Federico Oliveri*

# *Diritti degli internauti, obblighi degli Stati, responsabilità delle piattaforme digitali: problemi regolativi in materia di odio online*

*Abstract:* The digitalization of everyday life is advancing rapidly, in parallel with the increasing penetration of the internet and social media among the world population. In this scenario, legal responsibilities and obligations of digital platforms, with respect to what is shared by their users, have increasingly become a matter of public relevance, closely linked to the democratic quality of political regimes and of society as a whole. This article aims to critically reconstruct the main regulatory problems posed by countering hate speech on social networks. The final aim of this reconstruction is to advance a new regulatory model that is, in many respects, an alternative to the current one, which has become a source of criticism and legal disputes. In particular, the article offers a new definition of offline and online hate speech, using categories inspired by philosophical anthropology and pragmatics of communication; a justification of the right to be protected from online hate speech and the related obligations in reference to the principle of equal social dignity, within the frame of digital citizenship; criteria for a fair balance between protection from online hate speech and other competing rights, such as freedom of information, expression, and association; an overall regulatory model based on the co-responsibility of the various actors, engaged in protecting users from online hate speech according to their different capacities. This model should mitigate the current risks of a "privatization of censorship", caused by the delegation to digital platforms of the power to detect and remove prohibited contents.

*Keywords:* Hate speech, digital citizenship, fundamental rights, digital platforms

Chi odia si sforza d'allontanare e far venir meno quanto egli abbia in odio.
(Baruch Spinoza, *Etica*, Parte III, Proposizione 13, Scolio)

## 1. L'odio online come problema regolativo e sociale

La penetrazione globale della rete e dei social media[1] ha trasformato le principali aziende tecnologiche, come *Alphabet*, *Meta Platforms* e *Microsoft*, in attori di prima grandezza nelle dinamiche economiche, politiche e culturali attuali.

* Ricercatore aggregato presso il Centro Interdisciplinare "Scienze per la Pace" dell'Università di Pisa e assegnista di ricerca in "Filosofia del diritto" presso la Scuola di Giurisprudenza dell'Università di Camerino – federico.oliveri@cisp.unipi.it o federico.oliveri@unicam.it

**Teoria e Critica della Regolazione Sociale**, n. 2, 2021 • Mimesis Edizioni, Milano-Udine Web: www.mimesisjournals.com/ojs/index.php/tcrs • ISSN (online): 1970-5476 • ISBN: 9788857597379 • DOI: 10.7413/19705476056

In particolare, la facoltà degli ecosistemi digitali di diffondere, senza intermediazioni o costi apparenti, messaggi dalla natura più diversa all'interno di una sfera pubblica con milioni di potenziali destinatari, così come la loro capacità di condizionare opinioni e comportamenti, attribuisce a *Facebook*, *Instagram*, *Twitter, YouTube* e *TikTok* un potere senza precedenti per dei soggetti di diritto privato con finalità commerciali. Anche per questo, la responsabilità legale e gli obblighi di intervento delle piattaforme rispetto a quanto condiviso dai propri utenti, insieme alle relative forme di regolazione, costituiscono da tempo materie di rilevanza pubblica: a esserne investite sono le basi stesse della pacifica convivenza umana e la qualità democratica dei regimi politici e della società nel suo complesso[2].

La domanda di regolazione in materia è cresciuta costantemente nel corso degli ultimi decenni, in parallelo alla crisi dell'originaria visione della rete come spazio orizzontale e intrinsecamente democratico, finalizzato a condividere informazioni, incrementare le conoscenze e promuovere la partecipazione diretta di cittadini e cittadine alle scelte collettive[3].

Oggi, in effetti, guardiamo alla rete con occhi sempre più disincantati[4]. L'offerta di servizi online è monopolizzata da grandi multinazionali orientate alla massimizzazione dei profitti, anche attraverso la mercificazione dei dati personali degli utenti[5]. Diversi Stati operano, a vario titolo e con varie modalità, restrizioni all'uso della rete[6]. L'ambiente digitale rischia di produrre sovraccarichi informativi, di alimentare dipendenza e disagio psicologico[7], di favorire attività criminali[8], di diffondere disinformazione[9] e incentivare ostilità verso alcuni gruppi identificati sulla base della 'razza', del sesso, dell'identità di genere, dell'orientamento sessuale, della religione o di altre condizioni personali[10].

1 Secondo l'ultima rilevazione di *DataReportal*, pubblicata a gennaio 2022, gli utenti di internet sono 4,95 miliardi, pari al 62,5% della popolazione mondiale, con un incremento del 4% nell'ultimo anno. Il 58,4% della popolazione mondiale è iscritto a uno o più social, con un incremento nell'ultimo anno del 10,1%, e vi trascorre mediamente 2 ore e 27 minuti al giorno. I dati completi e aggiornati sono accessibili su https://datareportal.com.

2 van Dijk, Hacker 2018; Sunstein 2017; Dal Lago 2017; Monti 2019b.

3 Per una ricostruzione delle "utopie libertarie" all'origine della rete, si vedano Formenti 2000 e 2009. Sul ruolo che la rete e i social possono svolgere nei movimenti di protesta, si veda per tutti Earl, Kimport 2011.

4 Per una difesa della "neutralità" della rete, contro la sua demonizzazione in quanto tale, si veda soprattutto Ziccardi 2016.

5 Hindman 2018; Zuboff 2020.

6 Per una panoramica generale, si veda Kaye 2019. Sull'anti-terrorismo come argomento ricorrente, spesso pretestuoso, con cui gli Stati controllano la rete, si veda ancora Ziccardi 2016.

7 Lavenia 2012; Kuss, Pontes 2019.

8 Jaishankar 2011 (ed.)

9 Tra i molti riferimenti, si vedano almeno Mintz 2012 (ed.) e Quattrociocchi, Vicini 2016.

10 Sul nesso tra gli algoritmi dei motori di ricerca e la diffusione del razzismo, si veda Noble 2018. Sulla correlazione tra odio online e offline, si vedano gli studi empirici di Müller, Schwarz 2021 e si consultino le "mappe dell'intolleranza" accessibili su http://www.voxdiritti.it. Per una critica della correlazione lineare tra odio online e offline, si veda invece Mchangama 2015.

Tra i contenuti pericolosi che gli internauti producono o incontrano in rete, intendo qui concentrare l'attenzione sui cosiddetti "discorsi d'odio" o *hate speech*, secondo la fortunata terminologia anglosassone[11]. Questa scelta non registra soltanto l'intensità con cui il tema è oggi dibattuto, anche fuori dalle cerchie specialistiche, ma risponde soprattutto alla persistente insoddisfazione per i risultati raggiunti dalle attuali forme di regolazione nel contrastare e prevenire i discorsi d'odio, e nel tutelare le persone che ne subiscono gli effetti[12]. Tale insoddisfazione è riemersa di recente in due eclatanti controversie, legate al modo con cui *Facebook* avrebbe gestito (o, meglio, avrebbe omesso di gestire) l'*hate speech*.

Il caso più vicino nel tempo risale all'inizio di dicembre 2021. È stato reso noto che un gruppo di rifugiati Rohingya ha avviato nel Regno Unito e negli Stati Uniti cause legali contro *Facebook*, ritenuto responsabile di aver consentito e amplificato l'incitamento all'odio e alla violenza contro di loro. La società non avrebbe rimosso i contenuti ostili indirizzati contro la minoranza musulmana in Myanmar, né avrebbe bloccato account pericolosi per la loro incolumità, non solo per le carenze linguistiche del proprio sistema automatico e umano di rimozione dei discorsi d'odio, ma anche per non rallentare la propria penetrazione nel paese.

Il secondo caso risale alla fine di ottobre 2021, quando vari giornali statunitensi ed europei hanno divulgato i cosiddetti *Facebook Papers*: un *leak* di oltre 10.000 pagine, contenente documenti interni della società raccolti dall'ex dipendente Frances Haugen[13]. I documenti includono segnalazioni di vari operatori, secondo cui in molti casi la dirigenza ha dato priorità agli interessi economici rispetto alla sicurezza degli utenti, sottovalutando o ignorando la diffusione di *fake news* e *hate speech*, soprattutto in paesi a rischio di violenze contro le proprie minoranze. Nei *Papers* vengono citate anche ricerche interne, che mostrano come i meccanismi di funzionamento del *social network* favoriscano la diffusione della disinformazione e dell'incitamento all'odio, e come i dispositivi di intelligenza artificiale predisposti per rimuovere precocemente l'*hate speech* siano ancora poco performanti. Nel ribattere alle critiche, la società ha rivendicato l'esistenza di regole precise per la rimozione dei "contenuti nocivi", la significativa riduzione dell'incitamento all'odio sulla piattaforma e la periodica pubblicazione dei risultati raggiunti, ribadendo il proprio consenso a una regolamentazione pubblica della materia.

Anche volendo sospendere il giudizio sui risultati raggiunti da *Facebook* nel contrasto dell'odio online[14], resta il dato relativo all'elevato livello di conflittualità che accompagna la questione: da qui la necessità di ripensare il quadro regolativo complessivo e la sua applicazione pratica.

11 Per un'ampia e accurata ricostruzione dei dibattiti in materia, rinvio a Di Rosa 2020.

12 Di "fragilità del diritto" rispetto ai fenomeni d'odio parla Villaschi 2021: 113.

13 I *Facebook Papers* sono accessibili su https://www.protocol.com/facebook-papers.

14 Nelle comunicazioni pubbliche la società afferma di rimuovere oltre il 90% dei discorsi d'odio ma, nelle comunicazioni interne riportate nei *Facebook Papers*, ammette che la cifra di rimozione non supera il 5%. Il primo dato si riferisce ai contenuti rimossi tra quelli rilevati in "modo proattivo" tramite dispositivi automatici. Il secondo dato si riferisce, invece, al numero di rimozioni sulle segnalazioni ricevute dagli utenti.

In linea generale ritengo che, per ottenere risultati meno controversi e più duraturi sul fronte del contrasto, della prevenzione e della tutela delle vittime di *hate speech*, non sia sufficiente potenziare i dispositivi di intelligenza artificiale[15], né rendere più stringenti le attuali procedure di controllo, accompagnandole eventualmente con sanzioni più severe. Ritengo, invece, che occorra ripensare criticamente l'odio online come problema regolativo e sociale.

La teoria del diritto può costituire un valido supporto in questa direzione, sollevando tre domande fondamentali a cui, nei successivi paragrafi, cercherò di dare brevemente risposta: le definizioni di *hate speech* oggi in uso sono adeguate alla realtà del fenomeno, sia offline che online (§ 2)? Qual è il fondamento del diritto degli internauti a non subire discorsi d'odio online, e a chi spetta l'obbligo di garantire tale diritto, decidendo quali contenuti rimuovere e quali account disattivare (§ 3)? Chi e come può garantire il diritto a essere protetti dall'odio online, senza comprimere in modo irragionevole le libertà di informazione, di manifestazione del pensiero e di associazione in rete (§ 4)?

Quella che segue è una riflessione di tipo teorico-normativo: i riferimenti a documenti giuridico-politici o a sentenze non potranno, dunque, essere analitici ma saranno selezionati rispetto alle esigenze dell'argomentazione.

## 2. Discorsi d'odio: problemi di definizione e peculiarità dell'ambiente digitale

Negli studi giuridici, sociologici e linguistico-computazionali è comune lamentare l'assenza di una definizione condivisa di *hate speech*[16]. Tale assenza può dispiacere ma non deve sorprendere. Da una parte, essa rispecchia la natura complessa, ambivalente e situata di ogni comunicazione umana: tale natura emerge con forza proprio nei discorsi d'odio, ad esempio nei frequenti tentativi di negare o dissimulare l'ostilità sociale per non incorrere nella disapprovazione o nella censura[17]. Dall'altra parte, l'assenza di una definizione condivisa segnala la difficoltà di stabilire, una volta per tutte, quali espressioni ostili siano accettabili e quali no rispetto ai principi costituzionali dei diversi ordinamenti[18] o rispetto ai valori morali diffusi nelle varie società[19], anche tenendo conto che le forme dell'odio mutano in modo significativo nel tempo e nello spazio[20].

15 Si è sviluppato negli ultimi anni un intenso dibattito intorno ai dispositivi di "intelligenza artificiale" utilizzabili per individuare i discorsi d'odio e procedere alla loro cancellazione dalle piattaforme digitali ancor prima che vengano visualizzati. Per una ricostruzione della letteratura sul tema, con particolare attenzione ai limiti di tali dispositivi allo stato attuale del loro sviluppo, si veda Jahan, Oussalah 2021.

16 Sellars 2016; MacAvaney *et al.* 2019; Di Rosa 2020; Faloppa 2020; Röttger *et al.* 2021.

17 Su questo specifico aspetto si veda, da ultimo, Pintore 2021.

18 Sulla diversità di approccio regolativo al tema dell'*hate speech* tra Stati Uniti e paesi europei, si vedano almeno Kiska 2012 e Ziccardi 2020.

19 A titolo d'esempio, le legislazioni di alcuni paesi a maggioranza musulmana includono tra i discorsi d'odio le espressioni classificate come "blasfeme". Sul punto si veda ancora Ziccardi 2016.

20 Quando è esplosa la pandemia, ad esempio, si è registrata la diffusione di attacchi verbali e fisici a persone o gruppi accusati di diffondere il Covid-19. Si veda, per il Regno Unito, Gray, Hansen 2021.

Naturalmente, ciò non significa che la mancanza di consenso sulla definizione di *hate speech* sia irrilevante. In primo luogo, nel momento in cui si intendono perseguire penalmente le espressioni d'odio, il "principio di determinatezza" richiede che gli elementi del crimine siano definiti nel modo più preciso possibile, così da mantenere l'applicazione e l'interpretazione dei giudici e delle corti entro ambiti non arbitrari, e consentire un equo bilanciamento degli interessi in gioco. In secondo luogo, vista la dimensione globale delle piattaforme digitali e la natura transnazionale della rete, la diffusione di espressioni d'odio online richiede una definizione comune anche minima del fenomeno, per poter coordinare gli approcci regolativi dei diversi spazi giuridici nazionali e sovranazionali.

In ambito europeo, si è cercato di rimediare all'assenza di una definizione giuridica condivisa di *hate speech* facendo riferimento soprattutto a tre fonti sovranazionali di natura assai diversa, per altro significativamente disomogenee per ampiezza e contenuti.

La Raccomandazione n. 20 del 1997 del Comitato dei Ministri del Consiglio d'Europa definisce discorso d'odio "qualunque forma di espressione che diffonda, inciti, promuova o giustifichi l'odio razziale, la xenofobia, l'antisemitismo o altre forme di odio basate sull'intolleranza, incluse l'intolleranza espressa attraverso il nazionalismo aggressivo e l'etnocentrismo, la discriminazione e l'ostilità contro le minoranze, i migranti e le persone di origine migrante".

La Decisione quadro 2008/913/GAI, sulla lotta contro talune forme ed espressioni di razzismo e xenofobia mediante il diritto penale, definisce come discorso d'odio "ogni comportamento consistente nell'istigazione pubblica alla violenza o all'odio nei confronti di un gruppo di persone o di un suo membro, definito in riferimento alla razza, al colore, alla religione, all'ascendenza o all'origine nazionale o etnica". Gli Stati membri dell'Unione Europea sono obbligati a prevedere norme penali per sanzionare comportamenti intenzionali individuati sulla base di questa definizione. Ad essa fanno riferimento anche il Codice di condotta contro l'illecito incitamento all'odio online, stipulato nel 2016 tra la Commissione Europea e le principali piattaforme digitali, e la Direttiva UE 2018/1808 sui servizi di media audiovisivi, che ha agganciato la definizione dei "motivi" dell'odio alle caratteristiche protette da discriminazione ai sensi dell'art. 21 della Carta dei Diritti fondamentali dell'Unione Europea[21].

La Raccomandazione n. 15 adottata l'8 dicembre 2015 dalla Commissione europea contro il razzismo e l'intolleranza (ECRI) istituita presso il Consiglio d'Europa, definisce discorso dell'odio "il fatto di fomentare, promuovere o incoraggiare, sotto qualsiasi forma, la denigrazione, l'odio o la diffamazione nei confronti di una persona o di un gruppo, nonché il fatto di sottoporre a soprusi, insulti, stereotipi negativi, stigmatizzazione o minacce una persona o un gruppo e la giustificazione di tutte queste forme o espressioni di odio testé citate, sulla base della 'razza', del colore

21 L'articolo 21 della Carta, al primo capoverso, recita: "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, gli handicap, l'età o le tendenze sessuali".

della pelle, dell'ascendenza, dell'origine nazionale o etnica, dell'età, dell'handicap, della lingua, della religione o delle convinzioni, del sesso, del genere, dell'identità di genere, dell'orientamento sessuale e di altre caratteristiche o stato personale".

Nell'ordinamento italiano, una definizione di discorso d'odio *illegale* è desumibile dall'articolo 604-*bis* del Codice Penale[22]. Tra le condotte sanzionabili, il legislatore prevede il fatto di propagandare "idee fondate sulla superiorità o sull'odio razziale o etnico" e il fatto di istigare a commettere atti di discriminazione, di violenza o di provocazione alla violenza "per motivi razziali, etnici, nazionali o religiosi". Questa definizione può essere utilmente integrata con quanto disposto dal Decreto Legislativo 9 luglio 2003, n. 215 che, all'articolo 2, comma 3, definisce "molestie" quei "comportamenti indesiderati, posti in essere per motivi di razza o di origine etnica, aventi lo scopo o l'effetto di violare la dignità di una persona e di creare un clima intimidatorio, ostile, degradante, umiliante e offensivo".

Con la loro caledoscopica varietà, queste fonti esemplificano bene le difficoltà con cui, nello spazio giuridico europeo, si è provato a individuare le condotte (propaganda, istigazione all'odio, alla discriminazione, alla violenza, diffamazione, ecc.), le "caratteristiche protette" di gruppi e persone ('razza', appartenenza etnica, religione, sesso, orientamento sessuale, disabilità, ecc.) e le modalità specifiche (insulti, stereotipi negativi, minacce, ecc.) che dovrebbero definire l'*hate speech* vietato, tenendo conto anche dei suoi effetti (violazione della dignità personale, ecc.).

Un interrogativo cruciale che sorge, a questo punto, è se sia possibile stabilire in modo chiaro e rigoroso, nonché più complessivo e dinamico, gli elementi costitutivi dei discorsi d'odio, così da offrire agli operatori del diritto (e non solo) una base più solida per contrastare il fenomeno e difenderne le vittime. Ritengo che ciò sia possibile, a condizione di riflettere criticamente su alcuni errori logici che hanno impedito, finora, di raggiungere una definizione condivisa.

Il primo errore consiste nel definire l'*hate speech* a partire da documenti giuridico-politici già esistenti, per altro spesso formulati in chiave criminologica, invece di analizzare il fenomeno in sé come fenomeno di comunicazione, per poi stabilirne gli elementi costitutivi e le soglie di gravità secondo cui configurare eventuali condotte illecite. In questo modo si evita di schiacciare la definizione ricercata sulle manifestazioni più evidenti o estreme, non si perde di vista l'insieme delle possibili espressioni d'odio e si supera la tautologia che definisce i discorsi vietati come quei "discorsi d'odio illeciti".

Il secondo errore consiste nel definire l'*hate speech* sulla base dei comportamenti ritenuti meritevoli di sanzione, messi in atto secondo varie modalità, ai danni di

22 Si tratta di una norma di lontana derivazione internazionale (la Convenzione internazionale sull'eliminazione di ogni forma di discriminazione razziale del 1965, recepita nel nostro ordinamento con la Legge 13 ottobre 1975, n. 654), il cui ambito applicativo resta a oggi limitato ai soli motivi etnico-razziali o religiosi. L'articolo 604-*bis*, insieme all'articolo 604-*ter* contenente la cosiddetta "aggravante della discriminazione razziale", è stato significativamente inserito dal Decreto Legislativo 1 marzo 2018, n. 21 all'interno del Capo III della nuova Sezione I-bis del Codice Penale, dedicata ai "delitti contro l'uguaglianza". Per una ricostruzione evolutiva della fattispecie, prima dell'inserimento nel Codice Penale, si veda Citterio 2006.

gruppi e persone con "caratteristiche protette" identificate non si sa bene secondo quali criteri. La via corretta per arrivare a una definizione è quella di partire dalla causa (o dalle cause) dell'odio e dalle ragioni che indirizzano l'odio verso determinate categorie di persone piuttosto che verso altre, per poi individuare le condotte eventualmente punibili, tenendo conto delle possibili modalità di esecuzione. In questo modo si evita di confondere la causa dei discorsi d'odio con le loro manifestazioni esteriori e con le loro conseguenze.

Per superare questi errori occorre provare a definire i discorsi d'odio come fenomeno di comunicazione sociale, sia offline che online, rimandando a un secondo momento l'applicazione di tale definizione in ambito politico-giuridico.

A questo scopo si può utilizzare, da una parte, l'antropologia filosofica sviluppata da Baruch Spinoza nella sua *Etica* per comprendere la natura e le cause dell'odio come moto fondamentale dell'animo umano[23] e, dall'altra parte, gli "elementi della comunicazione" individuati da Roman Jakobson[24], la teoria degli "atti linguistici" formulata da John Searle[25] e gli "assiomi della comunicazione" elaborati dalla Scuola di Palo Alto[26] per ricostruire, nella sua complessità, la peculiare forma di comunicazione umana costituita dai discorsi d'odio.

Nella sua famosa deduzione geometrica delle passioni, Spinoza assegna all'odio un posto centrale definendolo come "tristezza associata all'idea d'una causa esterna", ove la tristezza è intesa, a sua volta, come il "passaggio da una maggiore a una minore capacità di agire" (*potentia agendi*). A questo moto dell'animo è connaturato lo "sforzo" (*conatus*) di "allontanare e far venir meno" quanto si abbia in odio. Più precisamente, "di ciò che si abbia in odio, ci si sforza d'affermare quanto s'immagina che gli faccia provare tristezza, e di negare, invece, quanto s'immagina che gli faccia provare gioia". Spinoza inoltre, cogliendo la dinamica imitativa della mente umana, afferma che "per il fatto d'immaginare che altri […] abbiano in odio qualcosa […] anche noi l'avremo in odio", ma anche che "ognuno si sforz[a] affinché venga […] da tutti odiato ciò che lui ha in odio".

Le intuizioni dell'*Etica* spinoziana sono confermate da recenti studi di psicologia sociale, che individuano la causa dell'odio nel fatto di attribuire "costanti intenzioni malvagie al bersaglio della propria ostilità, in base a valutazioni astratte di pericolo e a un senso generale d'impotenza nei suoi riguardi"[27]. Da qui lo scopo o l'effetto, connaturati all'odio, di distruggere *psicologicamente* (attraverso l'umiliazione, la svalutazione, la degradazione, il disgusto, la derisione), *socialmente* (attraverso la discriminazione, l'esclusione o la subordinazione) o *fisicamente* (attraverso l'allontanamento, la tortura o l'uccisione) la causa (immaginata) dei propri "sentimenti tristi". Viene, infine, riconosciuto il ruolo essenziale dell'immaginazione nella costruzione dell'odio. Non abbiamo bisogno di conoscere direttamente le persone che odiamo, anzi: odiamo le persone per il solo fatto di immaginarle dota-

23 Spinoza 2009 [1677].
24 Jakobson 1966 [1963].
25 Searle 2009 [1969].
26 Watzlawick, Beavin, Jackson 1971 [1967].
27 Fischer *et al.* 2018.

te di una certa "natura", che deriva loro dal fatto di appartenere a un determinato gruppo dal quale, per definizione, ci attendiamo del male.

Gli "elementi della comunicazione" studiati da Roman Jakobson possono illustrare proficuamente come usiamo il linguaggio per "agire verbalmente" l'odio, mettendo in luce le relazioni intersoggettive alla base di ogni *hate speech.*

Propongo di denominare "triangolo odioso" la relazione comunicativa fondamentale che si stabilisce tra un *emittente* (l'autore del messaggio d'odio), un *ricevente primario* (la persona o il gruppo bersaglio del messaggio d'odio) e un *ricevente secondario* (il pubblico, costituito dagli altri membri del gruppo cui l'autore del messaggio si ascrive, ma anche dagli altri membri del gruppo cui è ascritto il bersaglio, o da altri potenziali gruppi-bersaglio). Il *messaggio* della comunicazione contiene le convinzioni dell'emittente sulla "natura" dei soggetti del "triangolo odioso" e le prescrizioni relative a ciò che, in conseguenza di tale natura, i tre soggetti devono/possono fare o non fare. Il *referente* (ciò cui il messaggio si riferisce) è costituito da una "realtà" che si presume confermi ciò che l'emittente crede vero o ritiene doveroso fare. Il *contesto* è costituito dalle circostanze specifiche in cui si svolge e acquista senso tale comunicazione: per definizione, si tratta di un contesto pubblico. Il *canale* definisce i media (stampa, TV, radio, internet, social networks) attraverso cui si veicola il messaggio. Il *codice*, infine, definisce le forme (lingua, segni, gesti, immagini, video, ecc.) usate per trasmettere il messaggio.

La teoria degli "atti linguistici" di John Searle consente di esplicitare e articolare ulteriormente ciò che l'autore del messaggio d'odio vuole "fare con le parole". Con un uso *assertivo* del linguaggio, si può affermare l'esistenza di un certo stato di cose, ad esempio l'identità della persona o del gruppo individuata come causa del proprio o altrui male, e come giustificazione del proprio odio. Con un uso *espressivo* del linguaggio, si possono condividere i propri sentimenti di ostilità, disprezzo, disgusto, rabbia associati a quello stato di cose, ovvero i propri sentimenti di gioia a fronte di un male che colpisce l'altra persona o l'altro gruppo. Con un uso *commissivo* del linguaggio, ci si può impegnare a compiere determinate azioni (allontanare, escludere, punire, annientare, ecc. la persona o il gruppo odiato) conseguenti con lo stato di cose affermato e con lo stato d'animo espresso. Con un uso *direttivo* del linguaggio, si possono incitare, in primo luogo, i membri del proprio gruppo a imitarci, compiendo determinate azioni (odiare, allontanare, escludere, discriminare, eliminare la persona o il gruppo odiato) e, in secondo luogo, si possono indurre la persona o il gruppo odiato a compiere azioni corrispondenti (odiare sé stesso, autoescludersi, allontanarsi, esercitare atti di autolesionismo, suicidarsi).

Gli "assiomi della comunicazione", messi a punto da Paul Watzlawick e da altri studiosi della Scuola di Palo Alto, consentono infine di studiare ulteriori implicazioni relazionali dei discorsi d'odio. In particolare, il quinto assioma può essere utilizzato per individuare la finalità fondamentale di tali discorsi: ribadire o instaurare una relazione gerarchica di dominio tra l'autore (e il suo gruppo di appartenenza) e il destinatario (e il suo gruppo di appartenenza) del messaggio,

ovvero contestare la pari dignità sociale e la parità di diritti tra l'autore del discorso d'odio e il suo destinatario[28].

Alla luce di queste osservazioni, propongo di definire il discorso d'odio come segue: "ogni forma [*codice*] di comunicazione pubblica [*contesto*] diffusa attraverso qualsiasi mezzo [*canale*], con cui una persona o un gruppo di persone [*emittente*] identificano unilateralmente un altro gruppo, o una persona ritenuta appartenere a quel gruppo [*ricevente primario*], come la causa di un proprio male [*messaggio assertivo*] esclusivamente in base all'identità loro attribuita [*referente*], così da giustificare sentimenti di ostilità [*messaggio espressivo*], comportamenti discriminatori e azioni violente contro di loro [*messaggio commissivo*], da incitare gli altri membri del proprio gruppo a provare analoghi sentimenti, ad assumere analoghi comportamenti e a compiere analoghe azioni [*messaggio direttivo primario*], da indurre nella persona o nel gruppo ricevente sentimenti di disagio, inferiorità, paura, comportamenti di autoesclusione, azioni autolesioniste [*messaggio direttivo secondario*], allo scopo finale di affermare la superiorità propria e del proprio gruppo rispetto all'altra persona o all'altro gruppo [*messaggio assertivo, assioma gerarchico*]".

Questa definizione comporta, a mio avviso, almeno quattro vantaggi per chi voglia comprendere e contrastare i discorsi d'odio ricorrendo anche, ma non solo, a strumenti regolativi.

Un primo vantaggio è quello di relativizzare la questione, che spesso ha occupato i legislatori e le corti, se un discorso d'odio sia o meno idoneo a causare atti discriminatori o violenti, ovvero se configuri una "minaccia credibile" o un "pericolo concreto": la comunicazione odiosa costituisce in sé stessa una discriminazione e una violenza. Eventualmente è il contesto, con la conflittualità sociale in esso presente e il richiamo a un determinato retaggio storico, unito alla dimensione pubblica del messaggio, che può determinare la maggiore o minore probabilità che un discorso d'odio induca qualcuno all'azione.

Un secondo vantaggio è quello di ritenere poco rilevante la ricerca, nell'autore del messaggio ostile, di una chiara intenzione di incitare all'odio o di danneggiare la persona o il gruppo odiato. Si tratta di spostare l'attenzione dall'autore dei discorsi d'odio a coloro che ne sono i destinatari, valutando il danno che tali discorsi producono sull'autostima, sulla salute, sul benessere, sulla libertà delle persone[29]. La definizione fornita, inoltre, non specifica le singole "caratteristiche protette", col rischio di dar luogo a liste incomplete o chiuse di soggetti vulnerabili, ma mette in evidenza il fattore comune a tutti i discorsi d'odio: il fatto di attaccare una persona o un gruppo esclusivamente a causa della loro identità, reale o presunta.

28 Il quinto assioma della comunicazione umana viene così riassunto dagli autori: "tutti gli scambi di comunicazione sono simmetrici o complementari, a seconda che siano basati sull'eguaglianza o sulla differenza" (Watzlawick, Beavin, Jackson 1971: 62). Ai fini della sua applicabilità ai discorsi d'odio, tale assioma va riferito a un caso particolare di comunicazione complementare: quello in cui la "differenza" tra gli interlocutori viene intesa dal soggetto dominante come una differenza di valore, dando così luogo a una comunicazione gerarchica.

29 Adottando una prospettiva giuridico-penale *harm-based*, Marwick, Miller 2014 e McGowan 2019 richiamano l'attenzione sul vissuto personale dei destinatari d'odio.

Un terzo vantaggio è quello di spostare l'attenzione dalle singole parole o espressioni (che possono essere mascherate d'ironia, essere usate all'interno di controdiscorsi volti a criticare l'odio o essere sovvertite e riappropriate dai destinatari), al tipo di relazione che i discorsi d'odio intendono costruire tra le persone e i rispettivi gruppi: una relazione fondata sulla negazione dell'eguaglianza e sull'affermazione del dominio di una parte sull'altra, spinta nei casi estremi fino alla distruzione morale e fisica dell'altro/a[30].

Un quarto vantaggio, infine, è quello di evidenziare la dimensione pubblica in cui i discorsi d'odio acquistano senso, autorizzando a considerare più gravi quelle espressioni che possono raggiungere, per il canale utilizzato o per la maggiore visibilità e autorevolezza dell'autore, un pubblico più vasto, con maggiori possibilità di influenzarne il modo di pensare e di agire[31].

Quest'ultima considerazione consente di riflettere meglio sulle differenze tra l'odio online e quello offline, concludendo che il primo è dotato potenzialmente di una maggiore pericolosità dovuta alle specifiche caratteristiche del canale utilizzato.

L'UNESCO, in un rapporto nel 2015, ha evidenziato alcuni dei caratteri distintivi del discorso d'odio online: il maggiore impatto sociale, determinato dalla rapidità della diffusione, dall'ampiezza della platea raggiungibile, dalla possibilità che diventi "virale" e assuma una dimensione transnazionale; la permanenza nel tempo e il "ritorno imprevedibile", attraverso la condivisioni degli altri utenti o attraverso gli screenshot, che consentono di archiviarlo e diffonderlo successivamente, in privato e in pubblico, anche su piattaforme diverse; la percezione degli autori di essere protetti dall'anonimato[32].

Altre specificità della comunicazione sui social media possono spiegare la peculiare genesi dell'odio online, fornendo indicazioni utili a comprendere e contrastare il fenomeno[33].

In primo luogo, pesa l'effetto disinibente della comunicazione mediata dallo schermo e dalla tastiera, rispetto a quella in presenza. Privata degli elementi non verbali e para-verbali, la comunicazione online porta spesso a esprimere posizioni più radicali di quelle che si assumerebbero dal vivo, con modalità più secche e senza freni di alcuni tipo. Gli autori di discorsi d'odio tendono, inoltre, a non associare conseguenze dirette ai propri atti e percepiscono meno l'impatto negativo dei loro messaggi sulle persone.

30 Adottano, di fatto, questa chiave di lettura i giudici di Cassazione quando ritengono che l'aggravante razzista non rilevi nell'espressione "Italiano di merda", in quanto "l'etnia italiana maggioritaria nel nostro paese non sarebbe idonea a subire una situazione di inferiorità o subire una discriminazione". Si veda Cassazione Penale n. 9381/2006.

31 Per un'interessante ricostruzione del trattamento differenziato che vari personaggi pubblici (consiglieri comunali, leader di partito, parlamentari, scrittori, giornalisti, vignettisti) hanno ricevuto in Italia in sede giudiziaria, rispetto all'accusa di aver prodotto e diffuso discorsi d'odio razzisti, si veda Monti 2015.

32 UNESCO 2015.

33 Sulla capacità dell'ambiente digitale di generare anche anticorpi contro l'*hate speech*, spingendo alcuni utenti a manifestare dissenso verso gli *haters* e solidarietà verso i destinatari dell'odio, richiama opportunamente l'attenzione Bello 2021: 254.

In secondo luogo, nello spazio online si sviluppa un tipo di comunicazione spesso meno formale, che porta ad attenuare le convenzioni sociali: da qui il ricorso, più frequente che nella vita quotidiana, a un linguaggio offensivo, a un tono irrispettoso, a esagerazioni, frasi derisorie, sarcastiche, minacciose, incendiarie.

In terzo luogo, per effetto di una minore esposizione a contenuti di tipo diverso, nelle cosiddette "camere dell'eco" le posizioni tendono a radicalizzarsi e polarizzarsi, perdendo la capacità di interagire in modo nonviolento con posizioni diverse[34].

Queste tendenze, inoltre, possono essere ulteriormente rinforzate dalla peculiare accelerazione delle interazioni sui social media: il minor tempo di reazione comporta un minor tempo di riflessione, con l'effetto di enfatizzare la dimensione emotiva e reattiva della comunicazione.

Tali dinamiche vengono strumentalizzate e ulteriormente potenziate nella comunicazione politica online. Sui social il contegno e la moderazione hanno da tempo lasciato il posto a strategie di *engagement* fondate sulla spettacolarizzazione, in cui sentimenti di ostilità e pratiche di demonizzazione degli oppositori sono utilizzate per galvanizzare i *followers*, fomentando aggressività, rabbia, malcontento e odio. Attraverso stili comunicativi diretti e focalizzati sulle emozioni alcuni politici perseguono due obiettivi: accrescere la propria visibilità online in modo spregiudicato, puntando sulla capacità dei messaggi fortemente provocatori di stimolare l'attivazione degli utenti; ridurre le distanze con la "gente comune", che si ritiene possa identificarsi più facilmente con nozioni semplificate che non con ragionamenti astratti e complessi[35].

Da ultimo, occorre sempre essere consapevoli che lo scopo finale dei social media è di natura economico-commerciale: le piattaforme sono costruite per massimizzare il tempo di permanenza online e sfruttare l'attenzione degli utenti, in modo da esporli più a lungo e più intensamente possibile agli annunci pubblicitari. Come varie ricerche stanno mostrando, i contenuti che suscitano una reazione estrema, come i discorsi d'odio, hanno maggiori probabilità di stimolare l'*engagement* degli utenti e di prolungare la loro permanenza online[36]. In questo modo, si comprendono meglio sia l'iniziale reticenza delle piattaforme a occuparsi di *hate speech*, sia il loro successivo impegno nella "moderazione" dei contenuti[37], ma soprattutto si giustifica la necessità di una regolazione pubblica dell'ambiente digitale, focalizzata sulla tutela dei diritti fondamentali delle persone.

34 Quattrociocchi, Vicini 2016.

35 Per un'efficace e sintetica ricostruzione di queste dinamiche, rimando a Dal Lago 2017 e Barberis 2020. Per un recente studio empirico sulla polarizzazione indotta dai social media, si veda Levy 2021.

36 Munn 2020; Acemoglu *et al.* 2021.

37 In un modello di business basato sulla pubblicità, le piattaforme tendono a rimuovere i contenuti solo se ciò aumenta il tempo che alcuni utenti trascorrono online, incrementando la loro possibile interazione con gli annunci. Sulla base di questo assunto, recenti ricerche spiegano il cambio di strategia delle piattaforme digitali e il loro impegno nella rimozione dei discorsi d'odio. Si veda, a riguardo, Jiménez Durán 2022.

## 3. Il diritto alla protezione dall'odio online: fondamenti, bilanciamenti, obblighi di tutela

Le lotte sociali del dopoguerra in Italia hanno portato, in buona misura con successo, "la Costituzione nelle fabbriche". Con lo *Statuto dei lavoratori* "la fabbrica, in quanto luogo di lavoro, cessa di essere un luogo privato, una semplice proprietà immobiliare del datore di lavoro e diviene un luogo pubblico. E acquista la dimensione pubblica anche il rapporto di lavoro, entro il quale il lavoratore cessa di essere una merce e diviene oggetto di diritti fondamentali"[38].

In un tempo come quello presente, in cui la vita, le istituzioni, il lavoro, la produzione sono "sussunti" e ristrutturati dallo spazio digitale, si è posto un problema analogo: portare "la Costituzione nella rete", ossia costituzionalizzare le azioni e le interazioni che avvengono online, per dare consistenza a una "cittadinanza digitale". Si tratta di una necessità ineludibile, se è vero che la "rivoluzione informatica" non si limita ad aggiungere uno spazio virtuale a uno spazio reale, ma modifica l'idea stessa di realtà, dando vita a una nuova "ontologia" ovvero a nuovi criteri e nuove procedure secondo cui è possibile stabilire cosa sia "reale", cosa sia "esistente", cosa voglia dire "agire". Come ha affermato Maurizio Ferraris, la rete "è reale prima che virtuale, ossia non è una semplice estensione immateriale della realtà sociale, ma si definisce come lo spazio elettivo per la costruzione della realtà sociale"[39].

L'esigenza di *costituzionalizzare la rete* è tanto più pressante quanto più siamo consapevoli del potere che le grandi società informatiche detengono nel plasmare "la realtà" di ciò che sperimentiamo in rete, guidate prevalentemente da logiche di profitto. Tale potere si afferma in forme apparentemente non autoritarie: si esercita con il nostro consenso e attraverso la nostra volontaria collaborazione. Ma, come ogni potere, anche questo richiede di essere "democratizzato". Se non si vogliono liquidare le conquiste dello Stato democratico di diritto, il modello di regolazione della rete non può essere quello del mercato, neanche nella versione liberale del "*marketplace of ideas*"[40], ma piuttosto quello di un "bene comune" orientato a criteri di pubblica utilità e di universalità[41].

In questa prospettiva acquista senso l'idea di una cittadinanza digitale, da intendere come lo statuto personale di cui ogni essere umano gode per il solo fatto di "esistere", agire e interagire in rete. Tale statuto, così come la cittadinanza "predigitale", si sostanzia nel reciproco riconoscimento di una serie di diritti fondamentali, necessari al pieno e libero sviluppo della persona, cui devono logicamente

38 Ferrajoli 2001.

39 Ferraris 2020.

40 Questa espressione, tratta dalla *dissenting opinion* del giudice di Corte suprema Oliver Wendell Holmes Jr. nel caso *Abrams v. United States*, fa riferimento alla convinzione per cui la verità di un'affermazione dipende dall'esito della "concorrenza sul mercato delle idee" e non dall'opinione di un censore, sia esso rappresentato dal legislatore, dal governo o da qualche altra autorità pubblica. Su questa base, in nome del Primo emendamento della Costituzione, nel contesto statunitense si assegna un forte peso alla libertà di espressione rispetto alla difesa di altri diritti e interessi concorrenti.

41 Rodotà 2014.

corrispondere altrettanti obblighi di non lesione e di prestazione da parte di chi detiene, a vario titolo, il potere nello spazio digitale. Come ribadito nella *Dichiarazione dei diritti in internet*, approvata in Italia nel 2015, "la garanzia di questi diritti è condizione necessaria affinché sia assicurato il funzionamento democratico delle istituzioni e si eviti il prevalere di poteri pubblici e privati, che possano portare a una società della sorveglianza, del controllo e della selezione sociale"[42].

Il diritto a essere protetti dai discorsi d'odio online, se esiste, deve trovare il suo fondamento nella cittadinanza digitale. Si tratta di stabilirne l'esistenza, e di precisarne contenuto e peso, rispondendo alle seguenti domande: quali sono i beni giuridici che tale diritto riconosce come meritevoli di protezione? La rilevanza di tali beni, rispetto all'esercizio della cittadinanza digitale, è tale da giustificare l'intervento del diritto penale in caso di violazione? Con quali altri diritti può entrare in conflitto la tutela dai discorsi d'odio online? Secondo quali criteri è possibile operare un equo bilanciamento degli interessi in gioco?

La cittadinanza digitale, così come quella pre-digitale, si fonda su un principio di pari dignità sociale: se vogliono convivere con le proprie diversità e regolare democraticamente le proprie relazioni utilizzando il diritto, i membri di una collettività anche online non possono fare a meno di riconoscersi uguali in termini di dignità sociale e, conseguentemente, in termini di diritti fondamentali. Jeremy Waldron ha mostrato in modo convincente che l'esposizione ai discorsi d'odio, dentro e fuori la rete, mina esattamente il "bene pubblico" della pari dignità sociale, senza il quale viene meno la possibilità stessa di una convivenza civile e democratica[43].

La dignità, afferma il filosofo del diritto statunitense, non si esaurisce in un'astratta "aura kantiana" ma è qualcosa di molto concreto. È la base di quel reciproco riconoscimento che consente a tutti/e e a ciascuno/a di sentirsi parte a pieno titolo della collettività: è ciò che garantisce l'aspettativa di essere trattati alla pari degli altri, in qualsiasi circostanza e in ogni ambito della società. I discorsi d'odio hanno, invece, lo scopo di negare la pari dignità sociale: compromettono la reputazione e l'autostima delle persone, comunicando a loro e agli altri che non meritano di essere trattate equamente a causa della loro identità – razziale, etnica, nazionale, sessuale, ecc. – che le rende in qualche modo inferiori. Non si tratta di una semplice offesa o di una critica violenta a un determinato comportamento, ma di un attacco alla persona in quanto tale e al gruppo cui la si ascrive: un attacco al diritto di essere la persona che si è, e di condurre il tipo di vita che si ha liberamente deciso di condurre.

Come tutti i diritti, anche la tutela dai discorsi d'odio va bilanciata, in questo caso con i diritti di accesso alla rete, alla libertà di informazione, di espressione e di associazione, nonché con la protezione riconosciuta all'anonimato e con la libertà economica delle società tecnologiche. In linea di principio, la tutela del bene pubblico costituito dalla "pari dignità sociale", così come definito sopra, può godere

42 La *Dichiarazione*, giuridicamente non vincolante ma politicamente significativa, è stata elaborata dalla Commissione per i diritti e i doveri relativi ad Internet, istituita presso la Camera dei Deputati e presieduta da Stefano Rodotà. Il documento è l'esito di un percorso di consultazione pubblica e di audizioni, che si sono concluse il 14 luglio 2015 con la sua approvazione finale.

43 Waldron 2012. Si veda anche Ansuátegui Roig 2017.

di un primato rispetto ai beni coperti dagli altri diritti, facendoli retrocedere e ammettendo una loro compressione. Il rango elevato del bene protetto autorizza a ricorrere in caso di violazioni allo strumento del diritto penale, l'uso del quale attiva a sua volta le garanzie tipiche del "giusto processo".

Il diritto di accesso e il diritto di associazione entrano in gioco nella misura in cui la "sanzione" per la diffusione di discorsi d'odio potrebbe arrivare, nelle forme più severe, alla disattivazione temporanea o permanente del profilo della persona o di una pagina collettiva sul social network (senza contare i possibili procedimenti penali, nei casi più gravi). Data la rilevanza di questi diritti, è necessaria una regolamentazione "garantista" che protegga dall'espulsione dai social. Tale protezione implica l'obbligo di decidere la sanzione rispetto a norme di condotta pubbliche e chiare, attraverso procedure e decisioni trasparenti, tempestivamente comunicate, adeguatamente motivate e suscettibili di ricorso, secondo una graduazione della "pena" proporzionata alla violazione compiuta e al danno arrecato.

Il diritto alla libertà di informazione e il diritto alla libertà di espressione entrano in gioco nella misura in cui la "sanzione" per la diffusione di discorsi d'odio può consistere nella rimozione di singoli contenuti online. Anche in questo caso si tratta di diritti rilevanti, seppure la portata della sanzione sia minore rispetto alla disattivazione del profilo o della pagina (sempre senza contare possibili procedimenti penali). Serve, comunque, anche per questi casi, una regolamentazione garantista analoga a quella precedente.

La protezione dell'anonimato, in base alla quale ogni persona accede alla rete e comunica elettronicamente usando dispositivi che proteggono la sua identità ed evitano la raccolta di dati personali, può essere limitata nel rispetto di principi di necessità e proporzionalità per tutelare rilevanti interessi pubblici. Come ribadito anche nella *Dichiarazione dei diritti su Internet*, già ricordata, nei casi di violazione della dignità e dei diritti fondamentali, l'autorità giudiziaria, con provvedimento motivato, può senz'altro disporre l'identificazione dell'autore della comunicazione segnalata.

Il problema della censura e, in generale, dell'ingiusta limitazione dei diritti di cui sopra, si pone soprattutto per tutelare chi esprime ostilità nei confronti del potere, muovendo da posizioni di minoranza, vulnerabilità o dissenso politico: sarebbe quanto meno paradossale limitare il diritto alla presa di parola critica proprio di quei soggetti che, generalmente, sono silenziati e subiscono odio sociale, discriminazioni istituzionali e attacchi violenti[44].

## 4. Dalla privatizzazione della censura alla co-responsabilità nella tutela della pari dignità

Una volta stabiliti *in astratto* diritti e obblighi connessi alla tutela dall'odio online, e fissati i principi generali per un equo bilanciamento con altri diritti potenzialmente

44 Per una discussione di questi rischi, soprattutto in riferimento al contesto statunitense, si veda Keats Citron 2018. Su questi aspetti, si veda anche Pintore 2021.

in conflitto, resta da chiarire l'aspetto forse più complesso e delicato della questione: a chi incombono *in concreto* gli obblighi previsti a fronte dei diritti riconosciuti?

La complessità e la delicatezza della questione derivano dal fatto che lo spazio digitale è popolato da una molteplicità di attori, dotati di poteri diversi e portatori di differenti interessi.

Nel caso dei social media gli attori sono di almeno sei tipi: 1) le piattaforme digitali, 2) gli utenti, singoli o collettivi, che accedono all'infrastruttura comunicativa fornita dalle piattaforme, 3) gli inserzionisti che acquistano spazi pubblicitari dalle piattaforme, 4) gli ordinamenti statali, con i rispettivi principi costituzionali e le eventuali norme di settore, 5) le corti competenti e le autorità amministrative indipendenti, a cui gli utenti possono accedere al fine di veder garantiti i propri diritti lesi, alla luce dei principi costituzionali e delle eventuali norme di settore, 6) le organizzazioni della società civile impegnate nel contrasto e nella prevenzione dei discorsi d'odio, nonché nella tutela delle vittime.

Secondo alcuni studiosi, nella definizione e nell'attuazione delle disposizioni contro l'odio online esiste oggi uno squilibrio di potere a vantaggio delle piattaforme, che rischia di avere come effetto una "privatizzazione sostanziale della censura"[45]. Ci si riferisce, con legittima preoccupazione, al fatto che negli ultimi anni le autorità pubbliche abbiano di fatto delegato ai social networks il potere di decidere quali contenuti rimuovere sulla base degli "standard di comunità" fissati dalle piattaforme stesse e accettati dagli utenti al momento dell'iscrizione. Si tratta di riflettere su come sia possibile invertire questa tendenza, riequilibrando i rapporti tra i diversi attori a vantaggio dei soggetti portatori di un interesse pubblico, allo scopo di garantire un più equo e trasparente bilanciamento tra il rispetto della pari dignità sociale, da una parte, e le libertà di informazione, espressione e associazione, dall'altra.

Per risolvere il problema, propongo di applicare ai social media e alla loro regolazione un modello di "sovranità condivisa". Secondo tale modello i diversi attori coinvolti, in primis le piattaforme digitali e le pubbliche autorità, si riconoscono come co-responsabili nella garanzia dei diritti fondamentali propri della cittadinanza digitale e nell'adempimento degli obblighi connessi. Si tratta di una posizione che tiene conto della realtà del web e delle sue dinamiche concrete, ma anche delle esigenze normative proprie di un quadro costituzionale democratico: da una parte, occorre ammettere che la mole di dati circolanti quotidianamente sui social media eccede le capacità di controllo (ma anche l'interesse) degli attori pubblici, siano essi autorità di garanzia o di pubblica sicurezza; dall'altra parte, occorre che le regole di funzionamento dei vari social media e, soprattutto, le procedure di monitoraggio e censura dei contenuti che diffondono odio, siano uniformi e "garantiste", ovvero rispondano ai principi di uno Stato democratico di diritto quanto alle procedure e al bilanciamento dei diritti fondamentali in gioco, operando in una prospettiva di "diritto penale minimo"[46].

45 Monti 2019a.

46 Per un'esposizione analitica e completa del concetto di "diritto penale minimo", rimando a Ferrajoli 2014.

Secondo questo modello, la co-responsabilità nella protezione dall'odio online e nel rispetto delle libertà di informazione, espressione ed associazione andrebbe declinata su due livelli.

Il primo livello prevede l'esercizio condiviso di quattro "poteri normativi primari" orientati rispettivamente a: 1) definire criteri per identificare con sufficiente precisione i discorsi d'odio e i loro diversi gradi di gravità, in base alle considerazioni definitorie svolte sopra; 2) associare ai diversi gradi di gravità dei discorsi non ammessi una adeguata gradazione di "sanzioni", dalla rimozione del contenuto vietato alla sospensione dell'account o della pagina, fino alla loro cancellazione; 3) stabilire procedure trasparenti ed eque di "moderazione" secondo cui valutare i discorsi alla luce dei criteri stabiliti e decidere le "sanzioni" adeguate, precisando ruolo e limiti dei meccanismi automatizzati e preventivi di moderazione; 4) fornire agli utenti procedure accessibili ed eque di "ricorso" contro decisioni sanzionatorie ritenute errate, ingiuste o eccessive.

Il secondo livello prevede l'individuazione degli attori cui affidare l'effettiva applicazione dei criteri e delle procedure stabilite al primo livello, sulla base di un principio di economicità e di sussidiarietà. In base al principio di economicità, a parità di efficacia, è chiamato a intervenire quell'attore che può svolgere il proprio compito col minor dispendio di mezzi e di tempo possibile: in particolare, la rapidità dell'intervento è richiesta dalla stessa logica di funzionamento del web allo scopo, ad esempio, di prevenire che un contenuto vietato diventi virale. In base al principio di sussidiarietà, a parità di efficacia nella garanzia e nel bilanciamento dei diritti fondamentali, è chiamato a intervenire l'attore più prossimo ed accessibile agli utenti: nei casi in cui il rischio di una violazione della pari dignità sociale o di una compressione delle libertà di informazione, espressione ed associazione si faccia più marcato, ovvero il bilanciamento degli interessi in gioco sia particolarmente complesso, l'intervento va affidato all'attore che offre maggiori garanzie procedurali.

In concreto, questo modello a due livelli può configurarsi come segue. Al primo livello tutti gli attori interessati sono chiamati a elaborare un consenso sui principi, le regole e le procedure generali per la tutela dai discorsi d'odio. Gli "standard di comunità" e le "sanzioni" per le relative violazioni, le modalità di moderazione, segnalazione e rimozione dei contenuti vietati, le modalità di comunicazione e di ricorso, che attualmente ciascuna piattaforma digitale adotta e attua in regime di autoregolazione, dovrebbero essere verificate e armonizzate alla luce dei principi costituzionali e delle più significative pronunce giudiziarie dei vari paesi. Al secondo livello le piattaforme digitali sono "delegate" a intervenire sui discorsi d'odio più evidenti, meno complessi e con minori limitazioni dei diritti antagonisti: sono, infatti, gli attori che dispongono degli strumenti tecnici per intervenire più rapidamente e, nei casi più semplici, possono offrire agli utenti sufficienti garanzie di tutela. Gli attori pubblici, invece, intervengono nei casi meno evidenti, più complessi e dalle conseguenze civili e penali più rilevanti. Nello specifico: le autorità amministrative indipendenti, responsabili in materia di comunicazione e di media, potrebbero intervenire per stabilire se è opportuno o meno che il caso venga esaminato da un giudice; la magistratura ordinaria

interverrebbe nel merito del caso, su segnalazione delle autorità indipendenti o su iniziativa degli utenti, singoli o collettivi[47].

## 5. Conclusioni e prospettive

Ritengo che la definizione di *hate speech* e il modello di co-responsabilità regolatoria qui proposti abbiano migliori prospettive di successo rispetto ad altri approcci, sia sul piano dell'efficacia nella tutela dall'odio online che sul piano dell'equo bilanciamento dei diritti e degli interessi coinvolti[48].

Le violenze subite in Myanmar dai Rohingya, stimolate e amplificate dai discorsi d'odio che i responsabili di Facebook sono accusati di non aver rimosso, sollevano tuttavia un'obiezione rilevante: come operare in quei contesti in cui non vige uno Stato democratico di diritto di tipo occidentale, in cui principi costituzionali garantisti, seppur presenti, non vengono né attuati né rispettati, o in cui le autorità governative intervengono in modo autoritario sui social media e sul web?

In questi casi, le piattaforme digitali proprietarie dei social media dovrebbero condividere le proprie responsabilità regolatorie con i rappresentanti di qualificate e indipendenti organizzazioni per i diritti umani, avvalendosi del supporto di organismi sovranazionali come le organizzazioni internazionali regionali[49] o le stesse Nazioni Unite.

In ogni caso, occorrerà guardarsi da facili ottimismi, scambiando la riduzione del numero di espressioni d'odio online con il raggiungimento di una situazione

47 Alcune recenti vicende giudiziarie, seguite alla decisione di Facebook di disattivare pagine e account di movimenti ed esponenti di estrema destra in Italia, costituiscono un buon esempio di "garantismo": la libertà di espressione è stata presa tanto sul serio da considerarne la possibile violazione anche a danno di soggetti neofascisti. Per una sintetica ricostruzione comparata di queste vicende e dei loro diversi esiti, rimando ancora a Villaschi 2021.

48 Il *Codice di condotta* stipulato tra la Commissione Europea e le principali società informatiche costituisce un caso esemplare di auto-regolamentazione da parte dei soggetti privati, più che di co-regolamentazione. Esso contiene una serie di impegni che le piattaforme assumono, su base volontaria, per contrastare e prevenire l'odio online. Con l'adesione al *Codice*, le società si impegnano a esaminare entro 24 ore la maggior parte delle richieste di rimozione dei discorsi d'odio per come da loro definiti; si impegnano, inoltre, a pubblicare periodicamente i risultati della propria azione di controllo e a collaborare con le organizzazioni della società civile impegnate nel contrasto dell'odio online. Insieme alla Commissione Europea, le piattaforme si impegnano anche a proseguire "l'elaborazione e la promozione di narrazioni alternative indipendenti e di sostegno a programmi educativi che incoraggino il pensiero critico" nell'uso della rete. La "coregolamentazione" del settore è, invece, tra le finalità della Direttiva UE 2018/1808 sulla fornitura di servizi di media audiovisivi.

49 Mi riferisco all'Unione Africana, all'Associazione delle Nazioni del Sud-est asiatico, al Dialogo per la cooperazione asiatica, all'Organizzazione degli Stati americani, all'Unione delle nazioni sudamericane che, proprio sul terreno del contrasto e della prevenzione dei discorsi d'odio e, in generale, del rispetto dei diritti fondamentali nell'ambiente digitale, potrebbero approfondire la propria cooperazione. Sul ruolo delle organizzazioni internazionali regionali nel mantenimento della pace e della sicurezza globali, faccio riferimento alla riflessione giuridico-filosofica di Habermas 2005 [2004]: 107ss.

ottimale. Non è sufficiente pensare di intervenire sul complesso fenomeno sociale dell'odio, dentro e fuori la rete, con i soli strumenti della deterrenza e della sanzione, sia pure temperata da adeguate garanzie procedurali: in una società pienamente democratica ci si dovrebbe astenere dai discorsi d'odio non soltanto per paura delle conseguenze, ma per convinta adesione a un modello nonviolento di comunicazione e relazione interpersonale.

È, dunque, necessario agire in termini di prevenzione sui meccanismi strutturali che, nelle nostre società e nell'attuale sistema economico-politico, alimentano l'odio verso alcune categorie di persone considerate prive della medesima dignità delle altre. Si tratta di sviluppare, in generale, strategie alternative alla mera repressione, come l'elaborazione e la promozione di narrazioni alternative, di metodologie contro-discorsive, di campagne pubbliche di sensibilizzazione, di programmi educativi e laboratori scolastici che incoraggino il pensiero autonomo e critico, il ricorso a forme nonviolente di comunicazione e risoluzione dei conflitti, la presa di consapevolezza sulle dinamiche della rete e dei social media[50].

Come ha affermato Alessandro Baratta, "sappiamo che sostituire il diritto penale con qualcosa di meglio potrà avvenire solo quando avremo sostituito la nostra società con una società *migliore*"[51]. Migliore – aggiungo – sotto il profilo della giustizia sociale, economica e ambientale. Ciò valeva prima dell'avvento della rete, vale ancora oggi, e varrà sempre di più nella società iper-digitalizzata del futuro, alle prese con le sfide della crisi climatica, delle migrazioni globali, di possibili altre pandemie, dei conflitti armati e del rischio nucleare.

In conclusione, nessuna politica contro l'odio online sarà sostenibile, nel medio-lungo periodo, in assenza di misure che incidano sulle cause strutturali dell'aggressività e della violenza sociale. In assenza di una reale transizione ecologica accompagnata da adeguate politiche sociali e dell'occupazione, di politiche migratorie non discriminatorie nei confronti delle popolazioni del Sud del mondo, di politiche sanitarie pubbliche capaci di garantire un equo accesso alle cure e ai vaccini su scala globale, di una politica di disarmo generalizzato, sarà sempre più alto il rischio che ansie e frustrazioni sociali vengano indirizzate contro vecchi e nuovi soggetti "vulnerabili". E che dalle parole d'odio online si passi a parole e pratiche d'odio offline, o che l'odio offline trovi nelle bolle digitali micidiali casse di risonanza.

## Bibliografia

Acemoglu D., Ozdaglar A., Siderius J., 2021, *Misinformation. Strategic sharing, homophily, and endogenous echo chambers*, Technical report, National Bureau of Economic Research.

50 Sulla *critical digital literacy* in generale, si veda almeno Pangrazio 2016. Per l'applicazione del concetto al contrasto e alla prevenzione dell'*hate speech* nel caso specifico dell'antiziganismo, rimando a Agapoglou *et al.* 2021.

51 Baratta 2019 [1982].

Agapoglou Th., N. Mouratoglou, K. Tsioumis, K. Bikos 2021, "Combating Online Hate Speech through Critical Digital Literacy: Reflections from an Emancipatory Action Research with Roma Youths", *International Journal of Learning and Development*, 11 (2): 105-120.

Ansuategui Roig F.J. 2017, "Liberta di espressione, discorsi d'odio, soggetti vulnerabili: paradigmi e nuove frontiere", *Ars interpretandi*, 1: 29-48.

Baratta A., 2019 [1982], *Criminologia critica e critica del diritto penale. Introduzione alla sociologia giuridico-penale*, Milano: Meltemi.

Barberis M., 2020, *Come Internet sta uccidendo la democrazia. Populismo digitale*, Milano: Chiarelettere.

Bello B.G., 2021, "I discorsi d'odio in rete", in Th. Casadei, S. Pietropaoli (a cura di), *Diritto e tecnologie informatiche. Questioni di informatica giuridica, prospettive istituzionali e sfide sociali*, Milano: Wolters Kluwer, 247-261.

Citterio C., 2006, "Discriminazione razziale: figure di reato e oscillazioni del rigore punitivo nel tempo", in S. Riondato (a cura di), *Discriminazione razziale, xenofobia, odio religioso. Diritti fondamentali e tutela penale*, Padova: Cedam.

Dal Lago A., 2017, *Populismo digitale. La crisi, la rete e la nuova destra*, Milano: Raffaello Cortina.

D'Amico M., C. Siccardi (a cura di) 2021, *La costituzione non odia. Conoscere, prevenire e contrastare l'*hate speech on line, Torino: Giappichelli.

Di Rosa A. 2020, Hate speech *e discriminazione. Un'analisi performativa tra diritti umani e teorie della libertà*, Modena: Mucchi.

Earl J., Kimport K., 2011, *Digitally Enabled Social Change: Activism in the Internet Age*, Boston: MIT Press.

Faloppa F., 2020, *#Odio. Manuale di resistenza alla violenza delle parole*, Torino: Utet.

Ferrajoli L., 2001, "Lo Statuto dei lavoratori: un mutamento di paradigma in senso pubblicistico del rapporto di lavoro", *Quaderni rassegna sindacale*, 2: 117-123.

Ferrajoli L., 2016, *Il paradigma garantista. Filosofia e critica del diritto penale*, Napoli: Editoriale Scientifica (seconda edizione ampliata).

Ferraris M., 2020, "Metafisica del Web", lezione tenuta presso il Centro Nexa, Politecnico di Torino, January 8. Available at: https://nexa.polito.it/mercoledi-126 (accessed: December 30, 2021).

Fischer A., E. Halperin, D. Canetti, A. Jasini 2018, "Why We Hate", *Emotion Review*, 10 (4): 309-320.

Formenti C., 2000, *Incantati dalla rete. Immaginari, utopie e conflitti nell'epoca di Internet*, Milano: Raffaello Cortina.

Formenti C., 2009, *Cybersoviet. Utopie postdemocratiche e nuovi media*, Milano: Raffaello Cortina.

Gray C., Hansen K,. 2021, "Did Covid-19 Lead to an Increase in Hate Crimes Toward Chinese People in London?", *Journal of Contemporary Criminal Justice*, 37 (4): 569-588.

Habermas J., 2005 [2004], *L'Occidente diviso*, Roma-Bari: Laterza.

Hindman M., 2018, *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy*, Princeton: Princeton University Press.

Jahan S., Oussalah M., 2021, "A systematic review of Hate Speech automatic detection using Natural Language Processing", *arXiv*, 2106.00742v1 [cs.CL], May 22.

Jakobson R., 1966 [1963], *Saggi di linguistica generale*, Milano: Feltrinelli.

Jaishankar K., (ed.) 2011, *Cyber Criminology: Exploring Internet Crimes and Criminal behavior*, Boca Raton FL: CRC Press, Taylor and Francis Group.

Jiménez Durán R., 2022, "The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter", March 11. Available at SSRN: https://ssrn.com/abstract=4044098 (accessed: April 15, 2022).

Kaye D., 2019, *Speech Police: The Global Struggle to Govern the Internet*, New York: Columbia Global Reports.

Keats Citron D., 2018, "Extremist Speech, Compelled Conformity, and Censorship Creep", *Notre Dame Law Review*, 93 (3): 1035-1071.

Kiska R., 2012, "Hate speech: a comparison between the European Court of Human Rights and the United States Supreme Court jurisprudence", *Regent University Law Review*, 25: 107-151.

Kuss D..J, H.M., Pontes 2019, *Internet Addiction*, Boston: Hogrefe Publishing.

Lavenia G. 2012, *Internet e le sue dipendenze. Dal coinvolgimento alla psicopatologia*, Milano: FrancoAngeli.

Levy R., 2021, "Social media, news consumption, and polarization: Evidence from a field experiment", *American Economic Review*, 111 (3): 831-870.

MacAvaney S., H-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder 2019, "Hate speech detection: Challenges and solutions", *PLoS ONE*, 14 (8).

Marwick A.E., R.W. Miller 2014, "Online Harassment, Defamation, and Hateful Speech: A Primer of the Legal Landscape", *Fordham Center on Law and Information Policy Report*, June 10.

McGowan M.C., 2019, *Just Words: On Speech and Hidden Harm*, Oxford: Oxford Scholarship Online.

Mchangama J., 2015, "The problem with hate speech laws", *The Review of Faith & International Affairs*, 13 (1): 75-82.

Mintz A., (ed.) 2012, *Web of Deceit: Misinformation and Manipulation in the Age of Social Media*, Medford: Information Today.

Monti M., 2019a, "Privatizzazione della censura e Internet platforms: la libertà di espressione e i nuovi censori dell'agorà digitale", *Rivista italiana di informatica e diritto*, 1: 35-51.

Monti M., 2019b, "Le *Internet platforms*, il discorso pubblico e la democrazia", *Quaderni costituzionali*, 4: 811-840.

Müller K., C. Schwarz 2021, "Fanning the Flames of Hate: Social Media and Hate Crime", *Journal of the European Economic Association*, 19 (4): 2131-2167.

Munn L., 2020, "Angry by design: toxic communication and technical architectures", *Humanities and Social Sciences Communications*, 7 (57): 1-11.

Noble S.U., 2018, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York: New York University Press.

Pintore A., 2021, *Tra parole d'odio e odio per le parole*, Modena: Mucchi.

Quattrociocchi W., A. Vicini 2016, *Misinformation. Guida alla società dell'informazione e della credulità*, Milano: FrancoAngeli.

Pangrazio L., 2016, "Reconceptualising critical digital literacy", *Discourse: Studies in the Cultural Politics of Education*, 37 (2): 163-174.

Rodotà S., 2014, *Il mondo nella rete. Quali i diritti, quali i vincoli*, Roma-Bari: Laterza.

Röttger P., B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, J.B. Pierrehumbert 2021, "HateCheck: Functional Tests for Hate Speech Detection Models", *arXiv*, 2012.15606 [cs.CL] May 27.

Searle J., 2009 [1969], *Atti linguistici. Saggi di filosofia del linguaggio*, Torino: Bollati Boringhieri.

Sellars A., 2016, *Defining Hate Speech*, Boston University School of Law, Public Law Research Paper, 20.

Spinoza B., 2009 [1677], *Etica*, Roma-Bari: Laterza.
Sunstein C.R., 2017, *#Republic: Divided Democracy in the Age of Social Media*, Princeton NJ: Princeton University Press.
UNESCO 2015, *Countering online hate speech*, UNESCO, Parigi.
van Dijk J.A., K.L. Hacker 2018, *Internet and Democracy in the Network Society*, New York: Routledge.
Villaschi P., 2021, "La (non) regolamentazione dei social network e del web", in M. D'Amico, C. Siccardi (a cura di), *La costituzione non odia. Conoscere, prevenire e contrastare l'*hate speech on line, Torino: Giappichelli, 113-126.
Waldron J., 2012, *The Harm in Hate Speech*, Cambridge MA: Harvard University Press.
Watzlawick D.D., P. Beavin, J.H. Jackson 1971 [1967], *Pragmatica della comunicazione umana. Studio dei modelli interattivi, delle patologie e dei paradossi*, Roma: Astrolabio.
Ziccardi G., 2016, *L'odio online. Violenza verbale e ossessione in rete*, Milano: Raffaello Cortina.
Ziccardi G., 2020, *Online Political Hate Speech in Europe: The Rise of New Extremisms*, Cheltenham: Edward Elgar Publishing.
Zuboff S., 2020, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, New York: PublicAffairs.