RIVISTA ITALIANA DI INFORMATICA E DIRITTO
PERIODICO INTERNAZIONALE DEL CNR-IGSG

ISSN 2704-7318
Fascicolo 1-2022
DOI: 10.32091/RIID0064
Articolo sottoposto a peer review

Moderazione automatizzata e discriminazione algoritmica: il caso dell'*hate speech*

Pietro Dunn

La necessità per gli intermediari digitali di moderare i contenuti pubblicati e diffusi in rete dagli utenti si è fatta negli anni sempre più pressante. A fronte della crescita vertiginosa del flusso informazionale digitale, peraltro, si è reso oggigiorno essenziale il ricorso a strumenti di moderazione algoritmica per la rilevazione dei contenuti da rimuovere. Anche la rilevazione dei discorsi d'odio (*hate speech*) si fonda attualmente su un utilizzo massiccio di sistemi di intelligenza artificiale e *machine-learning*: la letteratura, tuttavia, ha rilevato come tali sistemi siano sovente viziati da *bias* discriminatori che rendono particolarmente elevato il rischio di falsi positivi ai danni delle minoranze. Il presente contributo pone in luce come nel sistema costituzionale europeo il contrasto ai contenuti d'odio sia giustificato dall'esigenza di perseguire un'uguaglianza sostanziale di tutte le componenti sociali e come, pertanto, un'applicazione discriminatoria del divieto di *hate speech* sia in sé incoerente con il sistema di valori dell'Unione europea. Se, dunque, l'intelligenza artificiale rappresenta uno strumento essenziale e ineludibile per garantire un più sicuro e tollerante ecosistema digitale, un elevato margine di errore, in termini di falsi positivi, non risulta essere pienamente accettabile. Occorre, pertanto, un ripensamento delle strategie legislative nell'ottica di offrire più adeguate garanzie, sostanziali e procedurali, a tutela della libertà di espressione e del diritto di non discriminazione dei gruppi marginalizzati.

Moderazione automatica – *Hate speech* – Discriminazione algoritmica – Uguaglianza sostanziale – Libertà di espressione

1. Introduzione: il ruolo odierno della moderazione

Le tecnologie digitali, e Internet in particolare, hanno permesso la diffusione di nuovi ed eccezionali strumenti per il godimento di diritti e libertà fondamentali. Nella ormai celebre sentenza *Reno v. ACLU* (1997)[1], la Corte Suprema degli Stati Uniti già segnalava e celebrava il ruolo della rete quale facilitatrice del "libero mercato delle idee", in piena sintonia con la storica interpretazione del Primo Emendamento resa dal giudice Holmes in *Abrams v. United States*[2]. Allo stesso tempo, il ciberspazio ha tuttavia dato adito a nuove sfide e nuovi pericoli[3], tant'è che,

P. Dunn è dottorando di ricerca in Law, Science and Technology presso l'Alma Mater Studiorum – Università di Bologna (CIRSFID-AI) e presso l'Università del Lussemburgo (FDEF).

Questo contributo fa parte del numero speciale "La Internet governance e le sfide della trasformazione digitale" curato da Laura Abba, Adriana Lazzaroni e Marina Pietrangelo.

nel vecchio continente, la Corte Europea dei Diritti dell'Uomo (Corte EDU) ha ripetutamente posto in luce gli accresciuti rischi legati alla società dell'informazione, concludendo che ciò possa giustificare un intervento più marcato degli Stati contraenti sulla libertà di espressione in rete[4].

Per fare ordine della caotica massa di informazioni caricate quotidianamente in Internet, nonché per ridurre la quantità di "mali informazionali" (*information bads*[5]), gli intermediari digitali hanno ben presto sviluppato strategie di moderazione dei contenuti sempre più complesse e raffinate[6]. Come posto in luce da Gillespie, nonostante gli intermediari abbiano a lungo cercato di presentarsi come fornitori di servizi (per lo più servizi di hosting) meramente neutrali, l'ideale di una piattaforma priva di alcun controllo rappresenta un'utopia[7]. Tutte le piattaforme moderano: anzi, la moderazione sarebbe da intendersi precipuamente quale prodotto stesso della piattaforma, in quanto rappresenterebbe in ultima istanza ciò che garantisce all'utente consumatore un'esperienza più o meno positiva della rete[8]. Essa è cioè parte integrante del pacchetto offerto dai social media[9].

D'altro canto, è proprio sulla moderazione privata operata dagli intermediari digitali che è andato crescendo il *focus* delle scelte legislative e politiche degli ultimi anni. Soprattutto a partire dalla seconda metà degli anni 2010, si è invero assistito a un sempre maggiore ricorso a tecniche di regolazione della libertà di espressione "di nuova scuola" (*new-school speech regulation*[10]). L'elemento caratteristico di queste nuove forme di regolazione è la scelta di intervenire non tanto attraverso l'imposizione di restrizioni e sanzioni che investano la libertà di espressione dei singoli individui quanto, piuttosto, attraverso la diretta regolazione delle infrastrutture digitali, attraverso l'elaborazione cioè di forme di responsabilità sussidiaria a carico dell'intermediario per la presenza e diffusione di contenuti illeciti generati da terzi[11]. L'Unione europea, tra gli altri, sembra avere intrapreso tale strada negli ultimi anni[12].

La diffusione di tali nuovi strategie di governance, unita a un'accresciuta sensibilità del pubblico, hanno spinto gli intermediari digitali a farsi maggiormente carico del loro ruolo di moderatori. Per far fronte, tuttavia, alla crescita esponenziale del traffico quotidiano di dati e informazioni in rete, l'utilizzo di sistemi di intelligenza artificiale (IA) ha acquisito un maggior rilievo anche in questo settore[13]. Il ricorso a strumenti automatici di decisione per la gestione dei contenuti in rete solleva peraltro una serie di perplessità con riferimento alla protezione e garanzia di diritti umani e valori costituzionali, ivi inclusi la libertà di espressione e informazione e il principio di non discriminazione. Ciò, soprattutto, appare evidente con riferimento alla moderazione dei discorsi d'odio (*hate speech*)[14].

Il presente contributo fornisce uno sguardo sul crescente ruolo dei sistemi di IA e *machine-learning* nell'ambito della rilevazione dei contenuti d'odio (paragrafo 2) e sull'impatto discriminatorio che tali strumenti possono avere sulla libertà di espressione dei gruppi minoritari e/o marginalizzati (paragrafo 3). Il paragrafo 4 argomenta come un'interpretazione del contrasto al fenomeno dell'*hate speech* in una prospettiva di uguaglianza sostanziale richieda un ripensamento altresì delle strategie legislative e di policy sul piano europeo.

2. Moderazione algoritmica dei contenuti e rilevazione dei discorsi d'odio

Secondo Grimmelmann[15], la moderazione dei contenuti rappresenta l'insieme di quei meccanismi di governance che strutturano la partecipazione a una comunità online, al fine di favorire la cooperazione tra gli utenti e prevenire la commissione di abusi. In senso lato, essa comprende due diversi aspetti. Il primo si riferisce alla rimozione dei contenuti contrari alle condizioni d'uso del servizio, nonché all'imposizione di sanzioni (ad esempio, la sospensione o cancellazione del profilo) a carico di chi li abbia postati: in tal senso, si può parlare di moderazione "in senso stretto" o di *hard moderation*[16].

Il secondo aspetto, invece, si riferisce all'organizzazione, distribuzione e disseminazione dei contenuti stessi, attraverso una loro gerarchizzazione atta a migliorare l'esperienza degli utenti. A questi sono offerti, infatti, i contenuti che più possano loro interessare: si parla, con riferimento a tale attività, di "cura dei contenuti" (*content curation*[17] o *soft moderation*[18]). La cura dei contenuti, che si basa generalmente sull'uso di sistemi automatizzati, quali i sistemi di raccomandazione, mirano a massimizzare l'engagement degli utenti[19] e, di conseguenza, i profitti del prestatore di servizi: la letteratura ha rilevato come ciò possa andare a discapito di importanti valori democratici, quali la protezione del pluralismo mediatico e di pensiero, alimentando da un lato la creazione di camere dell'eco e la polarizzazione del dibattito democratico e impattando dall'altro lato la capacità di diffusione dei contenuti prodotti da gruppi minoritari[20]. Nonostante tale significativo impatto della cura dei contenuti sull'ecosistema informazionale digitale, il *focus* del presente contributo

sarà posto in modo particolare sulla moderazione dei contenuti "in senso stretto".

Da un punto di vista pratico, le tecniche di moderazione possono adottare strategie differenti[21]. Una prima distinzione può essere fatta, sulla base del criterio temporale, tra moderazione *ex ante* e moderazione *ex post*, a seconda che il controllo venga esercitato prima o dopo la pubblicazione del contenuto. A sua volta, la moderazione ex post può essere proattiva, laddove l'intermediario si occupi attivamente di individuare i contenuti da rimuovere, o reattiva, quando invece si limiti a ricevere e valutare segnalazioni altrui (ad esempio, da parte di altri utenti del servizio).

Sotto un diverso profilo, la moderazione può essere operata da esseri umani (moderazione umana o manuale), da sistemi di IA (moderazione automatica o algoritmica) oppure attraverso una combinazione dei due (moderazione ibrida). In quest'ultimo caso, la funzione dei sistemi di IA è principalmente quella di operare una scrematura preventiva dei contenuti pubblicati dagli utenti e di rimettere al moderatore umano soltanto i casi più ambigui, istituendo tra l'altro un ordine di priorità rispetto all'ordine di revisione[22]. I sistemi ibridi hanno acquisito negli ultimi anni un rilievo sempre maggiore, soprattutto per le possibilità che gli strumenti di IA offrono agli intermediari di effettuare controlli su larghissima scala. Al tempo stesso, il ricorso all'algoritmo consente di ridurre l'esposizione dei moderatori umani a contenuti potenzialmente dannosi per il loro benessere psicofisico[23].

La ricerca relativa allo sviluppo, perfezionamento e aggiornamento dei sistemi automatizzati di moderazione si è dimostrata particolarmente feconda. Attualmente, gli intermediari digitali godono di una vasta gamma di strumenti algoritmici a loro disposizione, che possono essere variamente combinati a seconda della tipologia di *information bad* che si voglia filtrare e sulla base del formato (testuale, visuale, audiovisuale etc.) che si voglia analizzare[24]. Particolarmente diffusi e utilizzati sono, attualmente, i sistemi di *machine-learning* basati su reti neurali[25]: con riferimento a tali tecnologie, un terreno di ricerca particolarmente fertile risulta essere quello del *natural language processing* (NLP), ovverosia quella branca dell'informatica che si occupa di sviluppare le capacità delle macchine di analizzare contenuti testuali, con il fine specifico di trarre conclusioni in merito al significato del testo stesso[26].

L'utilizzo di sistemi automatizzati di moderazione è andato aumentando drasticamente negli ultimi anni e ha fatto uno straordinario balzo avanti a seguito dello scoppio della pandemia di COVID-19. Se infatti, da un lato, piattaforme e intermediari digitali sono entrati in uno "stato di emergenza"[27] durante la crisi sanitaria, a causa soprattutto dell'aumento preoccupante nella diffusione di *hate speech*[28] e fake news[29], dall'altro lato, le piattaforme hanno nei primi mesi dovuto sviluppare adeguati sistemi di IA per far fronte alla riduzione di manodopera umana disponibile derivante dalla necessità di porre in atto le adeguate misure di contenimento del contagio[30].

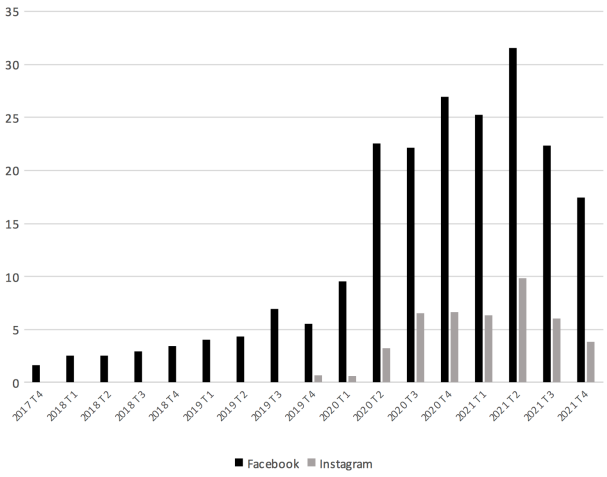

Figura 1: *Totale dei contenuti sanzionati da Instagram e Facebook come contenuti d'odio (in milioni)*

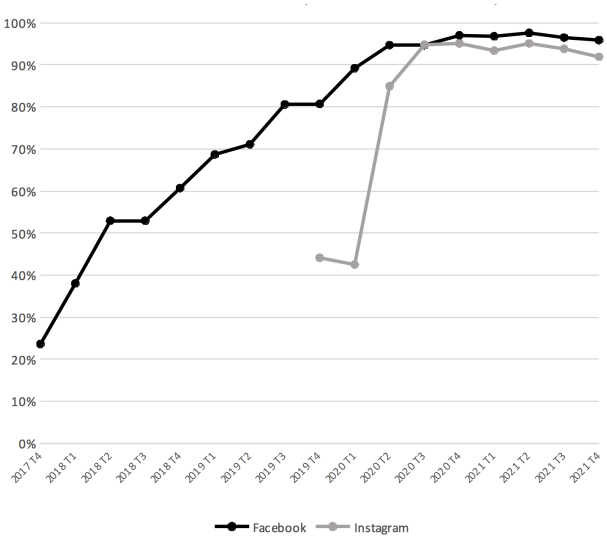

Figura 2: *Percentuale di contenuti d'odio rilevati tramite sistemi automatizzati sul totale dei contenuti d'odio*

I dati pubblicati da Facebook e Instagram nei loro report periodici sull'applicazione degli standard della comunità[31] confermano tali trend. Ciò emerge con particolare vigore con riferimento alla moderazione dei discorsi d'odio, sempre più automatizzata.

La Figura 1 mostra, in particolare, il numero di contenuti sanzionati, in quanto riconosciuti quali fattispecie di *hate speech* dalle piattaforme, nel periodo intercorrente tra l'ultimo trimestre del 2017 e il quarto trimestre del 2021, mentre la Figura 2 mostra, in percentuale, quanti di quei contenuti sono stati rilevati attraverso il ricorso a sistemi di intelligenza artificiale[32]. In tal senso, il numero di contenuti sanzionati da Facebook è aumentato vertiginosamente negli anni: un incremento particolarmente evidente si è avuto a seguito dello scoppio della pandemia, con un salto da 9,5 milioni di contenuti sanzionati nel primo trimestre del 2020 a 22,5 milioni nel trimestre successivo. In realtà, gli ultimi due trimestri del 2021 segnalano un'inversione di tendenza in tal senso, con una drastica riduzione dei numeri che, tuttavia, continuano tutt'ora a essere notevolmente più elevati rispetto all'epoca pre-pandemica[33].

Nel frattempo, un costante aumento si è avuto nell'utilizzo di sistemi di IA per la rilevazione di contenuti d'odio: infatti, se nell'ultimo trimestre del 2017 solo il 23,6% dei post non conformi al divieto di *hate speech* era rilevato proattivamente dagli algoritmi di Facebook, il dato è andato aumentando nel corso del tempo. Attualmente, i contenuti sanzionati dal social network in quanto ritenuti istiganti all'odio sono rilevati tramite sistemi automatizzati per il 96-97% circa. L'attuale CTO di Meta, Mike Schroepfer, ha celebrato questi risultati, sottolineando come l'utilizzo dei sistemi automatizzati di moderazione contribuisca a garantire un ecosistema digitale sicuro[34]. Del resto, secondo i dati pubblicati dalla stessa azienda, si è in effetti assistito a una diminuzione dallo 0,10-0,11% allo 0,03%[35] circa nel grado di "diffusione" di contenuti d'odio (ovverosia la percentuale stimata di visualizzazione di *hate speech* su tutti i contenuti visualizzati dagli utenti[36]).

Tuttavia, se è vero che l'avanzamento tecnologico nel settore rappresenta un fattore importante ed essenziale nella prospettiva di costruire un ciberspazio libero da intolleranze e violenze, è pur vero che le informazioni rese dal gruppo di Facebook rivelano un solo lato della medaglia. Come sottolineato in *Wired*, non è chiaro fino a che punto gli algoritmi di rilevazione dei discorsi d'odio siano andati effettivamente perfezionandosi[37]. Si tenga conto, in particolare, che i dati riportati sono meramente quantitativi: poco ci dicono sulla qualità delle scelte prese dai sistemi di IA o sulla percentuale di errori di rilevazione[38]. Tale ambiguità e opacità rispetto alla qualità e correttezza delle decisioni prese dall'algoritmo risultano essere particolarmente preoccupanti laddove si consideri il rischio di output discriminatori.

3. Margini di errore e *bias* discriminatori

Invero, i sistemi automatizzati di classificazione si basano su fondamenti statistico-probabilistici che, in quanto tali, rendono sempre inevitabile un più o meno elevato margine di errore. Gli errori, in particolare, possono tradursi in falsi negativi o in falsi positivi: se i primi minano l'efficacia di un sistema automatizzato di moderazione, i secondi possono invece essere dannosi per l'esercizio della libertà di espressione online degli utenti. Peraltro, le due tipologie di errore sono in generale inversamente proporzionali[39]: laddove, cioè, si implementi un sistema più "permissivo", vi sarà un rischio minore di falsi positivi a fronte di un maggior numero di contenuti illeciti, o contrari alle condizioni di utilizzo, rimasti impuniti; mentre un sistema più "severo" sarà, all'opposto, più difficilmente eludibile ma più probabilmente esposto al rischio di falsi positivi. In tal senso, il riferito incremento di contenuti sanzionati in quanto ricondotti alla sfera dell'*hate speech* da parte di sistemi di IA ha comportato (e comporta) un verosimile e proporzionale incremento nel numero di falsi positivi.

Se ciò è vero, l'implementazione di sistemi automatizzati di moderazione si traduce in sostanza in un bilanciamento tra due, talora contrastanti, esigenze: da un lato, la necessità di ridurre la diffusione di "mali informazionali"; dall'altro lato, l'esigenza di tutelare la libertà di espressione e il pluralismo di pensiero. Bilanciamento, peraltro, operato sempre più direttamente dalle piattaforme e dagli intermediari digitali. Se, dunque, la scelta di utilizzare tali sistemi richiede l'applicazione di un principio di proporzionalità che tenga conto di tali esigenze, il problema di fondo consiste nell'individuazione della soglia entro la quale il margine di errore (nel senso di falso positivo) sia da ritenersi "accettabile" a fronte del vantaggio sociale determinato dalla riduzione del grado di inquinamento dell'ecosistema informazionale digitale[40]. L'individuazione di tale soglia, tuttavia, può variare a seconda della tipologia di *information bad* che si voglia combattere. A tal proposito, con riferimento al fenomeno dei discorsi d'odio, alcuni fattori richiedono di essere tenuti in considerazione.

Un primo elemento di complicazione è determinato dalla nozione stessa di *hate speech*, tutt'altro che condivisa e ben definita[41]. A seconda della giurisdizione di riferimento, le condotte ascrivibili ai discorsi d'odio penalmente rilevanti possono variare notevolmente. Allo stesso modo, le piattaforme e gli intermediari digitali tendono a definire autonomamente il concetto di *hate speech* sanzionabile ai sensi dei loro termini e condizioni d'utilizzo: molto sovente, per di

più, le nozioni adottate da tali attori risultano essere notevolmente più ampie e aperte rispetto alle fattispecie considerate dai sistemi giuridici statali[42]. Il campo di applicazione di tali standard privati rischia in tal senso di risultare estremamente lato e, talora, pericolosamente indefinito.

Un secondo rilevante aspetto è dettato dal fatto che, come rilevato dalla giurisprudenza e dal dibattito internazionale, la possibilità di ascrivere una determinata forma espressiva alla classe dei discorsi d'odio è strettamente dipendente dalla ricostruzione del contesto all'interno della quale essa si inserisce. L'identità dell'autore e dei componenti dell'audience, per esempio, così come elementi contestuali quale il tempo e il luogo in cui un determinato contenuto sia stato pubblicato o condiviso, sono fattori potenzialmente dirimenti per comprendere lo scopo e i possibili effetti che una certa modalità espressiva può avere: tali fattori richiedono sempre un'attenta ricostruzione al fine di evitare eccessive e sproporzionate interferenze a danno della libertà di espressione individuale[43]. Eppure, ciò rappresenta una sfida rilevante per il moderatore algoritmico, in quanto le macchine, nonostante il loro straordinario potere computazionale[44] e la loro efficienza a livello di comprensione simbolico-sintattica, pongono ancora oggi dei problemi per quanto concerne la capacità di comprensione semantica[45]. È infatti difficile, per un sistema automatizzato, rilevare l'ironia o la satira nascoste dietro un particolare contenuto. Tra l'altro, tale compito è complicato notevolmente dalle modalità espressive caratteristiche della comunicazione in rete, le quali mescolano sovente elementi testuali, visivi e audiovisivi: si pensi, per esempio, ai cosiddetti "meme", contenuti multimodali che si caratterizzano per un'alta viralità e per il fatto di richiedere, ai fini della comprensione, una vera e propria "meme literacy" dell'audience[46].

Le difficoltà prodotte da tali sfide sono esse stesse alla base di quello che è il terzo fattore di complicazione. Come evidenziato da ormai consolidata letteratura, il margine di errore connesso all'utilizzo di sistemi automatizzati di rilevazione delle fattispecie di *hate speech* tende a impattare significativamente proprio sulle comunità tradizionalmente marginalizzate e discriminate[47]. Sempre più studi sono attualmente dediti alla ricerca di tecniche di *debiasing* dei moderatori automatici[48], ma il problema è ancora lungi dall'essere risolto.

Così, per esempio, si è da più parti rilevato come i contenuti pubblicati da membri della comunità afroamericana[49] o della comunità LGBTQIA+[50] siano maggiormente soggetti a subire sanzioni ingiustificate per violazione del divieto di *hate speech* o "*toxic*" *speech*. Le cause di tali risultati discriminatori sono plurime. Accade, per esempio, che i *dataset* utilizzati per allenare l'algoritmo non siano qualitativamente ottimali, soprattutto perché non rappresentativi del gergo e degli usi comunicativi tipici dei gruppi minoritari. In molti casi, i gruppi marginalizzati sviluppano la tendenza a utilizzare termini ed espressioni in sé stessi insultanti e discriminatori (si pensi alla *n-word*) con la doppia finalità, tuttavia, di riappropriarsi di tali termini svuotandoli della loro carica negativa (è questo il caso della parola *queer*, inizialmente utilizzata quale insulto per le persone LGBTQIA+ e facente oggi parte della sigla stessa) e di aiutare i membri della loro stessa comunità a "farsi la pelle dura"[51]. L'incapacità della macchina di cogliere tali sfumature di intenti e di significato rende così particolarmente elevato il rischio di falsi positivi, tant'è che, tra gli attivisti afro-americani, è rapidamente invalso il ricorso al neologismo "zucked" a indicare le frequenti sanzioni loro imposte dalle piattaforme di Meta ogniqualvolta essi pubblichino contenuti che discutano il tema del razzismo[52].

L'applicazione discriminatoria delle regole di una comunità online da parte dei sistemi di moderazione algoritmica è peraltro dettata altresì dagli stessi utenti. Nel 2016-2017, il genocidio e le persecuzioni a danno della comunità musulmana Rohingya in Myanmar sono stati incentivati, da un lato, dal fallimento da parte di Facebook nel ridurre effettivamente la diffusione di *hate speech* avente ad oggetto la minoranza, e, dall'altro lato, dalla ripetuta censura di contenuti di denuncia pubblicati da attivisti Rohingya: in effetti, come sottolineato da Suzor, in molti casi l'algoritmo della piattaforma teneva conto delle ripetute e numerose segnalazioni effettuate da utenti birmani, facenti parte della maggioranza, rispetto a tali contenuti[53].

Tali effetti si riscontrano, del resto, anche a livello di *content curation*: la letteratura ha sottolineato come l'architettura algoritmica delle piattaforme, incentrata a massimizzare l'engagement degli utenti della rete, tenda a premiare in termini di visibilità i contenuti pubblicati dalle categorie demografiche di maggioranza[54], relegando a spazi di nicchia o imponendo un vero e proprio *shadowban*[55] a carico dei gruppi marginalizzati.

4. La moderazione dell'*hate speech* in un'ottica di uguaglianza sostanziale

A fronte di tali rilievi, occorre chiedersi, anche in un'ottica normativa e di *policy-making*, se il margine di errore caratterizzante gli strumenti di moderazione automatizzata dei discorsi d'odio sia effetti-

vamente accettabile o meno a fronte della necessità di garantire agli utenti un ciberspazio maggiormente tollerante e sicuro. In tal senso, sembra essere ineludibile un riferimento all'ormai risalente dibattito concernente la domanda se sia o meno opportuno combattere il fenomeno dell'*hate speech* attraverso l'imposizione di restrizioni alla libertà di espressione, pur nella consapevolezza che tale dibattito non si è sviluppato negli anni con riferimento alla relazione intercorrente tra individuo e intermediario digitale (rapporto tra soggetti privati) ma, piuttosto, con riferimento a quella intercorrente tra persona fisica e istituzioni dello Stato (rapporto tra un soggetto privato e un soggetto pubblico).

Come è noto, il dibattito sulla punibilità dei discorsi d'odio ha condotto, in prospettiva comparata, a soluzioni ben diverse tra loro[56]. Così, se negli USA vige il primato del Primo Emendamento e della tutela del "libero mercato delle idee"[57], con la conseguenza che una normativa volta a limitare la diffusione di *hate speech* debba essere sottoposta a un severissimo scrutinio (*strict scrutiny*) di legittimità costituzionale, quasi sempre fatale[58], il vecchio continente ha dimostrato una ben maggiore apertura a simili restrizioni. Invero, a differenza del Primo Emendamento, sia l'art. 10 della Convenzione europea per la salvaguardia dei diritti dell'uomo e delle libertà fondamentali (CEDU) sia l'art. 11 della Carta dei diritti fondamentali dell'Unione europea (Carta di Nizza) ammettono l'imposizione di restrizioni e limitazioni alla libertà d'espressione se previste dalla legge e se necessarie in una società democratica per il perseguimento di un fine legittimo, quale è, tra gli altri, la protezione della reputazione o dei diritti altrui[59].

La scelta di ostacolare, anche per mezzo del diritto, la diffusione dei discorsi d'odio può essere ascritta a una pluralità di ragioni tra loro complementari. In primo luogo, la proibizione e punibilità dell'*hate speech* rappresenta uno strumento per proteggere e tutelare gli individui appartenenti ad una classe discriminata dal perpetuarsi e aggravarsi degli episodi di discriminazione e violenza nei loro confronti. Se, come sottolineato dalla Commissione per l'eliminazione della discriminazione razziale (CERD), il discorso razzista e il discorso d'odio possono porre seri pericoli e rischi a medio-lungo termine[60], la loro limitazione rappresenta uno strumento essenziale per la riduzione di reati e illeciti di matrice discriminatoria: in tal senso, l'*hate speech* costituisce una condotta pericolosa in quanto potenzialmente capace di produrre conseguenze dannose per una società democratica[61].

In secondo luogo, è stato da più parti rilevato come l'atto del discorso d'odio sia in sé dannoso per l'integrità psicofisica dei suoi destinatari, nonché per l'esercizio dei loro diritti e delle loro libertà costituzionali. Secondo Matsuda, esponente della *critical race theory* statunitense, le vittime di *hate speech* e *hate propaganda* soffrono in percentuali più alte di sintomi e disturbi quali: sensazioni di panico; aumento del battito cardiaco; difficoltà respiratorie; incubi; disturbi da stress post-traumatico (PTSD); ipertensione; psicosi; suicidio[62]. In *Beizaras e Levickas c. Lituania*[63], la Corte EDU ha recentemente confermato che l'*hate speech*, relativo, nel caso di specie, all'orientamento sessuale dei ricorrenti, rappresenta in sé e per sé un attacco all'integrità fisica e mentale di coloro che ne sono i destinatari. Le vittime, inoltre, vengono attraverso l'*hate speech* ristrette nelle loro libertà, in quanto l'esigenza di sottrarsi a messaggi d'odio le porta a modificare le proprie abitudini di vita e, in molti casi, a rinunciare a esprimere le loro personali opinioni e idee[64]. In ultima istanza, come magistralmente posto in luce da Waldron[65], al cuore delle normative di contrasto al fenomeno in oggetto vi è la necessità di tutelare l'eguale dignità delle comunità vittime e dei singoli individui che ne fanno parte[66].

In altre parole, la regolazione delle espressioni d'odio è mossa sia dall'esigenza di contenere il rischio di ordine pubblico legato a un incremento dell'attività criminosa di matrice discriminatoria sia, soprattutto, da quella di garantire alle categorie "protette" la possibilità di esercitare liberamente i propri diritti e libertà in una condizione di uguaglianza rispetto al resto della popolazione. In questo senso, l'intervento normativo volto a ridurre la diffusione di *hate speech* rappresenta uno strumento volto ad affermare e concretizzare l'uguaglianza sostanziale, e non solo formale, dei gruppi demografici marginalizzati. Lo scopo dell'imposizione di limitazioni alla libertà di espressione per ridurre la diffusione di odio ha quindi come fine ultimo l'empowerment di quei soggetti che l'*hate speech* mira a colpire. Del resto, come sottolineato da Fredman, il perseguimento dell'uguaglianza sostanziale richiede esso stesso un approccio multidimensionale al fenomeno della discriminazione che implichi anche la garanzia che ai gruppi minoritari o comunque discriminati sia concesso partecipare attivamente alla vita comunitaria, pubblica e politica[67].

Se, dunque, la *ratio* ultima del contrasto ai discorsi d'odio è legata al perseguimento dell'uguaglianza sostanziale, anche nella sua dimensione partecipativa, appare evidente che, nel caso della moderazione automatizzata di *hate speech*, la soglia di accettabilità dell'errore, soprattutto se dettato da *bias* di carattere discriminatorio, debba essere particolarmente elevata. Un'applicazione inconsistente ed iniqua tradisce lo stesso spirito originario della moderazio-

ne dell'*hate speech*, svuotando tale attività del suo significato egualitario e rendendola, anzi, controproducente rispetto agli interessi della collettività. Tra l'altro, il silenziamento di quelle categorie di persone che costituiscono le vittime tipiche dei discorsi d'odio rischia di depotenziare fortemente il ruolo, ritenuto da più parti fondamentale, della controarrazione[68].

Per evitare tale cortocircuito, potenzialmente aggravato dal ricorso a tecniche di *new-school speech regulation*, appare pertanto auspicabile un ripensamento, da parte delle istituzioni europee, delle strategie politiche e legislative di settore. Ciò non tanto nell'ottica di una demonizzazione del moderatore algoritmico, il quale costituisce invece uno strumento essenziale e utilissimo per il contrasto all'*hate speech*[69], quanto piuttosto nella prospettiva, da un lato, di incentivare i programmatori di tali sistemi a tenere in adeguata considerazione le esigenze legate al rispetto dei principi dell'uguaglianza sostanziale[70] e, dall'altro lato, di fornire maggiori tutele individuali, sostanziali e soprattutto procedurali[71], a quegli utenti della rete che siano maggiormente esposti ai rischi della discriminazione algoritmica.

In realtà, la Commissione europea ha dato segno negli ultimi anni di una maggiore consapevolezza dei rischi per la libertà di espressione degli utenti che sono ineludibilmente legati a una più massiccia e generalizzata moderazione dei contenuti da parte degli intermediari digitali. Il Regolamento (UE) 2021/784[72] prevede per esempio all'art. 5 che un fornitore di servizi, il quale sia stato riconosciuto come esposto a contenuti terroristici, debba predisporre misure specifiche volte a contrastarne la diffusione: nell'applicare tali misure, tuttavia, il fornitore dovrà tenere pienamente conto dei diritti e degli interessi legittimi degli utilizzatori (ivi inclusa la libertà di espressione e di informazione) e, nel contempo, agire in maniera diligente e, soprattutto, non discriminatoria. È inoltre disposta, all'art. 10, la predisposizione di meccanismi di reclamo a tutela degli utenti i cui contenuti siano stati rimossi, con l'obbligo per il fornitore di rendere decisioni motivate e fatto salvo l'eventuale ricorso all'autorità amministrativa o giudiziaria dello Stato.

A sua volta, la proposta di regolamento per il *Digital Services Act* (DSA) contiene alcune norme di rilievo in tal senso, richiedendo all'art. 12 che gli intermediari applichino le condizioni generali dei loro servizi in modo «equo, trasparente, coerente, diligente, tempestivo, non arbitrario, non discriminatorio e proporzionato», nonché rispettoso dei diritti e degli interessi legittimi delle parti coinvolte (compresi i diritti fondamentali previsti dalla Carta di Nizza), nonché imponendo, alle piattaforme online, di predisporre sistemi interni di gestione dei reclami da attuarsi «in modo tempestivo, non discriminatorio, diligente e non arbitrario» (art. 17). In quest'ultimo caso, peraltro, si prevede espressamente che sia data la possibilità per gli utenti di contattare un interlocutore umano al momento della presentazione del reclamo e che la nuova decisione non possa essere presa solamente attraverso sistemi automatizzati: a tal fine, è fatto inoltre obbligo ai fornitori di servizi di dotarsi di personale qualificato. Dal testo approvato in prima lettura dal Parlamento europeo il 20 gennaio 2022[73] traspare tra l'altro una ancor maggiore consapevolezza del potenziale impatto del DSA sui diritti degli utenti: gli emendamenti proposti, per esempio, includono numerosi riferimenti al principio di non discriminazione.

Peraltro, si è da più parti rilevato come le soluzioni adottate rappresentino in ultima istanza poco più che petizioni di principio, in quanto sovente non corredate da un apparato applicativo e procedurale ben definito e sufficientemente sviluppato. Non è del tutto chiaro, per esempio, se l'art. 12 del DSA implichi la possibilità di opporre qualsiasi diritto ricompreso nella Carta di Nizza oppure soltanto quella ristretta cerchia di diritti per i quali la Corte di Giustizia dell'UE abbia dichiarato la sussistenza di un'efficacia orizzontale[74]. È stato, in generale, posto in luce come il sistema introdotto dal DSA incentiverebbe un ulteriore incremento nell'utilizzo su vasta scala di sistemi di moderazione automatizzati, senza tuttavia la previsione di adeguati rimedi a tutela dell'individuo[75].

Inoltre, se il DSA, pur nell'apprezzabilissima ottica di armonizzazione e riduzione della frammentarietà del quadro normativo sugli intermediari digitali, mira a introdurre una disciplina quadro generale e orizzontale, tale approccio, se non accompagnato da interventi normativi più specifici, ha tuttavia l'inevitabile effetto di appiattire le peculiarità tipiche connesse alla moderazione di ciascun *information bad*. Così, per quanto concerne la rimozione dei contenuti d'odio, non sembra essere presente, nell'attuale testo della proposta di regolamento, la consapevolezza dei rischi tipici, ai danni del principio di uguaglianza sostanziale, che sono inevitabilmente connessi alla rilevazione automatizzata dell'*hate speech*. Se, da un lato, la proposta di regolamento si preoccupa di tutelare i gruppi marginalizzati da contenuti dannosi quali l'"illecito incitamento all'odio" e i "contenuti discriminatori illegali"[76], non sufficiente attenzione è prestata al collaterale, ed altrettanto dannoso, rischio di un'iniqua rimozione degli stessi.

In una prospettiva normativa, risulta pertanto essenziale tenere in maggiore considerazione le specificità e le finalità tipiche del contrasto ai discorsi

d'odio, ovverosia l'uguaglianza sostanziale dei gruppi tradizionalmente marginalizzati e discriminati. Uno strumento promettente sembra essere stato introdotto, per esempio, dall'emendamento del Parlamento europeo volto a introdurre un paragrafo 1-*bis* all'art. 19. Se nel testo originario della Commissione la figura del "segnalatore attendibile" rilevava soltanto ai fini della premoderazione dei contenuti[77], il nuovo testo richiederebbe alle piattaforme online di adottare le misure, tecniche e organizzative, atte a permettere ai segnalatori attendibili di emettere notifiche di rettifica in caso di errore di moderazione: tali notifiche, volte al ripristino di informazioni e contenuti, dovranno essere trattate e decise in via prioritaria e senza indugio. In altre parole, si darebbe la possibilità a segnalatori attendibili, indipendenti ed esperti in materie quali, per l'appunto, il contrasto al fenomeno dei discorsi d'odio, di corroborare le richieste di correzione delle decisioni prese dalla piattaforma. Peraltro, allo stato attuale non risulta chiaro se, quando e con quali modalità sarà possibile per gli utenti richiedere direttamente un simile intervento del segnalatore attendibile.

5. Conclusioni

Il costante incremento del flusso informativo in rete ha reso sempre più essenziale il ruolo degli intermediari digitali nella moderazione dei contenuti postati dagli utenti, al fine di ridurre la commissione di condotte illecite e la diffusione di materiali dannosi o illeciti in Internet. La necessità di tale attività si evince del resto dal crescente numero di iniziative politiche e legislative da parte delle istituzioni pubbliche, nazionali e sovranazionali, volte a delineare sistemi di *new-school speech regulation*. A fronte, tuttavia, della mole straordinaria di contenuti postati quotidianamente online, il ricorso da parte degli intermediari digitali a forme di moderazione automatizzata si è fatto negli anni massiccio.

Sebbene tali sistemi siano sempre più avanzati e raffinati, il loro utilizzo non è esente da criticità: un certo margine di errore è, di fatto, ineludibile. Ciò risulta essere particolarmente evidente in quei casi ove la rilevazione del "male informazionale" si fondi sulla comprensione semantica del contesto e dell'intenzione dell'autore del contenuto, quale è il caso dei discorsi d'odio. In questi casi, come evidenziato da ampia letteratura, risulta particolarmente elevato il rischio di falsi positivi, soprattutto a carico delle minoranze e dei gruppi marginalizzati o discriminati. Il concreto e significativo rischio che la moderazione automatizzata di *hate speech* si traduca in un silenziamento delle categorie discriminate, piuttosto che in una loro tutela, implica la necessaria pretesa di una più esigente soglia di accettabilità dell'errore. In caso contrario, la moderazione dei contenuti d'odio si svuoterebbe di significato, tradendo la *ratio* di fondo che ne giustifica il contrasto: la promozione del principio di uguaglianza sostanziale.

Sotto il profilo di *policy-making*, risulta pertanto auspicabile da parte del legislatore, nazionale ma soprattutto eurounionale, una maggiore attenzione ai rischi "collaterali" connessi a un quadro normativo che incentivi la moderazione (algoritmica) dei discorsi d'odio senza garantire al contempo un apparato adeguato di tutela delle libertà individuali e del diritto di non discriminazione degli utenti. Tale esigenza appare ancor più pressante nell'attuale contesto post-pandemico e, soprattutto, con riferimento alla discussione in corso relativa all'emanazione del *Digital Services Act*.

Note

[1] *Reno v. American Civil Liberties Union*, 521 US 844 (1997).

[2] *Abrams v. United States*, 250 US 616 (1919). Sul punto, si vedano tra gli altri L.C. Bollinger, *The Tolerant Society: Freedom of Speech and Extremist Speech in America*, Oxford University Press, 1988, 304 p., p. 59-61; M. Rosenfeld, *Hate Speech in Constitutional Jurisprudence: A Comparative Analysis*, in "Cardozo Law Review", vol. 24, 2003, n. 4, p. 1523-1567, spec. p. 1533-1535.

[3] Cfr. D. Lupton, *Digital risk society*, in A. Burgess, A. Alemanno, J.O. Zinn et al. (eds.), "Routledge Handbook of Risk Studies", Routledge, 2016, p. 301-309.

[4] Si vedano, *ex multis*, Corte EDU, *Stoll c. Svizzera*, 10 dicembre 2007, ric. 69698/01; *K.U. c. Finlandia*, 2 dicembre 2008, ric. no. 2872/02; *Pravoye Delo e Shtekel c. Ucraina*, 5 maggio 2011, ric. 33014/05. Cfr. O. Pollicino, *Judicial protection of fundamental rights on the Internet: A road towards digital constitutionalism?*, Hart, 2021, XXIV+235 p.

[5] G. Sartor, A. Loreggia, *The impact of algorithms for online content filtering or moderation. "Upload filters"*, studio richiesto dal Comitato JURI del Parlamento europeo, n. PE 657.101), 2020.

[6] J. Grimmelmann, *The Virtues of Moderation*, in "Yale Journal of Law and Technology", 2015, n. 17, p. 42-109.

[7] T. Gillespie, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, 288 p., a p. 5. Si veda anche, sul punto, N. Helberger, J. Pierson, T. Poell, *Governing online platforms: From contested to cooperative responsibility*, in "The Information Society", vol. 23, 2018, n. 1, p. 1-14.

[8] Così T. Gillespie, *op. cit.*, p. 13: «And moderation is, in many ways, the commodity that platforms offer. Though part of the web, social media platforms promise to rise above it, by offering a better experience of all this information and sociality: curated, organized, archived, and moderated».

[9] R. Wilson, M. Land, *Hate Speech on Social Media: Content Moderation in Context*, in "Connecticut Law Review", vol. 52, 2021, n. 3, p. 1029-1076, spec. p. 1054.

[10] J.M. Balkin, *Old-School/New-School Speech Regulation*, in "Harvard Law Review", vol. 127, 2013, n. 8, p. 2296-2342. Si veda, sulla regolazione degli intermediari digitali, G. Frosio

(ed.), *The Oxford handbook of online intermediary liability*, Oxford University Press, 2020, 782 p.

[11]Si vedano altresì J.M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, in "U.C. Davis Law Review", vol. 51, 2017, n. 3, p. 1149-1210; J.M. Balkin, *Free Speech Is a Triangle*, in "Columbia Law Review", vol. 118, 2018, n. 7, p. 2011-2056.

[12]In una prima fase, la Commissione europea faceva per lo più ricorso a strumenti di auto-regolazione dal basso. Tra questi, il rimando è, soprattutto, al *Codice di condotta dell'UE per contrastare l'illecito incitamento all'odio* (2016), nonché al *Codice di buone pratiche sulla disinformazione* (2019). Progressivamente, l'approccio della Commissione è tuttavia mutato nella direzione di una maggiore regolazione "dall'alto", attraverso il ricorso sempre più diffuso a strumenti di *hard law*. Si vedano, in particolare: la Direttiva (UE) 2018/1808 del Parlamento europeo e del Consiglio, del 14 novembre 2018, recante modifica della direttiva 2010/13/UE, relativa al coordinamento di determinate disposizioni legislative, regolamentari e amministrative degli Stati membri concernenti la fornitura di servizi di media audiovisivi (direttiva sui servizi di media audiovisivi), in considerazione dell'evoluzione delle realtà del mercato (2018) OJ L61/69; la Direttiva (UE) 2019/790 del Parlamento europeo e del Consiglio, del 17 aprile 2019, sul diritto d'autore e sui diritti connessi nel mercato unico digitale e che modifica le direttive 96/9/CE e 2001/29/CE (2019) OJ L130/92; e il Regolamento (UE) 2021/784 del Parlamento europeo e del Consiglio, del 29 aprile 2021, relativo al contrasto della diffusione di contenuti terroristici online (2021) OJ L172/79. Si veda, da ultimo, la proposta per il cosiddetto *Digital Services Act*: Proposta di Regolamento del Parlamento europeo e del Consiglio relativo a un mercato unico dei servizi digitali (legge sui servizi digitali) e che modifica la direttiva 2000/31/CE, COM(2020) 825 final.

[13]Si veda, tra gli altri, R. Gorwa, R. Binns, C. Katzenbach, *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, in "Big Data & Society", vol. 7, 2020, n. 1, p. 1-15.

[14]Tra le numerose definizioni che sono state offerte del termine *hate speech*, si veda in particolare European Commission against Racism and Intolerance (ECRI), *General Policy Recommendation no. 15 on Combating Hate Speech*, 21 March 2016, CRI(2016)15, p. 16: «Hate speech for the purpose of the Recommendation entails the use of one or more particular forms of expression – namely, the advocacy, promotion or incitement of the denigration, hatred or vilification of a person or group of persons, as well any harassment, insult, negative stereotyping, stigmatization or threat of such person or persons and any justification of all these forms of expression – that is based on a non-exhaustive list of personal characteristics or status that includes "race", colour, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity and sexual orientation».

[15]J. Grimmelmann, *op. cit.*, p. 47.

[16]R. Gorwa, R. Binns, C. Katzenbach, *op. cit.*, p. 3.

[17]E. Llansó et al., *Artificial intelligence, Content Moderation, and Freedom of Expression*, TWG, 26 February 2020, 32 p.

[18]R. Gorwa, R. Binns, C. Katzenbach, *op. cit.*, p. 3.

[19]In questo senso, Wu parla di controllo "positivo" della libertà di espressione. Si veda T. Wu, *Will artificial intelligence eat the law? The rise of hybrid social-ordering systems*, in "Columbia Law Review", vol. 119, 2019, n. 7, p. 2001-2028, a p. 2014.

[20]Si vedano, tra gli altri, E. Llansó et al., *op. cit.*; C.R. Sunstein, *#Republic: Divided Democracy in the Age of Social Media*, Princeton University Press, 2017, XIV+316 p.; E. Pariser, *The filter bubble: what the Internet is hiding from you*, Penguin, 2011, 294 p.; N. Helberger et al., *A freedom of expression perspective on AI in the media – with a special focus on editorial decision making on social media platforms and in the news media*, in "European Journal of Law and Technology", vol. 11, 2020, n. 3, p. 1-28; S. Milano et al., *Recommender systems and their ethical challenges*, in "AI & Society", vol. 35, 2020, n. 4, p. 957-967; N.P. Suzor, *Lawless: The Secret Rules That Govern Our Digital Lives*, Cambridge University Press, 2019.

[21]Si veda, in particolare, K. Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, in "Harvard Law Review", vol. 131, 2017, n. 6, p. 1598-1670.

[22]G. De Gregorio, *Democratising online content moderation: A constitutional framework*, in "Computer Law & Security Review", vol. 36, 2020, p. 1-17.

[23]Le condizioni lavorative dei moderatori umani, con particolare riferimento ai danni psichici da essi sovente riportati, sono state trattate in particolare in S.T. Roberts, *Behind the screen: Content moderation in the shadows of social media*, Yale University Press, 2019, 266 p. A p. 25, l'autrice sottolinea come i lavoratori del settore siano generalmente «poorly paid human beings who risk burnout, desensitization, and worse because of the nature of their work». Si veda altresì Cambridge Consultants, *Use of AI in Online Content Moderation*, Ofcom, 2019.

[24]Si vedano, sul punto, Cambridge Consultants, *op. cit.*; R. Gorwa, R. Binns, C. Katzenbach, *op. cit.*; G. Sartor, A. Loreggia, *op. cit.*

[25]Cfr. J. Burrell, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, in "Big Data & Society", vol. 3, 2016, n. 1, p. 1-12; F. Pasquale, *The black box society: the secret algorithms that control money and information*, Harvard University Press, 2015, 311 p.

[26]J. Eisenstein, *Introduction to Natural Language Processing*, MIT Press, 2019, 536 p. A p. 1 si definisce il *natural language processing* come «the set of methods for making human language accessible to computers». Si veda anche N. Duarte, E. Llansó, A. Loup, *Mixed messages? The limits of automated social media content analysis*, Center for Democracy & Technology, 2017, p. 9. Per quanto concerne l'applicabilità specifica dell'NLP per la moderazione dell'*hate speech*, si veda A. Schmidt, M. Wiegand, *A Survey on Hate Speech Detection using Natural Language Processing*, in L.W. Ku, C.T. Li (eds.), "Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media", ACL, 2017, p. 1-10. Peraltro, un campo di ricerca particolarmente fertile sembra essere, soprattutto ai fini della rilevazione di contenuti d'odio, quello della *sentiment analysis* (o *opinion mining*). Sull'argomento, si vedano B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, II ed., Cambridge University Press, 2020, XVIII+430 p.; F.A. Pozzi et al., *Challenges of Sentiment Analysis in Social Networks: An Overview*, in Id., "Sentiment Analysis in Social Networks", Morgan Kaufmann, 2017, p. 1-11.

[27]E. Douek, *Governing online speech: from "posts-as-trumps" to proportionality and probability*, in "Columbia Law Review", vol. 121, 2021, n. 3, p. 759-834.

[28]S. Agarwal, C.R. Chowdary, *Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19*, in "Expert Systems with Applications", 2021, n. 185, p. 1-9; United Nations, *Countering COVID-19 Hate Speech*, 2020.

[29]G. De Gregorio, O. Pollicino, P. Dunn, *Digitisation and the central role of intermediaries in a post-pandemic world*, in "MediaLaws", 2021; F. Tagliabue, L. Galassi, P. Mariani, *The "Pandemic" of Disinformation in*

COVID-19, in "SN Comprehensive Clinical Medicine", 2020, n. 2, p. 1287-1289.

[30] M. Lim, G. Alrasheed, *Beyond a technical bug: Biased algorithms and moderation are censoring activists on social media*, in "The Conversation", 2021.

[31] Facebook Transparency Center, *Community Standards Enforcement Report – Hate Speech*, Meta, 2022.

[32] I dati concernenti la moderazione dei discorsi d'odio su Instagram sono disponibili solo con riferimento al periodo successivo all'ultimo trimestre del 2019.

[33] Occorre sottolineare, peraltro, che nel momento in cui si scrive non è possibile ancora prevedere quali saranno gli effetti, a livello di moderazione dei contenuti d'odio, dello scoppio del conflitto russo-ucraino nel febbraio 2022.

[34] M. Schroepfer, *Update on Our Progress on AI and Hate Speech Detection*, Meta, 2021.

[35] Facebook Transparency Center, *Community Standards Enforcement Report*, cit.

[36] Per una definizione di "diffusione" si veda Facebook Transparency Center, *Prevalence*, Meta, 2021.

[37] T. Simonite, *Facebook's AI for Hate Speech Improves. How Much Is Unclear*, in "Wired", 2020.

[38] Peraltro, alcuni dati sono disponibili relativamente al numero di contenuti successivamente reintegrati sulle piattaforme. Tuttavia, soprattutto a seguito dello scoppio della pandemia, il ruolo dei reclami proposti dagli utenti sembra essere piuttosto marginale. Nel terzo trimestre del 2021, per esempio, a fronte di 22,3 milioni di contenuti sanzionati su Facebook, solo 1,1 milione di reclami sono stati proposti dagli utenti: di questi solo 90,7 mila sono stati accolti. Questo a fronte dei circa 30,3 mila contenuti reintegrati autonomamente da Facebook. In generale, tali dati appaiono essere poco rappresentativi del reale tasso di errore. Facebook Transparency Center, *Community Standards Enforcement Report*, cit.

[39] G. Sartor, A. Loreggia, *op. cit.*

[40] E. Douek, *op. cit.*

[41] A. Brown, *What Is Hate Speech? Part 2: Family Resemblances*, in "Law and Philosophy", vol. 36, 2017, n. 5, p. 561-613; P. Dunn, *Piattaforme digitali e moderazione dei contenuti d'odio: nodi giuridici e pratici*, in "MediaLaws", 2021.

[42] Si veda, in tal senso, R. Wilson, M. Land, *op. cit.* Per la (ampia) definizione utilizzata dalle piattaforme di Meta, si veda Facebook Transparency Center, *Hate speech*, Meta, 2021.

[43] Si veda, in particolare, il c.d. "Piano d'Azione Rabat" delle Nazioni Unite. Consiglio per i diritti umani delle Nazioni Unite, *Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred* (A/HRC/22/17/Add.4), 2013. Si veda altresì A. Weber, *Manual on hate speech*, Council of Europe Publishing, 2009, VI+98 p.

[44] M. Durante, *Potere computazionale. L'impatto delle ICT su diritto, società, sapere*, Meltemi, 2019, 397 p.

[45] L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo* (trad. it. M. Durante), Raffaello Cortina, 2017, XVIII+294 p., pp. 147-164.

[46] Per una comprensione del fenomeno del *meme* in Internet, si veda G. Marino, *Semiotics of spreadability: A systematic approach to Internet memes and virality*, in "Punctum", vol. 1, 2015, n. 1, p. 43-66, a p. 60.

[47] Per uno studio su come gli algoritmi utilizzati da piattaforme e intermediari digitali abbiano la tendenza a riprodurre e replicare *bias* discriminatori, soprattutto nei confronti delle donne afro-americane, si veda S.U. Noble, *Algorithms of oppression: how search engines reinforce racism*, New York University Press, 2018.

[48] J.H. Park, J. Shin, P. Fung, *Reducing Gender Bias in Abusive Language Detection*, in E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (eds.), "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing", ACL, 2018, p. 2799-2804; X. Zhou, M. Sap, S. Swayamdipta et al. (eds.), *Challenges in Automated Debiasing for Toxic Language Detection*, in P. Merlo, J. Tiedemann, R. Tsarfaty (eds.), "Proceedings of the Sixteenth Conference of the European Chapter of the Association for Computational Linguistics: Main Volume", ACL, 2021, p. 3143-3155.

[49] T. Davidson, D. Warmsley, M. Macy, I. Weber, *Automated Hate Speech Detection and the Problem of Offensive Language*, in "Proceedings of the Eleventh International AAAI Conference on Web and Social Media", vol. 11, 2017, n. 1, p. 512-515; T. Davidson, D. Bhattacharya, I. Weber, *Racial Bias in Hate Speech and Abusive Language Detection Datasets*, in T.S. Roberts, J. Tetreault, V. Prabhakaran, Z. Waseem (eds.), "Proceedings of the Third Workshop on Abusive Language Online", ACL, 2019, p. 25-35; M. Sap, D. Card, S. Gabriel et al., *The Risk of Racial Bias in Hate Speech Detection*, in A. Korhonen, D. Traum, L. Màrquez (eds.), "Proceedings of the Fiftyseventh Annual Meeting of the Association for Computational Linguistics", ACL, 2019, p. 1668-1678.

[50] T. Dias Oliva, D.M. Antonialli, A. Gomes, *Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online*, in "Sexuality & Culture", vol. 25, 2021, n. 2, p. 700-732. Il lavoro analizza come il software *Perspective*, sviluppato da Google per individuare contenuti ascrivibili alla classe del *toxic speech*, impatti rispettivamente i tweet pubblicati da celebri *drag queen* statunitensi e quelli pubblicati da altrettanto noti estremisti di destra. Secondo lo studio, i contenuti prodotti dalle prime sarebbero in molti casi rilevati come altrettanto tossici, e spesso come più tossici, dei secondi. Così, per esempio, il tweet «and I'm ... GAY. #HairsprayLive» della *drag queen* Mimi Imfurst risulterebbe essere tossico al 92,31% (p. 720), mentre il tweet del politico *alt-right* Richard Spencer «@hodgie2000 Of course, homosexuality is a naturally occurring phenomenon. But as another already said, so is cannibalism, addiction, suicide, self-harm, etc. The question is: what is the *cause* of this curious phenomenon, from evolutionary, genetic, social, or psychological perspectives», in cui l'omosessualità è di fatto paragonata al cannibalismo, alle dipendenze, al suicidio e all'autolesionismo, riporterebbe un grado di tossicità pari al solo 13,80% (p. 724).

[51] Con riferimento alla comunità LGBTQIA+, e in particolare alla sotto-comunità *drag*, si veda S. McKinnon, *"Building a thick skin for each other": The use of 'reading' as an interactional practice of mock impoliteness in drag queen backstage talk*, in "Journal of Language and Sexuality", vol. 6, 2017, n. 1, p. 90-127. L'autore, nell'occuparsi della pratica del *reading*, riconduce l'uso di espressioni apparentemente insultanti alla categoria della *mock impoliteness*, da intendersi come l'insieme di quelle «utterances, which could potentially be evaluated as genuine impoliteness outside the appropriate context, are positively evaluated by in-group members who recognize the importance of 'building a thick skin' to face a hostile environment from LGBT and non-LGBT people» (p. 90).

[52] J. Guynn, *Facebook while black: Users call it getting «Zucked», say talking about racism is censored as hate speech*, in "Usa Today", 2019. Tale fenomeno, inoltre, sembra colpire in modo ancora più intenso le donne: si veda in tal senso K.L. Gray, K. Stein, *"We 'said her name' and got zucked": Black Women Calling-out the Carceral Logics of Digital Platforms*, in "Gender & Society", vol. 35, 2021, n. 4, p. 538-545.

[53] N.P. Suzor, *op. cit.*, p. 128-129.

[54] A. Chakraborty, J. Messias, F. Benevenuto et al., *Who Makes Trends? Understanding Demographic Biases in*

Crowdsourced Recommendations, in "Proceedings of the Eleventh International AAAI Conference on Web and Social Media", vol. 11, 2017, n. 1, p. 22-31. Come posto in luce in E. Llansó et al., *op. cit.*, in numerosi casi gli algoritmi di *content curation* tendono, anzi, a premiare contenuti fortemente controversi quali *hate speech* e fake news. Ciò appare essere stato confermato dallo scandalo legato ai cosiddetti "Facebook Papers". The New York Times, *The Facebook Papers and their fallout*, 28 October 2021.

[55] C. Are, *The Shadowban Cycle: An autoethnography of pole dancing, nudity and censorship on Instagram*, in "Feminist Media Studies", 2021, p. 1-18.

[56] M. Rosenfeld, *op. cit.*

[57] J. Weinstein, *An Overview of American Free Speech Doctrine and Its Application to Extreme Speech*, in I. Hare, J. Weinstein (eds.), "Extreme Speech and Democracy", Oxford University Press, 2009, p. 81-91.

[58] L. Kendrick, *Content Discrimination Revisited*, in "Virginia Law Review", vol. 98, 2012, n. 2, p. 231-300, a p. 237.

[59] G. Pitruzzella, O. Pollicino, *Disinformation and hate speech*, Bocconi University Press, 2020, VI+168 p.

[60] Comitato per l'eliminazione della discriminazione razziale (CERD), *General recommendation No. 35: Combating racist hate speech*, CERD/C/GC/35, 2013.

[61] Così, per esempio, Comitato per i diritti umani, *Faurisson c. Francia*, 8 novembre 1996, Comunicazione n. 550/1993: «Since the statements made by the author, read in their full context, were of a nature as to raise or strengthen anti-semitic feelings, the restriction served the respect of the Jewish community to live free from fear of an atmosphere of anti-semitism».

[62] M.J. Matsuda, *Public Response to Racist Speech: Considering the Victim's Story*, in M.J. Matsuda et al. (eds.), "Critical Race Theory, Assaultive Speech, and the First Amendment", Westview, 1993, p. 17-51, a p. 24.

[63] Corte EDU, *Beizaras e Levickas c. Lituania*, ric. no. 41288/15, sent. del 14 gennaio 2020.

[64] Così M.J. Matsuda, *op. cit.*, p. 24-25: «Victims are restricted in their personal freedom. To avoid receiving hate messages, victims have to quit jobs, forgo education, leave their homes, avoid certain public places, curtail their own exercise of speech rights, and otherwise modify their behavior and demeanor. The recipient of hate messages struggles with inner turmoil. One subconscious response is to reject one's own identity as a victim-group member. As writers portraying the African-American experience have noted, the price of disassociating from one's own race is often sanity itself».

[65] Si veda J. Waldron, *The harm in hate speech*, Harvard University Press, 2012, VI+292 p., a p. 105-143.

[66] Così, del resto, Corte EDU, *Féret c. Belgio*, 16 luglio 2009, ric. 15615/07: «La tolérance et le respect de l'égale dignité de tous les êtres humains constituent le fondement d'une société démocratique et pluraliste. Il en résulte qu'en principe on peut juger nécessaire, dans les sociétés démocratiques, de sanctionner, voire de prévenir, toutes les formes d'expression qui propagent, encouragent, promeuvent ou justifient la haine fondée sur l'intolérance [...]».

[67] Così S. Fredman, *Substantive equality revisited*, in "International Journal of Constitutional Law", vol. 14, 2016, n. 3, p. 712-738, a p. 731-732: «The right to equality is concerned with two aspects of participation. The first is political. Given that past discrimination or other social mechanisms have blocked the avenues for political participation by particular minorities, equality laws are needed both to compensate for this absence of political voice and to open up the channels for greater participation in the future. [...] The second aspect of the participative dimension is to address the importance of community in the life of individuals. Rather than the universal, abstract individual of formal equality, substantive equality recognizes that individuals are essentially social. To be fully human includes the ability to participate on equal terms in community and society more generally».

[68] F. Faloppa, *#Odio. Manuale di resistenza alla violenza delle parole*, UTET, 2020, 304 p., a p. 199 definisce la controNarrazione come «una narrazione a breve termine, che nasce come risposta diretta e più immediata a uno specifico discorso, o a una specifica narrazione, d'odio» e aggiunge che essa «ha quindi l'obiettivo di evidenziare le incoerenze della narrazione che si vuole contrastare, tentando così di sfidarla sullo stesso piano, indebolirla portandone a galla i meccanismi, smantellarla e delegittimarla». Sul ruolo della controNarrazione nel contrasto all'*hate speech*, si veda European Commission against racism and intolerance (ECRI), *op. cit.* Si veda inoltre R. Cohen-Almagor, *Countering Hate on the Internet*, in "Annual Review of Law and Ethics", vol. 22, 2014, p. 431-443, a p. 435.

[69] G. Ziccardi, *Online Political Hate Speech in Europe: The Rise of New Extremisms*, Edward Elgar Publishing, 2020, p. 116-121.

[70] Sul punto si veda, in particolare, S. Wachter, B. Mittelstadt, C. Russell, *Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law*, in "West Virginia Law Review", vol. 123, 2020, n. 3, p. 735-790. Gli autori sottolineano come una prospettiva di uguaglianza sostanziale, anziché meramente formale, sia più in linea con il panorama normativo e i valori costituzionali caratterizzanti l'Unione europea. A tal fine, propongono il ricorso a quelli che loro definiscono *bias transforming metrics* (i quali tengono conto delle disuguaglianze storiche tra categorie di persone) quali parametri per valutare la *fairness* di un sistema di *machine learning*, anziché ai *bias preserving metrics* (che assumono invece una prospettiva di uguaglianza formale degli individui). «Put simply, developers have a choice between two types of metrics: (1) "bias preserving" metrics that take society as it currently exists as a neutral starting point or "level playing field" from which we can measure inequality and bias in machine learning; and (2) "bias transforming" metrics that acknowledge historical inequalities and start from the assumption that certain groups will have a worse starting point than others. [...] [O]ur choice of fairness metric can ensure machine learning applications do not exacerbate existing inequalities and fully acknowledge the extent and significance of existing inequalities. The choice of variables to condition on for fairness tests, thresholds for illegal disparity, and acceptable arguments to justify disparity are difficult political determinations» (p. 778).

[71] Cfr. G. De Gregorio, *op. cit.*

[72] Regolamento (UE) 2021/784, cit.

[73] Emendamenti del Parlamento europeo, approvati il 20 gennaio 2022, alla proposta di regolamento del Parlamento europeo e del Consiglio relativo a un mercato unico dei servizi digitali (legge sui servizi digitali) e che modifica la direttiva 2000/31/CE, P9_TA(2022)0014.

[74] N. Appelman, J. Quintais, R. Fahy, *Article 12 DSA: Will platforms be required to apply EU fundamental rights in content moderation decisions?*, in "DSA Observatory", 31 May 2021. Sul tema dell'efficacia orizzontale dei diritti fondamentali tutelati dalla Carta di Nizza si veda E. Frantziou, *The Horizontal Effect of the Charter: Towards an Understanding of Horizontality as a Structural Constitutional Principle*, in "Cambridge Yearbook of European Legal Studies", 2020, n. 22, p. 208-232.

[75] J. Barata, *The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations*, in "DSA Observatory", 27 July 2021.

[76] Cfr. considerando 12.

[77]L'art. 19 prevede che, in caso di segnalazioni pervenute da soggetti che, sulla base di criteri oggettivi, siano stati riconosciuti quali "segnalatori attendibili", le piattaforme online dovranno valutare tali segnalazioni in via prioritaria.

* * *

Automated content moderation and algorithmic discrimination: the case of hate speech

Abstract: The need for Internet intermediaries to moderate user-generated content has become more and more pressing. Besides, vis-à-vis the extraordinary increase in the quantity of daily online information, the resort to algorithmic tools for moderation is today essential. This is also true for the detection of hate speech acts, which is currently largely based on the use of AI and machine-learning techniques: however, scholarly literature has highlighted how such systems can often be vitiated by discriminatory biases which produce a high risk of false positives affecting minorities. The present contribution argues that, within the European constitutional framework, the fight against hateful contents finds its rationale in the goal of ensuring that all social groups can truly enjoy a substantive equality, and that, as a consequence, a discriminatory enforcement of hate speech bans is inconsistent with the value system of the EU. Therefore, although AI represents a fundamental and necessary tool to guarantee a safer and more tolerant digital ecosystem, a high rate of false positives is not fully acceptable when it comes to hate speech moderation. It is thus necessary to rethink the relevant political and legislative strategies, with a view to ensure that marginalised groups can enjoy appropriate substantive and procedural guarantees protecting their freedom of expression and their right to non-discrimination.

Keywords: Automated content moderation – Hate speech – Algorithmic discrimination – Substantive equality – Freedom of expression