H-ermes. Journal of Communication
H-ermes, J. Comm. 23 (2023), 23-76
ISSN 2284-0753, DOI 10.1285/i22840753n23p23
http://siba-ese.unisalento.it

Analisi normativa in tema di contrasto agli hate speech su Internet e i social media.

Pietro Falletta, LUISS Guido Carli

Regulatory analysis in the field of contrast to hate speech on the Internet and social media. *The rise of the Internet and the development of new media fostered the spread of Hate Speech. The present study moves in two directions: the first provides a comprehensive overview of regulations at international, European and national level; the second, on the other hand, defines guidelines, policies and sanctions introduced by digital platforms. The analysis of the regulatory framework concerns the differences between the legal systems, the approaches used to protect freedom of expression and their influence on the activities of online platforms. Moreover, four operators of online services were examined: Facebook, Twitter, Instagram and YouTube. The analysis starts with the explanation of these platforms' general guidelines and then develops into their practical application, impact and results*.
Keywords: hate speech, social media, digital platforms, regulatory frameworks, freedom of expression.

Introduzione

Internet e i nuovi mezzi di comunicazione hanno favorito il progressivo sviluppo del fenomeno dell'*hate speech*, specialmente attraverso i canali e le piattaforme di social network.

I social media nascono, infatti, come spazio di agevolazione del dibattito pubblico. Possono dunque ospitare, e talvolta enfatizzare discussioni offensive, persino violente, agendo da gigantesche casse di risonanza dell'odio.

Rispetto al dilagare del fenomeno dell'*hate speech* online, il presente studio intende indagare su due livelli di intervento nel contrasto ai discorsi d'odio. In una prima parte, si intende fornire una ricostruzione del quadro normativo, a livello internazionale, europeo e nazionale, per poi analizzare, nella seconda parte, linee guida, policy e misure sanzionatorie messe in campo dalle piattaforme digitali.

Per quanto concerne il quadro normativo, va evidenziata, anzitutto, una diversità di approccio tra ordinamento statunitense, da una parte, e ordinamento europeo e dei singoli Stati membri, dall'altra, circa la tutela da accordare alla libertà di espressione e, quindi, circa gli eventuali limiti apponibili ad essa, tra cui certamente rientra il contrasto all'*hate speech*. Tali differenze influiscono inevitabilmente sull'attività e sulla posizione delle piattaforme online, che hanno negli Stati Uniti il paese di origine, ma che operano e offrono servizi a livello globale. Ciò richiede quindi di verificare, altresì, la disciplina relativa alla

responsabilità dei provider rispetto ai contenuti diffusi e ospitati nelle rispettive piattaforme.

Nell'analisi delle policy e delle azioni poste in essere dai social media, si è scelto di studiare quattro gestori di servizi online, selezionati sulla base della loro popolarità ed eterogeneità: una piattaforma di social networking (Facebook), una di microblogging (Twitter) e due di condivisione foto e video (Instagram, YouTube).

Rispetto a tali piattaforme, sono dapprima selezionati i punti cardine delle linee guida generali di cui esse si sono dotate per definire i contenuti di incitamento all'odio; successivamente, viene analizzata la loro applicazione pratica attraverso misure operative quali policy e funzioni; infine, l'impatto di tali misure viene valutato mediante l'analisi dei risultati ottenuti.

Parte prima. *Mappatura, descrizione e analisi del quadro normativo a livello italiano e internazionale*

1. Il quadro normativo in materia di contrasto all'hate speech

È sempre più frequente, soprattutto nel contesto online, che la libera espressione del pensiero vada oltre i confini del *free speech* e sfoci nel cosiddetto *hate speech*, vale a dire in discorsi d'odio, espressioni che contengono elementi discriminatori nei confronti di individui o gruppi in ragione della loro origine razziale, etnica, religiosa, culturale, di genere o per l'orientamento sessuale.

Sebbene si tratti di manifestazioni che non nascono con i nuovi mezzi di informazione e di comunicazione, è evidente che esse hanno trovato nello spazio online canali con una potenzialità diffusiva e di propagazione certamente maggiori. Ciò è vero, anzitutto, per l'innegabile espansione delle forme di intolleranza che trovano nelle caratteristiche ontologiche della rete una evidente accentuazione. La diffusività del mezzo, la possibilità di agire in anonimato o di utilizzare profili falsi, a cui si aggiungono i fenomeni della polarizzazione di gruppo e della balcanizzazione delle idee, finiscono difatti per alimentare esponenzialmente la spirale d'odio.

Rispetto al contrasto all'*hate speech*, il quadro normativo si caratterizza per una disciplina fondata su diversi livelli di intervento: internazionale, europeo e nazionale.

Sul piano internazionale, il riferimento è, senza dubbio, alla Convenzione internazionale sull'eliminazione di ogni forma di discriminazione razziale, conclusa a New York il 21 dicembre 1965. La Convenzione, si legge già nell'art. 2, invita gli Stati a condannare qualsiasi forma di discriminazione razziale e a promuovere politiche volte alla eliminazione di ogni forma di discriminazione razziale e a favorire l'intesa fra le razze.

Un espresso divieto di incitamento all'odio è stato inserito nel Patto sui diritti civili e politici del 16 dicembre 1966, in cui si impone agli Stati membri di vietare "qualsiasi appello all'odio nazionale, razziale o religioso che costituisca incitamento alla discriminazione, all'ostilità o alla violenza" (art. 20).

La Convenzione europea dei diritti dell'uomo (CEDU) è poi uno dei primi documenti internazionali che contiene una esplicita previsione sul principio di non discriminazione, ove afferma che

> il godimento dei diritti e delle libertà riconosciuti nella presente Convenzione deve essere assicurato senza nessuna discriminazione, in particolare quelle fondate sul sesso, la razza, il colore, la lingua, la religione, le opinioni politiche e quelle di altro genere, l'origine nazionale o sociale, l'appartenenza a una minoranza nazionale, la ricchezza, la nascita o ogni altra condizione (art. 14).

L'art. 10, dedicato alla libertà di espressione, prevede che siano poste limitazioni a detta libertà se necessarie in una società democratica.

Con specifico riferimento al contesto online, vi è da tempo un'attenzione da parte delle istituzioni internazionali. Del 2004 è la Convenzione del Consiglio d'Europa sulla criminalità informatica, che fornisce le linee guida per gli Stati che vogliano dotarsi di una legislazione nazionale sul *cybercrime*. Con specifico riferimento alla lotta alla discriminazione nel contesto online, va in particolare considerato il Protocollo addizionale, che nasce proprio dalla necessità di "assicurare un buon equilibrio tra la libertà d'espressione e una lotta efficace contro gli atti di natura razzista e xenofoba" perpetrati tramite Internet (Considerando 12).

È chiaro come le risposte che i diversi ordinamenti hanno dato al contrasto all'*hate speech,* sia online che offline, variano a seconda dell'ampiezza accordata alla libertà di espressione. Ne costituiscono una piena evidenza le differenti scelte compiute in materia dall'ordinamento statunitense e da quello europeo.

È bene sin da subito sottolineare come i diversi approcci dei due modelli impattino necessariamente sull'attività e sulla posizione delle piattaforme online che hanno il luogo di origine e di stabilimento negli Stati Uniti e risentono, quindi, di un clima maggiormente favorevole alla circolazione dei contenuti; dal momento, però, che operano e offrono i loro servizi in Europa devono necessariamente confrontarsi con le relative regole.

1.1. L'hate speech online nell'ordinamento statunitense

Come noto, il *free speech* trova un'ampia tutela nell'ordinamento statunitense. Difatti, il Primo Emendamento della Costituzione degli Stati Uniti contiene un divieto, per il legislatore, di limitare la libertà di espressione. Si tratta di un divieto che si rivolge tanto al Congresso, quanto (in base alla *equal protection clause* del Quattordicesimo Emendamento) agli Stati federati.

L'impostazione statunitense si basa, in linea generale, sul *market place of ideas*, per cui al di fuori di casi eccezionali – e nella specie quando il *free speech* si traduca in un *clear and present danger* – anche le opinioni più impopolari e scabrose devono godere di protezione, perché parte di un libero mercato delle idee, dove la corretta informazione è destinata a emergere dal confronto tra opinioni contrastanti.

Inizialmente, la dottrina del *clear and present danger* conduce a una lettura conservatrice, per cui la pericolosità potenziale del discorso giustifica limitazioni alla libera espressione del pensiero. Successivamente, però, la nocività dell'espressione, e quindi la sua limitazione, viene legata alla concreta probabilità di realizzare l'evento dannoso. In tale contesto, spetta al giudice valutare nel caso di specie il nesso di causalità tra l'espressione e l'evento minacciato. Il Primo emendamento troverà poi una lettura più ampia da parte della giurisprudenza della Corte Suprema statunitense, che riterrà illegittimi tutti gli interventi che vadano a

limitare il puro pensiero e che non siano legati a un pericolo concreto e imminente.

Più di recente, tale impostazione è stata ridimensionata – con una conseguente maggiore limitazione del *free speech* – nei confronti di espressioni d'odio legate al terrorismo. In nome della sicurezza nazionale, il *market place of ideas* ha quindi subito un ridimensionamento. Si può citare, in proposito, una sentenza della Corte suprema del 2010 (*Holder v. Humanitarian Law Project*, 561 U.S. 1). Nel valutare la compatibilità con il Primo emendamento della legislazione introdotta per contrastare il terrorismo internazionale – nella specie la *Section* 23339 A e la *Section* 2239 B dell'*Antiterrorism and Effective Death Penality Act* – il giudice statunitense ritiene sufficiente la mera consapevolezza di svolgere attività in coordinamento con organizzazioni terroristiche straniere individuate dall'esecutivo come minaccia alla sicurezza interna, perché si possa prevedere delle limitazioni della libertà di espressione. Viene, quindi, punito il supporto materiale a dette organizzazioni, sebbene questo si estrinsechi in un'attività di puro pensiero, in quanto potrebbe dar luogo a ipotesi di *coordinated speech* vietato dalla legge federale.

In ogni caso, con particolare riguardo a quelle manifestazioni del pensiero che sfocino nell'*hate speech*, la giurisprudenza della Corte Suprema dimostra un approccio piuttosto permissivo financo nei confronti di espressioni legate alla religione, alla razza o all'etnia, particolarmente odiose, intolleranti e offensive.

Non è sufficiente, per tale giurisprudenza, una semplice istigazione all'odio. In un simile contesto si guarda con sospetto a qualsiasi forma di responsabilizzazione dei provider, giacché sistemi di filtraggio o qualunque intervento preventivo sui contenuti potrebbe dar luogo a forme di *collateral censoship.* Ciò spiega la scelta compiuta dal legislatore con il *Communication Decency Act* del 1996 di escludere una responsabilità, al pari di un editore, per il fornitore o l'utilizzatore di servizi digitali rispetto ai contenuti inseriti da terzi sulla propria piattaforma (art. 230, sect. C, 1 CDA).

Gli Stati Uniti hanno quindi inteso disciplinare la responsabilità delle piattaforme digitali con il fine di garantire la massima circolazione di contenuti − e così la massima libertà di espressione del pensiero – in linea con l'approccio di

valore quasi sacrale accordato al Primo emendamento. Si è esclusa, per tale via, una regolazione contenutistica di Internet, evitando di parificare le piattaforme a editori, mantenendo immuni i prestatori da responsabilità per la loro attività di moderazione. L'esenzione di responsabilità trova un limitato ambito di eccezione nel caso di violazione di norme penali federali, di diritti di proprietà intellettuale e di norme sulla riservatezza nelle comunicazioni elettroniche.

Tale impostazione è seguita anche dalla giurisprudenza. Si può citare, in proposito, la decisione della Corte del *Northen District of California* (Decisione del 18 novembre 2016, *Fields v. Twitter*) in cui sono stati respinti i ricorsi presentati nei confronti di Twitter e altri social network da parte dei parenti di alcune delle vittime di attentati terroristici di matrice islamica. Di fronte alle accuse mosse nei confronti dei social di dare spazio, tramite le loro piattaforme, alla propaganda islamista e al reclutamento di nuovi adepti, i giudici americani ritengono che il gestore della piattaforma non può essere considerato responsabile dei contenuti prodotti e diffusi dagli utenti.

1.2. L'impegno dell'UE nel contrasto ai discorsi d'odio in rete

L'ordinamento europeo dimostra nei confronti dell'*hate speech* un atteggiamento di minore tolleranza che deriva, come anticipato, dal grado di tutela accordato alla libertà di parola, la quale non gode nel vecchio continente della sacrale considerazione che l'ordinamento statunitense le riconosce in una società democratica e pluralista.

La libertà di espressione trova riconoscimento nell'art. 11 della *Carta dei diritti fondamentali dell'Unione europea* la quale ammette limitazioni all'esercizio dei diritti e delle libertà ivi riconosciute qualora siano necessarie e rispondano effettivamente a finalità di interesse generale riconosciute dall'Unione o all'esigenza di proteggere i diritti e le libertà altrui[1].

[1] L'art. 11 della Carta dei diritti fondamentali dell'Unione europea dedicato alla libertà di espressione e d'informazione prevede che "1. Ogni individuo ha diritto alla libertà di espressione. Tale diritto include la libertà di opinione e la libertà di ricevere o di comunicare informazioni o idee senza che vi possa essere ingerenza da parte delle autorità pubbliche e senza limiti di frontiera. 2. La libertà dei media e il loro pluralismo sono rispettati".

Per l'ordinamento europeo, inoltre, il principio di non discriminazione è un principio giuridicamente vincolante, sancito dall'art. 21 della Carta dei diritti fondamentali, secondo cui

> è vietata qualsiasi forma di discriminazione fondata, in particolare, sul sesso, la razza, il colore della pelle o l'origine etnica o sociale, le caratteristiche genetiche, la lingua, la religione o le convinzioni personali, le opinioni politiche o di qualsiasi altra natura, l'appartenenza ad una minoranza nazionale, il patrimonio, la nascita, la disabilità, l'età o l'orientamento sessuale.

Gli articoli 9 e 10 del *Trattato sul funzionamento dell'Unione europea* (TFUE) indicano, quali obiettivi delle politiche europee, la promozione di un elevato livello di istruzione e la lotta contro ogni tipo di discriminazione. L'art. 19 del medesimo *Trattato* ribadisce poi l'esigenza di contrastare attivamente ogni forma di discriminazione

A livello di diritto derivato, vanno considerate due direttive del 2000, che hanno introdotto rispettivamente il divieto di discriminazione in ragione dell'origine razziale o etnica (Direttiva 2000/43/CE), e il divieto di discriminazione per motivi religiosi, anagrafici, per l'orientamento sessuale, le convinzioni personali, per quanto concerne l'occupazione e le condizioni di lavoro (Direttiva 2000/78/CE).

Ancora più incisiva è la decisione quadro 2008/913/GAI sulla lotta contro il razzismo e la xenofobia che qualifica come reato, estendendolo espressamente anche all'online, l'istigazione pubblica alla violenza o all'odio nei confronti di un gruppo o di un singolo membro in ragione della razza, della religione e dell'ascendenza o dell'origine nazionale o etnica. Sulla decisione quadro è poi intervenuto il Parlamento europeo con la risoluzione del 14 marzo 2013, in cui si evidenzia la necessità di un intervento di revisione, al fine di includere anche le manifestazioni di antisemitismo, intolleranza religiosa, antiziganismo, omofobia e transfobia.

In accordo con quanto stabilito dalla decisione quadro è stato istituito, su iniziativa della Commissione europea, un Internet Forum che riunisce i Ministri degli Interni degli Stati membri dell'Unione Europea, oltre ai rappresentanti dei principali provider e delle istituzioni europee, al fine di individuare sistemi che

ostacolino la diffusione di contenuti che inneggiano all'odio, alla violenza o al terrorismo internazionale[2].

La centralità della lotta alla discriminazione è dimostrata, più di recente, dal *Piano d'azione dell'UE contro il razzismo 2020-2025*, che consiste in una serie di misure volte ad intensificare gli interventi, ad aiutare le persone appartenenti a minoranze razziali o etniche e a riunire i soggetti interessati nel contrasto efficace del razzismo.

Con riferimento specifico al contesto online, va segnalata la tendenza delle istituzioni europee ad incentivare, nel contrasto ai contenuti illeciti, forme di regolazione basate sul c.d. *soft law* e sulla cooperazione delle stesse piattaforme online.

D'altronde, l'Unione Europea ha mostrato – almeno fino agli interventi più recenti – di privilegiare forme di autoregolazione e co-regolazione, riconoscendo sempre più potere alle piattaforme in merito alla circolazione di contenuti illeciti, pur mantenendo invariata la disciplina sulla responsabilità dei provider. Quest'ultima stabilisce una generale esenzione di responsabilità e un divieto, per gli Stati membri, di prevedere un obbligo di sorveglianza da parte delle piattaforme sui contenuti inseriti, così come di ricercare attivamente fatti idonei ad indicare la presenza di attività illecite (art. 15, Direttiva 2000/31/CE). L'*internet service provider*, in qualità di *hosting* – come è il caso dei social network – è tenuto tuttavia a rimuovere prontamente il contenuto o a disabilitare l'accesso, non appena sia "effettivamente al corrente" del contenuto illecito (art. 14, Direttiva 2000/31/CE).

Nel senso di riconoscere un sempre maggior potere alle piattaforme, si è posta anche la giurisprudenza della Corte di giustizia dell'Unione Europea, in base alla quale ogni Stato membro può ordinare a Facebook di eliminare contenuti e limitare l'accesso ad essi a livello mondiale (CGUE, sentenza 3 ottobre 2019, causa C18/18). A ciò si aggiunge che la piattaforma è chiamata a eliminare, di propria iniziativa, i commenti "equivalenti" a quelli denunciati. È evidente come

[2] Comunicazione della Commissione al Parlamento europeo, al Consiglio, al Comitato economico e sociale europeo e al Comitato delle regioni, Agenda europea sulla sicurezza, COM(2015) 185 final, del 28 aprile 2015.

tale previsione apra la strada per riconoscere ai social network un dovere di sorveglianza dei contenuti immessi sulle proprie piattaforme, nonché una sempre maggiore discrezionalità, soprattutto per quanto riguarda i contenuti "equivalenti", in merito alla scelta sulla rimozione.

Proprio per contrastare la diffusione dell'incitamento alla discriminazione e all'odio in rete, nel maggio 2016 è stato elaborato il *Codice di condotta per la lotta contro le forme illegali di incitamento all'odio on line.*

Si tratta di un Codice che la Commissione europea ha concordato con i principali provider, inizialmente Facebook, Twitter, YouTube e Microsoft. In seguito, hanno aderito al Codice anche Instagram, Google+, Snapchat, Dailymotion e Jeuxvideo.com nonché, da ultimo, TikTok e LinkedIn.

Con la firma del Codice di condotta, le aziende informatiche hanno manifestato il loro impegno a proseguire nella lotta contro l'incitamento all'odio online. A tal fine, i firmatari assumono, in primo luogo, un impegno informativo nei confronti degli utenti al fine di "*precisare che sono vietate la promozione dell'istigazione alla violenza e a comportamenti improntati all'odio"*.

Si prevede, inoltre, la necessità di definire procedure di segnalazione interne agli *internet service provider*, in modo da rispondere entro ventiquattr'ore alle richieste di rimozione di contenuti d'odio e, se necessario, procedere alla loro cancellazione o disabilitare l'accesso al sito. Le parti si impegnano ad adottare delle Linee guida rivolte agli utenti, al fine di diffondere il divieto di ogni forma di istigazione all'odio e alla violenza.

Una particolare attenzione è dedicata, altresì, all'educazione e alla sensibilizzazione degli utenti in merito ai contenuti illeciti o comunque nocivi e sull'importanza di una loro segnalazione. Alcune disposizioni del Codice sono volte poi alla promozione della collaborazione tra le aziende informatiche e della costituzione di partenariati con le organizzazioni della società civile e le autorità nazionali. Infine, le aziende informatiche e la Commissione convengono di riesaminare gli impegni assunti a scadenze regolari, valutandone anche l'impatto in concreto.

È da rilevare, dal punto di vista più strettamente giuridico, il mancato valore normativo del testo, che contribuisce in ogni caso a una auotoregolamentazione e

a una sorveglianza dei contenuti inseriti sulle piattaforme, nell'ottica di una collaborazione tra le istituzioni e i soggetti che sulla rete gestiscono i flussi informativi.

A partire dal 2016 sono seguite, a cadenza annuale, le valutazioni relative all'attuazione del Codice. Nella quinta valutazione, pubblicata il 22 giugno 2020, si leggono risultati positivi, soprattutto se comparati con i primi dati del 2016.

Dalla valutazione risulta infatti che:

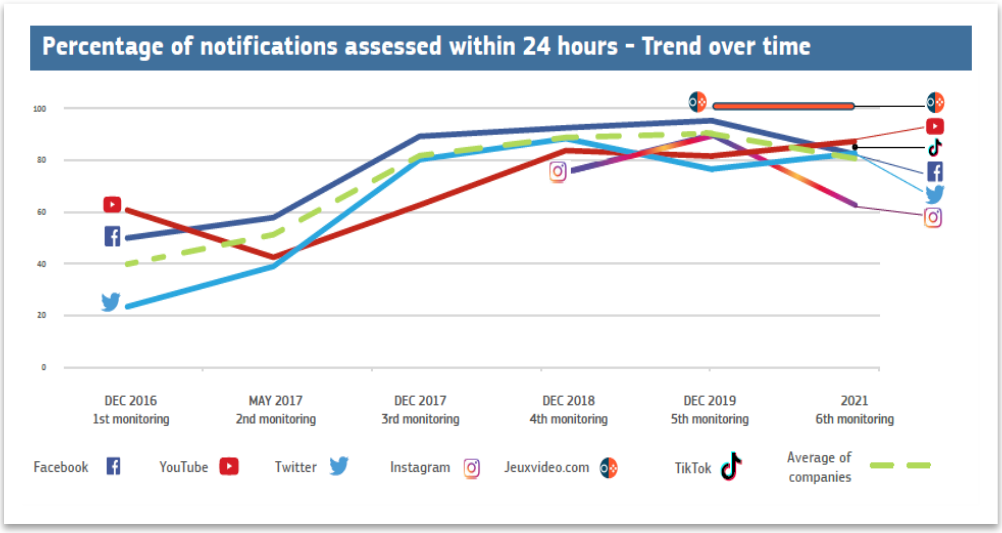
- il 90% dei contenuti segnalati è stato valutato dalle piattaforme entro 24 ore (a fronte del 40% di valutazioni svolte nel 2016);
- nel 2020 è stato rimosso il 71% dei contenuti ritenuti un illecito incitamento all'odio (contro il 28% del 2016);
- le rimozioni da parte delle piattaforme continuano a rispettare la libertà di espressione e ad evitare di rimuovere contenuti non necessariamente classificabili come illecito incitamento all'odio;
- le piattaforme hanno risposto e hanno fornito un *feedback* al 67,1 % delle segnalazioni ricevute. Si segnala, però, come solo Facebook informa sistematicamente gli utenti, mentre tutte le altre piattaforme dovranno apportare miglioramenti[3].

Meno incoraggianti, tuttavia, sono i dati da ultimo pubblicati nella sesta valutazione, relativa al 2021. Nella specie, si registrano risultati in calo rispetto all'anno precedente rispetto a:

- valutazione dei contenuti segnalati entro le 24 ore: sebbene rimanga alto (80%) scende di dieci punti percentuali rispetto al 2020 (fig.1);
- tasso di rimozione dei contenuti: scende al 62,5% (fig. 2);
- tasso di risposta e di fornitura di un *feedback* alle segnalazioni ricevute: scende al 60,3%. Si continua a registrare, sul punto, il primato di Facebook (con l'86,9%), seguito da Twitter (con il 54,1%) e Instagram (con il 41,9%). Risultati ancora poco soddisfacenti arrivano, invece, da YouTube (7,3%) (fig. 3).

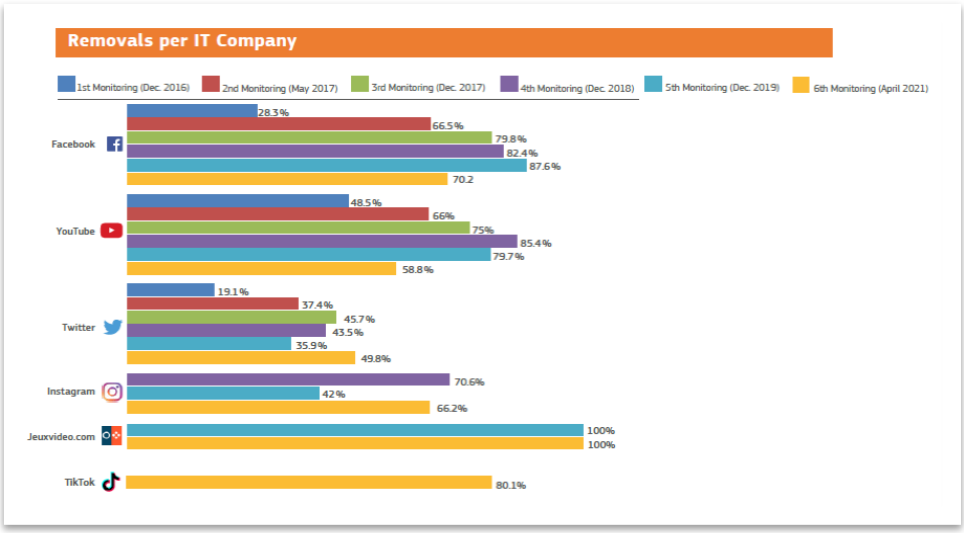
[3] Fonte: https://ec.europa.eu/commission/presscorner/detail/it/IP_20_1134 [ultimo accesso: 12/10/2021].

Fig.1



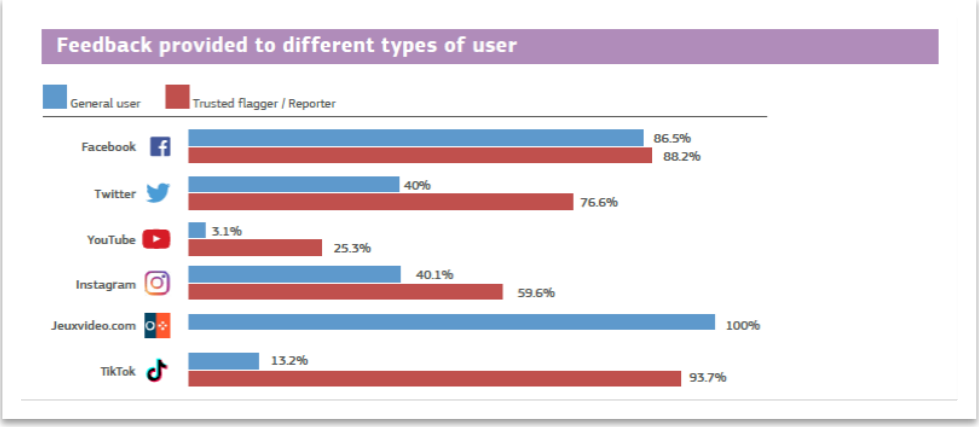
Fonte: European Commission, Countering illegal hate speech online 6th evaluation of the Code of Conduct - Factsheet 7 October 2021

Fig.2



Fonte: European Commission, Countering illegal hate speech online 6th evaluation of the Code of Conduct - Factsheet 7 October 2021

Fig.3



Fonte: European Commission, Countering illegal hate speech online 6th evaluation of the Code of Conduct - Factsheet 7 October 2021

Sempre a livello europeo, ma questa volta all'interno di atti di natura vincolante, è da segnalare la Direttiva 2018/1808/UE che, anche se limitatamente alle piattaforme online di condivisione di video (come è il caso di YouTube), richiede che siano adottate le misure appropriate per tutelare il grande pubblico dai contenuti che istigano alla violenza o all'odio nei confronti di un gruppo o di un membro di un gruppo[4], nonché da comunicazioni commerciali audiovisive discriminatorie[5].

L'attenzione dell'UE nei confronti del contrasto ai discorsi d'odio online emerge chiaramente anche nel "Piano d'azione per la democrazia europea", presentato dalla Commissione europea il 3 dicembre 2020. Con riferimento specifico all'*hate speech*, nel Piano si richiama l'impegno nella lotta contro l'incitamento all'odio online, considerato come uno dei motivi che può dissuadere le persone dall'esprimere le proprie opinioni e dal partecipare alle discussioni online. In tale contesto, la Commissione si impegna a proporre un'iniziativa al fine di estendere l'elenco dei reati previsti nel Trattato sul funzionamento dell'Unione europea (TFUE) all'art. 83, par. 1, ai crimini d'odio e all'incitamento

[4] V. considerando 47 e art. 6 della Direttiva.
[5] Ivi, art. 9, lett. b), ii).

all'odio, anche online[6]. A questo si aggiunge l'impegno a proseguire i lavori nell'ambito del Codice di condotta del 2016[7].

La modifica della norma del TFUE fornirebbe una base giuridica più sicura rispetto all'*hate speech* e permetterebbe alla Commissione di presentare proposte volte alla tutela di nuove categorie che sono vittime di discriminazione.

Le iniziative richiamate nel Piano sono, inoltre, considerate essenziali al fine di rafforzare la sicurezza dei giornalisti.

In un simile contesto, è destinata ad avere un rilevante impatto anche il *Digital Services Act* (DSA)[8], una proposta di Regolamento presentata dalla Commissione il 15 dicembre 2020 al fine di rafforzare e uniformare le responsabilità delle piattaforme online e dei fornitori di servizi informatici, a cui si accompagna una maggiore vigilanza sulle politiche relative ai contenuti delle piattaforme. Pur confermando, in linea generale, il regime di responsabilità limitata già delineato dalla Direttiva 2000/31/CE[9], il DSA prevede nuovi obblighi in capo ai provider, differenziandoli in quattro livelli, a seconda della dimensione e del servizio fornito.

La proposta prevede tre opzioni strategiche:

1. la previsione di una serie di obblighi procedurali per contrastare le attività illecite condotte dagli utenti e volti a garantire la trasparenza delle azioni poste dalle piattaforme e tutelare i diritti degli utenti;

2. la promozione di misure volontarie da parte dei prestatori di servizi per contrastare i contenuti illeciti e l'introduzione di misure di trasparenza in materia di pubblicità e sistemi di raccomandazione;

3. l'introduzione di obblighi più rigorosi per le piattaforme online di dimensioni molto grandi.

---

[6] Il testo attualmente in vigore prevede tra le sfere di criminalità in cui possono intervenire il Parlamento europeo e il Consiglio: terrorismo, tratta degli esseri umani e sfruttamento sessuale delle donne e dei minori, traffico illecito di stupefacenti, traffico illecito di armi, riciclaggio di denaro, corruzione, contraffazione di mezzi di pagamento, criminalità informatica e criminalità organizzata.

[7] Comunicazione della Commissione al Parlamento europeo, al Consiglio, al Comitato economico e sociale europeo e al Comitato delle regioni, sul piano d'azione per la democrazia europea, COM(2020) 790 final, del 3 dicembre 2020, p. 12.

[8] 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, del 15 dicembre 2020.

[9] In particolare, il riferimento è agli artt. 12-15 della Direttiva 2000/31/CE.

Rispetto all'obiettivo principale, ossia il contrasto alla diffusione online di contenuti illeciti, si introduce un sistema sanzionatorio nel caso in cui non avvenga una rimozione tempestiva del contenuto di cui le piattaforme stesse siano venute a conoscenza. A tal fine, vengono presi in considerazione due livelli: la natura del contenuto diffuso in rete e la dimensione delle piattaforme sulle quali esso viene veicolato, con la previsione di obblighi in capo ai diversi provider in base al loro ruolo, alla loro dimensione e al loro impatto sull'ecosistema digitale.

Tra le opzioni in esame, figurano l'introduzione obbligatoria di sistemi di notifica e azione e obblighi di segnalazione che imporranno alle piattaforme di fornire informazioni sulle modalità adottate per contrastare i contenuti illeciti, incluso l'incitamento all'odio.

Si favorisce, inoltre, il ricorso alla co-regolamentazione e agli strumenti volti a incrementare le misure di trasparenza della decisione algoritmica, con il fine precipuo di contrastare l'opacità dello spazio digitale e garantire la costruzione di una solida democrazia digitale.

La vera novità della proposta è da ravvisare nell'intento di procedimentalizzare l'intervento delle piattaforme al fine di garantire la massima trasparenza.

1.3. Il contrasto ai discorsi d'odio nell'ordinamento italiano

Nel nostro ordinamento la manifestazione del pensiero è tutelata in maniera piuttosto ampia dall'art. 21 della Costituzione, da leggere come espressione di fiducia del Costituente sul libero confronto delle idee[10]. Si tratta comunque di una libertà non assoluta, tale per cui al di là dell'unico limite espresso del buon

[10] L'art. 21 Cost. così dispone: "Tutti hanno diritto di manifestare liberamente il proprio pensiero con la parola, lo scritto e ogni altro mezzo di diffusione.
La stampa non può essere soggetta ad autorizzazioni o censure.
Si può procedere a sequestro soltanto per atto motivato dell'autorità giudiziaria nel caso di delitti, per i quali la legge sulla stampa espressamente lo autorizzi, o nel caso di violazione delle norme che la legge stessa prescriva per l'indicazione dei responsabili.
In tali casi, quando vi sia assoluta urgenza e non sia possibile il tempestivo intervento dell'autorità giudiziaria, il sequestro della stampa periodica può essere eseguito da ufficiali di polizia giudiziaria, che devono immediatamente, e non mai oltre ventiquattro ore, fare denunzia all'autorità giudiziaria. Se questa non lo convalida nelle ventiquattro ore successive, il sequestro s'intende revocato e privo d'ogni effetto.
La legge può stabilire, con norme di carattere generale, che siano resi noti i mezzi di finanziamento della stampa periodica.
Sono vietate le pubblicazioni a stampa, gli spettacoli e tutte le altre manifestazioni contrarie al buon costume.
La legge stabilisce provvedimenti adeguati a prevenire e a reprimere le violazioni".

costume, ulteriori limiti impliciti si ricavano dalla necessità di tutelare altri beni, parimenti garantiti dalla Costituzione.

È in particolare dalla tutela della pari dignità sociale dei cittadini (art. 3 Cost.), che si ricava un limite ai giudizi che si risolvono in espressioni di indegnità nei confronti dell'altro diverso da sé, in modo tale per cui simili manifestazioni del pensiero non sono da ritenere protette dall'art. 21 Cost. Da qui si esclude il contrasto con il quadro costituzionale dei divieti posti dal legislatore ordinario alla diffusione di idee fondate sulla superiorità o sull'odio razziale, nonché di incitamento a commettere, per le stesse motivazioni, atti di violenza o di provocazione alla violenza.

In recepimento agli obblighi derivanti dall'ordinamento internazionale, e nella specie la Convenzione di New York sull'eliminazione di ogni forma di discriminazione razziale, il legislatore italiano ha previsto alcune fattispecie di discorso d'odio. In particolare, è con la l. 13 ottobre 1975, n. 654 (c.d. legge Reale) che viene introdotta la fattispecie criminosa di diffusione di idee fondate sulla superiorità o sull'odio razziale, nonché di incitamento a commettere atti di violenza o di provocazione di violenza nei confronti di soggetti appartenenti a un gruppo nazionale, etnico o razziale (art. 3). Tali fattispecie sono poi state estese anche alla tutela degli appartenenti alle minoranze linguistiche (art. 18-*bis* della l. 482/1999, introdotto dalla l. 38/2001).

Ancora più incisivi sono stati due interventi successivi: nel 1993, con la c.d. legge Mancino (l. 205/1993, di conversione del d.l. 122/1993), vengono ridimensionate le conseguenze sanzionatorie e, successivamente, con la l. 85/2006, oltre ad un'ulteriore diminuzione della pena, vengono modificati i termini della condotta penalmente rilevante, per cui il verbo diffonde viene sostituito con il termine "propaganda", e il riferimento all'incitamento cede il posto all'istigazione. Una modifica, quest'ultima, che non sembrerebbe limitarsi al piano lessicale, dato che, dal punto di vista concettuale, le nuove fattispecie si riferirebbero ad un ambito assai più ristretto. Propaganda e istigazione richiamano infatti, rispettivamente, non una semplice espressione di idee ma un'influenza sul comportamento altrui e un incoraggiamento all'azione.

Più di recente, e in attuazione della decisione quadro del 2008, la legge n. 115 del 16 giugno 2016 ha aggiunto all'art. 3 della legge Reale un ulteriore comma che prevede un aggravante di pena nel caso in cui la propaganda, l'istigazione e l'incitamento "si fondano in tutto o in parte sulla negazione della Shoah o dei crimini di genocidio, dei crimini contro l'umanità e dei crimini di guerra" come definito dalla Corte penale internazionale.

Su tale quadro è, infine, intervenuto il d.lgs. 21/2018 che ha abrogato l'art. 3 della l. 654/1975 e ha trasposto le fattispecie in essa previste nel codice penale, con l'introduzione di due nuove disposizioni, gli artt. 604-*bis* e 604-*ter* c.p., che puniscono, rispettivamente, la propaganda e l'istigazione a delinquere per motivi di discriminazione razziale, etnica e religiosa e la rispettiva aggravante[11].

Per quanto riguarda le più recenti iniziative del legislatore italiano, va citato il *ddl Zan* (AS 2005), recante misure di prevenzione e contrasto della discriminazione e della violenza per motivi fondati sul sesso, sul genere, sull'orientamento sessuale, sull'identità di genere e sulla disabilità. La proposta di legge (bocciata al Senato nel 2021) si pone nel senso di estendere alle discriminazioni basate su detti motivi le fattispecie penali previste dagli artt. 604-*bis* e 604-*ter,* prevedendo un'aggravante nel caso in cui le condotte in rilievo siano commesse a mezzo Internet, oltre che in occasione di manifestazioni pubbliche o aperte al pubblico.

Al di là del diritto penale, che dovrebbe comunque rappresentare l'*extrema ratio*, la normativa interna vieta i comportamenti discriminatori in vari ambiti. Si citano in proposito il d.lgs. 215/2003 di recepimento della direttiva 2000/43/CE,

[11] In base all'art. 604-*bis* c.p. "salvo che il fatto costituisca più grave reato, è punito: a) con la reclusione fino ad un anno e sei mesi o con la multa fino a 6.000 euro chi propaganda idee fondate sulla superiorità o sull'odio razziale o etnico, ovvero istiga a commettere o commette atti di discriminazione per motivi razziali, etnici, nazionali o religiosi; b) con la reclusione da sei mesi a quattro anni chi, in qualsiasi modo, istiga a commettere o commette violenza o atti di provocazione alla violenza per motivi razziali, etnici, nazionali o religiosi. È vietata ogni organizzazione, associazione, movimento o gruppo avente tra i propri scopi l'incitamento alla discriminazione o alla violenza per motivi razziali, etnici, nazionali o religiosi. Chi partecipa a tali organizzazioni, associazioni, movimenti o gruppi, o presta assistenza alla loro attività, è punito, per il solo fatto della partecipazione o dell'assistenza, con la reclusione da sei mesi a quattro anni. Coloro che promuovono o dirigono tali organizzazioni, associazioni, movimenti o gruppi sono puniti, per ciò solo, con la reclusione da uno a sei anni.
Si applica la pena della reclusione da due a sei anni se la propaganda ovvero l'istigazione e l'incitamento, commessi in modo che derivi concreto pericolo di diffusione, si fondano in tutto o in parte sulla negazione, sulla minimizzazione in modo grave o sull'apologia della Shoah o dei crimini di genocidio, dei crimini contro l'umanità e dei crimini di guerra, come definiti dagli articoli 6, 7 e 8 dello statuto della Corte penale internazionale".

la l. 67/2006, concernente misure per la tutela giudiziaria delle persone con disabilità vittime di discriminazioni.

L'attenzione del nostro ordinamento rispetto al dilagare di forme di odio e violenza è dimostrato, altresì, dalla istituzione della Commissione straordinaria per il contrasto dei fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza, insediata presso il Senato della Repubblica il 30 ottobre 2019[12].

Con riferimento specifico all'*hate speech* online, il nostro ordinamento, in linea con l'approccio adottato dall'Unione europea, favorisce forme di autoregolamentazione. Il riferimento è, in particolare, al *Regolamento in materia di contrasto all'hate speech* (Delibera 157/19/CONS) adottato dall'Autorità per le Garanzie nelle Comunicazioni (AGCOM) nel maggio del 2019. Se per i contenuti audiovisivi si riconosce una potestà di intervento diretto dell'Autorità, per quanto riguarda il contesto online, si prevede la promozione, da parte della stessa AGCOM, di forme di co-regolamentazione e di Codici di condotta da elaborare con il coinvolgimento delle piattaforme web, in modo da definire misure volte a contrastare la diffusione in rete, e in particolare sui social media, di contenuti in violazione dei principi sanciti a tutela della dignità umana e per la rimozione dei contenuti d'odio. Tali misure dovrebbero favorire sistemi efficaci di individuazione e segnalazione degli illeciti e dei loro responsabili, oltre che promuovere campagne di sensibilizzazione al fine di favorire l'inclusione e la coesione sociale[13].

1.3.1. La decisione dei giudici italiani di fronte all'oscuramento di pagine e profili di movimenti di estrema destra da parte di Facebook e Instagram

Rispetto alla repressione dell'odio online nel nostro ordinamento, va richiamato un caso che ha visto protagonisti i social network Facebook e

[12] Per maggiori dettagli sui lavori fino ad oggi svolti dalla Commissione, si rimanda al seguente link: https://www.senato.it/notes9/Web/18LavoriNewV.nsf/OdGSpecConvCommWebLeg143?ReadForm&10/2021/18.
[13] V. art. 9 del Regolamento AGCOM.

Instagram e due movimenti di estrema destra, Forza Nuova e Casa Pound. I social hanno oscurato profili ufficiali dei due movimenti, profili di leader ed esponenti locali, nonché la pagina ufficiale di Casa Pound Italia, in ragione della presenza negli stessi di contenuti d'odio. Per le piattaforme, tale provvedimento si porrebbe in linea con le regole da esse introdotte negli ultimi anni, in applicazione, nella specie, della norma delle proprie policy dedicata a "Persone e organizzazioni pericolose"[14].

Sul caso è intervenuto il giudice di merito, con tre pronunce successive e contrastanti nelle conclusioni. In particolare, il Tribunale di Roma ha dapprima riconosciuto a Facebook una "pozione speciale", tale da considerarlo uno strumento di cittadinanza, circostanza che porta ad escludere una assimilazione del rapporto tra il social e gli utenti a un normale rapporto tra privati (Trib. Roma, ord. 12 dicembre 2019). Da detta "specialità" dovrebbe discendere, secondo tale prospettiva, l'obbligo per Facebook di rispettare, nei contratti con gli utenti, "principi costituzionali e ordinamentali" che costituiscono, per il social, condizione e limite della sua attività. Date queste premesse, per il giudice di prime cure nell'oscurare la pagina dell'Associazione di Promozione Sociale Casapound e del profilo personale dell'amministratore della stessa, Facebook avrebbe violato il diritto al pluralismo da parte della piattaforma social, a danno di un'Associazione che non potrebbe, così, esprimere i propri messaggi politici. Per il giudice, il social ha l'obbligo di ospitare ogni opinione accettata nell'agone politico, ancor più se si tratta di opinioni espresse da rappresentanti di forze che si sono presentate alle elezioni. Il Tribunale di Roma ha così imposto a Facebook di riattivare gli account sospesi.

La sentenza successiva riconduce invece il rapporto tra il social e gli utenti a un rapporto privatistico (Trib. Siena, sez. unica civile, ord. 19 gennaio 2020). In tale pronuncia, viene pertanto riconosciuto in capo a Facebook "il buon diritto, di

[14] In base a tale previsione, Facebook espressamente sancisce che "per impedire e interrompere atti di violenza reali, non permettiamo la presenza su Facebook di organizzazioni o individui che proclamano missioni violente o che sono coinvolti in azioni violente. Questo include organizzazioni o individui coinvolti nelle seguenti attività: Terrorismo; Odio organizzato; Omicidio di massa o seriale; Traffico di esseri umani; Violenza organizzata o attività criminale.
Rimuoviamo inoltre contenuti che esprimono supporto o elogio di gruppi, leader o individui coinvolti in queste attività". Per un dettaglio delle policy di Facebook si rimanda *infra*, par. 2.2

origine contrattuale, a procedere alla disattivazione della pagina e del profilo" di un utente che aveva violato la policy del social attraverso comprovati e ripetuti episodi di *hate speech*. Il provvedimento chiarisce, a questo riguardo, che Facebook non può essere seriamente paragonato ad un soggetto pubblico nel fornire un servizio di indubbia rilevanza sociale e socialmente diffuso ma che rimane, comunque, prettamente privatistico.

Più di recente, ancora una volta il Tribunale di Roma fornisce una diversa decisione non più fondata, come la precedente, sulla garanzia del pluralismo politico e informativo (Trib. Roma, ord. 24 febbraio 2020). Secondo il giudice, Facebook avrebbe dovuto (e non solo potuto) cancellare gli account riconducibili a Forza Nuova, dato che nel contenere messaggi d'odio contrastavano con l'ordinamento sovranazionale e italiano (e non semplicemente con le policy del social).

Si legge infatti nell'ordinanza: "Facebook non solo poteva risolvere il contratto grazie alle clausole contrattuali accettate al momento della sua conclusione, ma aveva il dovere legale di rimuovere i contenuti, una volta venutone a conoscenza, rischiando altrimenti di incorrere in responsabilità".

Parte seconda

Struttura ed efficacia delle policy attuate dalle principali piattaforme social (twitter, facebook, instagram, youtube)

1. Principi e linee guida delle piattaforme social: cosa viene etichettato come odio online e perché

Ogni social media, operando come piattaforma aziendale e fornitore di servizi, definisce cosa costituisce *hate speech* e determina le proprie regole di condotta per contrastarlo.

Di seguito si illustrano le linee guida generali di cui Twitter, il gruppo Meta e YouTube si sono dotati per definire cosa sia etichettabile come "incitamento all'odio" e cosa no.

Emerge fin da subito come le piattaforme abbiano adottato una definizione pressoché convergente di tali contenuti "d'odio" ma si distinguano per il grado di specificità con cui questi sono affrontati nelle norme di condotta e nelle linee

guida generali, nonché – come vedremo – nelle singole funzioni che nel tempo si sono date per contrastarli, peculiari per ogni piattaforma.

Particolarmente dettagliata e ricca di esempi sul fronte delle linee guida risulta la definizione suggerita da Facebook e YouTube (con particolare attenzione ai video), mentre Twitter enfatizza la necessità di operare una distinzione dei contenuti basata sul contesto.

1.1. Twitter: l'importanza del contesto

Twitter si fonda sul principio di agevolazione della conversazione pubblica, impegnandosi a "dare a tutti la possibilità di creare e condividere idee e informazioni e di esprimere opinioni e convinzioni senza barriere". Negli anni, tuttavia, questa piattaforma di microblogging ha registrato sistematiche violazioni delle sue linee guida che colpiscono sproporzionatamente le categorie protette.

La piattaforma dichiara:

> Sappiamo che la possibilità di esprimere sé stessi su Twitter può essere compromessa quando si subisce un abuso. Le ricerche rilevano come alcuni gruppi di persone siano colpiti dagli abusi online in maniera sproporzionata. Per le persone che si identificano con più gruppi sottorappresentati, l'abuso può essere più frequente, di natura più grave e con un impatto maggiore.

In linea generale Twitter definisce come incitamento all'odio, e dunque vieta, un contenuto che rispecchi i seguenti criteri, definiti nelle norme di condotta della piattaforma:

> Non puoi promuovere la violenza contro altre persone, attaccarle o minacciarle sulla base di razza, etnia, origine nazionale, casta, orientamento sessuale, genere, identità di genere, religione, età, disabilità o malattia grave. Non puoi utilizzare immagini o simboli che incitano all'odio nella tua immagine o intestazione del profilo. Inoltre non puoi utilizzare il tuo nome utente, il nome visualizzato o la bio del profilo per commettere abusi, come molestare qualcuno o esprimere odio nei confronti di una persona, un gruppo o una categoria protetta.

La casistica è sufficientemente dettagliata: si va dalle Minacce di violenza ("Sono proibiti i contenuti che includono minacce di violenza contro un obiettivo identificabile. [...] Gli utenti ritenuti responsabili della condivisione di minacce di violenza vedranno sospeso il proprio account con effetto immediato e

permanente"), all'augurio, speranza o invocazione di un serio danno a una persona o a un gruppo (*"*Sono proibiti i contenuti che augurano, sperano, promuovono, incitano o auspicano morte, lesioni corporali gravi o serie malattie a un'intera categoria protetta e/o a persone che potrebbero farne parte").

In questo caso si riportano esempi concreti:

> È proibito prendere di mira persone con contenuti intesi a incitare la paura o diffondere stereotipi paurosi su una categoria protetta, incluso l'affermare che i membri di una categoria protetta hanno più probabilità di prendere parte ad attività pericolose o illegali, ad esempio 'Tutti i [membri di un gruppo religioso] sono terroristi'. Contenuti umilianti ripetuti e/o non consensuali come insulti, epiteti, metafore razziste e sessiste o simili. [...] È proibito prendere di mira persone con insulti, luoghi comuni o altri contenuti ripetuti volti a disumanizzare, degradare o rafforzare stereotipi negativi o dannosi riguardo a una categoria protetta. Vietiamo altresì la deumanizzazione di un gruppo di persone sulla base della rispettiva religione, casta, età, disabilità, malattia grave, origine nazionale, razza o etnia".
> Una sezione a parte è dedicata alle immagini: sono considerate immagini che incitano all'odio "tutti i loghi, i simboli o le immagini il cui obiettivo è promuovere l'ostilità e la cattiveria contro altre persone sulla base di razza, religione, disabilità, orientamento sessuale, identità di genere o etnia/origine nazionale.

Twitter elenca dunque una serie di esempi, tra questi ci sono "simboli storicamente associati a gruppi violenti (ad esempio la svastica nazista)", ma anche "immagini lesive della dignità umana [...] immagini modificate per inserire caratteristiche animali" o "alterate in modo da includere simboli di odio o riferimenti a genocidi".

Non sono consentiti contenuti con immagini che incitano all'odio nei video in diretta, nelle bio dell'account, nelle foto del profilo o immagini d'intestazione.

Twitter conferisce particolare importanza al contesto e lo cita direttamente nelle sue linee guida: "Alcuni Tweet che possono sembrare incitanti all'odio se visti singolarmente - dichiara - potrebbero non esserlo nel contesto di una conversazione più ampia. Ad esempio i membri di una categoria protetta possono fare riferimento l'uno all'altro utilizzando termini che sono generalmente considerati denigratori. Se questi termini vengono utilizzati in modo consensuale, non manifestano l'intento di offendere. Al contrario, costituiscono un mezzo per rivendicare parole storicamente impiegate per sminuire le persone".

È la stessa piattaforma a individuare le difficoltà nell'esame e nella distinzione tra contenuti che possono delinearsi come incitamento all'odio e altri che non hanno lo scopo di offendere:"Per aiutare i nostri team a capire il contesto, qualche volta abbiamo bisogno di parlare direttamente con la persona presa di mira per assicurarci di avere tutte le informazioni necessarie prima di prendere qualsiasi provvedimento".

Affinché siano presi provvedimenti, comunque, gli utenti non devono necessariamente far parte di una specifica categoria protetta. La piattaforma si impegna a non chiedere ai suoi iscritti di dimostrare o confutare l'appartenenza a una categoria protetta, né a indagare su tali informazioni.

1.2. Facebook: gli "Standards della community" tra feedback degli utenti e contributo degli esperti

Il gruppo Facebook proibisce l'incitamento all'odio, definito: "attacco diretto rivolto alle persone sulla base di quelle che sono note come le loro 'categorie protette' - razza, etnia, nazionalità di origine, disabilità, religione, casta, orientamento sessuale, genere, identità di genere e malattie gravi".

Gli "attacchi" si configurano come: "discorsi violenti o disumanizzanti, stereotipi nocivi, dichiarazioni di inferiorità, espressioni di disprezzo, disgusto o rifiuto, imprecazioni e incitazioni all'esclusione o alla segregazione".

Negli *Standards della Community*[15], Facebook definisce cosa sia consentito o meno sulla piattaforma, basandosi sui consigli di esperti in settori quali tecnologia, sicurezza pubblica e diritti umani, ma soprattutto sui *feedback* ricevuti dalla sua community, una ricchezza che ha deciso di sfruttare più di altre piattaforme.

Per quanto riguarda la violenza e i comportamenti criminali, si pone l'obiettivo di "impedire possibili atti di violenza offline che potrebbero essere correlati a contenuti su Facebook. [...]". Da qui la necessità di considerare linguaggio e contesto per distinguere le dichiarazioni casuali da contenuti che

[15] Meta, Standard della Community, https://transparency.fb.com/it-it/policies/community-standards

possano costituire una minaccia reale alla sicurezza pubblica o personale. “Per determinare se una minaccia è credibile” - scrive Facebook – “potremmo anche prendere in considerazione altre informazioni come la visibilità pubblica di una persona e i rischi per la sua sicurezza fisica. In alcuni casi, rileviamo minacce ipotetiche o auspicate rivolte a terroristi e altri soggetti colpevoli di atti violenti (ad es. “I terroristi meritano di essere uccisi”) e riteniamo che tali minacce non siano credibili in assenza di prove specifiche del contrario”.

Per quanto riguarda vere e proprie organizzazioni pericolose, Facebook ne vieta la presenza sulla piattaforma: “Questo include [...]: terrorismo, odio organizzato, omicidio di massa (compresi i tentativi) o omicidio plurimo, traffico di esseri umani, violenza organizzata o attività criminale”.

Rimuove inoltre contenuti “che esprimono supporto o elogio di gruppi, leader o individui coinvolti in queste attività”. Vieta poi espressamente di “agevolare, organizzare, promuovere o tollerare determinate attività criminali o lesive che prendono di mira persone, aziende, proprietà o animali [...]”.

Pur riconoscendo come essenziale l’impegno verso la libertà di espressione, nei suoi Standards Facebook chiarisce che tale libertà può essere limitata per tutelare uno o più valori, tra cui la dignità.

La piattaforma fa inoltre riferimento a tematiche specifiche e di particolare attualità (soprattutto sulle sue pagine), come l’immigrazione, argomento che spesso accende controverse discussioni degli utenti. Amnesty International ha rilevato, ad esempio, che tra il 15 giugno e il 30 settembre 2020, i cinque post su Facebook ad aver generato maggiore incidenza di *hate speech* sono tutti incentrati sui temi “immigrazione” e “minoranze religiose”. “Proteggiamo rifugiati, migranti, immigrati e richiedenti asilo dagli attacchi più gravi” – dichiara Facebook – “pur consentendo di commentare e criticare le politiche sull’immigrazione”.

Anche Facebook si occupa in modo esplicito delle immagini, con l’obiettivo di rimuovere “i contenuti che promuovono la violenza o celebrano la sofferenza o l’umiliazione di altre persone”. Consente immagini anche forti che favoriscano la sensibilizzazione su determinati temi ma decide di evidenziarne i contenuti avvertendo l’utente:

> Sappiamo che le persone apprezzano la possibilità di discutere di temi importanti come le violazioni dei diritti umani o gli atti di terrorismo, sappiamo anche che le persone hanno sensibilità diverse riguardo ai contenuti forti e violenti. Per questo motivo, aggiungiamo un'etichetta di avviso ai contenuti particolarmente forti o violenti per evitare che siano visualizzati dai minori di diciotto anni.

Facebook ha infine una sezione dedicata ai Contenuti che esprimono crudeltà e insensibilità, "che si rivolgono a vittime di gravi crudeltà fisiche o emotive", ha un approccio meticoloso alla definizione dei contenuti da non pubblicare e fa un'ulteriore distinzione in tre livelli di gravità, molto dettagliati, di cui riportiamo alcuni stralci significativi:

- Livello 1

Contenuti rivolti a una persona o a un gruppo di persone che contengono:

> Discorsi di incitamento o sostegno alla violenza; Discorsi o immagini disumanizzanti sotto forma di confronti, generalizzazioni o dichiarazioni comportamentali non classificate in relazione a: insetti, animali culturalmente percepiti come intellettualmente e fisicamente inferiori, sporcizia, batteri, malattie e feci, predatori sessuali, individui subumani, criminali violenti e sessuali, altri criminali (compresi, a titolo esemplificativo e non esaustivo, "ladri", "rapinatori di banca" o affermazioni in cui si etichettano tutte le categorie protette o semiprotette come "criminali"), affermazioni che ne negano l'esistenza, derisione del concetto, degli eventi o delle vittime dei crimini di odio, anche se nell'immagine non è presente alcuna persona reale.
>
> Confronti disumanizzanti designati, generalizzazioni o dichiarazioni comportamentali [...] che comprendono: persone di colore e scimmie o creature simili, persone di colore e attrezzature agricole, caricature di persone di colore sotto forma di blackface, ebrei e ratti, ebrei che governano il mondo o controllano le principali istituzioni, ad esempio le reti di social media, l'economia o il governo, informazioni negazioniste o distorte sull'Olocausto, musulmani e maiali, musulmani e relazioni sessuali con capre o maiali, messicani e creature simili ai vermi, donne come oggetti domestici o il riferimento alle donne come proprietà o "oggetti", persone transgender o non-binarie a cui si fa riferimento come se fossero oggetti, dalit, caste riconosciute o persone delle caste inferiore, come i lavoratori più umili.

- Livello 2

Contenuti rivolti a persone o a gruppi di persone sulla base delle loro caratteristiche protette:

> Generalizzazione che affermano l'inferiorità [...]: per carenze fisiche relative a igiene e aspetto fisico; per carenze mentali relative a capacità intellettive, istruzione e salute

mentale; per carenze morali relative a tratti caratteriali percepiti negativamente a livello culturale, termini dispregiativi correlati all'attività sessuale;
[...] espressioni relative all'essere inadeguato, alla superiorità/inferiorità di un'altra categoria protetta, all'essere diversi dalla norma;
Espressioni di disprezzo: ammissione di intolleranza basata su categorie protette, espressioni secondo cui non dovrebbe esistere una categoria protetta, espressioni di odio o di rifiuto;
Espressioni di disgusto, che suggeriscono che il bersaglio causi nausea, repulsione o disgusto;
Imprecazioni: riferimento a qualcuno citando genitali o ano, termini o frasi volgari con l'intenzione di insultare, termini o frasi che augurano l'interazione in attività sessuali o il contatto con genitali, ano, feci o urina.

- Livello 3

Contenuti rivolti a persone o a gruppi di persone sulla base delle loro categorie protette che contengono uno dei seguenti elementi:

segregazione [...], esclusione esplicita, intesa come l'espulsione di determinati gruppi o il dichiararli non ammessi, esclusione politica, intesa come la negazione del diritto alla partecipazione politica, esclusione economica, intesa come la negazione dell'accesso ai diritti economici e la limitazione della partecipazione al mercato del lavoro, esclusione sociale, intesa come la negazione dell'accesso agli spazi (fisici e online) e ai servizi sociali, contenuti che descrivono o attaccano persone con insulti, ovvero parole intrinsecamente offensive e comunemente usate per insultare le persone per le caratteristiche elencate sopra.

1.2.1. Instagram: le maggiori difficoltà nei messaggi diretti

Facebook e Instagram appartengono al medesimo gruppo, condividono gli stessi principi e policy in termini di definizione e di contrasto all'incitamento all'odio.

Ad aprile 2021 la piattaforma pubblica un articolo nel suo blog[16] in cui ribadisce la volontà di essere "un luogo in cui le persone possono connettersi con ciò che amano" pur nella consapevolezza che "proprio come nel mondo offline, ci sarà sempre chi abusa degli altri [...]".

Instagram dichiara di aver rafforzato le proprie regole contro l'incitamento all'odio, ribadendo di non tollerare "attacchi alle persone in base alle loro caratteristiche, inclusa razza o religione". Sottolinea di aver "rafforzato queste

[16] Instagram (2021), *An update on our work to tackle abuse on Instagram*

regole l'anno scorso, vietando forme più implicite di incitamento all'odio, come i contenuti che raffigurano Blackface e gli stereotipi antisemiti comuni".

"Agiamo ogni volta che veniamo a conoscenza di incitamento all'odio - scrive Instagram - e miglioriamo continuamente i nostri strumenti di rilevamento per individuarlo più velocemente"[17].

È la piattaforma stessa, quindi, a segnalare l'abuso più frequente registrato tra gli utenti: quello che avviene nei messaggi diretti (DM), definiti *"*più difficili da affrontare rispetto ai commenti su Instagram" poiché "conversazioni private", per le quali, spiega Instagram, "non utilizziamo la tecnologia per rilevare in modo proattivo contenuti come incitamento all'odio o bullismo come facciamo in altri luoghi".

1.3. YouTube: esempi pratici e attenzione ai video

Anche YouTube, così come Facebook, definisce con estrema precisione quali siano i contenuti etichettati come incitamento all'odio[18], ovvero quelli

> che incitano alla violenza o all'odio nei confronti di individui o gruppi sulla base di una qualsiasi delle seguenti caratteristiche: età, casta, disabilità, etnia, identità ed espressione di genere, nazionalità, razza, condizione di immigrato, religione, sesso/genere, orientamento sessuale, condizione di vittima di un grave evento violento e di famigliare di una vittima, condizione di veterano […] Contenuti che disumanizzano individui o gruppi definendoli subumani, paragonandoli ad animali, insetti, parassiti, malattie o altre entità non umane, che promuovono o celebrano la violenza [...].

Tra questi anche "l'uso di insulti e stereotipi razziali, religiosi o di altro tipo che promuovono o incitano all'odio".

Tali contenuti possono essere nella forma di

> discorsi, testi o immagini che promuovono gli stereotipi in questione o li trattano come verità oggettive; contenuti che affermano che degli individui o gruppi sono fisicamente o mentalmente inferiori, [...], contenuti che sostengono la superiorità di un gruppo rispetto a quelli che presentano le caratteristiche di cui sopra allo scopo di giustificare violenza, discriminazione, segregazione o esclusione nei loro confronti;

[17] *Ibidem*
[18] YouTube, Norme sull'incitamento all'odio

> teorie complottiste secondo le quali alcuni individui o gruppi sono malvagi, corrotti o dannosi; contenuti che promuovono la sottomissione o l'oppressione di individui o gruppi o che negano che eventi violenti e ben documentati abbiano avuto luogo; attacchi rivolti all'attrazione emotiva, romantica e/o sessuale di una persona nei confronti di un'altra; contenuti che includono propaganda suprematista di incitamento all'odio, compreso il reclutamento di nuovi membri o la richiesta di sostegno economico per la propria ideologia.

Sono inclusi anche video musicali "che promuovono il suprematismo e incitano all'odio nel testo, nei metadati o nelle immagini".

Per chiarire concretamente quali contenuti siano inammissibili, YouTube elenca una serie di esempi pratici come: "Sono felice che questo [evento violento] sia accaduto", oppure "Esci e dai un pugno a [Gruppo di persone che possiedono le caratteristiche di cui sopra]", "[Gruppo di persone che possiedono le caratteristiche di cui sopra] sono una minaccia per la nostra esistenza, quindi dobbiamo cacciarli via appena possibile" e molti altri.

2. Misure operative delle piattaforme: policy e sanzioni, nuove funzioni e risultati

I principali erogatori di servizi online hanno sviluppato i propri strumenti di *governance* per limitare i contenuti dannosi sulle loro piattaforme.

Oltre alla pubblicazione degli "standard di comunità" o delle "politiche di contenuto" che abbiamo esaminato, alcune piattaforme web hanno introdotto organismi di supervisione per governare la moderazione dei discorsi d'odio online. Questi organismi possono essere composti da esponenti di organizzazioni della società civile, esperti legali, accademici, membri di ONG o di organizzazioni per i diritti delle minoranze. Come ha dichiarato nel 2019 Brent Harris – Director of Global Affairs di Facebook – "non ci sentiamo di avere questa responsabilità da soli"[19].

Esempi di questi organismi sono: il Trust and Safety Council di Twitter (annunciato nel 2016) e l'Oversight Board di Facebook (istituito nel 2019). YouTube, al contrario, non si è dotato di un comitato di supervisione.[20] Mentre

[19] Brent Harris, Director of Global Affairs di Facebook all'evento "Who should regulate free speech online?", presso Chatham House, Londra, 27 Giugno 2019.
[20] Council of Europe, *Models of Governance of Online Hate Speech*, 2020.

l'organismo di supervisione di Twitter ha principalmente un ruolo di consulenza, gli esperti indipendenti di Facebook hanno il potere di emettere sentenze vincolanti nei casi in cui gli utenti facciano appello alla rimozione dei loro contenuti dai social. A gennaio 2021 l'Oversight Board di Facebook ha annunciato le sue prime cinque decisioni, selezionate tra 20,000 episodi di ricorso, ribaltando il giudizio di Facebook in quattro casi. La piattaforma ha avuto sette giorni per ripristinare i contenuti ritenuti conformi alle regole. In questa occasione, il Board ha dichiarato che "troppe decisioni di moderazione dei contenuti oggi sono incoerenti e opache"[21]. Tra i casi più critici esaminati dall'organismo di sorveglianza, la scelta di Facebook di bannare Donald Trump in seguito ai disordini del 6 gennaio 2021; il Board ha confermato il *ban*, stabilendo tuttavia che non potesse essere definitivo e fissandone la durata a due anni.

Di seguito analizzeremo le policy e le nuove funzioni di cui le quattro piattaforme (Twitter, Facebook, Instagram, YouTube) si sono dotate per "tradurre" in azioni pratiche le proprie linee guida generali e mettere in atto le conseguenti sanzioni.

2.1. Policy di Twitter: azioni su singolo Tweet, messaggio diretto, account. Il caso Trump

Le policy messe in atto da Twitter si aggiornano di continuo[22]. Regole e opzioni di applicazione vengono modificate nel tentativo di rispondere alle esigenze di controllo e di sicurezza di un mondo in costante cambiamento: il Web. Per soddisfare tali necessità, Twitter si avvale di ricerche sulle tendenze del comportamento online, utilizza i *feedback* degli utenti e i contributi di enti e personalità esterne, quali appunto i membri del *Trust & Safety Council*.

Queste alcune delle principali novità introdotte negli ultimi anni relative ai discorsi d'odio:

- 2013: è stata aggiunta tra le Regole di Twitter – nella categoria "Abuso e Spam" – una sezione che vieta esplicitamente le "molestie mirate";
- 2015: è stata lanciata la sezione "Comportamento abusivo";

[21] Facebook Oversight Board (2021) https://www.oversightboard.com/decision/
[22] Twitter (2020) Updating our rules against hateful conduct

- 2016: sono stati inseriti maggiori dettagli sulla condotta odiosa ed è stato aggiornato l'elenco dei comportamenti abusivi, che vietano: le avances sessuali indesiderate, la pubblicazione o la condivisione di foto o video intimi di qualcuno che sono stati prodotti o distribuiti senza il loro consenso, desideri o speranze di danno, le minacce di esporre o *hackerare* qualcuno.
- 2019: sono state ampliate le fattispecie di condotta odiosa e le politiche dei media per includere i nomi utente abusivi e le immagini odiose;
- 2020: sono state aggiornate le regole contro l'*hate speech* per includere il linguaggio che disumanizza sulla base dell'età, della disabilità o della malattia; a dicembre, la piattaforma annuncia di aver nuovamente inasprito le proprie regole di condotta contro i contenuti che "disumanizzano le persone sulla base di razza, etnia o origine nazionale".
- 2021: il 1° settembre Twitter ha annunciato un test su "Safety Mode", una funzione che blocca temporaneamente gli account per sette giorni proprio per l'uso di un linguaggio potenzialmente dannoso, come insulti o commenti di odio[23].

La piattaforma di microblogging intraprende azioni su uno specifico contenuto (ad esempio, un singolo Tweet o un Messaggio Diretto), su un account, o su una combinazione di queste opzioni.

Le misure a livello di Tweet sono:

a) Etichettare un Tweet che può contenere informazioni contestate o fuorvianti ("possiamo aggiungere un'etichetta al contenuto per fornire un contesto e informazioni aggiuntive").
b) Limitare la visibilità dei Tweet "'Rende i contenuti meno visibili su Twitter, nei risultati di ricerca, nelle risposte e nelle timeline").
c) Chiedere la rimozione di un Tweet ("Richiediamo al trasgressore di rimuoverlo prima che possa twittare di nuovo. Inviamo un'email di notifica al trasgressore identificando il Tweet in violazione e quali policy

[23] Twitter, Introducing Safety Mode, 2021: https://blog.twitter.com/en_us/topics/product/2021/introducing-safety-mode [ultimo accesso: 01/09/2021]

sono state violate. Dovrà quindi seguire il processo di rimozione o fare appello alla nostra revisione se ritiene che abbiamo commesso un errore").

d) Nascondere un Tweet in violazione in attesa della sua rimozione ("Nel periodo intermedio tra quando Twitter intraprende un'azione esecutiva e la persona rimuove il Tweet, nascondiamo quel Tweet dalla vista pubblica e sostituiamo il contenuto originale con un avviso che dichiara che il Tweet non è più disponibile perché ha violato le nostre regole. Questo avviso sarà disponibile per 14 giorni dopo la rimozione del Tweet*").*
e) Avviso di eccezione di interesse pubblico ("In rari casi, possiamo determinare come sia nell'interesse pubblico che un Tweet che altrimenti violerebbe le nostre regole rimanga accessibile [...] Un avviso spiega l'eccezione e ti dà la possibilità - se lo desideri - di visualizzare il Tweet. Disattiveremo le risposte, i retweet e i like").

Le misure a livello di Messaggi Diretti (DM) consistono invece nell' interruzione delle conversazioni tra un trasgressore segnalato e l'account del segnalante, oppure nell'attivazione di un avviso (in una conversazione di messaggi diretti di gruppo, il messaggio diretto violato può essere segnalato con un avviso "per assicurarsi che nessun altro nel gruppo possa vederlo di nuovo").

Ci sono infine le misure a livello di Account:

a) Richiesta di modifiche: "Se il profilo o il contenuto multimediale di un account non è conforme alle nostre politiche, possiamo renderlo temporaneamente non disponibile e richiedere che il trasgressore modifichi i media o le informazioni nel suo profilo per renderlo conforme".
b) Modalità di sola lettura: "La persona può leggere la sua timeline e sarà solo in grado di inviare messaggi diretti ai suoi seguaci. La durata di questa azione esecutiva può variare da 12 ore a 7 giorni, a seconda della natura della violazione".
c) Verifica della proprietà dell'account: "Per garantire che i trasgressori non abusino dell'anonimato che offriamo e molestino gli altri sulla piattaforma, potremmo richiedere al proprietario dell'account di verificare la proprietà con un numero di telefono o un indirizzo email. Questo ci

aiuta anche a identificare i trasgressori che gestiscono più account per scopi abusivi e prendere provvedimenti su tali account".

d) Sospensione permanente: "La nostra azione più severa. Lo rimuove dalla visualizzazione globale e il trasgressore non potrà creare nuovi account. Informiamo le persone che sono state sospese per violazioni di abuso e spieghiamo quale politica o quali politiche hanno violato e quale contenuto era in violazione".

Twitter incoraggia i suoi utenti a segnalare i contenuti che violino le norme sulla condotta che incita all'odio, sia all'interno di tweet che nei messaggi diretti.

La sanzione più pesante introdotta da Twitter è dunque la rimozione dell'account dalla piattaforma. Sanzione, quest'ultima, comminata persino a un Presidente degli Stati Uniti: Donald Trump. Il 6 Gennaio 2021, infatti, una folla di rivoltosi del gruppo complottista di estrema destra "QAnon" ha fatto irruzione nel Campidoglio degli Stati Uniti mentre il Congresso stava per certificare i risultati delle elezioni presidenziali. Quello stesso giorno Twitter ha sospeso momentaneamente l'account di Trump, dopo che il presidente aveva "giustificato" tale assalto in un suo tweet. Trascorse 12 ore, Twitter ha restituito l'account a Trump, avvertendolo che ulteriori violazioni delle sue policy ne avrebbero comportato una sospensione permanente, anche in vista della diffusione, dentro e fuori Twitter, di "piani per future proteste armate, inclusa una proposta per un attacco al Campidoglio il 17 gennaio". L'account di Trump è stato poi sospeso definitivamente a causa di alcuni tweet che, secondo la piattaforma, potevano essere interpretati (in rete) come una delegittimazione delle elezioni presidenziali, apparendo un "sostegno" a coloro che avevano commesso atti violenti a Capitol Hill".

Foto: Twitter

Quella adottata da Twitter, va detto, è stata una misura straordinaria: per anni il social network aveva rifiutato di censurare i contenuti di Trump, ritenuti di interesse pubblico. A seguito dell'attacco di gennaio 2021, invece, il social ha aggiornato le proprie policy introducendo una sezione relativa all'incitamento all'odio e all'organizzazione di attacchi e proteste violente e alla condivisione di informazioni deliberatamente fuorvianti sull'esito delle elezioni. A seguito di questa decisione sono stati sospesi definitivamente oltre 70 mila account dedicati alla propaganda del gruppo complottista di estrema destra QAnon.

L'attacco al Campidoglio è solo uno degli episodi che lega l'*hate speech* sui social all'aumento di atti di violenza reale. Ad esempio:

a) in Germania è stata trovata una correlazione tra i post su Facebook anti-rifugiati del partito di estrema destra "Alternative für Deutschland" e gli attacchi ai rifugiati. Gli studiosi Karsten Muller e Carlo Schwarz hanno osservato un aumento di tali attacchi, come incendi dolosi e aggressioni, a seguito di picchi nei post di odio.
b) In Myanmar, leader militari e nazionalisti buddisti avrebbero usato i social media per demonizzare la minoranza musulmana Rohingya. Sebbene i Rohingya costituissero solo il 2% della popolazione, gli etnonazionalisti affermavano che essi avrebbero presto soppiantato la maggioranza buddista. Su questo fronte la commissione d'inchiesta delle Nazioni Unite ha affermato: "Facebook è stato uno strumento utile per coloro che cercano di

diffondere l'odio, in un contesto in cui, per la maggior parte degli utenti, Facebook è Internet"[24].

2.1.1. Contrasto all'odio online: i risultati di Twitter

Dal 2012 Twitter pubblica il suo *Transparency Report* biennale. Gli ultimi dati [25], relativi al periodo luglio - dicembre 2020, hanno rilevato complessivamente i seguenti contenuti (Tab. 1)

Tab.1

Categoria di contenuto	**Account che hanno subito un intervento**	**Account sospesi**	**Contenuti rimossi**	**Differenza rispetto al report precedente**
Abuso/molestie	964,459	86,202	1,448,418	+142% account che hanno subito un intervento
Condotta odiosa	1,126,990	157,815	1,628,281	+77% account che hanno subito un intervento
Terrorismo/ estremismo violento	58,750	58,750	0	-35% account che hanno subito un intervento
Violenza	49,146	34,829	59,933	+106% account che hanno subito un intervento

I contenuti, nei quali si è rilevata una condotta odiosa, che hanno subito un intervento da parte della piattaforma sono aumentati addirittura del 77%, segno di un'attenzione crescente e di un efficace sistema di monitoraggio.

[24] https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf [ultimo accesso: 1 settembre 2021]
"Report of the independent international fact-finding mission on Myanmar", Human Rights Council; 10–28 September 2018
[25] Fonte: https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec.

Un'altra ricerca di DataMediaHub e KPI – che ha monitorato le conversazioni su Twitter Italia dal 25 aprile al 17 giugno 2020 – conferma sostanzialmente tale dato: il report ha identificato 679mila tweet e 263mila condivisioni di contenuti di incitamento all'odio da parte di 148mila utenti unici. Si tratta di un numero complessivamente marginale, pari al 3,7% dei tweet postati sulla piattaforma nel periodo, da parte dell'1,4% degli utenti unici presenti su Twitter. La maggior incidenza proviene da insulti "generici" (due terzi del totale) e legati all'ideologia politica (un quarto del totale).

2.2. Policy Gruppo Facebook[26]: dalle segnalazioni degli utenti all'azione (e ai limiti) dei moderatori

L'intento e il contesto, principalmente regionale e linguistico, sono i criteri più difficili da definire per i moderatori di Facebook. La piattaforma prende ad esempio la parola "frocio": "questa potrebbe essere considerata un discorso d'odio se diretta a una persona, ma, in paesi come l'Italia, la parola 'frocio' è utilizzata dagli attivisti LGBT per denunciare l'omofobia". Facebook fa dunque ammenda dei propri errori di valutazione, osservando: "Se non riusciamo a rimuovere il contenuto che segnalate perché pensate sia un discorso d'odio, sembra che non stiamo rispettando i valori dei nostri standard comunitari. Quando rimuoviamo qualcosa che avete postato e che credete sia una visione politica ragionevole, può sembrare una censura".

Il Gruppo Facebook chiarisce dunque fin da subito che nell'applicazione delle sue policy e nelle relative sanzioni, si porrà l'obiettivo di considerare che "le parole hanno diversi significati e conseguenze per le persone in base a comunità locale, lingua o contesto culturale".

La piattaforma dichiara pertanto la volontà di impegnarsi "per tenere conto di queste sfumature e applicare allo stesso tempo in modo coerente ed equo le nostre normative alle persone e al loro diritto di espressione". Un'operazione, questa, che in alcuni casi potrebbe non riuscire. Facebook ne è consapevole: "In alcuni casi - scrive Facebook - ciò implica che potremmo non rilevare contenuti e

[26] Meta, https://transparency.fb.com/it-it/policies/

comportamenti contrari alle nostre violazioni, mentre in altri l'applicazione degli standard potrebbe limitarsi ai casi in cui ci vengano forniti contesto e informazioni ulteriori".

Una delle peculiarità di Facebook è quella di prevedere la possibilità per gli utenti di segnalare i contenuti che ritengono dannosi, sia per sé stessi che per persone o categorie terze. Questa azione diventa quindi complementare all'attività di monitoraggio dell'intelligenza algoritmica. "Le persone possono segnalare contenuti potenzialmente in violazione, tra cui pagine, gruppi, profili, singoli contenuti e commenti. Offriamo anche alle persone la possibilità di controllare la propria esperienza, consentendogli di bloccare, non seguire più o nascondere persone e post".

In alcuni casi i contenuti che incitano all'odio sono ammessi sulla piattaforma se propedeutici a campagne di sensibilizzazione.

Le diverse casistiche sono governate dalle seguenti regole:

a) Ripubblicazione di *hate speech* al fine di condannarlo

"Riconosciamo che le persone in alcuni casi condividono contenuti che incitano all'odio di cui non sono autori allo scopo di condanna o sensibilizzazione. [...] Le nostre normative sono pensate per lasciare spazio a questi tipi di discorsi, ma chiediamo alle persone di chiarire le proprie intenzioni. Quando l'intenzione non è chiara, possiamo rimuovere il contenuto".

b) Azione di interesse pubblico

> In alcuni casi, consentiamo contenuti per sensibilizzare l'opinione pubblica che in altri casi sarebbero contrari ai nostri Standard della community, se si tratta di contenuti rilevanti e di pubblico interesse. Procediamo in questo modo solo dopo aver soppesato l'interesse pubblico rispetto ai potenziali danni e ci basiamo sugli standard internazionali in materia di diritti umani, come indicato nella nostra Normativa aziendale sui diritti umani, per effettuare queste valutazioni. Ad esempio, abbiamo consentito contenuti che illustravano in modo esplicito la guerra o le sue conseguenze laddove importanti per il dibattito pubblico.

In caso di reiterazione delle violazioni sulla piattaforma, il Gruppo Facebook si riserva di limitare le pubblicazioni da parte dell'utente, disabilitarne il profilo e, in casi gravi, allertare le forze dell'ordine.

> Le conseguenze per la violazione degli Standard della community dipendono dalla gravità della violazione e dai precedenti della persona sulla piattaforma. Ad esempio,

> nel caso della prima violazione, potremmo solo avvertire la persona, ma se continua a violare le nostre normative, potremmo limitare la sua capacità di pubblicare su Facebook o disabilitare il suo profilo. Potremmo anche informare le forze dell'ordine quando, a nostro avviso, sussiste la possibilità reale di seri rischi di danno fisico o minacce dirette alla sicurezza pubblica.

Per valutare i risultati di queste policy è necessario considerare che i moderatori di Facebook non sono specificamente formati per affrontare l'incitamento all'odio, un tema più complesso rispetto ad altri tipi di contenuti vietati, come nudità o violenza (l'azienda sta ora sperimentando la specializzazione in incitamento all'odio). I moderatori incaricati di revisionare l'incitamento all'odio, ad esempio, non sono autorizzati a vedere il contesto di un post (come commenti, foto di accompagnamento o un'immagine del profilo) che li aiuterebbe meglio a capire l'intenzione di un commento. L'azienda esclude il contesto per proteggere la privacy degli utenti, ostacolando la capacità dei moderatori di far rispettare le proprie politiche.

Nel 2019 Facebook ha iniziato a consentire agli algoritmi di rimuovere automaticamente i contenuti di incitamento all'odio senza l'ulteriore controllo di un revisore umano. Tale software era in grado di rilevare in modo proattivo solo il 65% dei commenti che l'azienda considerava incitamento all'odio. Percentuale che oggi, secondo Facebook, sarebbe salita al 95%.

L'impegno della piattaforma guidata da Mark Zuckerberg è quello di riuscire a utilizzare un'intelligenza artificiale (IA) in grado di isolare i contenuti dannosi. La vera sfida per questi sistemi automatizzati è "individuare un discorso d'odio in slang, o scritto male intenzionalmente, farlo in una frazione di secondo e su miliardi di persone". Facebook ha quindi presentato un nuovo pacchetto di IA, tra cui *Linformer*, una nuova architettura che permette di "addestrare" modelli di intelligenze artificiali su testi più lunghi e complessi.

Ci sono ancora però altre lacune: mentre Facebook oggi risulta efficace nel rimuovere gli insulti razziali e le offese esplicite, i commenti più insidiosi potrebbero essere autorizzati a rimanere. Monitorando l'*hate speech* sulla piattaforma, la rivista accademica "First Monday" ha rilevato come i contenuti che richiedono un'interpretazione contestuale non vengano rimossi: ad esempio un commento che chiama Oprah [Oprah Winfrey n.d.r.] una "stronzetta del

cotonificio" ("cotton patch bitch") non viene eliminato se i moderatori non lo legano agli afroamericani discendenti degli schiavi nelle piantagioni di cotone[27]. Allo stesso modo, commenti come "vite nere spiaccicate" (*Black Lives Splatter*) non sono stati considerati da Facebook come discorsi d'odio, probabilmente perché non contenevano un insulto nella sua accezione comune. Da una recente analisi di Amnesty International, inoltre, risulta che un commento su dieci pubblicato sul colosso di Zuckerberg sia "offensivo, discriminatorio o hate speech", con una forte incidenza di contenuti islamofobi, sessisti e anti negazionisti.

I problemi sorgono anche quando l'intelligenza artificiale delle piattaforme risulta inefficace con particolari lingue locali e le aziende hanno investito poco nell'assunzione di personale che le conosca bene. Nel 2015, in Myanmar, secondo Reuters, Facebook aveva impiegato solo due parlanti birmani. Dopo una serie di violenze anti-musulmane iniziata nel 2012 gli esperti hanno lanciato l'allerta sul terreno fertile che i monaci buddisti ultranazionalisti avevano trovato su Facebook per diffondere discorsi d'odio al grande pubblico. La piattaforma ha ammesso di aver fatto troppo poco, e nell'agosto 2018 ha bandito i funzionari militari dalla piattaforma e si è impegnata ad aumentare il numero di moderatori che parlano correntemente la lingua locale.

Dall'introduzione dell'Oversight Board, casi delicati come questo sono spesso assegnati agli esperti indipendenti. Nel caso 2020-003-FB-UA, ad esempio, i membri del'Oversight Board hanno valutato se, nell'ambito di un conflitto armato, Facebook avesse ragione di rimuovere un post che presumibilmente conteneva un insulto odioso. Il board ha commissionato un'analisi linguistica indipendente e concluso che tale linguaggio aveva un obiettivo disumanizzante e dunque violava gli Standard di Comunità[28].

"Ora stiamo lavorando ad una nuova IA in grado di analizzare contemporaneamente testi, immagini e video", ha svelato Mike Schroepfer, Chief Technical Officer della piattaforma.

[27] C. RING CARLSON – H. ROUSSELLE, *Report and Repeat: Investigating Facebook's hate speech removal process*, in *First Monday*, 27 gennaio 2020.
[28] Facebook Oversight Board, 2021: https://oversightboard.com/news/436612660860568-oversight-board-upholds-facebook-decision-case-2020-003-fb-ua/.

> Una foto innocua con un testo sotto altrettanto innocuo se preso singolarmente, possono avere un significato completamente diverso se messi assieme. [...] Sono sottigliezze che si consideravano fuori portata per un'intelligenza artificiale. Ecco, stiamo superando questo ostacolo, stiamo passando dall'analisi del dettaglio a quello del contesto[29].

Con gli utenti Facebook che pubblicano contenuti in più di 160 lingue, la piattaforma punta ad avere un'unica IA in grado di isolare i contenuti d'odio in tutte le lingue e i dialetti. L'algoritmo Xlm-r, utilizzato dal 2019 per combattere l'*hate speech*, analizza i testi e riesce a trasferire l'esperienza fatta in una lingua, in altri idiomi. Nonostante i progressi nello sviluppo di sistemi multilingua, però, resta una maggiore efficienza in lingua inglese. Secondo l'organizzazione non governativa Avaaz, molto attiva su questi temi, Facebook etichetta i contenuti falsi che non sono in lingua inglese con circa una settimana di ritardo rispetto a quelli in inglese: 24 giorni per i testi in inglese contro i 30 giorni necessari per quelli in altre lingue. Questo approccio "America First" danneggia i consumatori europei, principalmente italiani, francesi e portoghesi.

Nel 2019, il sito di notizie statunitense "The Verge" ha pubblicato il leakage di una conversazione intercorsa tra Mark Zuckerberg e i dipendenti di Facebook nel luglio di quello stesso anno. Le parole del CEO sembravano suggerire che i contenuti ospitati sulla piattaforma fossero troppo numerosi perché l'azienda potesse moderare i singoli commenti, nonostante ampi investimenti nei moderatori. Per questo motivo, la piattaforma aveva optato per etichettare commenti come "i bianchi sono stupidi", o "gli uomini sono maiali" al pari di insulti antisemiti o razzisti.

Secondo una recente inchiesta del quotidiano statunitense *The Washington Post*, Facebook avrebbe tuttavia rivisto l'approccio generalista, intraprendendo una significativa revisione dei suoi algoritmi che rilevano l'incitamento all'odio. Lo scopo della piattaforma sarebbe invertire anni di pratiche "race-blind" che

[29] Avaaz (2021) Left Behind: How Facebook is neglecting Europe's infodemic. https://secure.avaaz.org/campaign/en/facebook_neglect_europe_infodemic/ ultimo accesso: 05/08/2021

avrebbero favorito la rimozione di offese rivolte a utenti bianchi e al contrario segnalato alcuni post innocui di persone di colore.

La revisione, nota come *Progetto WoW*, prevederebbe la "riprogettazione dei sistemi di moderazione automatizzata di Facebook per migliorare il rilevamento e l'eliminazione automatica del linguaggio d'odio, includendo con esso insulti diretti a neri, musulmani, persone di più di una razza, comunità LGBTQ ed ebrei"[30]. Nello specifico, secondo i documenti visionati dal *The Washington Post*, Facebook assegnerebbe dei punteggi numerici, ponderati in base al danno percepito, a diversi tipi di attacchi online.

> Nella prima fase del progetto [...] gli ingegneri hanno affermato di aver cambiato i sistemi dell'azienda per de-prioritizzare i commenti sprezzanti contro "bianchi", "uomini" e "americani". Facebook considera ancora tali attacchi come incitamento all'odio e gli utenti possono ancora segnalarli all'azienda. Tuttavia, la tecnologia della piattaforma ora li etichetta come "a bassa sensibilità" - o minor probabilità di essere dannosi - affinché non vengano più eliminati automaticamente dagli algoritmi dell'azienda. Ciò significa che ogni giorno vengono eliminati circa 10 mila post in meno, secondo i documenti interni all'azienda[31].

La portavoce di Facebook, Sally Aldous, ha affermato:

"Nell'ultimo anno, abbiamo aggiornato le nostre politiche per catturare discorsi di odio più impliciti, come contenuti che raffigurano Blackface, stereotipi sugli ebrei e contenuti negazionisti"[32].

Poiché descrivere le esperienze di discriminazione può comportare la critica dei bianchi, gli algoritmi di Facebook spesso hanno rimosso automaticamente quel contenuto, dimostrando che anche l'intelligenza artificiale avanzata può non essere così efficace quando si tratta di cogliere le sfumature.

Oltre a cancellare i commenti che protestavano contro il razzismo, l'approccio di Facebook ha a volte portato a un netto contrasto tra le sue rimozioni automatiche e le segnalazioni effettive degli utenti sui discorsi di odio. Al culmine delle proteste a livello nazionale a giugno 2020 per l'uccisione di George Floyd,

[30] The Washington Post (2020), Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show: https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/
[31] *Ibidem*
[32] *Ibidem*

un uomo di colore disarmato, ad esempio, i primi tre termini dispregiativi rimossi dai sistemi automatizzati di Facebook furono: "spazzatura bianca", un insulto gay e "cracker", secondo un grafico interno ottenuto da *The Washington Post* e riportato per la prima volta da *NBC News*. Durante quel periodo gli insulti rivolti a persone appartenenti a gruppi emarginati, inclusi neri, ebrei e persone transgender, sarebbero stati rimossi meno frequentemente.

Facebook è stato anche aspramente criticato dai suoi stessi revisori indipendenti in un rapporto sui diritti civili che ha riscontrato come le politiche di incitamento all'odio della piattaforma rappresentassero una "tremenda battuta d'arresto" quando si trattava di proteggere i suoi utenti di colore. Più di una dozzina di dipendenti si sono dimessi per protestare contro le politiche dell'azienda sull'incitamento all'odio. Gli inserzionisti di Facebook hanno organizzato un boicottaggio su questioni relative ai diritti civili, per migliorare il trattamento dell'azienda nei confronti dei gruppi emarginati.

A ottobre 2021 la ex dipendente di Facebook Frances Haugen ha denunciato che, in seguito alle elezioni statunitensi del 2020, la piattaforma social avrebbe allentato la censura dei messaggi d'odio e i contenuti che promulgavano disinformazione sul risultato elettorale. Secondo Haugen, il motivo sarebbe stato prettamente economico: "Facebook guadagna di più quando si consumano più contenuti. Le persone si divertono a interagire con cose che suscitano una reazione emotiva. E più a rabbia vengono esposti, più interagiscono e più consumano"[33]. La ex product ha già presentato otto denunce alla Securities and Exchange Commission e condiviso alcuni documenti interni con il quotidiano "*The Wall Street Journal*", sostenendo che gli algoritmi introdotti nel 2018 al fine di aumentare l'engagement avessero premiato contenuti nocivi sulla piattaforma. Dalle carte emerge che Facebook sarebbe a conoscenza delle carenze nel suo sistema di moderazione dei contenuti, così come dei potenziali effetti negativi che questa scelta avrebbe avuto sulla salute mentale degli utenti.

[33] Cbs, *60 Minutes*, ottobre 2021; https://www.youtube.com/watch?v=_Lx5VmAdZSI

La piattaforma ha definito tali accuse fuorvianti. La società ha ribadito la sua volontà di modificare la politica dei contenuti, oltre che il suo impegno nell'aumentare la diversità nelle assunzioni e nella leadership.

2.2.1. Instagram: l'autotutela degli utenti e le nuove funzioni "anti-odio"

Instagram ha intensificato i suoi sforzi per contrastare l'incitamento all'odio e gli abusi online sulla sua piattaforma, ampliando la gamma di strumenti che aiutano gli utenti a proteggersi. Questi ultimi, infatti, possono auto-tutelarsi disattivando i commenti sotto i propri post o applicando dei filtri che governino il flusso delle risposte.

Ad aprile 2021, Instagram ha lanciato una funzione per impedire agli utenti di visualizzare messaggi privati (Direct Messages, DM) potenzialmente offensivi filtrando parole, frasi ed emoji offensive sull'app di condivisione di foto. Insieme all'opzione di filtro per i messaggi diretti offensivi, la nuova funzione renderebbe anche più difficile per le persone bloccate dagli utenti ricontattarle tramite nuovi account. Il filtro, attivabile su Instagram nelle impostazioni sulla privacy, può essere personalizzato per includere parole, frasi ed emoji che si desidera bloccare o evitare di ricevere nelle richieste di messaggi. Gli utenti possono segnalare, eliminare o aprire i messaggi che verranno ordinati in una cartella delle richieste nascoste.

> Assicurarsi che le persone non vedano contenuti odiosi nei messaggi diretti è più impegnativo, dato che si tratta di conversazioni private. Gli account aziendali e dei creatori, che tendono ad avere volumi elevati di follower e ricevono i messaggi più offensivi da persone che non conoscono, hanno la possibilità di disattivare i DM da persone che non seguono. Abbiamo iniziato a estendere questi controlli agli account personali in molti paesi e speriamo di renderli presto disponibili a tutti. Le persone possono anche scegliere di disattivare i tag o le menzioni di chiunque non conoscano o bloccare chiunque gli invii messaggi indesiderati[34].

Gli utenti che inviano messaggi privati offensivi vengono puniti con la disattivazione dell'account. Instagram assicura inoltre uno stretto monitoraggio di potenziali nuovi account creati dagli stessi per aggirare il blocco.

[34] https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse

> Adotteremo misure più severe quando verremo a conoscenza di persone che infrangono le nostre regole nei messaggi diretti. Attualmente, quando qualcuno invia messaggi diretti che infrangono le nostre regole, vietiamo a quella persona di inviare altri messaggi per un determinato periodo di tempo. Ora, se qualcuno continua a inviare messaggi in violazione, disattiveremo il suo account. Disattiveremo anche i nuovi account creati per aggirare le nostre restrizioni sui messaggi e continueremo a disabilitare gli account che troviamo creati esclusivamente per inviare messaggi offensivi[35].

Instagram sta inoltre aggiornando le misure per proteggere gli utenti dai contatti indesiderati che hanno già bloccato:

> Stiamo rendendo più difficile per qualcuno che hai già bloccato di contattarti di nuovo attraverso un nuovo account. Con la funzione 'blocca questo account e quelli che potrebbe creare', ogni volta che decidi di bloccare qualcuno su Instagram, avrai la possibilità di bloccare sia il suo account che bloccare preventivamente i nuovi account che la persona potrebbe creare.

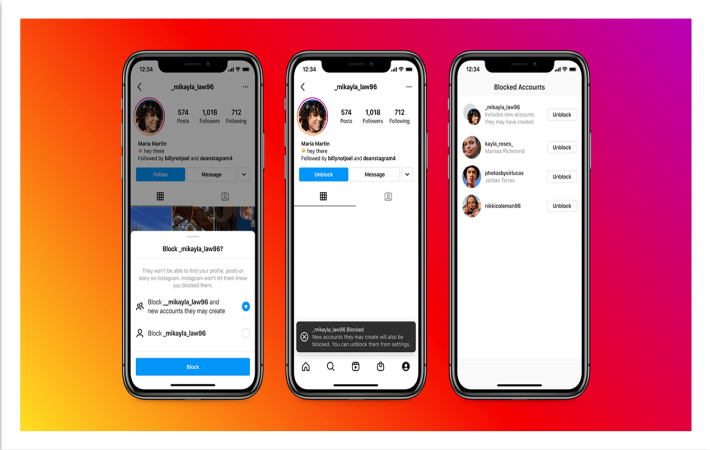


Foto: Instagram

In seguito alle offese circolate sui social network durante le Olimpiadi e gli Europei di calcio, nell'agosto 2021, Instagram ha deciso di introdurre due nuove funzioni per proteggere gli utenti da messaggi e commenti di odio sulla piattaforma[36]. La prima si chiama *Limits* e permette di oscurare automaticamente

[35] https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse

[36] Instagram, *Introducing New Ways to Protect Our Community from Abuse: https://about.instagram.com/blog/announcements/introducing-new-ways-to-protect-our-community-from-abuse*

le richieste di messaggi e i commenti di persone che non seguono il nostro profilo o che hanno iniziato a seguirlo da poco tempo. Dal 10 agosto 2021 Limits è disponibile per tutti gli utenti Instagram e può essere attivata o disattivata nelle impostazioni privacy.

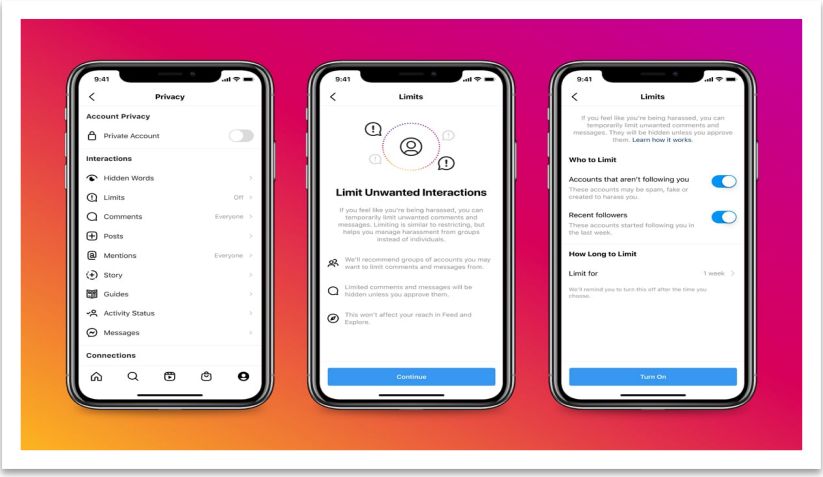


Foto: Instagram

La seconda funzione si chiama *Hidden Words*, parole nascoste: consente di filtrare emoji, parole e keyword offensive inserendole automaticamente in una cartella nascosta inaccessibile. Se si sceglie di aprire la cartella, il testo del messaggio sarà nascosto in modo da non trovarsi di fronte al linguaggio offensivo, a meno che non si tocchi per scoprirlo; a quel punto, l'utente ha la possibilità di accettare la richiesta del messaggio, cancellarla o segnalarla. Tale funzione, inoltre, riesce a rilevare le richieste di messaggi di scarsa qualità o che potrebbero essere spam.

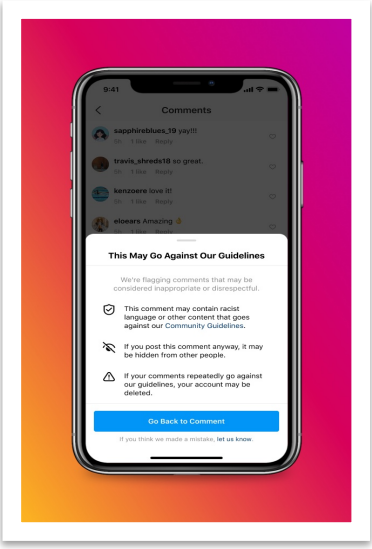


Foto: Instagram

Instagram ha introdotto inoltre Avvisi più evidenti per scoraggiare le molestie:

> Mostriamo già un avviso quando qualcuno tenta di pubblicare un commento potenzialmente offensivo. Se tentano di pubblicare più volte commenti potenzialmente offensivi, mostriamo un avvertimento ancora più forte, avvisandoli che potremmo rimuovere o nascondere il loro commento se procedono. Ora, invece di aspettare il secondo o il terzo commento, mostreremo questo messaggio più forte la prima volta. Abbiamo scoperto che questi avvertimenti scoraggiano davvero le persone dal pubblicare contenuti offensivi.

2.2.2. Contrasto all'odio online: i risultati di Facebook e Instagram

Negli ultimi tre mesi del 2020, il 97% dei discorsi d'odio eliminati dal Gruppo Facebook è stato individuato dai sistemi automatici prima della segnalazione di un utente in aumento rispetto al 94% del trimestre precedente e all'80,5% della fine del 2019. Alla fine del 2017, prima che l'algoritmo Xlm-r fosse impiegato, questo dato era stagnante al 24%.

In termini di contenuti rimossi, gli algoritmi di Intelligenza Artificiale di Facebook hanno permesso, nel trimestre da aprile a giugno del 2021, di eliminare 31,5 milioni di contenuti di incitamento all'odio, in confronto ai 25,2 milioni del primo trimestre dell'anno. I dati sono incoraggianti anche su Instagram: nel medesimo periodo sono stati rimossi 9,8 milioni di post, rispetto ai 6,3 milioni del periodo gennaio-marzo.

2.3. Le Policy di YouTube: approccio graduale e "segnalatori attendibili"

In caso di contenuti offensivi o di incitamento all'odio, YouTube procede gradualmente limitando le funzionalità per tali contenuti (commenti, video consigliati, "mi piace", idoneità agli annunci) e segnalandoli agli utenti con un messaggio di avviso introduttivo. La piattaforma può inoltre rimuovere i contenuti o imporre altre sanzioni al creatore nella seguente casistica[37]:

> Quando un creator istiga ripetutamente gli spettatori ad adottare comportamenti illeciti; prende di mira, insulta e offende ripetutamente un gruppo sulla base delle caratteristiche di cui sopra in più caricamenti; espone un gruppo che possiede le caratteristiche di cui sopra al rischio di subire danni fisici in base al contesto sociale o politico locale; crea contenuti che danneggiano l'ecosistema di YouTube in quanto insiste nel fomentare ostilità nei confronti di un gruppo che possiede le caratteristiche di cui sopra per ricavarne un utile finanziario personale.

In caso di violazioni gravi o reiterate, YouTube si riserva la facoltà di chiudere il canale o l'account del soggetto interessato. A tal fine, ha attivato un sistema di avvertimenti:

> Se i tuoi contenuti violano queste norme, li rimuoveremo e ti invieremo un'email per informarti della nostra decisione. Se è la prima volta che violi le nostre Norme della community, è probabile che tu riceva soltanto un avviso, senza alcuna sanzione al tuo canale. In caso contrario, emetteremo un avvertimento nei confronti del tuo canale. Se ricevi 3 avvertimenti nell'arco di 90 giorni, il tuo canale YouTube verrà chiuso.

Dietro la crescita di YouTube "c'è un algoritmo che crea playlist personalizzate. YouTube afferma che questi consigli guidino oltre il 70% del suo tempo di visualizzazione, rendendo l'algoritmo tra i più grandi decisori di ciò che la gente guarda"[38]. Secondo YouTube, le modifiche apportate nel 2019 a questo algoritmo di raccomandazione hanno dimezzato durante quell'anno le

[37] Youtube, Norme sull'incitamento all'odio: https://support.google.com/youtube/answer/2801939?hl=it&ref_topic=9282436. Ultimo accesso: 31/08/2021
[38] The Wall Street Journal, Febbraio 2018, *How YouTube Drives People to the Internet's Darkest Corners*

visualizzazioni di video ritenuti "contenuti limite" per la diffusione di disinformazione.

Negli anni, infatti, sono stati apportati diversi miglioramenti volti a contrastare l'odio online.

> Questo lavoro si è concentrato su quattro pilastri: rimozione di contenuti violativi , raccolta di contenuti autorevoli , riduzione della diffusione di contenuti limite e ricompensa per i creatori di fiducia .
> [...] È fondamentale che i nostri sistemi di monetizzazione premino i creator affidabili che aggiungono valore a YouTube. Abbiamo da tempo linee guida relative agli inserzionisti che vietano la pubblicazione di annunci su video che includono contenuti che incitano all'odio e le applichiamo rigorosamente. [...] I canali che violano ripetutamente le nostre norme sull'incitamento all'odio verranno sospesi dal Programma partner di YouTube, il che significa che non possono pubblicare annunci sul proprio canale o utilizzare altre funzionalità di monetizzazione come Superchat[39].

YouTube si avvale inoltre di una rete di "segnalatori attendibili". Al momento vi fanno parte 180 istituti accademici, partner governativi e ONG. YouTube invita anche i singoli utenti che segnalano un elevato volume di video con un alto tasso di accuratezza a partecipare al programma Segnalatore attendibile. I video identificati dai segnalatori attendibili hanno priorità di revisione ma non vengono rimossi in automatico, bensì vengono sottoposti alla revisione da parte di persone fisiche.

La piattaforma, tuttavia, è stata contestata per i suoi sforzi insufficienti. Ad esempio, alcuni critici hanno osservato come in certi casi invece di rimuovere video che avevano provocato molestie omofobe, YouTube si sia limitato a impedire all'utente di condividere le entrate pubblicitarie.

Un caso che fece molto discutere, qualche anno fa, fu quello del conflitto tra gli YouTuber Carlos Maza e Steven Crowder, quest'ultimo accusato dal primo di molestie razziste e omofobe. Maza disse di aver provato per anni a chiedere l'intervento della piattaforma e chiese ai suoi follower di segnalare i video a YouTube. Crowder, da parte sua, dichiarò che i suoi epiteti erano da considerare scherzosi.

[39] YouTube (2019) *Our ongoing work to tackle hate;* https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate. Ultimo accesso: 31/08/2021

YouTube, dopo aver indagato sui video di Crowder, dichiarò che non avevano violato le politiche sulle molestie della piattaforma ma annunciò di aver comunque rimosso la monetizzazione degli annunci pubblicitari dai video di Crowder. La piattaforma scrisse:

> Ci sono due politiche chiave in gioco qui: molestie e incitamento all'odio […] Per essere chiari, l'uso di epiteti razziali, omofobi o sessisti da soli non violerebbe necessariamente nessuna di queste politiche. Ad esempio, il linguaggio osceno o offensivo è spesso usato nelle canzoni e negli spettacoli comici.
> Se dovessimo eliminare tutti i contenuti potenzialmente offensivi, perderemmo un discorso prezioso, un discorso che consente alle persone di tutto il mondo di alzare la voce, raccontare le loro storie[40].

Impegnandosi ad aggiornare le proprie policy su *hate speech* e molestie, You Tube annunciò:

> Anche se il contenuto di un autore non viola le nostre linee guida comunitarie, daremo un'occhiata al contesto e all'impatto in modo più ampio, e se il loro effetto danneggerà la comunità in senso ampio, interverremo. [...] Nel caso del canale di Crowder, un'analisi approfondita ha rilevato che singolarmente i video segnalati non violavano le nostre Norme della community. Tuttavia abbiamo visto il danno diffuso alla community e abbiamo preso la decisione di sospendere la monetizzazione[41].

È evidente come rilevare un discorso d'odio e agire di conseguenza sia spesso molto difficile per queste piattaforme.

2.3.1. Contrasto all'odio online: i risultati di Youtube

Nel primo trimestre del 2021, il 99.4% dei contenuti rimossi per violazione delle norme di Youtube sono stati rilevati dai sistemi di segnalazione automatica, un dato sostanzialmente in linea con quello del 2021 (99.6%). Le Norme della community di YouTube vengono applicate in modo uniforme in tutto il mondo. Il paese con più video rimossi per violazione delle norme è l'India, seguita dagli Stati Uniti. Nel 2021 l'Italia si trovava al ventinovesimo posto nella classifica globale dei contenuti rimossi.

[40] Youtube (2019) Taking a harder look at harassment: https://blog.youtube/news-and-events/taking-harder-look-at-harassment/
[41] *Ibidem*

Tra gennaio e marzo 2021, YouTube ha rimosso 33.871 canali offensivi o che incitavano all'odio, pari all'1.5% di tutti i canali rimossi nel periodo, e in aumento rispetto all'anno precedente.

Il numero totale di video rimossi per incitamento all'odio da gennaio a marzo 2021 è 85.247, pari allo 0.9% di tutti i contenuti rimossi. Numero in diminuzione rispetto al medesimo periodo dell'anno precedente, quando erano stati rimossi 107.1704 video, pari all'1.8% di tutti i contenuti rimossi.

I commenti rimossi per incitamento all'odio nel primo trimestre del 2021 sono 43.601.586, pari al 4.2% di tutti i commenti rimossi. Anche qui si registra un calo rispetto all'anno precedente, quando i commenti rimossi per incitamento all'odio erano 43.539.615, il 6.3% del complessivo.

Considerazioni conclusive

L'attuale *governance* dell'*hate speech* online si caratterizza per la mancanza di armonizzazione definitoria tra governi nazionali, organizzazioni intergovernative, piattaforme Internet e organizzazioni della società civile.

La difficoltà nel definire cosa costituisca *hate speech* riflette la sua percezione in termini di dualismo tra controllo dell'incitamento all'odio e libertà di parola. Difatti, come visto, alla mancanza di un significato condiviso dei discorsi d'odio si affianca la diversa sensibilità degli ordinamenti rispetto al tema, che deriva dalla differente propensione a bilanciare la libertà di espressione con la dignità umana.

In tale contesto, vengono chiamati in causa il ruolo e la responsabilità delle piattaforme web, nella specie i social network, che rappresentano ormai il mezzo principale di veicolo dei discorsi d'odio. Operando come un fornitore di servizi, ogni social network ha avuto spazio di manovra per adottare una definizione di *hate speech* più omogenea e precisa, come nel caso di Youtube o Facebook, o più circostanziale, come nel caso di Twitter.

In proposito va segnalato, a livello di Unione Europea, la sussistenza di un quadro normativo sulla responsabilità degli *internet service provider* in parte non più adeguato alle modifiche che hanno subito in questi anni le piattaforme; a questo si è affiancato il recente approccio delle istituzioni europee (di livello sia

legislativo sia giudiziario) volto a riconoscere maggiore potere di intervento ai prestatori di servizi internet sui contenuti che circolano sulle loro piattaforme. A ciò si aggiungono disposizioni di *soft law,* attraverso un coinvolgimento diretto dei principali provider (v. Codice di condotta per lottare contro le forme illegali di incitamento all'odio *online*), nonché le policy dei social che, da soggetti privati, impongono ulteriori e specifiche regole agli utenti che le sottoscrivono, per iscriversi e partecipare al dibattito che si instaura sulle piattaforme.

Tale circostanza ha portato a una sempre maggiore responsabilizzazione delle piattaforme, a cui è seguita una sorta di "privatizzazione" nella tutela dei diritti, nel momento in cui si rimette al soggetto privato la decisione di cosa deve restare in rete e, quindi, dell'ampiezza da accordare alla libertà di parola piuttosto che alla dignità umana. È evidente la necessità di un coinvolgimento delle piattaforme, che sono i primi soggetti che possono venire a conoscenza della presenza di contenuti illeciti o comunque nocivi e che sono in grado, quindi, di apprestare prontamente misure di tutela per gli utenti minacciati. Al tempo stesso, non si possono sottovalutare i rischi di lasciar fuori un intervento pubblico in materia e demandare la regolazione del web esclusivamente ai soggetti privati che governano la rete.

Si può ora avanzare qualche considerazione conclusiva, in merito all'efficacia delle misure poste in essere sin qui – sul piano istituzionale e da parte delle stesse piattaforme – al fine di limitare la spirale d'odio nel contesto online.

I risultati positivi che si sono registrati soprattutto nella valutazione del Codice di condotta relativa al 2020, oltre che delle nuove funzionalità previste dalle piattaforme in questi ultimi anni, danno conto di un sempre maggiore impegno di tali soggetti sul tema della discriminazione e dell'odio online. Dall'introduzione del Codice di condotta, infatti, le piattaforme co-firmatarie hanno attuato un diffuso rafforzamento degli strumenti forniti agli utenti per segnalare i contenuti; si nota, inoltre, una crescente enfasi verso la tutela della messaggistica privata.

Tuttavia, la strada da percorrere è ancora lunga. L'ultima valutazione relativa al 2021 mostra dati meno rassicuranti rispetto all'anno precedente, in merito all'attività svolta dalle piattaforme nella lotta all'*hate speech*. In particolare, la Commissione ha evidenziato lacune per quanto concerne i *feedback* agli utenti e i

sistemi di trasparenza dei social. Va sottolineato, in proposito, come la recente proposta contenuta nel *Digital Serivces Act* evidenzia proprio l'importanza della trasparenza e di sistemi di notifica agli utenti, su cui le piattaforme saranno molto probabilmente chiamate ad operare in maniera più attiva.

Su tale quadro sono intervenute, da ultimo, le rivelazioni sull'attività della piattaforma Facebook da parte di Frances Haugen, che hanno evidenziato la tendenza dell'azienda a privilegiare il profitto ai danni della sicurezza e dei diritti degli utenti, oltre che una poca trasparenza relativamente alle pratiche nei confronti delle autorità di regolamentazione. Per ciò che interessa principalmente in questa sede, dai documenti resi pubblici risulta, inoltre, che l'azienda fosse a conoscenza del fatto che la modifica dell'algoritmo del 2018, pensato per migliorare l'esperienza degli utenti, avesse in realtà condotto a una crescita dell'*hate speech* sulla piattaforma social.

Va quindi tenuto conto delle lacune che ancora presentano i sistemi algoritmici e dell'inefficacia che tali strumenti hanno spesso dimostrato nella rimozione di contenuti inneggianti all'odio. Nello specifico, l'intento e il contesto, principalmente regionale e linguistico, si confermano i criteri più complicati da definire per i moderatori delle piattaforme considerate. Nonostante l'ambizione, comune ai quattro erogatori di servizi online, di sviluppare sistemi multilingua adatti ad una clientela globale, i contenuti che non sono espressi in lingua inglese risultano sistematicamente penalizzati.

Dalla recente introduzione di organismi di supervisione da parte del gruppo Facebook e di Twitter, i contenuti più delicati sono spesso assegnati ad alcuni esperti indipendenti al fine di complementare l'attività di moderazione logaritmica. Nel caso di Facebook, tuttavia, l'azione dell'organismo è risultata sin qui poco tempestiva e di corto raggio. Youtube, infine, non si è dotato di un comitato di supervisione.

Tutto ciò dimostra il rischio, prima accennato, di lasciare la *governance* in materia esclusivamente, o anche solo prevalentemente, alle stesse piattaforme digitali.

Rispetto al quadro così delineato, risulterà fondamentale – a fianco del ruolo che intenderanno svolgere le piattaforme – l'intervento del legislatore europeo, il

quale ha già posto le basi verso una nuova fase di regolamentazione della responsabilità dei fornitori di servizi online. Come si è visto, con la recente proposta del *Digital Services Act*, si pone nuovamente al centro l'*hard law* rispetto al ruolo primario finora svolto dalla giurisprudenza e dall'autoregolamentazione, al fine di garantire una maggiore procedimentalizzazione delle controversie tra piattaforme e utenti, nonché un più incisivo controllo da parte delle istituzioni "tradizionali" sulla diffusione dell'odio in rete.

Riferimenti bibliografici

ABBONDANTE F., 2018, *Il ruolo dei social network nella lotta all'hate speech: un'analisi comparata fra l'esperienza statunitense e quella europea*, in G.L. Conti - M. Pietrangelo - F. Romeo, *Social media e diritti. Diritto e social media*, Napoli, 41 ss.

BARAKEREZ D. - SCHARIA D., *Freedom of Speech, Support for Terrorism, and the Challenge of Global Constitutional Law*, in *Harvard National Security Journal*, 2/2011.

BASSINI M., 2019, *Internet e libertà di espressione*, Roma.

CARMI G.E., 2008, *Dignity Versus Liberty: The Two Western Culture of Free Speech*, 22 agosto 2008, in https://ssrn.com/abstract=1246700

CASTORINA E., 2012, *Manifestazione del pensiero e messaggi di "odio sociale" nel cyberspazio. Una regolamentazione multilivello ancora incompiuta*, in M. Villone - A. Ciancio - G. De Minico - G. Demuro - F. Donati (a cura di), *Nuovi mezzi di comunicazione e identità. Omologazione o diversità?,* Roma, 105 ss.

COLE D., 2012, *The First Amendment's Borders: The Place of Holder v. Humanitarian Law Project in First Amendment Doctrine*, in *Hardvard Law and Policy Review*, 6/2012, 148 ss.

DUNN P., *Il contrasto europeo all'hate speech online: quali prospettive future?*, in *medialaws.eu*, 20 gennaio 2021.

FALLETTA P., 2020, *Controlli e responsabilità dei social network sui discorsi d'odio online*, in *MediaLaws - Riv. dir. media*, 3/2020, 146 ss.

FALLETTA P., 2021, *Il contrasto all'*hate speech, in M. Mensi - P. Falletta, *Il diritto del web,* Padova, III ed., 2021, 151 ss.

GRANDINETTI O., 2021, *Facebook vs. CasaPound e Forza Nuova, ovvero la disattivazione di pagine social e le insidie della disciplina multilivello dei diritti fondamentali*, in *MediaLaws - Riv. dir.media,* 1/2021, 173 ss.

MAGNANI C., 2019, *Il Regolamento dell'Agcom sull'*hate speech*: una prima lettura*, in *Forum Quad. cost.*, 14 giugno 2019.

MAZZOLAI B., *Hate speech e comportamenti d'odio in rete: il caso Forza Nuova c. Facebook*, in *Dir. inform.*, 3/2020, 552 ss.

MELZI D'ERIL C. - VIGEVANI G.E., 2019, *Facebook e l'odio in rete serve più trasparenza su ciò che è ammesso*, in *Ilsole24ore*, 10 settembre 2019.

MULLER K. - SCWARZ C.,2021, *Fanning the Flames of Hate: Social Media and Hate Crime*., in *Journal of the European Economic Association*, agosto 2021, 2131 ss.

SPADARO I., 2020, *Il contrasto allo hate speech nell'ordinamento costituzionale globalizzato*, Torino.

STRADELLA E., 2011, *Odio razziale e libera manifestazione del pensiero negli Stati Uniti*, in D. Tega (a cura di), *Le discriminazioni razziali ed etniche. Profili giuridici di tutela*, Roma, 2011, 118 ss.

ZICCARDI G., 2016, *L'odio online. Violenza verbale e ossessioni in ret*e, Milano.

Sitografia

ALLAN R., *Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?*, 2017, https://about.fb.com/news/2017/06/hard-questions-hate-speech/ [ultimo accesso: 10/10/2021].

AMNESTY INTERNATIONAL, *Barometro dell'odio e intolleranza pandemica,* 2021, https://www.amnesty.it/barometro-dellodio-intolleranza-pandemica/ [ultimo accesso: 08/10/2021].

AVAAZ, *Left Behind: How Facebook is neglecting Europe's infodemic*, 2021, https://secure.avaaz.org/campaign/en/facebook_neglect_europe_infodemic/ [ultimo accesso: 05/08/2021].

DATAMEDIAHUB - KPI, *Rapporto sull'Hate Speech in Italia su Twitter*, 2020, http://www.datamediahub.it/2020/06/22/rapporto-sullhate-speech-in-italia/ [ultimo accesso: 10/10/2021].

EUROPEAN COMMISSION, *The EU Code of conduct on countering illegal hate speech online. The robust response provided by the European Union* https://ec.europa.eu/commission/presscorner/detail/it/IP_20_1134 [ultimo accesso : 12/10/2021]

FACEBOOK, *Community Standards enforcement*, 2021, https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021/ [ultimo accesso: 22/10/2021].

FACEBOOK, *Facebook Oversight Board*, 2021, https://www.oversightboard.com/decision/. [ultimo accesso: 22/10/2021].

FACEBOOK AI, *How Facebook uses super-efficient AI models to detect hate speech,* 2020, https://ai.facebook.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/ [ultimo accesso: 05/08/2021].

FACEBOOK AI, *Update on Our Progress on AI and Hate Speech Detection*, 2021, https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/[ultimo accesso: 05/08/2021].

HUMAN RIGHTS COUNCIL, *Report of the independent international fact-finding mission on Myanmar,* 10-28 settembre 2021, https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf [ultimo accesso: 01/09/2021].

INSTAGRAM, *An update on our work to tackle abuse on Instagram*, 2021, https://about.instagram.com/blog/announcements/an-update-on-our-work-to-tackle-abuse-on-instagram [ultimo accesso: 22/10/2021].

SENATO DELLA REPUBBLICA, Commissione straordinaria per il contrasto dei fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza, https://www.senato.it/notes9/Web/18LavoriNewV.nsf/OdGSpecConvCommWebLeg143?ReadForm&10/2021/18 [ultimo accesso 22.10.2021].

YOUTUBE, *Norme sull'incitamento all'odio*, https://support.google.com/youtube/answer/2801939?hl=it&ref_topic=9282436 [ultimo accesso: 31/08/2021].

YOUTUBE, *Our ongoing work to tackle hate,* 2019*,* https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate [ultimo accesso: 31/08/2021].

YOUTUBE, *Taking a harder look at harassment,* 2019*,* https://blog.youtube/news-and-events/taking-harder-look-at-harassment/[ultimo accesso: 31/08/2021].

YOUTUBE, *Transparency Report*, 2021, https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=it&policy_removals=period:2021Q1&lu=policy_removals [ultimo accesso: 31/08/2021].

THE VERGE, *Why you can't say 'men are trash' on Facebook,* 2019*,* https://www.theverge.com/interface/2019/10/3/20895119/facebook-men-are-trash-hate-speech-zuckerberg-leaked-audio[ultimo accesso: 01/09/2021].

THE WALL STREET JOURNAL, *Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show,* 14 settembre 2021, https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739?mod=article_inline [ultimo accesso: 04/10/2021].