ciò che è frutto di una falsificazione operata da avversari politici o da terzi.

La preoccupazione ha spinto, già negli Stati Uniti alla presentazione di numerosi Bill che appunto tendono ad affrontare il problema da varie prospettive.

Nel 2018, ad esempio, il Senatore Benn Sasse ha presentato– il Malicious Deep Fake Prohibition Act – che appare essere più legge manifesto che avere una reale portata sul piano dell'effettività.

La punibilità del creatore di deep fake si scontra con la difficoltà di identificare l'autore, dalla transnazionalità (connaturata alla struttura della rete) che impedisce di individuare luogo e dunque la legge da applicare.

[19] Supreme Court of the United States, *United States v. Alvarez* (617) F. 3d 1198.

La possibilità di ancora una volta prevenire la propagazione di queste nuove forme di manipolazione sofisticata della verità verrebbe lasciata all'uso di I.A. e quindi algoritmi ancora più sofisticati – in grado di impedire la propalazione di queste nuove forme di falsi, peraltro ancora in fase sperimentale[20] oppure l'uso della tecnologia blockchain o della firma digitale[21].

Questo comporterà un sempre maggior utilizzo di tecniche invasive quanto alla privacy, monitoraggi costanti e continui da parte delle piattaforme mettendo ancora più a rischio la quasi inesistente tutela della riservatezza.

D'altro canto, ancora una volta saranno le piattaforme ad avere un ruolo fondamentale per arginare il fenomeno. Nel Bill prima citato il meccanismo di coinvolgimento dei provider è analogo a quello sperimentato per la tutela del Copyright il cd. 'notice and takedown'[22] in

[20] Si v. ad esempio il progetto, ancora in fase sperimentale, della Technical University of Munich, la Federico II di Napoli e l'University of Erlangen-Nuremberg, dimostra che questo approccio è incoraggiante. Attraverso l'accumulazione immagini o video veri e le loro controparti falsificate (per esempio con un faceswap) è possibile addestrare una rete neurale in grado di classificare una nuova immagine o video come vera oppure contraffatta.

[21] Per una critica sull'uso di tale strumento si v. lo studio del Parlamento Europeo, *Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States*, in http://www.europarl.europa.eu, specificamente p. 122.

[22] Section 512 del Digital Millenium Copyright Act: La procedura prevista dalla legge peraltro articolata richiedendo, infatti, la sussistenza di presupposti formali affinché l'internet service provider possa validamente agire. È richiesta, infatti, che: 1) la notifica sia scritta che contenga la firma di un soggetto in grado di rappresentare il titolare del diritto d'autore; 2) deve contenere l'identificazione del o dei lavori che il titolare del diritto ritiene siano stati violati; 3) l'indicazione del sito in cui esso è stato caricato; 4) i contatti di colui il quale ha attivatola procedura ; 5) una dichiarazione in cui il reclamante dichiari ha motivo di credere in buona fede che il materiale sia stato utilizzato in modo illecito e infine una dichiarazione sotto la sanzione di spergiuro che le informazioni offerte sono accurate e che l'autore o autorizzato. Una volta ricevuto il reclamo il 'provider' comunica all'utente che avrà modo di instaurare un procedimento di counter notification contentente 1) la firma; 2) informazioni circa il materiale rimosso e una dichiarazione sotto la sanzione di spergiuro che l'utilizzatore ha motive di credere, in buona fede, che il materiale è stato rimosso o disabilitato per errore; 3) fornirei dati personali e una dichiarazione che acconsente l'uso per le Corti Federali di utilizzo dei propri dati nonché l'accettazione del processo. L'onere tuttavia di essere tempestivi e diligenti non comporta però a carico dei provider di un obbligo generalizzato di monitoraggio.

virtù del quale le piattaforme rimuovono il contenuto rapidamente sulla base della segnalazione puntuale e precisa degli utenti o di soggetti della società civile a tanto deputati e in caso di omessa o tardiva rimozione o blocco possono essere soggetti a forme di risarcimento del danno.

Per evitare forme di regolazione 'hard' gli Over the Top stanno immaginando possibili forme di autoregolazione che "*tranquillizzino i Governi*" ed evitare qualsiasi forma di responsabilità anche indiretta.

Twitter ha di recente introdotto una proposta (contenuta in una bozza di policy pubblicata dal social sul proprio blog) in cui il social potrebbe avvertire gli utenti se stanno guardando o condividendo video, audio o foto 'manipolati', ma senza rimuoverli.

Interessante notare che è la stessa piattaforma ad avere avviato una sorta di consultazione pubblica con i propri utenti che fino al 27 novembre possono postare i propri commenti.

Peraltro, sempre il social ha precisato che i video verrebbero rimossi solo se considerati minacciosi per l'incolumità fisica di qualcuno o se possono portare a 'danni gravi'. Le motivazioni addotte dalla piattaforma sull'introduzione di una simile policy è impedire "*tentativi deliberati di fuorviare o confondere le persone con media manipolati mettono a rischio l'integrità della conversazione*".

Analogamente a quanto già osservato da autorevole dottrina la questione della 'self regulation' in tema di rimozione di contenuti pericolosi pone un problema di censura privata. I Giganti della rete non solo definiscono in modo autonomo la fattispecie lesiva ma agiscono, in molti casi, quanto ai rimedi in maniera oscura e con valutazioni che

---

Anche il meccanismo di 'notice and take down' viene considerato oramai obsoleto dalla dottrina statunitense che tende a sottolineare gli effetti negativi sia per la tutela del copyright sia per i diritti fondamentali, entrambi non garantiti sufficientemente,anche in presenza di una responsabilità indiretta. Sul punto si vedano le osservazioni di K. Fisher, *Facebook Shoots First, Ignores Questions Later; Account Lock-Out Attack Works*, in *Ars Technica*, 28 Aprile, 2011 a proposito del *social* più diffuso al mondo. Rischi non minori sono relativi alla tutela della privacy. Si registra fra l'altro un'asimmetria nella procedura di blocco o rimozione del contenuto che può avvenire nel giro di pochissime ore mentre il ripristino nel caso di errore può richiedere giorni che produce il cd. 'chilling effect' e quindi un'implicita restrizione al free speech. Più in generale su tali questioni ancora J. M. Urban, J. Karagani, B. Schofield, *Notice and Takedown in Everyday Practice* in *UC Berkeley Public Law Research*, 2755628, 2017.

non possono essere oggetto di contestazione da parte di colui il quale vede rimosso il contenuto.

D'altro canto, è evidente il paradosso di affidare la scelta del blocco e della rimozione a chi poi investe ingenti quantità di capitale in quelle tecnologie che producono deep fake e che sono contemporaneamente ci profila. Controllante e controllore divengono lo stesso soggetto[23].

E allora siamo destinati al mondo descritto da Philip Dick?

Forse dobbiamo convincerci che la semplice regola giuridica non basta.

Le significative trasformazioni tecnologiche hanno sempre messo in crisi vecchi sistemi per crearne nuovi. E se vale ancora la riflessione di Carnelutti che chi conosce solo il diritto non conosce il diritto, l'avvento di Internet impone non solo la comprensione di linguaggi e comportamenti sconosciuti ma il (ri)appropriarsi di un vocabolario che attinga dalla filosofa, dalla letteratura, dall'etica nuova linfa in grado di fornire chiavi di letture più ampie e più profonde dei fenomeni: prima umani e poi tecnici. Se la tecnica trasforma la fattualità della vita umana non ne cambia, purtroppo o per fortuna, l'essenza.

[23] Si v. sul punto E. Lehner, *Fake news e democrazia*, in *MediaLaws*, 1, 2019, p. 12; ed ancora di recente P. Nemitz, *Constitutional Democracy and Technology in the age of Artificial Intelligence*, 18 August 2018, in https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3234336, p. 7; M. Moore, *Tech Giants and Civic Power* in *King's College London*, April 2016, https://www.kcl.ac.uk/sspp/policyinstitute/CMCP/Tech-Giants-and-Civic-Power.pdf.