Piattaforme online e controllo dei contenuti pericolosi

Carla Bassu

Il rapporto tra Internet e democrazia rappresenta una frontiera del diritto contemporaneo. In particolare, con riferimento al tema oggetto di questo intervento, si nota come le piattaforme online gestite da privati esercitano di fatto e, forse, loro malgrado un ruolo nella dinamica democratica perché si rivelano strumentali allo sviluppo e alla manifestazione di libertà fondamentali quali la manifestazione del pensiero, l'espressione della personalità individuale e la libertà di informazione. Nello specifico, i *social network* maneggiano e veicolano contenuti delicati che, in un'epoca segnata dalla minaccia incombente del terrorismo e dal dilagare di diverse forme di istigazione all'odio si prestano a esercitare un impatto sulla dinamica democratica e sulle scelte degli ordinamenti[1]. Le piattaforme entrano in gioco nel momento in cui diffondono materialmente contenuti che gli ordinamenti ritengono pericolosi per la sicurezza per motivi differenti: perché istigano alla violenza, inneggiano al terrorismo, fanno proselitismo per cause estremiste etc.

La questione è delicata e riguarda l'annosa e complicata tematica del rapporto tra libertà e sicurezza, con particolare riferimento alla libertà di espressione di cui, come si è detto, la rete e i social media in particolare sono ormai strumenti fondamentali[2].

Ora, è lecito e, se si, fino a che punto limitare la libertà di espressione per proteggere la sicurezza?[3] Quale è il confine tra legittima manifestazione del proprio pensiero e

[1] *Ex multis* cfr. A. Vedaschi, *The dark side of counterterrorism: arcana imperii and salus rei publicae*, in *The American Journal of Comparative Law*, 66(4), 2018, 877 ss.; G. De Minico, *Costituzione. Emergenza e terrorismo*, Napoli 2016; C. Bassu, *Terrorismo e costituzionalismo. Percorsi comparati*, Torino 2010; T. Groppi (a cura di), *Democrazia e terrorismo*, Napoli 2006; P. Bonetti, *Terrorismo, emergenza e costituzioni democratiche*, Bologna 2006; T.E. Frosini, *Il diritto costituzionale alla sicurezza*, in *forumcostituzionale.it*, 2006; G. de Vergottini, *Guerra e Costituzione*, Bologna 2004

[2] I. Gagliardone - D. Gal - T. Alvez - D. Martinez, *Countering online hate speech*, Unesco Series on Internet Freedom, Paris, 2015. Cfr. T.E. *Frosini, Liberté, Egalité, Internet*, Napoli 2019; sull'impatto delle tecnologie digitali sul regime di riconoscimento e garanzia della libertà di espressione cfr., *ex multis*, J. M. Balkin, *The future of Free Expression in a digital age*, in *Pepperdine Law Review*, 36, 2008, 212 ss.; P. Costanzo, *Il fattore tecnologico e le sue conseguenze*, in Aa- Vv., *Costituzionalismo e globalizzazione*, Atti del XXVI convegno annuale, Salerno, 22-24 novembre, Napoli 2012.

[3] Cfr. *ex multis*, O. Pollicino - G. De Gregorio, *Hate speech, una prospettiva di diritto costituzionale comparato*, in *Giornale di Diritto amministrativo*, 4, 2019, 421 ss.; I. Spigno, *Discorsi d'odio: modelli costituzionali a confronto*, Milano 2018; O. Pollicino - G. Pitruzzella - S. Quintarelli, *Parole e potere: libertà di espressione, hate speech e fake news*, Milano, 2017; F. Di Tano, *Hate speech online: scenari, prospettive e criticità giuridiche del fenomeno*, in *Ciberspazio e diritto*, 51(2-3), 2014, 413 ss.; L. Scaffardi, *Oltre i confini della libertà di espressione: l'istigazione all'odio razziale*, Padova 2009; J. Waldron, *The Harm in hate speech*, Cambridge, 2012; M. Rosenfeld, *Hate speech in Constitutional Jurisprudence. In the Content and Context of Hate Speech*, Cambridge, 2012 V. Zeno-Zencovich, *Freedom of expression: A Critical Comparative and Analysis*, London, 2008.

diffusione di messaggi eversivi e potenzialmente pericolosi per la pubblica sicurezza?[4] Quale è il ruolo degli operatori online?

La sensibilità del tema è dimostrata dalla presa di posizione di aziende del calibro di Facebook, Twitter, Microsoft e YouTube che si sono affiancate agli Stati nella predisposizione di piani di azione volti a ottenere la rimozione dalle proprie piattaforme di contenuti ritenuti pericolosi. Si colloca in questa prospettiva di intervento integrato il *Global Internet Forum to Counter Terrorism* (*GIFCT*)[5], un programma orientato a impedire agli estremisti violenti l'utilizzo dei servizi di *hosting* delle piattaforme aderenti al progetto. Strutturato sulla base di realtà come lo *Ue Internet Forum*, il foro ambisce ad agevolare la collaborazione tra le grandi aziende promotrici e le imprese tecnologiche di dimensioni più ridotte definendo anche percorsi di interazione con la società civile, il mondo accademico e le istituzioni nazionali, sovranazionali come l'Ue e internazionali come l'ONU. Dal punto di vista operativo, nell'ambito del forum vengono proposte ed elaborate soluzioni coordinate, finalizzate a sviluppare e perfezionare sistemi di rilevamento e classificazione dei contenuti e definire procedure utili allo scopo di garantire la trasparenza sulla rimozione dei materiali terroristici. Un aspetto importante del *GIFCT* è rappresentato dalla condivisione delle competenze tecnologiche sfruttabili nella lotta al terrorismo da parte delle grandi compagnie promotrici e interlocutori pubblici e privati. In particolare, rileva l'accordo di cooperazione con lo *UN Security Council Counter-Terrorism Executive Directorate* (*UN CTED*) insieme con il progetto *ICT4Peace*, che consente la creazione di una rete di partnership con le aziende più piccole, aiutate nell'adozione e nello sviluppo delle tecnologie necessarie a controllare i contenuti estremisti online e sostenute negli investimenti sulla pratica del *counterspeech*, ovvero la replica ai contenuti violenti. L'impegno dei colossi del *web* che hanno ideato e realizzato l'iniziativa si colloca in una prospettiva di responsabilità etica ma non si può trascurare il fatto che la carenza di strumenti di controllo sui contenuti postati sulle piattaforme ha determinato un forte impatto di immagine ed economico prodotto dalla diffusione di *fake news* e contenuti violenti, con conseguenze sulle inserzioni pubblicitarie che, come è noto, costituiscono una significativa, se non la principale, fonte di introiti. Si pensi, per esempio, a quanto accaduto allorché su YouTube spot commerciali sono stati associati a pubblicazioni pro-ISIS inneggianti alla violenza terroristica, mettendo in luce la necessità di una sostanziale riconsiderazione dello strumentario di *brand safety*. Anche Facebook ha dovuto prendere atto della fallacia del proprio sistema di controllo emersa in modo lampante in occasione della elaborazione da parte del *social media* dei video animati che raccolgono il meglio dell'anno appena trascorso dall'utente. Se l'iscritto al *social network* è un simpatizzante dell'Isis può infatti accadere che tale collage metta in fila una carrellata di immagini inneggianti al terrorismo. La questione è particolarmente delicata perché denuncia la possibilità di un ruolo «proattivo» del *social network* nella diffusione di messaggi violenti perché nei video sopra menzionati, assemblati tramite algoritmo, il sistema si complimenta per il successo

[4] A. Tsesis, *Dignity and Speech: The Regulation of Hate Speech in a Democracy*, in *Wake Forest Law Review*, 44, 2009, 497 ss.

[5] Cfr. anche *gifct.org*.

in termini di *like* di immagini raffiguranti violenza e morte sul fronte dell'Isis[6].
Facebook non è fino a ora riuscito a realizzare l'obiettivo di evitare la diffusione di contenuti violenti, incitanti alla violenza razzista o terroristica[7]. Il monitoraggio effettuato sulle pagine create da circa tremila utenti risultanti in qualche modo collegati a organizzazioni estremiste classificate dal governo degli Stati Uniti ha messo in luce la presenza di contenuti brutali (ad esempio: video di esecuzioni; immagini di teste decapitate; tributi a "martiri" jihadisti) sfuggiti alle maglie del controllo. Il sistema di vigilanza non è pienamente efficace. L'azione di supervisione è svolta prevalentemente in via automatica, tramite l'impiego di algoritmi studiati per riconoscere e rimuovere i messaggi violenti o riconducibili a gruppi terroristici, cui si affiancano più o meno trentamila operatori umani che effettuano un ulteriore screening. È vero che risulta che una percentuale vicina al cento per cento dei contenuti pericolosi viene effettivamente eliminata dal sistema[8] ma il dato si riferisce solo ai contenuti individuati e non è chiaro quale sia la percentuale di materiale terroristico che la piattaforma non riesce a identificare.
La centralità della libertà di espressione nell'impianto costituzionale statunitense[9], sancita chiaramente dal primo emendamento della Costituzione, è alla base della storica esitazione da parte delle autorità nordamericane al riconoscimento di forme di responsabilizzazione degli intermediari digitali, che potrebbero determinare attività di censura collaterale[10]. La normativa primaria federale di riferimento è il *Communication Decency Act* che, alla *Section 230*, esclude che chi fornisce servizi interattivi digitali possa essere dichiarato responsabile di contenuti prodotti e diffusi da altri. In questo senso si collocano le decisioni delle Corti Usa che hanno respinto i ricorsi presentati dai parenti di alcune vittime di attentati compiuti di estremisti islamici i quali accusavano *social network* del calibro di Twitter e Facebook di aver agevolato l'attività dei terroristi che si sono serviti delle piattaforme *social* per fare propaganda e reclutare nuovi adepti[11]. Tuttavia, nonostante la tendenza sia ancora quella di negare la responsabilità degli operatori del web per la diffusione di contenuti espressi da altri, il fenomeno dell'*hate speech* e l'abnorme potenzialità di divulgazione di discorsi d'odio o *fake news* pubblicati online, soprattutto tramite i *social network*, anche negli Usa spinge a riflettere su forme di controllo e contenimento delle pubblicazioni online. In particolare, si discute

[6] V. in particolare il caso del video generato da Facebook il 17 giugno 2018, riportato nel dossier *How Facebook Helps Terrorists and Hate Groups Networks on its Website. Radical Connections*, Rev. 3, April 29, 2019.

[7] A. Al Azm - K.A. Paul, *How Facebook Made It Easier Than Ever to Traffic Middle Eastern Antiquities*, in *World Politics Review*, August 2018.

[8] Cfr. Press Release from Monika Bickert, Director of Global Policy Mgmt. and Brian Fishman, Counterterrorism Policy Mgmt., *Hard Questions: How Effective Is Technology in Keeping Terrorists off Facebook?*, Facebook (Apr. 23, 2018).

[9] V. *ex multis*, D.M. Fraleigh - J.S. Tuman, *Freedom of Expression in the marketplace of ideas*, Los Angeles, 2019. Cfr. O. Pollicino, *La prospettiva costituzionale sulla libertà di espressione nell'era di Internet*, in *questa Rivista,* 1, 2018, 50, confronta la natura «tollerante» dell'approccio americano alla «democrazia militante impegnata nella strenua difesa di un sistema valoriale che l'esercizio degli stessi diritti costituzionalmente tutelati rischia di mettere a repentaglio», richiamando in questo senso K. Loewenstein, *Militant Democracy and Fundamental Rights*, in *American Political Science Review*, 31, 1937, 417 ss.

[10] L'espressione "*collateral censorship*" si deve a J. M. Balkin, *Old School/New School Speech Regulation*, in *Harvard Law Review,* 2014, 2296 ss.

[11] V. *U.S. District Court of Northern California, Fields v. Twitter*, November 18, 2016

sull'opportunità di modificare il *Communication Decency Act* per imporre ai *provider* (che attualmente agiscono solo su base volontaria) di rimuovere contenuti riferibili alla propaganda terroristica o in generale di istigazione alla violenza. Diversa è la situazione nell'area di influenza europea dove, nel contesto internazionale, la CEDU ammette espressamente limitazioni alla libertà di espressione finalizzate a salvaguardare i principi della democrazia, prevedendo in generale che l'esercizio di una libertà trovi un limite invalicabile nella garanzia del diritto altrui che non può subire una eccessiva, indebita compressione[12]. Legittimi sono dunque nello spazio europeo interventi normativi statali che prevedano sanzioni per chi si renda responsabile di incitamento all'odio o alla violenza, sebbene massima cura debba sempre essere prestata a che il diritto individuale a manifestare il proprio pensiero sia salvaguardato da eccessiva incisività.
Ma il nodo è sempre nella difficoltà di individuare il grado legittimo di limitazione applicabile alla libertà di pensiero dal momento che il criterio è flessibile e inevitabilmente caratterizzato da una certa discrezionalità decisionale da parte del decisore pubblico e dell'operatore giuridico che interpreta e applica la normativa. Così dubbi sorgono circa la possibilità che la repressione di determinate manifestazioni del pensiero effettuata allo scopo di tutelare un assunto sentimento comune determini in effetti una indiretta gerarchizzazione delle sensibilità interne a una comunità, privilegiando di fatto le correnti di pensiero "dominanti" a svantaggio dei gruppi minoritari. Ancora, la limitazione della libertà di espressione intesa come strumentale a interessi collettivi quali la tutela della sicurezza e la salvaguardia dei principi cardine della democrazia, potrebbe rivelarsi rischiosa nel momento in cui venisse intesa come mezzo di censura. Da qui si comprende la mancata imposizione da parte della CEDU e, in generale, delle istanze del diritto ultranazionale della previsione di sanzioni penali imposte a livello nazionale applicabili in caso di incitamento all'odio o alla violenza terroristica.
A livello sovranazionale[13] si registra comunque un ambizioso piano di prevenzione della divulgazione di contenuti riconducibili al terrorismo attraverso la rete, nell'ambito del quale si colloca la proposta della Commissione europea sottoposta all'attenzione dei leader dei Paesi membri Ue, riuniti a Salisburgo il 19 e 20 settembre 2018[14].
Per quanto riguarda le discipline nazionali, si segnala la legge tedesca entrata in vigore il 1° gennaio 2018[15], definita dai media «Legge Facebook» e mirata a contrastare l'*hate speech* online, categoria flessibile che comprende minacce, insulti e discriminazioni via web. Il provvedimento è rivolto soprattutto ai *social network*, chiamati a vigilare sul rispetto delle severe norme su diffamazione, incitamento all'odio e minacce vigenti in Germania. Chi ometterà di rimuovere i contenuti diffamatori entro i tempi stabiliti sarà

[12] V. J.F. Flauss, *The European Court of Human Rights and the Freedom of Expression*, in *Indiana Law Journal,* 84(3), 2009, 809 ss.; cfr. H. Cannie - D. Voohroof, *The abuse Clause and Freedom of Expression in the European Human Rights Convention: An Added Value for Democracy and Human Rights Protection?*, in *Netherlands Quarterly of Human Rights,* 29(1), 2011, 54 ss.

[13] Per una prospettiva comparata sulle limitazioni alla libertà di espressione in seno all'Unione europea v. O. Pollicino - M. Bassini, *Free speech, defamation and the limits to freedom of expression in the EU: a comparative analysis*, in A. Savin - J. Trzaskowski (eds.), *Research Handbook On EU Internet Law*, Cheltenham – Northampton, 2014, 508 ss.

[14] V. *Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online. A contribution from the European Commission to the Leaders' meeting in Salzburg on 19-20 September* 2018 COM (2018) 640 final 12.9.2018.

[15] *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken*.

sanzionato da multe fino a 50 milioni di euro. La legge non è esente da criticità rappresentate soprattutto dalla vaghezza della definizione dei messaggi censurabili e dallo spazio di azione lasciato alle società private che gestiscono i *social media* e che, di fatto, sono chiamate ad agire in sede applicativa con una certa discrezionalità consentita dalla connotazione vaga del parametro normativo. Non sempre è facile individuare contenuti effettivamente offensivi, basti pensare ai messaggi satirici o "spiritosi": a questo proposito si sono registrati casi eclatanti come la sospensione dell'account Twitter del magazine satirico Titanic a seguito della pubblicazione di messaggi che prendevano di mira scherzosamente una esponente del partito di estrema destra Alternative für Deutschland (AfD) la quale aveva criticato la polizia di Colonia per aver pubblicato alcune comunicazioni in lingua araba[16]. In UK, nell'aprile 2019 il Ministro per Digitale, cultura, media e sport, Jeremy Wright, ha presentato, in collaborazione con il Ministero dell'Interno, un libro bianco che contempla la creazione di un codice di condotta e di un'autorità indipendente chiamata a vigilare sul comportamento delle società di Internet, con potere di comminare sanzioni amministrative e disporre l'oscuramento dei siti nell'ipotesi in cui si configurino violazioni. Particolarmente significativo è il caso della normativa spagnola antiterrorismo, approvata nel 2015, divenuta nota con l'eloquente denominazione di "Ley Mordaza", ossia legge bavaglio, che interviene in senso restrittivo con riferimento a diversi aspetti della libertà di manifestazione del pensiero. Stringente è la sanzione del reato di apologia del terrorismo e della denigrazione delle vittime del terrorismo e dei loro parenti[17] in base al quale negli ultimi anni sono state effettuate numerose incriminazioni legate a contenuti espressi via *web*. Anche in Francia è all'attenzione del Parlamento una proposta di legge che obblighi i gestori dei motori di ricerca e dei *social network* a rimuovere tassativamente entro 24 ore dalla richiesta (a prescindere da chi è inviata) ogni contenuto riconducibile prima facie all'incitamento all'odio, alla violenza o alla discriminazione[18]. La sanzione in caso di non ottemperanza potrà arrivare fino al 4 per cento del fatturato mondiale annuo dell'operatore interessato, una cifra enorme, simile alla multa più severa imposta dal GDPR, la nuova normativa europea sulla privacy. Incaricato di verificare e diffidare formalmente il responsabile della mancata rimozione tempestiva dei contenuti illeciti sarebbe l'organismo francese analogo alla nostra Autorità indipendente per le garanzie nelle telecomunicazioni: il Consiglio dell'audiovisivo[19]. Ancora, una legge australiana del 2019 prevede di multare i social media per una somma pari fino al 10 per cento del loro fatturato annuale e di condannare i dirigenti fino a tre anni di carcere, nel caso in cui non provvedano a rimuovere tempestivamente i contenuti violenti condivisi sulle rispettive piattaforme[20].

---

[16] L'account, che vanta 464mila followers è stato sospeso e sbloccato pochi giorni dopo.

[17] V. art. 578 del codice penale spagnolo

[18] La proposta di legge in tema di "*Lutte contre la haine sur internet*" è reperibile al sito *assemblee-nationale.fr*.

[19] Cfr. F. Rossi, *La penalizzazione della propaganda jihadista online in Francia*, in *Diritto Penale Contemporaneo*, 5, 2019, 125 ss.

[20] V. *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019*; cfr. H. Leung, *Australia has passed a Sweeping Law to punish Social Media Companies for not Policing Violent Content. Here's what to know*, in *The Time*, April 5, 2019.

Certamente il momento storico risulta favorevole alla promozione di misure che tutelino la sicurezza, anche a scapito delle libertà individuali. Non si può trascurare che il libro bianco britannico così come la nuova normativa australiana risalgono a pochi giorni dopo che in Nuova Zelanda si era consumato il massacro di Christchurch, puntualmente reso disponibile alla visione universale grazie a una diretta streaming via Facebook. Il fatto che il video sia stato rimosso dalla piattaforma dopo ben dodici minuti ha suscitato clamore e spinto l'amministratore delegato Zuckeberg, in un editoriale pubblicato sul Washington Post, a reclamare dai governi regole stringenti in ordine a quali siano i contenuti accettabili e quali no perché il peso della discrezionalità non può ricadere su aziende private. Nel tentativo di tutelarsi da richiami a responsabilità dirette, Facebook ha assunto l'impegno a bloccare le pubblicazioni inneggianti al nazionalismo. In Italia, per esempio, si registra la sospensione degli account di alcuni componenti della formazione politica di estrema destra Casa Pound, che avevano pubblicato materiale considerato riconducibile alle categorie di apologia del fascismo e incitamento all'odio e alla violenza.

La tendenza che si evince dall'osservazione delle proposte all'esame delle istituzioni nazionali è una responsabilizzazione degli operatori che sono chiamati a vagliare nel merito i contenuti e a procedere alla eventuale rimozione in base a una sostanziale discrezionalità. Importante sarebbe ottenere una riduzione significativa degli automatismi nella individuazione dei contenuti e delle immagini ritenuti pericolosi, investendo su sistemi di monitoraggio a tappeto che, caso per caso e con attenta valutazione del contesto di riferimento, considerino la oggettiva carica di rischio in termini di incitamento all'odio e alla violenza.