XVIII Legislatura

Senato della Repubblica

Commissione straordinaria per il contrasto dei fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza

ANALISI COMPARATIVA SUL FENOMENO DELL'ISTIGAZIONE ALL'ODIO ONLINE

luglio 2022

La lettura dell'Analisi comparativa sul fenomeno dell'istigazione all'odio online conferma la validità del lavoro svolto dalla Commissione straordinaria per il contrasto dei fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza nel suo intenso anno di attività, affidando a tutti noi riflessioni importanti e il compito di affrontare con determinazione un fenomeno così complesso, impegno che inevitabilmente dovrà essere trattato nella XIX Legislatura.

Voglio ringraziare, a nome della presidente Segre e mio personale, tutto l'Ufficio di Presidenza per il lavoro importante fatto insieme.

Lasciamo agli atti di questa legislatura, al dibattito pubblico e alle attività del prossimo Parlamento un lavoro di grande rilievo.

Il fatto che, a differenza della istituzione della Commissione, la relazione conclusiva sull'indagine conoscitiva sia stata votata all'unanimità, rimarca l'importanza e il successo politico di questa Commissione.

Sen. Francesco Verducci

Analisi comparativa sul fenomeno dell'istigazione all'odio online

Ricerca per la Commissione antidiscriminazioni del Senato della Repubblica

OFFICIAL DRAFT

Re-Imagine Europa

Hanno collaborato:
Luca De Biase
Erika Widegren
Manlio De Domenico
Pierluigi Sacco
Grigoris Neofytou
Gabriele Balbi
Matteo Nicolosi

INDICE

INDICE 5
INTRODUZIONE 6
Evidenze principali 6
Interpretazioni 7
Prospettive 8
Sezione I: DEFINIRE L'INCITAMENTO ALL'ODIO 9
Introduzione 9
Cosa s'intende per "Incitamento all'odio"? 9
Difficoltà e preoccupazioni 12
Conclusioni 14
Sezione II: QUANTIFICARE L'INCITAMENTO ALL'ODIO ONLINE 15
Sezione III: EVIDENZE 17
Mappatura dell'ecosistema dell'odio online 17
Indicatori fondamentali dell'odio online 17
Collegare comportamenti non autentici, disinformazione ed odio durante la pandemia di COVID-19 18
Metodi avanzati basati su analisi di rete e grandi basi di dati 21
Metodi basati sull'analisi di dati classica e piccole basi di dati 21
Osservazioni conclusive e raccomandazione politica 24
Sezione IV: AZIONI NORMATIVE A CONFRONTO 26
Introduzione 26
La situazione 27
Raccomandazioni e insegnamenti appresi 31
Raccomandazioni per amministratori e politici 32
CONCLUSIONI 36
BIBLIOGRAFIA 37

INTRODUZIONE

Questo rapporto ha lo scopo di valutare la rilevanza dell'istigazione all'odio online. Si tratta di un compito complesso per la natura dell'argomento che non si presta a essere definito univocamente ma che si rivela assolutamente necessario per il numero di persone in qualche modo coinvolte e per la pericolosità delle sue conseguenze sulla qualità della convivenza civile e sulla stessa tenuta della democrazia.

Una sezione preliminare del lavoro è dedicata alle definizioni alternative dei "discorsi di odio" che, in quest'epoca profondamente segnata dalle comunicazioni sui social network, converge spesso con la questione dell'istigazione all'odio online. Esistono molte definizioni anche perché esistono molte motivazioni per enunciarle: alcune sono dedicate all'indagine approfondita sul piano giuridico o sociologico, altre sono operazionali e servono a guidare l'azione di chi sulle piattaforme digitali compie scelte pragmatiche ogni giorno per eliminare i messaggi d'odio dalla rete. Le definizioni pragmatiche sono spesso usate anche dagli scienziati che analizzano grandi quantità di dati per arrivare a osservazioni originali e per scoprire dinamiche di valore generale sul fenomeno dell'istigazione all'odio online.

Una grande parte parte del lavoro è dedicata alla raccolta ragionata di ciò che si conosce del fenomeno dal punto di vista quantitativo. Le difficoltà metodologiche causate dalla pluralità di definizioni in materia non frenano la crescita delle ricerche scientifiche. Le ricerche sono molto numerose a causa dell'evidente rilevanza strategica dell'argomento e dall'abbondanza di dati che si possono raccogliere sui social network ed elaborate con le tecniche della data science. Pur nella varietà delle definizioni del fenomeno dell'istigazione all'odio online, si dimostra una rilevanza crescente del fenomeno.

Naturalmente una conoscenza molto più profonda si potrebbe ottenere se ci fosse la piena collaborazione delle principali piattaforme dei media sociali che dispongono dei dati relativi al comportamento degli utenti. Queste piattaforme, per le loro politiche interne, cancellano - automanticamente e non - un'enorme quantità di messaggi che secondo le loro definizioni contengono messaggi di odio e violenza. Sicché i ricercatori che studiano dall'esterno il fenomeno si devono accontentare di osservare i messaggi che non sono stati eliminati dalle piattaforme.

L'ultima parte del rapporto è dedicata a un'analisi comparativa delle esperienze di policy relative al fenomeno dell'istigazione all'odio online nei diversi paesi europei. Proprio per l'importanza dei media sociali digitali, una policy di grande respiro non può che darsi una prospettiva almeno europea. E la Commissione europea sta in effetti lavorando sull'argomento. Questo però non toglie che la profondità dell'azione necessaria per affrontare correttamente ed efficacemente un fenomeno così complesso e rilevante ha bisogno della partecipazione informata dei cittadini di tutti i paesi e delle istituzioni che li rappresentano, oltre che della sensibilità di tutte le diverse culture che compongono la comunità degli europei.

Evidenze principali

L'esistenza dell'istigazione all'odio online è percepita dalla maggioranza della popolazione che usa i social network e questo è un motivo per astenersi dal dibattito per quasi la metà della popolazione. Una ricerca Eurobarometer ha mostrato che il 75% delle persone impegnate nei social network in Europa ha visto o subito abusi, minacce o odio online (European Commission 2016a)[1]. Quindi, quando si pensa alla libertà di espressione delle persone che odiano occorre pensare anche alla libertà di espressione delle persone che temono di essere colpite dalla violenza degli odiatori.

[1] Eurobarometer: «75% of respondents who are engaging in social media discussions have seen or experienced online abuse, hate speech or threats; and almost half of them (48%) say that they are discouraged from engaging in debates as a result»

In Italia, secondo una ricerca SWG, l'86% della popolazione che usa i social network afferma che nei social network sono molto o abbastanza diffusi contenuti che rimandano a posizioni volgari, offensive, violente o che incitano all'odio. E il 32% dichiara di entrare in contatto diretto con messaggi violenti quando è online. Le vittime dell'odio online secondo gli intervistati sono i migranti, i gay, le donne e i politici per circa 30% degli intervistati, mentre con percentuali inferiori sono citati i musulmani, altre minoranze, i bambini, i disabili e le persone famose. Per l'88% degli utenti dei social network è una pratica grave e socialmente pericolosa. La generazione più critica nei confronti di questo fenomeno è quella dei millennials con il 91% degli intervistati che si dichiara profondamente irritato dalla violenza che si è diffusa online (SWG 2017).

I poli nei quali si concentrano le maggiori percentuali di messaggi di istigazione all'odio online sul totale dei messaggi pubblicati sono le grandi città. Una ricerca ha verificato questo fenomeno su Twitter, trovando che a Roma, Milano, Torino, Napoli, la percentuali di messaggi di istigazione all'odio online sul totale supera il 18,5%. Le autrici della ricerca mostrano una correlazione forte tra ineguaglianza economica e quantità di discorsi d'odio (Denti & Faggian 2021).

Una ricerca di Twitter ha dimostrato che gli utenti che usano l'algoritmo di raccomandazione offerto dalla piattaforma sono più impegnati in discorsi politici radicali di quelli che usano la piattaforma senza tener conto delle indicazioni dell'algoritmo (PNAS 2021). Secondo molte ricerche, gli algoritmi di raccomandazione sviluppati dalle piattaforme per selezionare automaticamente i messaggi che probabilmente interessano i singoli utenti in base ai loro comportamenti passati, hanno una potenziale rilevanza nell'amplificazione dei messaggi più coinvolgenti: il che in certi casi si traduce in una maggiore diffusione dei discorsi di odio, nella possibilità che le persone si ritrovino in echo-chamber caratterizzate da opinioni uniformi e, infine, nella potenziale tendenza alla radicalizzazione delle opinioni.

Il bilanciamento dei diritti umani che insistono sulla questione che qui si sta trattando è la sola strada per garantire un'azione politica corretta ed efficace in materia. La libertà di espressione e il diritto a essere difesi dalle violenze sono dimensioni irrinunciabili dei principi che guidano e ispirano la comunità degli europei. Questo quadro ispira ogni azione politica in materia, compresa quella che tende a rafforzare le difese della società dal fenomeno dell'istigazione all'odio online. Il tema dell'odio in rete, in effetti, può essere un problema per la qualità della democrazia e per l'ordine pubblico. La Commissione europea ha avviato un processo normativo per arrivare a definire certe manifestazioni di odio come un "crimine" allo stesso livello del terrorismo (European Commission 2021).

Secondo il Consiglio d'Europa, la maggior parte dei paesi europei ha creato un meccanismo di raccolta di informazioni e di supporto per le vittime del cyberbullismo, dei discorsi di odio online e dei crimini motivati dall'odio, con il contribuito delle autorità nazionali e di Organizzazioni non governative (Council of Europe 2022a)[2]. La sempre migliore conoscenza del fenomeno e la costruzione di organizzazioni che consentano di trovare aiuto a tutti i cittadini che sono vittime dell'istigazione all'odio online sono percorsi necessari per la qualità della convivenza civile.

Interpretazioni

La pluralità delle dimensioni del fenomeno dell'istigazione all'odio online si riflette in una vasta varietà di interpretazioni. L'argomento infatti non può essere compreso se non con un approccio consapevole della sua complessità. Non esistono ricette semplici e lineari per comprenderlo e tantomeno per contrastarne le conseguenze.

[2] Council of Europe: «Most European countries have established national reporting mechanisms and support for victims of cyber bullying, hate speech and hate crime, provided by national authorities and NGOs»

La definizione del fenomeno è a sua volta complessa. Il lavoro definitorio genera conseguenze diverse nei diversi contesti disciplinari: giuridico, scientifico, etico, politico, tecnologico, e così via. Nella sezione del rapporto dedicata a questo argomento si raccolgono le principali concettualizzazioni in materia. Ma una macro-classificazione si può proporre in chiave sintetica anche in questa sede introduttiva.

Da una parte, si distinguono forme di istigazione all'odio online che sono chiaramente illegali e altri discorsi d'odio culturalmente aberranti ma legali. In questo quadro si inseriscono anche le questioni che riguardano il bilanciamento tra i diversi diritti delle persone. Evidentemente, le forme di espressione dell'odio non illegali possono godere della protezione del più generale diritto alla libertà di espressione. D'altra parte, nel contesto dei media sociali, le stesse forme espressive possono essere ingigantite dalle logiche algoritmiche e dai modelli di comportamento tipici dei media sociali. E nel caso che per esempio che la comunicazione di opinioni sconfini nell'invito all'azione, fosse anche soltanto l'azione di rilanciare quelle opinioni, si possono generare in qualche caso anche conseguenze sull'ordine pubblico.

Dall'altra parte, si distinguono forme di istigazione all'odio online motivate da un intento politico e altre che riflettono la sofferenza - sociale o psicologica - di persone che si sentono trattate ingiustamente dalla vita e che tendono a ribaltare all'esterno quella loro sofferenza. E del resto le due dimensioni dell'odio possono convergere quando certe forze politiche considerano la sofferenza dei cittadini come uno strumento per manipolarli.

Prospettive

I metodi di analisi e i risultati dei ricercatori che si dedicano all'analisi del fenomeno dell'istigazione all'odio online sono discussi nella sezione dedicata all'argomento di questo rapporto.

L'impressione generale è quella di un aumento dell'importanza e della diffusione del fenomeno. E per questo sorge l'esigenza di pensare a forme di azione tese alla salvaguardia della qualità della convivenza civile e, in qualche caso, addirittura alla difesa della democrazia.

Questo pone problemi strategici legati alla necessità di bilanciare i diritti dei cittadini che evidentemente devono potersi esprimere e nello stesso tempo devono poter vivere con una certa sicurezza. Una strategia concentrata sulla persuasione più che sulla repressione, sull'educazione più che sulla criminalizzazione, sarà probabilmente analizzata per le dimensioni del fenomeno che non hanno un diretto riflesso sulll'ordine pubblico, ma generano sofferenze e divisioni nella società, tenendo peraltro presente, certamente, la libertà di espressione di chi odia ma anche quella di chi essendo vittima di discorsi violenti tende a rinunciare alla propria libertà di espressione.

Qualsiasi sviluppo politico in materia di contrasto all'istigazione all'odio online dovrà avere una prospettiva di collaborazione tra le istituzioni e gli Stati europei da una parte e, dall'altra, una co-progettazione del futuro con le piattaforme esistenti e con quelle che verranno.

E tutto questo sarà tanto più efficace quanto più potrà contare su una vasta e profonda ricerca interdisciplinare sul fenomeno dell'istigazione all'odio online. L'investimento nella ricerca sui fenomeni di malessere sociale è il primo passo della cura.

Sezione I:
DEFINIRE L'INCITAMENTO ALL'ODIO

"How can we ban [hate speech] if nobody agrees about what it is?"
"Come possiamo vietare l'incitamento all'odio se non ci accordiamo su cosa sia?"

- *Dott. Alexander Brown, Lettore di Teoria Politica e Giuridica*
(Brown 2017, p.605)

Introduzione

Sebbene l'attenzione in merito all'incitamento all'odio online nella letteratura scientifica sia in costante aumento, si sa sorprendentemente poco circa la prevalenza, le cause o le conseguenze delle diverse forme di linguaggio nocivo su differenti piattaforme. Per giunta, soltanto di recente i ricercatori hanno cominciato ad esaminare l'efficacia dei diversi approcci per contrastare l'odio online, e la nostra comprensione dei costi collaterali di tali interventi è particolarmente limitata. Questa sezione esamina lo stato dell'arte della ricerca, intesa come ricerca scientifica ma anche ricerca accademica in ambito legale e documenti politici, nel definire l'incitamento all'odio (sia online che offline). La sezione si concentra in particolare sui dibattiti in corso ed esamina le limitazioni nel quadro degli approcci attuali che si propongono di definire e rilevare l'incitamento all'odio online.

Cosa s'intende per "Incitamento all'odio"?

Ad oggi, non esiste consenso nella comunità scientifica su una singola definizione di incitamento all'odio, sia esso online o offline (vedi anche Papucnova et al. 2021, p. 2). Di norma, l'incitamento all'odio si intende come linguaggio offensivo motivato da pregiudizi, dannoso e di natura ostile, che sia rivolto a una persona o un gruppo a causa delle loro caratteristiche innate, effettive o percepite (Faris et al. 2016; Cohen-Almagor 2011).

Ad un'estremità dello spettro si trovano quelle definizioni che includono un'ampia varietà di espressioni dirette contro un individuo o un gruppo specifico o facilmente identificabile, sulla base di caratteristiche arbitrarie o irrilevanti dal punto di vista normativo. Alla parte opposta dello spettro troviamo definizioni che richiedono l'intenzione di ferire ed implicano che l'incitamento all'odio debba consistere in "espressioni pericolose", che possano essere collegate a fomentare violenza di massa o danni fisici diretti a gruppi terzi (Benesch, 2013). Di conseguenza, l'incitamento all'odio può essere inteso in tre distinti contesti: quello degli strumenti politici di matrice europea, quello accademico e quello pratico.

A. Definizioni che derivano dalle politiche di regolamentazione europee

In primo luogo, gli strumenti giuridici europei hanno ripetutamente stabilito il significato dell'incitamento all'odio come segue:

- Recentemente, la raccomandazione del Consiglio sulla lotta all'incitamento all'odio, adottata nel maggio 2022, afferma che:

"Per incitamento all'odio si intendono tutti i tipi di espressioni che incitano, promuovono, diffondono o giustificano la violenza, l'odio o la discriminazione nei confronti di una persona o di un gruppo di persone, o che li denigrano, in ragione di caratteristiche personali o status reali o attribuiti come "razza", colore, lingua,

religione, nazionalità, origine nazionale o etnica, età, disabilità, sesso, identità di genere e orientamento sessuale."

- ECRI General Policy Recommendation No. 15:

"Messaggi che violano le norme legali esistenti e richiedono una regolamentazione governativa, ovvero messaggi condivisi pubblicamente, incitano, promuovono o giustificano odio, discriminazione o ostilità verso un gruppo e/o un individuo specifico, sulla base di determinati attributi, come razza o etnia origine, religione, disabilità, genere, età, orientamento sessuale/identità di genere." (ECRI 2016)

- Council Framework Decision 2008/913/JHA:

"Incitamento pubblico alla violenza o all'odio, diretto contro un gruppo di persone o un membro di tale gruppo definito sulla base di razza, [colore], discendenza, religione o credo, o origine nazionale o etnica" e "condonare, negare o banalizzare pubblicamente crimini di genocidio, crimini contro l'umanità e crimini di guerra [come definiti nel diritto dell'UE], quando la condotta si sviluppa in modo tale da incitare alla violenza o all'odio contro tale gruppo o un membro di tale gruppo." (EUR-Lex 2008; Commissione europea 2016b; Sellars 2016)

- Council of Europe Recommendation No. R (97) 20:

''L'incitamento all'odio comprende tutte le forme di espressione che diffondono, incitano, promuovono o giustificano l'odio razziale, la xenofobia, l'antisemitismo o altre forme di odio basate sull'intolleranza, tra le quali: intolleranza espressa dal nazionalismo aggressivo e dall'etnocentrismo, discriminazione e ostilità contro minoranze, migranti e persone di origine immigrata.'' (Consiglio d'Europa, 1997)

- A livello globale, l'articolo 20 della Convenzione Internazionale sui Diritti Civili e Politici (ICCPR) afferma che:

''(...) qualsiasi difesa dell'odio nazionale, razziale o religioso che costituisca incitamento alla discriminazione, all'ostilità o alla violenza è vietata dalla legge.'' (Assemblea Generale delle Nazioni Unite, 1966)

- L'organizzazione per la libertà di espressione "Articolo 19" ha sviluppato i Principi di Camden sulla Libertà di Espressione e l'Uguaglianza (Camden Principles on Freedom of Expression and Equality), fornendo una guida interpretativa per gli articoli della Convenzione Internazionale sui Diritti Civili e Politici (ICCPR) e cercando di dissuadere gli agenti dall'abusare dell'articolo 20 specificando questioni relative all'"istigazione", nonché a ciò che costituisce "discriminazione", "ostilità" e "violenza". (Articolo 19, 2009)

- In merito all'istigazione all'odio, è importante anche la Convenzione Internazionale per l'Eliminazione di Tutte le Forme di Discriminazione Razziale (International Convention on the Elimination of All Forms of Racial Discrimination - ICERD), che include una clausola più rigorosa dell'articolo 20 della citata ICCPR, in quanto non richiede intenzionalità o "apologia dell'odio" ed include la disseminazione nella lista delle pratiche punibili (Ufficio delle Nazioni Unite per la prevenzione del genocidio e la responsabilità di proteggere - UN's Office on Genocide Prevention and the Responsibility to Protect 2021, p.3).

- A complemento di questi principi, il Piano d'azione di Rabat propone un "test soglia in sei parti" per giustificare le restrizioni alla libertà di espressione, considerando il contesto socio-politico, lo stato di chi parla, l'intenzione di incitare all'antagonismo, il contenuto del discorso, l'estensione della diffusione e la probabilità di danno (Ufficio delle Nazioni Unite per la prevenzione del genocidio e la responsabilità di proteggere - UN's Office on Genocide Prevention and the Responsibility to Protect 2021, p.3).

Come illustrato fin qui, istigazione all'odio e crimini d'odio sono pienamente riconosciuti a livello internazionale come azioni criminali (si veda anche Organizzazione per la sicurezza e la cooperazione in Europa – OSCE, 2009). È inoltre interessante notare che, nel dicembre del 2021, la Commissione europea ha

pubblicato una Comunicazione al Parlamento Europeo ed al Consiglio d'Europa allo scopo di stimolare una decisione del Consiglio che includesse istigazione all'odio e crimini d'odio nell'attuale lista dei cosiddetti "Crimini EU" ("EU crimes") come previsto dall'articolo 83 del Trattato sul Funzionamento dell'Unione europea (Commissione europea 2021; TFEU 2009). Una simile decisione permetterebbe alla Commissione, in un secondo momento, di rafforzare il quadro giuridico per contrastare l'incitamento all'odio in tutta l'Unione europea.

B. Definizioni accademiche

In secondo luogo, l'argomento dell'istigazione all'odio è stato ampiamente dibattuto anche da studiosi accademici ed esperti del settore.

Un particolare punto di vista considera che l'incitamento all'odio "esprime, incoraggia, fomenta o incita all'odio contro un gruppo di individui caratterizzati da una caratteristica particolare o da un insieme di caratteristiche come razza, etnia, genere, religione, nazionalità e orientamento sessuale [...] ed è spesso (ma non necessariamente) espresso in un linguaggio offensivo, irato, abusivo ed insultante" (Parekh 2012, p. 39). Lo stesso autore identifica poi i tre principali aspetti della sua teoria sull'incitamento all'odio:

(1) diretto a individui o gruppi di individui specifici,
(2) cui sono attribuite qualità non desiderabili, che portano a stigmatizzarli,
(3) portando alla discriminazione, perché l'attribuzione di attributi indesiderabili incoraggia a vedere l'obiettivo stesso come indesiderabile (Parekh 2012, p. 39)

Andando nello specifico, quest'ultimo aspetto è enfatizzato dallo studioso Gelber che sostiene che la disuguaglianza è un elemento centrale dell'incitamento all'odio manifestato da una relazione "autorità subordinata vs. autorità superiore". Secondo l'autore, l'emozione dell'odio non deve essere evocata affinché l'incitamento all'odio possa causare danni e l'incitamento all'odio non deve solo riguardare o influenzare una minoranza o un gruppo sociale svantaggiato, come invece si pensava in precedenza (Gelber, 2019).

Un'altra prospettiva accademica spiega che l'incitamento all'odio ha l'obiettivo di deumanizzare, molestare, intimidire, degradare, vittimizzare o incitare alla brutalità contro il proprio bersaglio (Cohen-Almagor 2011).

Rispetto all'incitamento all'odio online, esistono approcci simili. L'Anti-Defamation League lo definisce come qualsiasi rappresentazione di idee che promuovono l'odio, la discriminazione o la violenza contro qualsiasi individuo o gruppo di persone, sulla base di aspetti quali razza, colore, origine etnica, nazionalità o etnia e religione attraverso i media digitali (Anti-Defamation League 2016). Altri studiosi considerano il cyberhate come un comportamento di denigrazione, minaccia, esclusione e provocazione che può incitare alla violenza attraverso i media (Wachs & Wright 2019; Hawdon et al. 2017).

Allo stesso modo, Watanabe et al. classificano il cyberhate come uso di linguaggio violento, offensivo o aggressivo, indirizzato verso un gruppo specifico contraddistinto da caratteristiche comuni, che potrebbero essere di religione, razza, identità sessuale e di genere o affiliazione politica attraverso l'uso di Internet e dei Social Networks, basato su qualche sorta di squilibrio di potere, proposto in modo sistematico e incontrollabile, attraverso i media digitali e spesso motivato da ideologie a cui aderiscono individui e gruppi, che derivano da comportamenti che possono essere considerati atti di comunicazione deviante in quanto possono violare standard culturali condivisi, regole o norme di interazione sociale in contesti di gruppo (Watanabe et al. 2018).

C. Definizioni pratiche

In terzo luogo, alcune definizioni possono essere reperite sulle piattaforme digitali (ovvero i social media) e sugli strumenti di rilevamento attualmente esistenti, che hanno sviluppato le proprie definizioni di istigamento all'odio per garantire che gli utenti si attengano alle regole e come parte delle loro regolamentazioni interne (per esempio i termini del servizio o gli standard della comunità)

- Ad esempio, la sezione "incitamento all'odio" delle Norme della community di YouTube afferma che "rimuoviamo i contenuti che promuovono la violenza o l'odio contro individui o gruppi in base a uno dei seguenti attributi basati su razza, religione, disabilità, identità ed espressione di sesso/genere, età, nazionalità, o lo status di veterano, o il cui scopo principale è incitare all'odio o incoraggiare la violenza sulla base di queste caratteristiche fondamentali" (YouTube 2022).

- I termini di servizio di Twitter stabiliscono che la compagnia proibisce la "condotta d'odio" incluso "[promuovere] la violenza o l'attacco diretto o la minaccia a persone terze sulla base di razza, etnia, paese d'origine, orientamento sessuale, genere ed identità di genere, affiliazione religiosa, età, disabilità o malattia." La società sottolinea inoltre che non consente account il cui scopo principale è incitare a danneggiare terzi "sulla base di queste categorie" (Twitter 2022).

- La definizione di incitamento all'odio di Facebook include l'incitamento alla violenza verbale come inteso dalle piattaforme sopra citate, identificando invece l'incitamento all'odio come contenuto che attacca direttamente "le persone in base alla loro razza; etnia; origine nazionale; appartenenza religiosa orientamento sessuale; sesso, genere o identità di genere; o disabilità o malattie gravi" (Facebook 2022).

Di conseguenza, sulla base di un'analisi del contenuto dei vari tipi di definizioni, Fortuna e Nunes propongono una loro interpretazione teorica, probabilmente una delle più complete:

"L'incitamento all'odio è un atto verbale che attacca o sminuisce, che incita alla violenza o all'odio contro i gruppi, sulla base di caratteristiche specifiche come l'aspetto fisico, la religione, la discendenza, l'origine nazionale o etnica, l'orientamento sessuale, l'identità di genere o altro, e può verificarsi anche con stili ed in forme sottili o quando si usa l'umorismo'' (Fortuna & Nunes 2018).

Difficoltà e preoccupazioni

Da un punto di vista metodologico, identificare l'istigazione all'odio online a scopo di ricerca è un problema caratterizzato da difficoltà crescenti (Ufficio delle Nazioni Unite per la prevenzione del genocidio e la responsabilità di proteggere - UN's Office on Genocide Prevention and the Responsibility to Protect 2021, p.2), incluse:

i. La varietà delle comunità digitali e di forme d'incitamento all'odio online (tipo di linguaggio, immagini, ecc.)

ii. Le definizioni usate per inquadrare il problema

L'assenza di una definizione di istigazione all'odio che sia comunemente accettata non crea certamente condizioni favorevoli ad una più efficace regolamentazione del fenomeno (Topidi 2021, p.257). Ciò è il risultato di due ragioni principali: la diversa interpretazione della libertà di parola, prevalentemente tra paesi o regioni, e le differenziazioni interconnesse nella concettualizzazione del danno (Alkiviadou 2018, pp.5-6). Entrando nel dettaglio, quando è il momento di tradurre la legge ed i principi internazionali in leggi a livello nazionale, ogni paese ha un'approccio leggermente differenziato rispetto alla definizione di istigazione all'odio in termini di come viene espresso, chi siano le potenziali vittime, e quali danni si debbano verificare

affinché un attacco verbale possa essere considerato come offensivo (Ufficio delle Nazioni Unite per la prevenzione del genocidio e la responsabilità di proteggere - UN's Office on Genocide Prevention and the Responsibility to Protect 2021, p.4).

Similmente, il Dangerous Speech Project di Susan Benesch (DSP 2021) suggerisce che ci siano due principali difficoltà con il termine "istigazione all'odio". Prima di tutto, "odio" è un termine vago che può definire vari livelli d'intensità e sottendere a diverse conseguenze – "Il termine 'odio' nell'incitamento all'odio significa che l'oratore odia, o cerca di persuadere gli altri a odiare, o vuole far sentire le persone odiate?" (Benesch 2021, p.7)

In secondo luogo, "incitamento all'odio" in sostanza significa che persone o gruppi sono attaccati per via della loro identità o appartenenza ad un gruppo. Questo significa che una legge o definizione dovrebbe necessariamente specificare se tutte le identità ed i gruppi sono da considerarsi inclusi, ed in caso contrario, quali tipi di gruppo sono inclusi. Il Dangerous Speech Project sostiene che leggi troppo vaghe possano essere utilizzate in modo improprio contro gruppi vulnerabili o di opposizione politica e civica, con il possibile risultato di danneggiare gli stessi gruppi che le leggi sull'incitamento all'odio mirano a proteggere. D'altra parte, si potrebbe anche sostenere che una definizione troppo restrittiva, che si concentri su gruppi ed identità specifici, potrebbe comportare l'esclusione legale o la mancanza di strumenti giuridici per affrontare il problema (Ufficio delle Nazioni Unite per la prevenzione del genocidio e la responsabilità di proteggere - UN's Office on Genocide Prevention and the Responsibility to Protect 2021, p.4).

iii. Contesti storici e sociali, sottigliezze linguistiche

Rimane inoltre aperto il problema della differente sensibilità culturale sull'utilizzo di alcune parole, che sono considerate come problematiche in alcune culture ma non in altre (Papcunová et al. 2021, p.11). Un altro problema è costituito dal fatto che la validità e quindi la genericità del gruppo di indicatori che definiscono l'incitamento all'odio è di solito dipendente dalla maniera in cui l'argomento (come per esempio LGBTQIA+, genere, dibattiti politici, ecc.) si relaziona con l'orizzonte culturale (Papcunová et al. 2021, p.13). Inoltre, la stragrande maggioranza degli studi è stata condotta utilizzando i dati dei social media in lingua inglese e quindi non rivela ovviamente abbastanza rispetto ad altre piattaforme o contesti culturali. Un'ulteriore complicazione è data dal fatto che definizioni e proposte per rilevare l'istigazione all'odio sono altamente politicizzate, in particolare nell'ambito dei regimi autoritari e dei conflitti (Siegel 2020, p.76).

Dal punto di vista tecnologico, l'incitamento all'odio è un argomento di studio complesso a causa della scarsa affidabilità dei sistemi di rilevazione, della natura opaca degli algoritmi proprietari, della carenza di accesso ai dati in mano alle società e così via (Ufficio delle Nazioni Unite per la prevenzione del genocidio e la responsabilità di proteggere - UN's Office on Genocide Prevention and the Responsibility to Protect 2021, p.2).

A proposito delle tre piattaforme di social media di cui si è discusso in precedenza, emergono tre principali questioni. Per prima cosa, YouTube vieta uno spettro maggiore di attività in quanto include nei divieti anche l'umiliazione delle persone. Ciò nonostante, YouTube menziona il divieto di attacchi verbali contro un gruppo determinato anziché contro una persona che appartenga a quel gruppo. Questo deve forse essere inteso in modo che, se una particolare espressione fosse diretta contro un individuo che apparteneva a un gruppo a causa delle caratteristiche di quel gruppo (piuttosto che contro il gruppo nella sua interezza), tale discorso non sarebbe proibito? Considerando nella loro interezza i contenuti del Codice di Condotta stabilito fra la Commissione europea e le compagnie Informatiche (Commissione europea 2016b), è più probabile che il problema sia dovuto ad un utilizzo improprio del linguaggio. Infine, nonostante il Codice di Condotta, Twitter sembra limitare il controllo dei contenuti sulla propria rete, a meno che non si tratti di una minaccia di violenza diretta e specifica (piuttosto che astratta e generalizzata). Questo approccio è più rigoroso rispetto ai documenti legali di cui sopra che comportano sanzioni penali (Alkiviadou 2018, p.6).

In ogni caso, è indispensabile fare maggiore chiarezza su come queste sfide possano essere affrontate per ottenere una migliore comprensione di come l'incitamento all'odio online emerga e si intensifichi, in modo da poter studiare e pianificare risposte più efficaci.

Conclusioni

Dal momento che l'istigazione all'odio online è diventato un problema sempre più evidente sulle piattaforme di social media, è stato messo al centro degli ordini del giorno nei dibattiti accademici, politici e legali. La discussione in merito ad una definizione di incitamento all'odio è tuttavia lontana dal concludersi. Sia in conseguenza del fatto che una definizione è ancora in discussione, che a causa della natura altamente specifica e in evoluzione dell'incitamento all'odio online, rilevare sistematicamente i contenuti che incitano all'odio è un compito estremamente difficile. Altrettanto impegnativo è però produrre una definizione accurata dell'incitamento all'odio data la sua natura multidimensionale e complessa (Siegel 2020, pp.75-76).

Le tecniche all'avanguardia che utilizzano l'apprendimento automatico, le reti neurali e l'incorporazione di funzionalità contestuali hanno migliorato la nostra capacità di misurare e monitorare l'incitamento all'odio, ma la maggior parte del lavoro empirico esistente è abbastanza frammentato, e sovente rileva una singola tipologia di incitamento, su una sola piattaforma, per un tempo limitato. Sarebbe necessaria una corretta comprensione multidisciplinare del fenomeno per contrastarlo efficacemente. Pertanto, l'incitamento all'odio dovrebbe essere affrontato e perseguito in base alla sua operatività (Papcunová et al. 2021, p.2).

Sezione II:
QUANTIFICARE L'INCITAMENTO ALL'ODIO ONLINE

Ad oggi, non esiste consenso nella comunità scientifica su una definizione che consenta di misurare l'incitamento all'odio online (online hate speech – OHS) (Castaño-Pulgarin et al. 2021; Chetty et al. 2018). Per misurare qualsiasi fenomeno, in particolare se riferito all'ambito sociale, è necessario per prima cosa definirlo e circoscrivere l'ambiente nel quale si sviluppa. A tal proposito, è estremamente importante comprendere innanzitutto che l'OHS è un fenomeno complesso, che deve necessariamente essere approcciato con strumenti adatti ad indagare ed analizzare sistemi complessi. È importante riconoscere gli agenti, siano essi individui o gruppi, che mettono in atto comportamenti definiti, così come l'ambiente nel quale operano: le piattaforme di comunicazione online. È solo l'interazione fra agenti mediata da tali piattaforme, nonché le interazioni fra gli agenti e le medesime piattaforme, a definire un ecosistema digitale diviso in carnefici e vittime, mitigatori ed amplificatori. Questi ruoli possono essere assunti di volta in volta da individui o istituzioni, e sono spesso intercambiabili a seconda del contesto.

È doveroso notare come il fenomeno non dovrebbe essere studiato in connessione ad una particolare piattaforma digitale, come per esempio uno specifico social media network, bensì rispetto all'intera gamma di piattaforme che consentono la comunicazione fra utenti. Infatti il fenomeno è chiaramente trasversale, nonostante le singole piattaforme vengano spesso sfruttate dagli incitatori di odio online per affinare strategie che sfruttino le loro peculiari caratteristiche. È proprio la possibilità di scambiare informazioni attraverso canali rapidi, siano essi anonimi o non anonimi e con potenziale esposizione a un pubblico limitato o illimitato, che innesca il comportamento scorretto di un individuo o di un gruppo.

In questo senso, la complessa natura dell'ecosistema responsabile dell'incitamento all'odio online non può essere pienamente compresa sulla sola base delle attività umane nell'ambiente digitale. Di fatto, l'incitamento all'odio online spesso riproduce azioni o atteggiamenti scorretti propri del mondo reale, nel quale le comunicazioni vengono scambiate ad una velocità inferiore a causa dei tradizionali vincoli sociali e comportamentali. Ogni tentativo di esaminare l'incitamento all'odio online non può pertanto prescindere da un'analisi approfondita dei meccanismi basilari che si nascondono dietro i comportamenti ostili nello spazio fisico ed in quello digitale, ed alle relazioni fra i diversi contesti online ed offline.

Esistono diverse forme d'incitamento all'odio che sono state riconosciute come particolarmente rilevanti per via delle loro caratteristiche motivazionali, emozionali e cognitive. Fra queste, possiamo fare alcuni esempi:

- Incitamento all'odio contro individui o istituzioni che sono considerati estremamente influenti e privilegiati;
- Incitamento all'odio contro individui o istituzioni cui è riconosciuta una particolare competenza o prestigio scientifico;
- Incitamento all'odio contro minoranze etniche, culturali o sociali;
- Incitamento all'odio contro soggetti fragili perché affetti da problemi di salute, discriminazione o marginalizzazione;
- Incitamento all'odio contro individui ed istituzioni che sono visti come sostenitori di idee in opposizione a quelle del carnefice, quando il dibattito politico o ideologico è polarizzato.

Si potrebbero elencare moltissimi altri esempi, ma questa breve lista dimostra chiaramente come l'incitamento all'odio possa assumere caratteristiche differenti in ogni diverso contesto, prendendo di mira di volta in volta potenti o emarginati, persone istruite e meno istruite, e così via.

Allo stesso modo, si possono elencare molti differenti fattori motivazionali che alimentano l'incitamento all'odio:

- Il desiderio di approvazione sociale da parte di coetanei, colleghi o persone con idee simili;
- L'autosoddisfazione nell'affermare le proprie idee ed i propri giudizi;
- L'immersione narrativa e l'intrattenimento (sebbene possa apparire paradossale, esistono persone che incitano all'odio perché si divertono a discutere e litigare in una situazione nella quale si sentono relativamente al sicuro dalle conseguenze di tali azioni e pensano di non poter essere ritenute responsabili per il loro comportamento d'odio);
- Gestione dell'ansia e della dissonanza cognitiva, in questo caso, i bersagli dell'incitamento all'odio sono comodi capri espiatori per giustificare frustrazioni e fallimenti personali;
- Difesa "tribale" contro esterni che sono visti come portatori di tratti sociali e culturali considerati inaccettabili per gli standard ideologici e morali di riferimento per il gruppo interessato;
- Tratti oscuri della personalità: le persone con atteggiamenti psicopatici o narcisistici trovano particolarmente gratificante essere odiosi nei confronti degli altri per coltivare il loro (immotivato) senso di grandiosità e superiorità morale.

Di nuovo, anche dal punto di vista motivazionale possono esistere giustificazioni scatenanti molto diverse dietro questi comportamenti, che in alcune circostanze possono rafforzarsi fra loro.

In luce delle considerazioni menzionate, piuttosto che ragionare in termini di incitamento all'odio in generale, considerando motivazioni e bersagli, è indispensabile definire un contesto adeguato ad affrontare il problema specifico dell'incitamento all'odio che sia particolarmente rilevante per il fenomeno in esame.

Valutare l'incitamento all'odio online è un compito reso più complicato dall'impossibilità di effettuare misurazioni in condizioni controllate, come per esempio si potrebbe fare in un laboratorio, cosa che richiederebbe la progettazione di metodi innovativi per raccogliere informazioni in tempo quasi reale dalle piattaforme digitali. In casi specifici, nei quali un particolare evento scatena una risposta specifica e fortemente motivata d'incitamento all'odio in termini di linguaggio, contenuto e bersagli, esiste però la possibilità di effettuare esperimenti naturali che possono essere rivelatori rispetto ai meccanismi sottostanti ed alle condizioni abilitanti. Le piattaforme digitali consentono talvolta un accesso, seppur limitato, al contenuto pubblicato da utenti umani, come avviene per Twitter, oppure nessun tipo di accesso, come accade per esempio per WhatsApp. L'analisi di grandi moli di dati offre alcuni degli strumenti necessari a questo scopo, sebbene non sia tuttora semplice raccogliere volumi di dati tanto significativi in un così breve lasso di tempo. Comunque, la raccolta massiva di dati non è sufficiente da sola a permettere l'adeguata valutazione di un fenomeno tanto complesso: si rendono quindi necessari approcci interdisciplinari e transdisciplinari, che richiedono l'applicazione di tecniche olistiche.

È opportuno ribadire che difficoltà molto simili si verificano in ambiti scientifici altrettanto complessi, come quelli che si occupano di analisi delle reti sociali e dei sistemi biologici. Le *scienze sociali computazionali* (Lazer et al. 2009) sono probabilmente il campo interdisciplinare, situato al crocevia fra scienze relazionali, scienze sociali e comportamentali, informatica e matematica applicata, che sembra offrire il maggiore potenziale per fornire risposte soddisfacenti alle sfide qui delineate (Lazer et al. 2020).

Sezione III:
EVIDENZE

Mappatura dell'ecosistema dell'odio online

Come anticipato, è indispensabile riconoscere che i cluster di odio online non agiscono in isolamento l'uno dall'altro, ma costruiscono un ecosistema efficace e resiliente che collega il mondo fisico e quello digitale. Nell'ambito digitale, i gruppi su piattaforme diverse sono interconnessi e interdipendenti (Figura 1), in maniera affine a quelli che caratterizzano il mondo fisico. Questi gruppi sono strutturati in reti complesse e si adattano alle condizioni mutevoli dettate dalla regolamentazione (Johnson et al. 2019), rivelando una resilienza simile a quella degli ecosistemi biologici (May 2019). Questo rende difficile controllarli o smantellarli completamente con approcci convenzionali (Velasquez et al. 2021). Alla luce degli esempi citati di fattori motivazionali alla base dell'incitamento all'odio, è relativamente semplice capire perché tali fenomeni tendano ad essere così persistenti: sono generalmente stimolati piuttosto che rallentati dall'opposizione e dalle sanzioni sociali. In particolare, tali contromisure sono spesso interpretate come evidenti conferme della correttezza e necessità dei presupposti concettuali su cui si fonda il comportamento d'odio.

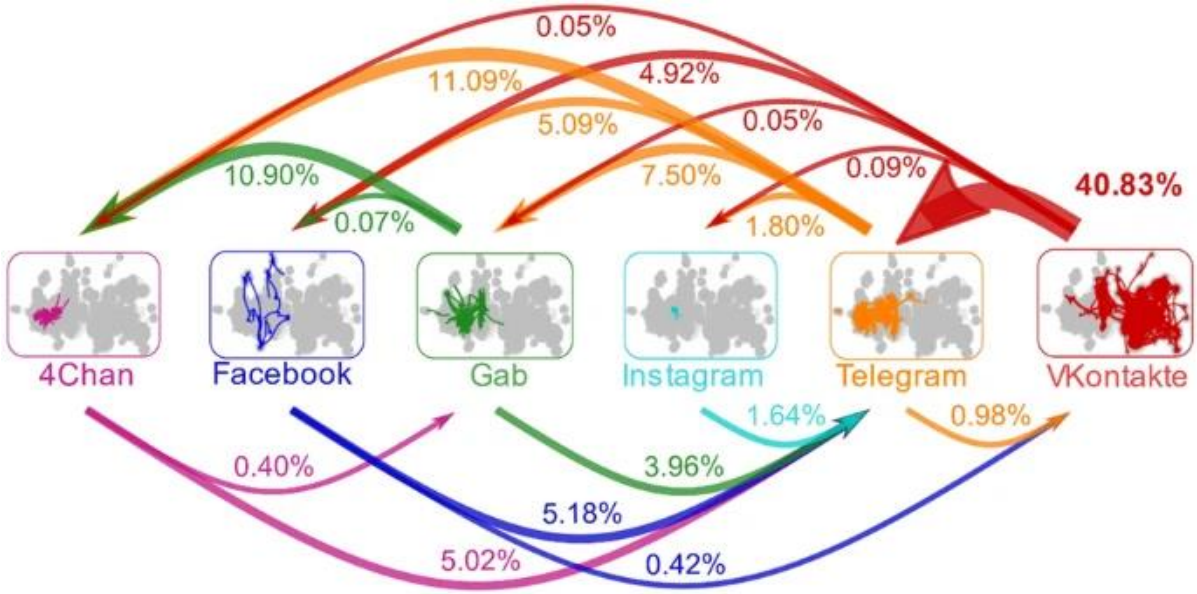

Figura 1. Il "multiverso" dell'odio online, una mappa di connessione empirica fra i cluster d'odio e alcune piattaforme digitali. Immagine pubblicata in Velazquez et al (2021).

Indicatori fondamentali dell'odio online

Un approccio ampiamente condiviso è quello di affidarsi agli stessi strumenti messi a disposizione dalle piattaforme digitali per quantificare il coinvolgimento: misurare la quantità di post d'odio e la loro frequenza, il numero di visualizzazioni e, in generale, il tipo di coinvolgimento e supporto che sono in grado di produrre. Per esempio: generano più risposte, che rappresentano il coinvolgimento, o più condivisioni, che quantificano il supporto?

L'ipotesi sottostante è che sia possibile condividere una scala comune per la misurazione, nella quale il valore più basso è lo zero che corrisponde a nessun coinvolgimento o condivisione, mentre valori maggiori rappresentano la presenza di comunicazione d'odio. Attraverso un'adeguata comparazione numerica, della

frequenza o del grado di comunicazioni d'odio rispetto a metriche prestabilite, è persino possibile calcolare la possibilità di essere esposti a crimini d'odio online.
Un simile approccio permette di stimare la quantità d'incitamento all'odio online a livello della popolazione, sebbene abbia due principali punti deboli:

1. Per essere efficace, richiede il riconoscimento semiautomatico dell'incitamento all'odio in contenuti testuali, sonori e visuali, un compito che è ancora difficile da portare a termine e richiede tecniche avanzate di apprendimento automatico a partire da basi di dati etichettate manualmente.
2. Non fornisce alcuna informazione a livello individuale, che richiede osservazioni più complesse, per esempio basate sull'analisi delle reti sociali.

Collegare comportamenti non autentici, disinformazione ed odio durante la pandemia di COVID-19

Nell'ecosistema dell'odio online, le vittime sono individui, gruppi o istituzioni, mentre i responsabili possono essere individui o gruppi che agiscono sia direttamente che indirettamente, per esempio attraverso agenti non umani come i social bots (Ferrara et al. 2016), in modo più o meno coordinato. È importante ribadire che il coordinamento dei comportamenti online può essere una proprietà emergente, ovvero uno stato collettivo che in certe circostanze può essere raggiunto nonostante non sia stato pianificato intenzionalmente. I social bots sono stati approfonditamente studiati per il loro ruolo nelle dinamiche sociali online, in special modo nel corso di eventi che attraggono l'attenzione collettiva come quelli di natura politica (Davis et al. 2016; Shao et al. 2018; Stella, Ferrara & De Domenico 2018; González-Bailón & De Domenico 2018; Gallotti et al. 2020). Recentemente, è stato proposto un processo metodologico per rivelare i legami tra l'incitamento all'odio e le attività dei bot durante la pandemia di COVID-19, rivelando un odio maggiore nelle comunità che sono più dense e più isolate dagli altri gruppi (Uyheng & Carley 2020), coerentemente con le motivazioni "tribalistiche" alla base dei comportamenti d'odio (Greene 2013). La maggiore attività online, dovuta per esempio alla recente pandemia, ha esposto gli utenti della nota piattaforma online Tik Tok, in particolare i bambini piccoli, a un aumento dell'odio online a causa dell'estremismo di destra presente in video, commenti, simboli e immagini (Weimann & Masri 2020). Tali post hanno generato centinaia di migliaia di visualizzazioni, stimolando solo interventi circoscritti da parte della piattaforma per mitigare il fenomeno. Il medesimo aumento nell'esposizione dei bambini alla violenza è stato calcolato su Twitter nello stesso periodo, nel corso del periodo di lockdown (Babvey et al. 2021) per mezzo del metodo illustrato nella Figura 2.

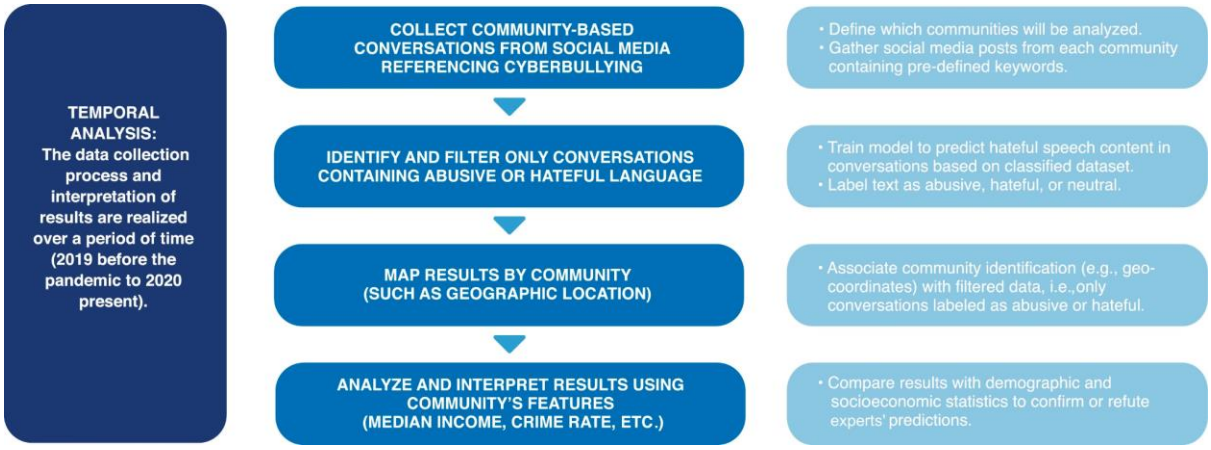

Figura 2. Procedura utilizzata per quantificare il cambiamento del cyberbullismo e dei contenuti offensivi o che incitano all'odio su Twitter durante la pandemia di COVID-19 rispetto al periodo pre-pandemia. Immagine pubblicata da Babvey et al (2021).

La pandemia di COVID-19 è stata un terreno fertile per la proliferazione dell'odio online in molti contesti e per una varietà di ragioni, aprendo una finestra d'opportunità unica per osservare e misurare il fenomeno su larga scala. La pandemia è un incubatore ideale per la comunicazione d'odio nel senso che attinge chiaramente a molti, se non tutti, i bersagli ed i fattori motivazionali che sono stati precedentemente citati. Istituzioni e scienziati sono presi di mira perché identificati come manipolatori e portatori di interessi ed obiettivi nascosti. Le minoranze possono essere identificate come diffusori preferenziali o persino come responsabili dei focolai pandemici. Le persone con fragilità possono essere accusate come indirette responsabili per le restrizioni e limitazioni alla libertà individuale e di movimento dei soggetti più sani. I sostenitori di vedute opposte rispetto alle misure e strategie di sanità pubblica più appropriate possono essere rappresentati come stupidi, prevenuti, o in collusione con i poteri e gli interessi costituiti. Per lo stesso motivo, la riduzione significativa delle opportunità d'interazione sociale durante una pandemia globale può rinforzare ulteriormente il bisogno di approvazione sociale delle persone e l'importanza della gratificazione dell'autostima che si ottiene difendendo il proprio interesse. Similmente, una pandemia genera alti livelli d'ansia che possono essere tenuti a freno attraverso comportamenti d'odio, anche attraverso la gratificazione ed il divertimento che alcuni soggetti sperimentano attraverso l'incitamento all'odio. In aggiunta, durante una pandemia il "tribalismo" è fortemente accentuato come spesso avviene nelle situazioni di pericolo, e molti soggetti che presentano tratti oscuri della personalità sentono un particolare bisogno di distinguersi dalle masse attraverso il rifiuto di sottostare alle regole e raccomandazioni delle autorità di sanità pubblica. Stimando la tossicità delle conversazioni online per mezzo di tecniche automatizzate, di nuovo su Twitter, è stato per esempio possibile identificare e mappare l'aumento dei contenuti tossici e fomentatori contro gli asiatici (Costello et al. 2021) (Figura 3) o contro le autorità di sanità pubblica come l'Organizzazione mondiale per la sanità (Majó-Vázquez et al. 2020) (Figura 4).

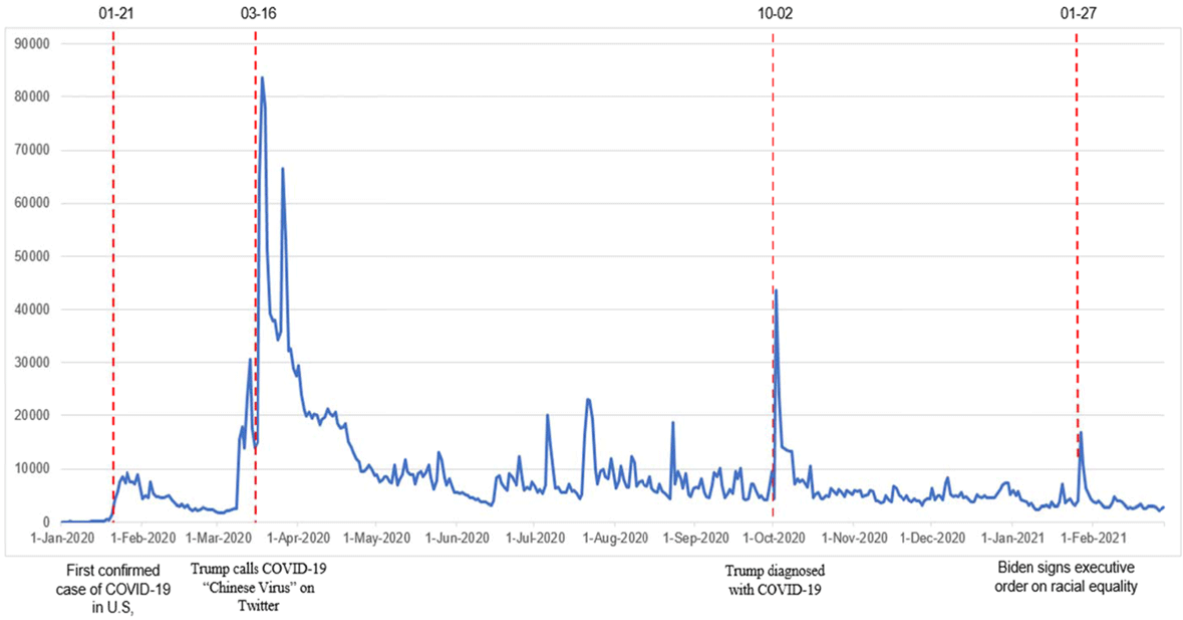

Figura 3. Volume giornaliero di post contenenti parole chiave connesse all'odio verso gli Asiatici su Twitter. nel corso del primo anno di pandemia da COVID-19. Analisi pubblicata da Costello et al (2021).

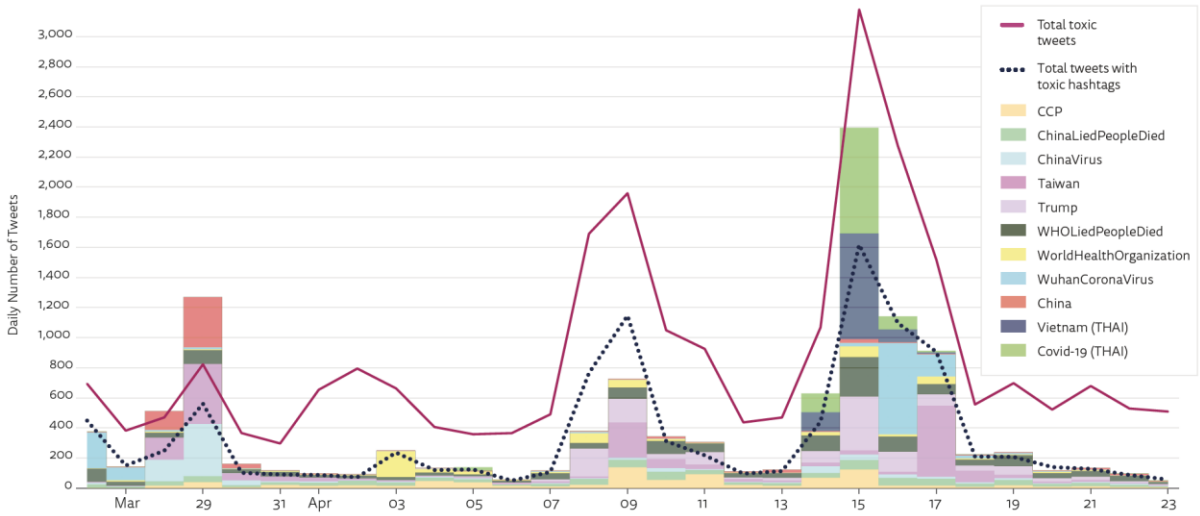

Figura 4. Volume giornaliero dei post su Twitter nei primi mesi della pandemia da COVID-19. La linea color rosso scuro indica il numero di post che si stima trasmettano un messaggio tossico. La linea di puntini blu rappresenta il conteggio dei tweet giornalieri che includano almeno uno dei 22 hashtag più tossici. Le barre a pila rappresentano il numero di tweet che includono ciascuno dei più popolari hashtag tossici.

In modo analogo, l'odio estremista ha trovato terreno fertile nella diffusione di disinformazione sul COVID-19, che ha aiutato ad incentivare vedute e comportamenti d'odio con queste modalità:

- Alimentando l'ostilità e rafforzando il risentimento (Schwarz & Holnburger 2019)
- Sfruttando l'opportunità di prendere per bersaglio gruppi specifici (Holbrook 2020)
- Accusando specifiche istituzioni politiche (Vegetti & Littyay 2022)

In aggiunta ed in connessione con la disinformazione, significative disparità di reddito possono ispirare la comunicazione d'odio, com'è stato recentemente registrato nel caso dell'Italia (Figura 5). Attraverso un sistema multi-fase (Musto et al. 2015) studiato per elaborare i tweet con riferimenti geografici, è stato possibile mappare la distribuzione dei contenuti d'odio rispetto alle diverse dimensioni dell'intolleranza, come l'omofobia, il razzismo, la violenza, la disabilità, l'antisemitismo e la discriminazione di genere (Denti & Faggian 2021).

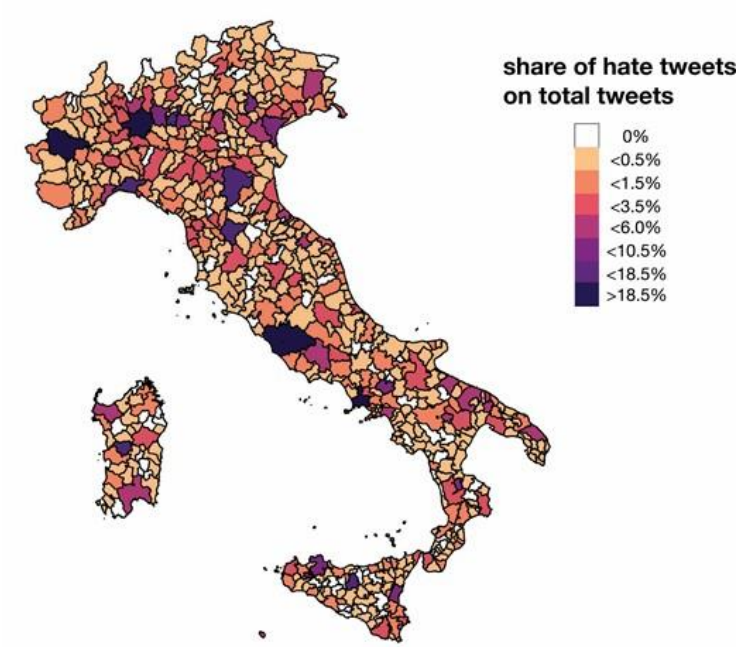

Figura 5. La mappa dei tweet d'odio nelle 611 aree locali del mercato del lavoro italiano (Local Labour Market Areas - LLMAs). Figura pubblicata da Denti & Faggian (2021)

Metodi avanzati basati su analisi di rete e grandi basi di dati

È possibile mappare i grandi volumi di dati raccolti dalle piattaforme digitali, come Twitter, con differenti scopi:

- Ricostruire le interazioni online nei social network, utilizzando gli account utente come nodi
- Arricchire la rete con metadati sui nodi ottenuti attraverso tecniche automatizzate o semiautomatizzate
- Analizzare le reti sia a livello individuale che di comunità

Una simile procedura è ben consolidata in discipline come le scienze delle reti sociali e le scienze sociali computazionali (Stella, Ferrara & De Domenico 2018), per via dell'ipotesi che il contenuto raccolto tramite i filtri (per esempio attraverso la scelta di hashtag e parole chiave) sia rappresentativo del fenomeno in esame. In questo contesto, le comunità caratterizzate da opinioni polarizzate, note come echo chambers, e il loro ruolo nelle dinamiche umane come la diffusione della menzogna e dell'odio (Lima et al. 2018) o la centralizzazione delle informazioni (Sacco et al. 2021), sono ancora soggette a numerose ricerche (Del Vicario et al. 2016a; Del Vicario et al. 2016b).
Questi metodi consentono di valutare ruolo ed importanza percepita di un individuo sia come autore che come bersaglio, e di identificare e caratterizzare comunità online con peculiari dinamiche di odio (Uyheng & Carley 2021). Le medesime metodologie sono recentemente state impiegate per produrre una meta-analisi della ricerca scientifica sull'odio online, registrando tendenze, hotspot e comunità di ricerca che sono efficaci nelle indagini su questi argomenti (Waqas et al. 2019).

Metodi basati sull'analisi di dati classica e piccole basi di dati

Oltre ai sofisticati approcci delle scienze sociali computazionali, possono essere impiegati anche approcci più tradizionali basati su metodi classici e analisi di basi di dati più piccole per evidenziare aspetti diversi e complementari dell'incitamento all'odio online. D'altronde, i metodi tradizionali di analisi trasversale e longitudinale possono essere d'aiuto nel comprendere diversi aspetti specifici del fenomeno. Inoltre, la ricerca qualitativa può essere molto utile per fornire un resoconto dettagliato rispetto a come si formano le convinzioni ed i valori, a come si sviluppa l'interazione sociale, e così via.
Alcune di queste analisi sono particolarmente utili per comprendere come le caratteristiche specifiche degli ambienti digitali possano incanalare l'esposizione di alcuni soggetti verso determinati tipi di contenuti provocatori o offensivi. Il tempo trascorso online e la disinvoltura nell'ambito digitale (ovvero, la sicurezza di sé sull'accesso e la navigazione nei contenuti online) possono determinare una maggiore esposizione a contenuti dannosi.
Tale disinibizione online è spesso associata alla giovane età e all'elevata alfabetizzazione digitale; dal momento che la giovane età può rendere i soggetti meno in grado di riconoscere e contrastare forme sottili di manipolazione, non è sorprendente che le piattaforme digitali offrano spesso significative opportunità di reclutamento di giovani istigatori all'odio (Harriman et al. 2020). Le strategie di vittimizzazione si concentrano frequentemente su questioni assai critiche per i giovani, rendendoli particolarmente suscettibili e fragili e quindi più soggetti a manipolazioni, come l'orientamento sessuale, l'aspetto fisico e l'etnia (Oksanen et al. 2014). L'orientamento sessuale è una questione particolarmente critica tra gli adolescenti LGBTQ+, che molto spesso sono vittime di incitamento all'odio. Le conseguenze di questi fatti possono essere estremamente gravi e comportare l'istigazione a pensieri suicidi, tanto da renderli paragonabili ad altre forme di reato che richiedono uno specifico intervento giudiziario. Purtroppo, la nostra società non è ancora sufficientemente attrezzata per difendere questi soggetti, com'è chiaramente riscontrabile attraverso i sondaggi e le interviste di follow-up (Keighley 2022).

Un recente sondaggio del Pew Research Center ha stabilito che quattro americani su dieci hanno sperimentato qualche sorta di molestie online. Sebbene l'argomento più delicato come pretesto per le molestie sia la politica, la crescente diffusione delle molestie di tipo sessuale e dello stalking desta serie

preoccupazioni (Pew Research Centre 2021a). Nonostante tutti i potenziali benefici, gli ambienti digitali sembrano sempre meno favorevoli al mantenimento delle norme sociali di civiltà e rispetto per le persone, e per via di questa mancanza di giudizio in merito alle complesse conseguenze psico-comportamentali della partecipazione digitale, l'incitamento all'odio può persino reclamare la legittimazione come forma di libertà di espressione (Pew Research Centre 2017). In ogni caso, la maggior parte degli americani si dice preoccupato rispetto alle problematiche di sicurezza personale in connessione alla partecipazione digitale.

D'altra parte, le diverse piattaforme presentano notevoli differenze in termini di gravità e frequenza dei comportamenti abusivi. Nella Figura 6 si possono osservare la distribuzione fra le varie piattaforme degli episodi d'odio e molestie (a sinistra), ed il tasso di molestie ricevute dagli utenti abituali (a destra), rilevati in un'indagine rappresentativa degli utenti americani (ADL 2019). Non sorprende che la maggior parte degli episodi di comportamento abusivo si verifichino sulle piattaforme più affollate. Ci sono però delle piattaforme che, a dispetto delle loro dimensioni relativamente limitate, sono particolarmente adatte a facilitare comportamenti abusivi, fra cui le piattaforme di gaming come Twitch, ma anche le piattaforme di discussione come Reddit, nelle quali le conversazioni tendono facilmente ad assumere caratteristiche istigatorie quando gli argomenti sono controversi. È interessante notare che piattaforme spesso accusate di scambi caustici e contenuti incendiari come Twitter siano a questo proposito relativamente meno pericolose in termini di tassi di molestie degli utenti abituali.

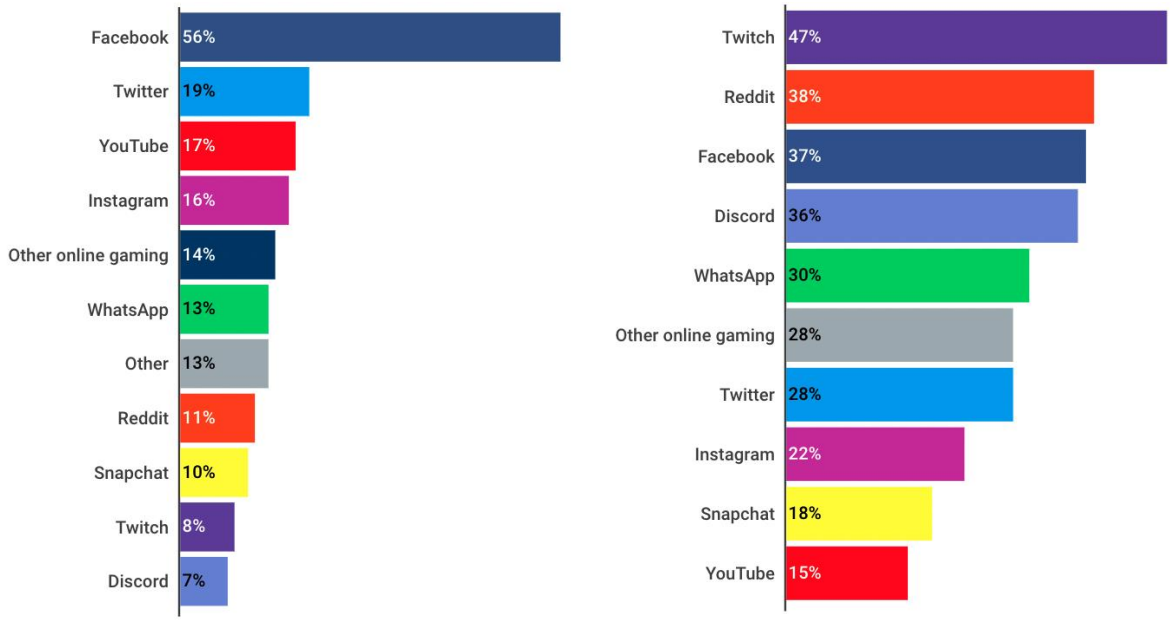

Figura 6: Tracciamento dell'odio e delle molestie online (riquadro di sinistra) e delle molestie agli utenti abituali delle piattaforme (riquadro di destra) misurate da un sondaggio rappresentativo degli americani a livello nazionale, condotto dal 17 dicembre 2018 al 27 dicembre 2018.

Come riportato dal Pew Research Institute, i social media in generale sono il luogo nel quale si riportano più comunemente le molestie (e dov'è dunque più probabile subirne), a prescindere dalla specifica tipologia di comportamento abusivo (Pew Research Centre 2021b), come si vede nella Figura 7. Le differenze con altre risorse digitali sono sorprendenti per quanto riguarda l'incidenza di tali comportamenti; questo fa riflettere sulla necessità di capire meglio come sviluppare forme specifiche di relazioni sociali digitali che possano trasformare i social media da un'area sostanzialmente senza legge in un vero spazio pubblico di discussione che le persone possono esplorare in modo sicuro e che permetta loro d'interagire con ragionevoli aspettative di sicurezza personale e di sviluppare relazioni civili.

Social media sites are the most commonly reported location for people's most recent online harassment encounter, regardless of the type of behavior

Among the 41% of U.S. adults who have personally experienced online harassment, % who say ___ was where their most recent experience occurred

		LOCATION						NUMBER OF LOCATIONS	
		Social media site	Forum/ discussion site	Texting/ messaging app	Online gaming	Personal email account	Online dating site/app	Multiple	One
	Total	75	25	24	16	11	10	41	55
Most recent experience included									
LESS SEVERE BEHAVIORS	Offensive name-calling	77	29	23	18	9	10	42	56
	Purposeful embarrassment	83	30	27	13	12	10	46	52
MORE SEVERE BEHAVIORS	Physical threat	80	37	40	22	19	18	60	36
	Sexual harassment	76	29	46	15	17	28	58	37
	Stalking	79	32	54	15	30	20	65	31
	Sustained harassment	81	44	47	25	26	22	67	29
	Any more severe behaviors	76	31	38	17	18	19	55	41
	Only less severe behaviors	76	23	15	13	6	5	31	66
	Multiple behaviors	81	35	37	19	15	16	57	41
	One behavior	73	21	17	13	8	7	31	65

Note: Multiple responses were allowed for the types of harassing behaviors experienced most recently and the locations those behaviors took place in. Some respondents did not indicate what occurred in their most recent incident. Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Sept. 8-13, 2020.

PEW RESEARCH CENTER

Figura 7. Incidenza differenziale delle molestie online in diversi tipi di risorse digitali. Fonte: Pew Research Center.

La letteratura scientifica in questo ambito sta aumentando a dismisura, ed un numero crescente di evidenze indica una criticità intrinseca dell'istigazione all'odio. È particolarmente importante notare che nelle interazioni tradizionali nel mondo fisico, le persone sono più efficacemente vincolate dai codici culturali e dalla pratica dell'interazione diretta. Le emozioni sociali come il senso di colpa o la vergogna, o la pressione e l'applicazione delle regole di civiltà da parte del prossimo, sono state accuratamente distillate attraverso millenni di evoluzione culturale, sebbene con significative differenze geografiche e culturali nel tempo e nello spazio. Al contrario, nelle interazioni online, non solo lo scambio sociale è spesso anonimo o almeno altamente mediato, ma diventa impossibile utilizzare i classici meccanismi di imposizione sociale come lo sguardo moralizzante, la postura del corpo e persino la disposizione e la progettazione degli spazi pubblici, le persone si sentono più autorizzate a indulgere in molestie e comportamenti abusivi da una robusta sensazione che non saranno ritenute responsabili (Tsesis 2002).

L'introduzione di internet è certamente troppo recente per pensare sia possibile lo sviluppo di un nuovo e specifico gruppo di convenzioni sociali o l'accumulo di un sufficiente capitale di relazioni sociali (Spottswood & Wohn 2020). D'altra parte, la crescente pervasività delle interazioni e della comunicazione digitale suggerisce una particolare urgenza nello sperimentare interventi pilota che possano guidare ed accelerare l'evoluzione culturale nel promuovere la socialità digitale.

L'effetto disinibitorio di cui beneficia l'incitamento all'odio online e la menzione ossessiva di tale fenomeno da parte dei media tradizionali e dei social media che cercano di coinvolgere gli utenti digitali facendo leva su emozioni negative (Jalonen 2014) e contenuti discutibili, possono a loro volta ripercuotersi sulle stesse interazioni nel mondo fisico, mettendo a rischio persino una parte del capitale di relazioni sociali che si è accumulato nel contesto predigitale. A tal proposito, persino media tradizionali e non interattivi possono

diventare amplificatori di attitudini e contenuti che istigano all'odio attraverso i loro stessi articoli (Sakki & Hakoköngäs 2022), raggiungendo un pubblico meno alfabetizzato digitalmente che potrebbe essere attirato nell'interazione online con capacità e preparazione insufficienti. Ciò potrebbe determinare un risultato paradossale: persone di mezz'età o anziane potrebbero diventare a loro volta il facile bersaglio degli istigatori all'odio, che utilizzerebbero più facilmente tutti gli strumenti manipolatori su un pubblico che è spesso disarmato di fronte alle loro strategie. Le differenze nei tassi di alfabetizzazione digitale tra le generazioni stanno anche alimentando nuove forme di conflitto intergenerazionale, ulteriormente esacerbato dalla pandemia (Meisner 2021).

In una società dalle conoscenze avanzate come la nostra, il ruolo rivestito dai social media e dagli istigatori all'odio rappresenta una questione di cruciale importanza nella delegittimazione delle competenze e della professionalità. Ciò è diventato particolarmente evidente nel corso della recente pandemia, durante la quale gli esperti si sono spesso trovati nella scomoda posizione di dover discutere con individui scientificamente incompetenti in possesso di sofisticate capacità di manipolazione, spesso nel contesto dei media di grande diffusione o dei dibattiti sui media digitali, con l'ulteriore preoccupazione di diventare il bersaglio di campagne d'odio coordinate. È un tema molto preoccupante in uno scenario globale in cui la nostra preparazione nel reagire a minacce importanti come una pandemia dipende in buona parte dalla capacità degli esperti di far sentire la loro voce e renderla credibile al grande pubblico (Science Magazine 2022). La natura benigna della pubblica informazione in questo caso determina insidiose situazioni di dilemma sociale in cui molte persone potrebbero avere interesse a sfruttare un simile circolo vizioso, creando e facendo circolare informazioni fuorvianti o false e imponendone la circolazione e la legittimazione attraverso incitamento all'odio online e molestie digitali. Tale situazione crea problemi nuovi, relativamente senza precedenti, di sicurezza personale e "censura invisibile" per gli scienziati, che potrebbero essere incentivati a valutare se la diffusione dei risultati della loro ricerca possa istigare reazioni online che potrebbero minacciare la loro stessa sicurezza e quella della loro famiglia (Nature Magazine 2021).

Osservazioni conclusive e raccomandazione politica

Il primo passo importante dal punto di vista politico è quello di sperimentare interventi che possano frenare le dinamiche sociali in ambito digitale che facilitano, e in una certa misura legittimano, l'incitamento all'odio online e le molestie. A questo scopo, è necessario depolarizzare gli argomenti più dibattuti e controversi e contrastare efficacemente i processi psicologici di dis-umanizzazione delle vittime di abusi online da parte dei responsabili. Sebbene la nostra comprensione della psicologia sociale dei processi digitali sia in forte espansione, è ancora limitata, e non siamo ancora in possesso di strumenti affidabili in termini metodologici e d'intervento per far fronte a questi problemi in modo costante ed efficace. In ogni caso, possiamo almeno contare su una prima base di risultati sistematici che consentano d'iniziare a lavorare alla definizione ed alla messa a punto di possibili protocolli di intervento (Windisch et al. 2021).

Per avere successo, questo nuovo approccio politico deve tuttavia fare affidamento su una complessa integrazione tra le tecniche e le intuizioni che derivano dalla letteratura delle scienze sociali computazionali e quelle che scaturiscono da forme più tradizionali di analisi e valutazione delle politiche, come quelle discusse nell'ultima sezione. Conseguire una buona comprensione di alcuni meccanismi sociopsicologici senza metterli in relazione con le proprietà della meccanica statistica degli ambienti digitali che riguardano modalità di interazione e canali di produzione e circolazione delle informazioni può essere fuorviante, e portare ad individuare raccomandazioni normative non pienamente efficaci. Al contrario, combinando tali informazioni diventa possibile sperimentare interventi che, pur essendo controintuitivi per gli amministratori politici predigitali, possono risultare molto efficaci (Artime et al. 2020).

Le emozioni morali umane sono state d'enorme aiuto per la società umana nello sviluppo di forme complesse e super articolate di cooperazione. Tali emozioni si sono però evolute per far fronte ad ambienti sociali nei quali le relazioni con il prossimo e l'accesso agli estranei erano mediati da complessi meccanismi

rituali il cui proposito era anche mantenere il conflitto sotto un livello ragionevole di controllo. Negli ambienti digitali, la componente rituale è stata rapidamente smantellata, nonostante i nuovi rituali digitali non siano generalmente associati al perseguimento e alla stabilizzazione di norme minime di civiltà (Smith 2022). Le stesse emozioni morali che hanno assicurato la coesione sociale in piccoli gruppi e garantito interazioni ritualizzate all'esterno del gruppo sono ora sostanzialmente responsabili dell'incitamento e della comunicazione d'odio. Oggi, le scienze sociali si trovano pertanto ad affrontare una sfida straordinaria e indifferibile, che non saremo in grado di superare senza una nuova, solida alleanza tra politica, scienza e società civile.

Sezione IV:
AZIONI NORMATIVE A CONFRONTO

"Hate is hate – and no one should have to put up with it."
"L'odio è odio, e nessuno dovrebbe sopportarlo"

Ursula von der Leyen, Presidente della Commissione europea
(State of the Union 2020)

Introduzione

Meno di dieci anni fa, quella relativa alla regolamentazione dei social media era una discussione di nicchia nella maggior parte delle democrazie. Sin dal 2016, politici, studiosi, giornalisti e società civile hanno criticato il modo di agire delle piattaforme di social media perché permettono che l'istigazione all'odio ed altri fenomeni dannosi come la disinformazione, le interferenze esterne nelle consultazioni elettorali, il cyber bullismo e la generale frammentazione della società continuino a perpetrarsi (TGW, 2020).

Sono in molti a collegare la crescente polarizzazione della società, la retorica sempre più aggressiva nel dibattito pubblico e l'aumento del numero di attacchi violenti con l'istigazione all'odio online. Questo spinge i governi a cercare risposte attraverso nuove leggi. Proprio per questo motivo, nel suo discorso sullo stato dell'Unione del 2020, Ursula von der Leyen ha chiesto di includere l'istigazione all'odio nella lista dei crimini contro l'Unione.

Per quanto riguarda la regolamentazione di social media e contenuti digitali, si tratta di un ambito potenzialmente vastissimo, che spazia dalle controversie nei dibattiti politici fino a contenuti indubbiamente illegali e pericolosi. Non sorprende dunque che i vari paesi stiano sperimentando le più disparate regolamentazioni per contrastare l'istigazione all'odio online, data la notevole difformità nelle definizioni (si veda la Sezione I) e le difficoltà nel rilevare e valutare l'istigazione all'odio stessa (si veda la Sezione III).

Considerando che stiamo affrontando la terza grande rivoluzione nella storia delle comunicazioni, considerando che la *scrittura* e la *stampa* siano state le prime due, è importante non dimenticare che ci sono voluti oltre mille anni per "addomesticare" la scrittura e quasi duecento per regolamentare la stampa attraverso lo sviluppo di nuove regole come i diritti d'autore e le leggi contro la diffamazione.

La regolamentazione dei contenuti illegali non è infatti alcunché di nuovo da parte dei governi, che affrontano simili problemi da moltissimo tempo. La maggiore difficoltà nell'ambito digitale deriva dalla mancanza di un'entità legale che possa essere ritenuta responsabile del contenuto che viene pubblicato. Le piattaforme di social media hanno fatto quanto era in loro potere per distinguersi dai media tradizionali (che sono rigorosamente regolati) e posizionarsi come intermediari o distributori di contenuti altrui.

In un ecosistema digitale, è necessario sviluppare idee nuove su come gestire l'incitamento all'odio online che possano fornire soluzioni efficaci senza essere in conflitto con i principi europei fondamentali come la libertà di parola. Oggi, questa è una delle sfide più importanti che i governi debbono affrontare, spesso partendo da cornici legislative proprie dell'era della stampa e della televisione (TWG 2020).

Questo capitolo ha lo scopo di fare una panoramica delle diverse proposte legislative allo studio per l'istigazione all'odio online, e fornire consigli su potenziali provvedimenti legislativi e pratiche efficaci.

La situazione

La generale considerazione di internet e le piattaforme di social media come un vantaggio per la democrazia è repentinamente cambiata nel 2013, quando lo Stato islamico di Siria e Iraq (ISIS) ha scioccato il mondo postando video violenti di esecuzioni e decapitazioni, allo scopo di utilizzare odio e violenza per trasformarli in discordia e paura.
L'efficacia dell'istigazione all'odio online (Online Hate Speech – OHS) nell'aumentare la frammentazione della società e la violenza ha portato il problema della regolamentazione dei social media al centro del dibattito politico, sollevando interrogativi importanti su come i governi dovrebbero agire.

Poiché l'odio online si è diffuso nel mondo reale attraverso azioni concrete e morte, la Commisisone europea ha proposto di includere l'istigazione all'odio ed i crimini d'odio nella lista dei reati inclusi nell'articolo 83 del Trattato sul Funzionamento dell'Unione europea (TFEU 2009)[3], che si articola così:

> 1. Il Parlamento europeo e il Consiglio, deliberando mediante direttive secondo la procedura legislativa ordinaria, possono stabilire norme minime relative alla definizione dei reati e delle sanzioni in sfere di criminalità particolarmente grave che presentano una dimensione transnazionale derivante dal carattere o dalle implicazioni di tali reati o da una particolare necessità di combatterli su basi comuni.
>
> Dette sfere di criminalità sono le seguenti: 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.
>
> In funzione dell'evoluzione della criminalità, il Consiglio può adottare una decisione che individua altre sfere di criminalità che rispondono ai criteri di cui al presente paragrafo. Esso delibera all'unanimità previa approvazione del Parlamento europeo

Questo permetterà alla Commissione, in un secondo momento, di rafforzare il Quadro legislativo per contrastare l'istigazione all'odio ed i crimini d'odio nell'intera Unione europea.

Fino ad un paio d'anni fa, si riteneva che le leggi sui crimini d'odio fossero sufficienti per risolvere il problema. Purtroppo, come in molte altre aree della legislazione, le leggi sono inadeguate anche a causa di viralità e la velocità con le quali l'istigazione all'odio si diffonde ed al significativo impatto che essa può avere non solo sull'individuo ma sull'intera società. Le evidenze mostrano che le norme sociali e morali che regolano efficacemente il mondo reale hanno effetti molto limitati nel contrastare l'odio online (Lowry et al. 2016). Queste sono le ragioni:

a) L'istigazione all'odio online è, per sua stessa definizione, un fenomeno digitale che riguarda persone che si concentrano in "echo chambers" (Himelboim et al. 2013), che si possono definire come sistemi chiusi nei quali stereotipi e pregiudizi si amplificano, rinforzandosi fra loro (Sunstein 2017).

b) C'è una generale accettazione acritica delle informazioni che si trovano su internet, il che rafforza ulteriormente il pregiudizio e l'identità sociale (Hall 2013).

[3] Treaty on the Functioning of the European Union (2009) tradotto in italiano

c) Gli odiatori digitali quando sono online sperimentano un "effetto disinibitorio" (Suler 2004) che diminuisce il loro senso di responsabilità e moderazione e abbatte le barriere sociali che nel mondo reale moderano i comportamenti aggressivi e radicali (ElSherief et al. 2018). Questo è ulteriormente esacerbato dalla maggiore percezione di anonimità.

Comunque, la novità del fenomeno, la difficoltà nell'accordarsi su una definizione e su come misurare l'istigazione all'odio online rende difficile formulare proposte che combattano efficacemente l'istigazione all'odio garantendo il rispetto dei valori fondamentali della società, come la libertà di espressione. Qui sotto, sono analizzate alcune delle iniziative che sono state sviluppate per gestire il problema, a partire da quelle che più tradizionalmente si concentrano sul contenuto fino a quelle più originali e sperimentali.

A. GERMANIA – Legge tedesca sui network (NetzDG law)

La legge tedesca sui network (NetzDG law)[4], entrata in vigore nel gennaio 2018, rappresenta un test chiave contro la lotta all'istigazione all'odio su internet. La nuova legge non ha imposto alcun obbligo alle piattaforme social media ma ha introdotto multe elevate per chi non rispetta gli obblighi legali esistenti.

Secondo la legge, un social media può essere multato fino a 50 milioni di euro per l'infrangimento dei propri obblighi. La legge sui network è piuttosto controversa ed è stata criticata per la sua incostituzionalità, particolarmente in merito alla libertà di parola ed al fatto che complica le cose per gli utenti rispetto a come effettuare reclami.

Per giunta, ci sono dubbi sul fatto che la NetzDG abbia ottenuto risultati significativi per il suo obiettivo dichiarato di prevenire l'istigazione all'odio. Le evidenze suggeriscono che le piattaforme riescano in parte a eludere la regolamentazione. Nel caso esemplare di Facebook, la piattaforma ha reso difficile presentare reclami sulla base NetzDG, preferendo citare i loro oscuri standard di comunità per rimuovere ampie quantità di contenuto. Da quando la legge è entrata in vigore, molti partiti politici hanno proposto di emendarla o rigettarla completamente.

La NetzDG tedesca ha evidenziato la natura difficoltosa di monitorare i contenuti digitali e la necessità di un approccio sistemico e efficiente alla legislazione. Gli sforzi per trovare soluzioni più innovative si basano più su proposte sistemiche che su regolamentare i contenuti.

B. REGNO UNITO - Online Safety Bill

L'Online Harms White Paper, pubblicato nel 2019, sottolinea alcuni dei problemi e delle difficoltà che devono essere contrastati per fare dell'UK "il posto più sicuro per essere online"[5]. Il termine "danno online" in questo contesto significa "comportamento online che può ferire una persona fisicamente o emotivamente. Si potrebbe trattare informazioni dannose pubblicate online o informazioni inviate a una persona".

Una bozza dell'Online Safety Bill è stata inclusa nel discorso della regina dell'11 maggio 2021[6] con l'intento di proporre una nuova regolamentazione per rendere l'ambiente digitale più sicuro. Specifica regole precise rispetto ai contenuti illegali ma suggerisce anche una cornice legale per contrastare contenuti legali ma dannosi.

---

[4] https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html
[5] https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper
[6] https://www.gov.uk/government/publications/draft-online-safety-bill

Contenuti illegali: è richiesto a tutte le piattaforme interessate dalla legge di rimuovere proattivamente i contenuti illegali prioritari. Questo include i contenuti legati a crimini d'odio insieme a contenuti dannosi per donne e ragazze, come quelli connessi a immagini sessuali (per esempio la pornografia di vendetta e quella estrema), e legiferazioni su molestie e cyberstalking. Oltre ai i reati prioritari, tutte le piattaforme dovranno garantire di avere sistemi e procedure efficaci per rimuovere rapidamente gli altri contenuti illegali una volta che questi siano segnalati o che le piattaforme li rilevino. È incluso il cyberflashing, che è stato dichiarato illegale proprio attraverso la legge in oggetto. Ciò contribuirà a impedire che i criminali utilizzino Internet come arma con cui minacciare e commettere violenze contro le donne.

Legale ma dannoso:. Le grandi aziende di social media dovranno mantenere le promesse fatte agli utenti, intervenendo contro i contenuti dannosi vietati dai loro termini di servizio. Il Regno Unito stabilirà, nella legislazione secondaria, una serie di categorie prioritarie per i contenuti "legali ma dannosi", tra cui è molto probabile che vi siano abusi misogini

Come nell'esempio precedente, l'Online Harms Bill si concentra su un modello di moderazione dei contenuti per affrontare l'incitamento all'odio e i contenuti dannosi online. Vista la recente introduzione di questo disegno di legge non è ancora possibile constatare l'efficacia di tale azione.

C. UNIONE EUROPEA - L'incitamento all'odio e l'UE

Il 9 dicembre 2021, la Commissione ha pubblicato la comunicazione 'A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime' che mira a stimolare una decisione del Consiglio che estenda ai crimini ispirati dall'odio e all'incitamento all'odio l'attuale elenco dei cosiddetti "crimini dell'UE" come stabilito dall'articolo 83 del TFEU. Tale decisione consentirebbe alla Commissione, in una fase successiva, di rafforzare il quadro giuridico sulla lotta all'incitamento all'odio e ai crimini ispirati dall'odio in tutta l'UE.

L'incitamento all'odio illegale è definito nel diritto dell'UE ai sensi della decisione quadro sulla lotta contro determinate forme ed espressioni di razzismo e xenofobia mediante il diritto penale[7] come segue:

'Incitamento all'odio'
Sono punibili come reati alcuni comportamenti, come di seguito delineati:

- *istigazione pubblica alla violenza o all'odio contro un gruppo di persone o un membro di tale gruppo definito sulla base di razza, colore, discendenza, religione o credo, o origine nazionale o etnica;*
- *il reato sopra menzionato, se realizzato mediante la diffusione o distribuzione pubblica di opuscoli, immagini o altro materiale;*
- *condonare, negare o banalizzare pubblicamente crimini di genocidio, crimini contro l'umanità e crimini di guerra come definiti nello Statuto della Corte penale internazionale (articoli 6, 7 e 8) e crimini definiti nell'articolo 6 della Carta del Tribunale militare internazionale, quando il comportamento è posto in essere in modo tale da incitare alla violenza o all'odio contro tale gruppo o un membro di tale gruppo.*

È punibile anche l'istigazione, il concorso o la complicità alla commissione dei suddetti reati.
In relazione ai reati elencati, i paesi dell'Unione europea devono accertarsi che siano punibili con:

- *sanzioni effettive, proporzionate e dissuasive;*
- *la reclusione, non inferiore nel massimo termine ad almeno un anno.*

Per quanto riguarda le persone giuridiche, le punizioni devono essere effettive, proporzionate e dissuasive e devono consistere in sanzioni penali o non penali. Inoltre, le persone giuridiche possono essere punite con:
- *esclusione dal diritto a benefici o aiuti pubblici;*

[7] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV:l33178

- *interdizione temporanea o permanente dall'esercizio o dall'attività commerciale;*
- *l'imposizione del controllo giudiziario;*
- *ordinanze di liquidazione giudiziaria.*

L'avvio di indagini o azioni penali per reati razzisti e xenofobi non deve dipendere dalla denuncia o dall'accusa della vittima.

'Reati d'odio'

In ogni caso, la motivazione razzista o xenofoba deve essere considerata come circostanza aggravante o, in alternativa, il giudice deve essere autorizzato a tenerne conto nella determinazione delle sanzioni da applicare.

L'incitamento all'odio e i crimini ispirati dall'odio hanno registrato un forte aumento in tutta Europa e sono diventati un fenomeno particolarmente grave e preoccupante, sia offline che online. È necessaria un'azione comune fra i Paesi dell'UE per affrontare questa sfida a livello europeo. Al momento non esiste però una base giuridica per rendere l'incitamento all'odio e i crimini ispirati dall'odio a livello dell'UE un reato federale. L'elenco esistente dei reati federali nel trattato sul funzionamento dell'Unione europea (TFUE) deve essere esteso per garantire norme minime comuni su come definire i reati e le sanzioni applicabili in tutti gli Stati membri dell'UE. La presente iniziativa è il primo passo nel processo di ampliamento dell'elenco dei reati dell'UE. Il successivo passaggio consisterebbe nell'approvazione dell'iniziativa da parte degli Stati membri, in modo che la Commissione possa poi presentare una proposta legislativa.

In ordine temporale, si tratta dell'ultima iniziativa sviluppata a livello europeo, parte di una serie che comprende:

- **2016: Codice di condotta dell'UE per contrastare l'illecito incitamento all'odio online** [8]

La Commissione europea ha lanciato il proprio Codice di condotta nel maggio 2016 insieme a quattro importanti società IT (Facebook, Microsoft, Twitter e YouTube), nel tentativo di rispondere alla proliferazione di incitamento all'odio razzista e xenofobo online. Nel corso del 2018 Instagram, Snapchat e Dailymotion hanno aderito al Codice di condotta, Jeuxvideo.com ha aderito a gennaio 2019 e TikTok a settembre 2020. Il 25 giugno 2021 anche LinkedIn ha annunciato la sua partecipazione.

Lo scopo del Codice è garantire che le richieste di rimozione dei contenuti siano gestite rapidamente. Quando le aziende ricevono una richiesta di rimozione dalla loro piattaforma online di contenuti ritenuti illegali, valutano tale richiesta rispetto alle loro regole e alle linee guida comunitarie e, ove necessario, alle leggi nazionali, che recepiscono il diritto dell'UE sulla lotta al razzismo e alla xenofobia. Le aziende si impegnano a riesaminare la maggior parte di queste richieste in meno di 24 ore, ed a rimuovere il contenuto, se necessario, nel rispetto del principio fondamentale della libertà di parola. Ad oggi, otto società hanno aderito al Codice, in particolare Facebook, YouTube, Twitter, Microsoft, Instagram, Dailymotion, Snapchat e Jeuxvideo.com.

Il Codice di condotta viene valutato attraverso il monitoraggio da parte di una rete di organizzazioni della società civile, dislocate in diversi paesi dell'UE. Utilizzando una metodologia comunemente concordata, queste organizzazioni testano il modo in cui le aziende IT applicano il Codice nella pratica. Questo si declina nell'inviare regolarmente alle società IT richieste di rimozione dei contenuti, registrando poi quanto tempo impiegano le società IT a valutare la richiesta, quali sono i risultati e il feedback che ricevono dalle aziende.

[8] https://ec.europa.eu/info/sites/default/files/code_of_conduct_hate_speech_en.pdf

- **2021: Relazione del Parlamento europeo sul rafforzamento della democrazia, della libertà e del pluralismo dei media nell'UE: l'uso indebito di azioni di diritto civile e penale per mettere a tacere giornalisti, ONG e società civile (2021/2036(INI))**[9]

La sezione di questo documento che riguarda l'incitamento all'odio afferma "l'importanza di standard europei comuni e un approccio coordinato per affrontare l'incitamento all'odio, in particolare nell'ambiente online".

È spiegato che "negli ultimi anni l'incitamento all'odio e la discriminazione nei media, sia online che offline, così come la cyberviolenza, sono diventati sempre più diffusi nei confronti di giornalisti, ONG, accademici, difensori dei diritti e altri attori della società civile, compresi quelli che difendono i diritti LGBTIQ, questioni di uguaglianza di genere, religione o credo, minacciando così la libertà dei media, la libertà di espressione, informazione e riunione, nonché la sicurezza pubblica; si ricorda che l'incitamento all'odio online può incoraggiare alla violenza offline; si ricorda inoltre la necessità di promuovere il codice di condotta della Commissione per contrastare l'incitamento all'odio illegale online, e si sottolinea che le giornaliste devono affrontare le stesse pressioni delle loro colleghe per questioni relative ai contenuti, ma sono più spesso vittime di violenze e molestie sessuali".

D. SVEZIA – L'agenzia di "Difesa Psicologica"[10]

Nel febbraio del 2022, la Svezia ha annunciato la creazione di una nuova agenzia per affrontare la minaccia, percepita o reale, portata dall'influenza indesiderata della Russia sulle imminenti elezioni nazionali svedesi. Questa mossa controversa riprende le tattiche dell'era della guerra fredda; pur non avendola concepita per far fronte all'incitamento all'odio, il dipartimento che l'ha istituita lavora per sviluppare e rafforzare la capacità complessiva della società di difesa psicologica.

A seguito dello sfruttamento dell'odio come arma nell'influenzare il comportamento delle persone, ad esempio durante l'assalto a Capitol Hill, la campagna per la Brexit o le due ultime elezioni presidenziali statunitensi, la Svezia ha deciso che la questione doveva essere affrontata da una prospettiva di sicurezza nazionale anziché da una " prospettiva basata sui diritti".

L'idea è di fornire sostegno alla popolazione, alle autorità, ai comuni, ai media, alle organizzazioni di volontariato di difesa e alla società civile in generale, nonché lavorare per un maggiore coordinamento tra questi attori. Un loro importante compito è essere responsabili dell'istruzione, dell'esercizio e dello sviluppo delle conoscenze, che in effetti include organizzare e disseminare la ricerca effettuata nell'ambito della difesa psicologica.

Lo scopo è creare resilienza nella società contro lo sfruttamento delle informazioni e della sfera pubblica da parte di attori stranieri, ma anche affrontare alcuni degli ostacoli principali in tema di incitamento all'odio.

Raccomandazioni e insegnamenti appresi

Una notevole quantità di idee, iniziative legislative e regolatorie sono emerse negli ultimi dieci anni per combattere l'incitamento all'odio online. Alcune si sono concentrate sul problema dal punto di vista dei diritti individuali, altre da quello dei crimini d'odio; qualcuna è partita dalla prospettiva del pluralismo dei media, altre da quella della sicurezza nazionale. L'eterogeneità di questi approcci evidenzia gli effetti divergenti dell'aumento dell'incitamento all'odio sulle nostre società e la complessità di sviluppare una

[9] https://www.europarl.europa.eu/doceo/document/A-9-2021-0292_EN.pdf
[10] https://www.mpf.se/en/

soluzione efficace. Anche se molte di queste iniziative sono molto nuove e il loro successo nell'affrontare efficacemente i fenomeni è ancora da valutare, ecco alcuni insegnamenti che se ne possono trarre:

a) La necessità di un approccio sistemico
Una parte delle azioni legislative è stata indirizzata al controllo di contenuti specifici e a garantire che le piattaforme rimuovano i contenuti caratterizzati come "incitamento all'odio". Ciò si è rivelato difficile, poiché le definizioni e la comprensione di ciò che intendiamo per incitamento all'odio non sono uguali per tutti. Quando le piattaforme hanno rimosso i contenuti, inoltre, non pare che questo abbia avuto alcun effetto generale sul fenomeno più ampio dell'incitamento all'odio. Infatti, durante il COVID-19 l'incitamento all'odio ha raggiunto nuovi livelli e sembra che la polarizzazione della società sia in continuo aumento. Inoltre, questo tipo di azioni solleva la questione di chi decide cosa dovrebbe essere censurato e cosa non dovrebbe esserlo.

È imperativo ricordare che la libertà di parola è un principio fondamentale della democrazia e deve essere protetta.

b) Problemi di autoregolamentazione
La pratica ha finora evidenziato l'insufficienza nell'autoregolamentazione da parte delle piattaforme. Certamente è difficile garantire velocità e coerenza attraverso questo metodo, poiché i contenuti si muovono rapidamente e in modo virale attraverso i social network. È pure molto difficile valutare l'effetto reale di queste azioni poiché l'accesso ai dati per la maggior parte delle piattaforme è limitato.

È opportuno notare che, fino ad ora, l'autoregolamentazione delle piattaforme si è concentrata principalmente sulla moderazione dei contenuti invece di guardare più in profondità gli effetti negativi delle scelte strutturali delle piattaforme, come quelle evidenziate in merito al pulsante "retweet" di Twitter. Uno degli ingegneri di Twitter che ha lavorato allo sviluppo di questa funzione, in seguito si è pentito di averla sviluppata ed ha affermato che "potremmo aver appena consegnato un'arma carica a un bambino di quattro anni" (Haidt 2022).

c) L'importanza di chiarire gli intenti
Come si può vedere sopra, sono state proposte differenti soluzioni che affrontano diverse sfaccettature del problema dell'incitamento all'odio. Come sottolineato nei capitoli precedenti, l'incitamento all'odio ha una serie di attributi ed effetti diversi che preoccupano gli amministratori politici. È importante definire il proprio obiettivo quando si sviluppano proposte legislative. Può darsi che, invece di una legislazione unica per l'incitamento all'odio, sia più utile mettere in atto una varietà di proposte legislative, diverse ma complementari che affrontino i diversi aspetti del fenomeno (vedere sotto).

d) Flessibilità ed adattabilità
Poiché la tecnologia si sviluppa a una velocità sempre più rapida, è imperativo sviluppare qualsiasi proposta legislativa o normativa tenendo presente che dovrebbe essere in grado di adattarsi alle mutevoli circostanze tecnologiche, sociali ed economiche. Sarebbe inoltre utile, data la scarsa conoscenza dell'ambito digitale e i ripetuti fallimenti di tentativi di regolamentare l'ecosistema dei media online, effettuare valutazioni periodiche e a campione per garantire che la proposta raggiunga l'obiettivo desiderato

Raccomandazioni per amministratori e politici

L'era digitale ha cambiato drasticamente la natura dello spazio pubblico. Sono emersi nuovi, potenti attori, che lo controllano in quanto hanno accesso agli strumenti più efficaci. Come questo rapporto evidenzia, sebbene lo studio dei meccanismi che modellano i social media sia agli inizi, questo fenomeno avrà un

impatto duraturo sulla società. C'è crescente consenso tra gli studiosi sul fatto che l'attuale ecosistema dell'informazione, incluso ovviamente l'incitamento all'odio, possa essere una minaccia alla coesione sociale e al nostro stile di vita democratico. Che ciò oggi sia chiaro agli amministratori politici di tutta Europa è evidente dall'importanza che gli è stata data fra le priorità politiche delineate dalla Commissione von der Leyen. La questione non sembra più essere SE, ma COME. La seguente lista è un tentativo di evidenziare alcune delle domande e delle priorità che gli amministratori politici dovrebbero tenere in mente nei prossimi anni:

1) Organizzare e regolare l'ecosistema dei media digitali online.

L'incitamento all'odio e la prevalenza di questo fenomeno in ambito digitale sono collegati al cambiamento dell'ecosistema informativo emerso dall'arrivo delle tecnologie digitali. Vista l'importanza dell'informazione per la società, un ecosistema diverso sarebbe essenziale per preservare e rafforzare la resilienza della società e promuovere le caratteristiche ed i valori della comunicazione aperta e dello scambio di idee che sono così vitali per la democrazia, come affermato dal presidente Barack Obama (Stern 2022). Come dimostrato da moltissime ricerche, un popolo che beneficia di una sana diversità di opinioni ed esperienze può rispondere in modo intelligente a domande difficili; questa capacità è tuttavia compromessa quanto più una società diventa polarizzata.

L'ecosistema della comunicazione gioca un ruolo fondamentale per ogni società. Oggi, il potere si sviluppa anche sui social media. Ciò crea la necessità di regolamentare questo nuovo spazio per garantire che gli aspetti positivi dei social media vengano rafforzati, invece di trasformarli in un nuovo mezzo per manipolare le società e imporre interessi specifici. Esistono molti esempi di come questo possa avvenire, come lo scandalo Facebook-Cambridge Analytica, ma anche il recente referendum irlandese.

È importante ricordare che il panorama dei media tradizionali è uno dei più regolamentati proprio per salvaguardare l'ecosistema dell'informazione, nel migliore interesse della società. Esistono regole diverse per i media tradizionali e per i nuovi media; le scelte che oggi facciamo su come regolamentare l'ambiente dei nuovi media saranno responsabili del modo in cui il nostro panorama politico si svilupperà nei decenni a venire.

Diversi studiosi hanno ipotizzato che per cambiare il modo di organizzare e regolare l'ecosistema dell'informazione, dovremmo cambiare il modo in cui affrontiamo la domanda stessa: modificare la narrativa. È innegabile che l'infrastruttura di comunicazione sia fondamentale nel plasmare la società, e che uno spazio pluralistico e trasparente, che consente ad una moltitudine di opinioni e prospettive di far breccia nei cittadini, sia una pietra angolare del nostro modo di organizzare la società. Attualmente, lo spazio in cui il pubblico si confronta con l'informazione è invece proprietà privata di conglomerati globali stranieri con i propri interessi e programmi. L'Europa dovrebbe pensare a come lo spazio pubblico potrebbe tornare a diventare un "bene pubblico", indipendente dagli interessi dei media, delle aziende e del governo cui si possa affidare il ruolo di preservare la diversità sociale che è essenziale per la democrazia.

2) Proporre regole di accesso ai dati più trasparenti e democratiche per garantire la piena tutela del diritto alla privacy e consentire il massimo utilizzo delle possibilità scientifiche e educative che derivano dai big data.

L'attuale organizzazione dell'ecosistema dei media crea un oligopolio ristretto e opaco da parte delle società che controllano i dati. Le limitazioni nell'accesso ai dati da parte di società private che oggi controllano i dati dell'Europa rendono difficilissimo valutare correttamente i meccanismi che modellano i social media e il loro impatto sulla società. Valutare le minacce portate dall'incitamento all'odio, dalle "guerre dell'informazione" e dalle altre pratiche manipolative è quasi altrettanto difficile. Tale questione solleva anche importanti questioni morali ed etiche. Misure legislative efficaci dovrebbero proporre nuovi modi di considerare la proprietà dei dati, che il GDPR ha già introdotto nell'arena normativa dell'UE e cui si sta lavorando nella costruzione del prossimo Data Act. Più che con soluzioni tecnologiche specifiche, le società e i governi

dovrebbero elaborare principi che ispirano la regolamentazione, per poi valutare l'applicazione di tali politiche rispetto ai principi guida, assicurandosi che non vengano traditi nella pratica e che non vi siano scappatoie chiaramente sfruttabili per aggirarli. Si tratta di un processo iterativo di tentativi, errori e aggiustamenti, il cui punto di partenza dovrebbe essere una dichiarazione d'ispirazione umanista sui diritti degli individui rispetto ai dati che producono.

In questo contesto, è fondamentale che le autorità pubbliche a livello nazionale e dell'Unione europea contribuiscano a salvaguardare gli spazi in cui dati ed elaborazioni possono essere organizzati come beni comuni in continua evoluzione, ricordando che il patrimonio comune non si configura solo nei dati, ma anche nell'ecosistema relazionale attorno ad essi ed al fatto che questo ecosistema patteggi e condivida pratiche e regole. Ciò è possibile solo a fronte di importanti investimenti per supportare le persone, le piccole imprese e le pubbliche amministrazioni a prendere coscienza dei dati che producono, e del fatto che a volte non possono essere a conoscenza né dei dati che producono né delle modalità di utilizzo degli stessi; si potrà realizzare se siamo in grado di trovare modi facili, convenienti e coinvolgenti per trasformare i dati in informazioni fruibili, e farvi seguito con azioni concrete. In effetti, è difficile immaginare una soluzione alla disinformazione e ai suoi effetti sulla società senza una profonda riflessione sulle architetture sociali e sui processi che portano a questo tipo di fenomeni.

3) Proporre alternative all'attuale modello di business.

L'attuale modello di business dei nuovi media è basato sulla pubblicità e implica una costante battaglia per l'attenzione, favorendo la sopravvivenza dei soggetti più rabbiosi. L'escalation militare, simile a una guerra fredda, di algoritmi, BOT, IA e altri strumenti e tecniche di "guerra informatica" riflette bene le insidie del nostro attuale ecosistema dei media digitali online: un mondo ansioso e polarizzato, incapace di affrontare la complessità della società. Questi sono fra i fattori fondamentali che rendono redditizio l'incitamento all'odio online.

"Il finanziamento di alternative decentralizzate alle attuali piattaforme mediatiche potrebbe essere molto più efficace che regolamentarle o cercare di prendere accordi con esse. Il decentramento delle piattaforme sociali dovrebbe essere incoraggiato da politiche attive, che porterebbero anche al superamento dei monopoli attuali. È indispensabile elaborare un modello di business in grado di aprire un mercato che attualmente non esiste. Accadrà in ogni caso, prima o poi, ma se la Commissione potesse accelerare questo processo, sostenendo e stimolando qualcosa che pare ancora lontano dall'essere creato dal mercato stesso, ciò costituirebbe una mossa strategica fondamentale. Per farlo, abbiamo bisogno di tutta la saggezza di cui siamo capaci, questa è davvero la migliore risorsa che possiamo mobilitare in Europa: il nostro senso di libertà e socialità, la nostra capacità di integrare le differenze, la nostra ricerca aperta e il pensiero critico. Il punto è proprio quello di fare delle piattaforme social decentralizzate un ingrediente chiave della cultura europea, di farne l'oggetto di un "desiderio sociale". C'è così tanto spazio, in questo senso, per implementare la politica dell'UE: le auto elettriche sono incentivate; quindi, perché non le piattaforme social decentralizzate?"[11]

4) Sviluppare un approccio europeo all'Intelligenza Artificiale (AI).

Oggi, la capacità dell'IA di contraffare i media è ampiamente maggiore di quella della stessa IA di identificare la contraffazione di tali media. Come evidenziato sopra, l'IA è da tempo attiva nell'ambito dei media digitali e sta già avendo un enorme impatto sull'economia, sull'ecosistema dei media e sulla società. È quindi estremamente urgente concepire un approccio europeo all'IA che sia basato su valori antropici ed umanistici.

[11] Expert interview with Piero Molino from the SMART Report.

5) Incoraggiare l'intelligenza collettiva e la collaborazione intersettoriale.

Soluzioni innovative dovrebbero essere progettate per poter affrontare la crescente divisione tra i cittadini, il bisogno di creare le condizioni per la solidarietà, l'empatia e la formazione di un ecosistema relazionale inclusivo e di alta qualità nella società digitale. Tali soluzioni possono essere pensate soltanto combinando diverse discipline e tecnologie. Sarebbe bene partire con l'ideazione e la progettazione di nuovi processi educativi, come professioni e ruoli per cittadini e professionisti che si dedichino a "riconnettere" la società, per ottenere maggiore qualità ed inclusività negli ecosistemi relazionali nelle città e nelle regioni. Per mezzo di azioni sociali culturali, utilizzando dati e algoritmi con modalità trasparenti, creando alternative per le attuali industrie "estrattive" dei dati informatici che aprano alla sfera pubblica, in modo che la società possa concordare regole, regolamenti e protocolli in modo aperto, inclusivo, accessibile agli esperti e comprensibile a tutti.

Le collaborazioni e le interazioni tra scienze, tecnologia, arte e design dovrebbero portare opportunità di immaginare innovazione sociale nella sfera pubblica, nella quale i risultati dell'innovazione scientifica e tecnologica non solo siano condivisi, ma fungano anche da stimolo per un approccio costruttivo ed inclusivo e favoriscano processi collettivi di costruzione di scenari futuri in grado di portare migliori relazioni, informazioni, comunicazione, conoscenze e occasioni di scambio.

Questo rapporto mostra che non sarà sufficiente sostenere gli strumenti tradizionali per contrastare la diffusione dell'incitamento all'odio e salvaguardare i nostri sistemi democratici. Il problema è molto più profondo e capillare, ed è necessario un approccio sistemico per comprendere i meccanismi che modellano i social media e il loro impatto sulla società.

CONCLUSIONI

L'analisi contenuta in questo rapporto consente di leggere la complessità del fenomeno. Le lezioni che si imparano analizzandone le componenti sono state brevemente riportate nelle sezioni dedicate ai vari argomenti. Non occorre un'ulteriore conclusione, ma la descrizione di un percorso necessario, qualsiasi sia l'orientamento della policy che si vorrà adottare. E si tratta di un percorso di ricerca.

La comprensione del fenomeno dell'istigazione all'odio online richiede un approccio interdisciplinare, consapevole delle opportunità offerte dall'analisi delle reti, dalla scienza dei dati, dalla fisica della complessità sociale: le regolarità che la lettura matematicamente avanzata dei dati consente di vedere emergere sono una dimensione irrinunciabile dell'analisi di miliardi di messaggi, di relazioni tra persone e di storie individuali. E nello stesso tempo, tutto questo non può che arricchirsi con la collaborazione di chi conduce le indagini sociologiche, psicologiche e antropologiche che servono a leggere nelle storie degli umani coinvolti una diversa dimensione della complessità. Infine l'analisi dei media considerati l'ecosistema nel quale si sviluppano le dinamiche culturali può condurre a comprendere il ruolo delle interfacce e degli algoritmi che governano la relazione tra le persone e le piattaforme, così come può indirizzare a studiare le grandi narrative che compongono il quadro interpretativo con il quale le persone comprendono la realtà e si fanno una ragione delle sue prospettive.

La strategia di ricerca deve insomma essere disegnata in una chiave interdisciplinare e può diventare esemplare di uno studio della complessità intesa non come un'astrazione ma come la reale dimensione nella quale si svolge la vita quotidiana delle persone e si giocano le speranze di migliorarla. La collaborazione tra gli scienziati dei dati e i ricercatori di storie risuonerà in una direzione necessaria alla comprensione di un fenomeno intorno al quale la società si gioca la sua coesione.

Le menti connesse grazie ai media sociali di fatto pensano insieme. E la qualità dei loro pensieri è influenzata dalla qualità delle connessioni. Le opportunità di collaborazione che l'avvento dei media digitali ha generato sono immense. Ma in qualche caso, i mezzi che servono alla rete delle relazioni sociali facilitano la diffusione di malesseri e sofferenze. Se i problemi del corpo sociale sono descritti in qualche caso ricorrendo alla metafora della malattia, anche le soluzioni possono apparire come una forma di cura.

La cura della qualità delle relazioni sociali è un compito della società e delle sue istituzioni. E l'investimento necessario a garantirla può diventare una scelta fondamentale della politica che una società decide di darsi. In questo quadro, l'investimento in ricerca è il primo passo. Da questo possono nascere le innovazioni capaci di cambiare il corso della storia.

BIBLIOGRAFIA

1. ADL (2019). Report 'Online Hate and Harassment: The American Experience'. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience
2. Alkiviadou (2018). Hate speech on social media networks: towards a regulatory framework?, Information & Communications Technology Law, DOI:10.1080/13600834.2018.1494417
3. Anti-Defamation League (2016). Responding to Cyberhate: Progress and trends. New York: ADL
4. ARTICLE 19 (2009). The Camden Principles on Freedom of Expression and Equality (2009), https://www.article19.org/data/files/pdfs/standards/the-camden-principles-on-freedom-of-expression-and-equality.pdf
5. Artime et al. (2020). Effectiveness of dismantling strategies on moderated vs. unmoderated online social platforms. Scientific Reports 10, 14392
6. Babvey et al. (2021). Using social media data for assessing children's exposure to violence during the COVID-19 pandemic. Child Abuse & Neglect, 116, 104747
7. Benesch (2013). Dangerous speech: A proposal to prevent group violence. Voices That Poison: Dangerous Speech Project proposal paper. https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf
8. Benesch (2021). Dangerous Speech: A Practical Guide. https://dangerousspeech.org/
9. Brown (2017). What is hate speech? Part 2: family resemblances. Law Philos 36(5):561–613. https://doi.org/10.1007/s10982-017-9300-x
10. Castaño-Pulgarin et al. (2021). Internet, social media and online hate speech. Systematic review. Aggression and Violent Behaviour, 58, 101608. DOI:10.1016/j.avb.2021.101608
11. Chetty et al. (2018). Hate speech review in the context of online social networks. Aggression and violent behaviour, 40, 108-118. https://doi.org/10.1016/j.avb.2018.05.003
12. Cohen-Almagor (2011). Fighting hate and bigotry on the Internet. Policy and Internet, 3(3), 1–26 https://doi.org/10.2202/1944-2866.1059
13. Costello et al. (2021). COVID-19: A Pandemic of Anti-Asian Cyberhate. Journal of Hate Studies, 17(1)
14. Council of Europe (1997). Recommendation No. R (97) 20 of the Committee of Ministers to member states on "Hate Speech". Retrieved from https://rm.coe.int/1680505d5b
15. Council of Europe (2022a). Reporting Hate Speech. https://www.coe.int/en/web/no-hate-campaign/reporting-hate-speech
16. Council of Europe (2022b). Recommendation CM/Rec16[1] of the Committee of Ministers to Member States on combating hate speech. https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=090000168a67955#_ftn2
17. Davis et al. (2016). Botornot: A system to evaluate social bots. In Proceedings of the 25th international conference companion on world wide web (pp. 273-274)
18. Del Vicario et al. (2016a). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3), 554-559
19. Del Vicario et al. (2016b). Echo chambers: Emotional contagion and group polarisation on facebook. Scientific reports, 6(1), 1-12
20. Denti & Faggian (2021). Where do angry birds tweet? Income inequality and online hate in Italy", Cambridge Journal of Regions, Economy and Society, Volume 14, Issue 3, November 2021, Pages 483–506, https://doi.org/10.1093/cjres/rsab016
21. DSP (2021). The Dangerous Speech Project by Susan Benesch. https://dangerousspeech.org/
22. ECRI (2016). General Policy Recommendation No. 15 on combating hate speech. https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/
23. ElSherief et al. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. Proceedings of the International AAAI Conference on Web and Social Media, 12: 42–51
24. EUR-Lex (2008). Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV:l33178

25. European Commision (2021). Press Release 'The Commission proposes to extend the list of 'EU crimes' to hate speech and hate crime' https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6561
26. European Commission (2016a). Press Release 'Promoting Media Pluralism and Democracy: Commission hosts second Annual Colloquium on Fundamental Rights'. https://ec.europa.eu/commission/presscorner/detail/en/ip_16_3690
27. European Commission (2016b). Code of conduct on countering illegal hate speech online. https://ec.europa.eu/info/sites/default/files/code_of_conduct_hate_speech_en.pdf
28. European Commission (2021). Communication from the Commission to the European Parliament and the European Council, 'A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime', COM(2021) 777 final, https://ec.europa.eu/info/sites/default/files/1_1_178542_comm_eu_crimes_en.pdf
29. European Parliament (2021). Report on strengthening democracy and media freedom and pluralism in the EU: the undue use of actions under civil and criminal law to silence journalists, NGOs and civil society (2021/2036(INI)). https://www.europarl.europa.eu/doceo/document/A-9-2021-0292_EN.pdf
30. Facebook (2022). Community Standards. https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/
31. Faris et al. (2016). Understanding harmful speech online. Berkman Klein Center Research Publication No. 2016-21 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882824
32. Ferrara et al. (2016). The rise of social bots. Communications of the ACM, 59(7), 96-104.
33. Fortuna & Nunes (2018). A survey on automatic detection of hate speech in text .ACM Comput Surv (CSUR) 51(4):1–30. https://doi.org/10.1145/3232676
34. Gallotti et al. (2020). Assessing the risks of 'infodemics' in response to COVID-19 epidemics. Nature Human Behaviour, 4(12), 1285-1293
35. Gelber (2019). Differentiating hate speech: a systemic discrimination approach. Crit Rev Int Soc Pol Phil. https://doi.org/10.1080/13698230.2019.1576006
36. Germany's Network Enforcement Act (2017). https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html
37. González-Bailón & De Domenico (2018). Bots are less central than verified accounts during contentious political events. Proceedings of the National Academy of Sciences, 118(11)
38. Greene (2013). Moral tribes. Emotion, reason and the gap between us and them. Penguin, New York
39. Haidt (2022). Why the Past 10 Years of American Life have been Uniquely Stupid. https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/
40. Haidt (2022). Why the Past 10 Years of American Life have been Uniquely Stupid. https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/
41. Hall (2013). Hate crimes, 2nd ed. Abingdon: Routledge.
42. Harriman et al. (2020). Youth exposure to hate in the online space: An exploratory analysis. International Journal of Environmental Research and Public Health 17(22), 8531
43. Hawdon et al. (2017). Exposure to online hate in four nations: A cross-national consideration. Deviant Behavior, 38(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985
44. Himelboim et al. (2013). Birds of a feather tweet together: integrating network and content analyses to examine cross-ideology exposure on Twitter, Journal of Computer-Mediated Communication, 18: 40–60
45. Holbrook (2020). The Challenge of Conspiracy Theories for Strategic Communications. The RUSI Journal, 165(1), 26-36
46. Jalonen (2014). Social media – An arena for venting negative emotions. Online Journal of Communication and Media Technology 4, 5704
47. Johnson et al. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. Nature, 573(7773), 261-265. DOI: 10.1038/s41586-019-1494-7

48. Keighley (2022). Hate hurts: Exploring the impact of online hate on LGBTQ+ young people. Women and Criminal Justice 32(1-2), 29-48
49. Lazer et al. (2009). Computational social science. Science, 323(5915), 721-723. DOI:10.1126/science.1167742
50. Lazer et al. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060-1062. DOI:10.1126/science.aaz8170
51. Lima et al. (2018). Inside the right-leaning echo chambers: Characterising gab, an unmoderated social system. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 515-522)
52. Lowry et al. (2016). Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model, Information Systems Research, 27: 962–986
53. Majó-Vázquez et al. (2020). Volume and patterns of toxicity in social media conversations during the COVID-19 pandemic
54. May (2019). Stability and complexity in model ecosystems. In Stability and Complexity in Model Ecosystems. Princeton university press.
55. Meisner (2021). Are you OK, Boomer? Intensification of ageism and intergenerational tensions on social media amid COVID-19. Leisure Sciences 43(1-2), 56-61
56. Musto et al. (2015). CrowdPulse: A framework for real-time semantic analysis of social streams. Information Systems, 54, 127-146
57. Nature Magazine (2021). 'I hope you die': how the COVID pandemic unleashed attacks on scientists. https://www.nature.com/articles/d41586-021-02741-x
58. Oksanen et al. (2014). Exposure to online hate among young social media users. In: Soul of society: A focus on the lives of children and youth. Emerald, Bingley, 253-273
59. OSCE (2009). Ministerial Council Decision No. 9/09 Combating Hate Crimes, MC(17) Journal No. 2, Agenda item 8. https://www.osce.org/files/f/documents/d/9/40695.pdf
60. Papcunová et al. (2021). Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. https://doi.org/10.1007/s40747-021-00561-0
61. Parekh (2012). Is there a case for banning hate speech? In: Herz M, Molnar P (eds) The content and context of hate speech: rethinking regulation and responses. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139042871.006
62. Pew Research Centre (2017). Online Harassment 2017. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/
63. Pew Research Centre (2021a). The State of Online Harassment. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/
64. Pew Research Centre (2021b). Online harassment occurs most often on social media, but strikes in other places, too. https://www.pewresearch.org/fact-tank/2021/02/16/online-harassment-occurs-most-often-on-social-media-but-strikes-in-other-places-too/
65. PNAS (2021). Algorithmic amplification of politics on Twitter - https://doi.org/10.1073/pnas.2025334119
66. Sacco et al. (2021). Emergence of knowledge communities and information centralization during the COVID-19 pandemic. Social Science & Medicine, 285, 114215
67. Sakki & Hakoköngäs (2022). A critical discursive psychological study of dialogical constructions of hate-speech in established media and online discussions. In: The far-right discourse of multiculturalism in intergroup interactions. Palgrave, London, 85-111
68. Schwarz & Holnburger (2019). Disinformation: what role does disinformation play for hate speech and extremism on the internet and what measures have social media companies taken to combat it? Hate Speech and Radicalisation Online
69. Science Magazine (2022). In the Line of Fire. https://www.science.org/content/article/overwhelmed-hate-covid-19-scientists-face-avalanche-abuse-survey-shows
70. Sellars (2016). Defining hate speech. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244

71. Shao et al. (2018). The spread of low-credibility content by social bots. Nature communications, 9(1), 1-9
72. Siegel (2020). 'Online Hate Speech', in Social Media and Democracy: The State of the Field, Prospects for Reform' (2020) https://doi.org/10.1017/9781108890960
73. Smith (2022). The internet is not what you think it is. A history, a philosophy, a warning. Princeton University Press, Princeton
74. Spottswood & Wohn (2020). Online social capital: Recent trends in research. Current Opinion in Psychology 36, 147-152
75. State of the Union Address (2020). https://ec.europa.eu/info/sites/default/files/soteu_2020_en.pdf
76. Stella, Ferrara & De Domenico. (2018). Bots increase exposure to negative and inflammatory content in online social systems. Proceedings of the National Academy of Sciences, 115(49), 12435-12440
77. Stern (2022). Obama: I Underestimated the Threat of Disinformation. https://www.theatlantic.com/ideas/archive/2022/04/barack-obama-interview-disinformation-ukraine/629496/
78. Stern (2022). Obama: I Underestimated the Threat of Disinformation. https://www.theatlantic.com/ideas/archive/2022/04/barack-obama-interview-disinformation-ukraine/629496/
79. Suler (2004). The online disinhibition effect. Cyber Psychology & Behaviour, 7: 321–326
80. Sunstein (2017). #Republic: Divided Democracy in the Age of Social Media. Princeton, NJ: Princeton University Press
81. Sweden's Psychological Defence Agency (2022). https://www.mpf.se/en/
82. SWG (2017). La diffusione dell'odio in rete, 2017 - https://www.swg.it/politicapp?id=syin
83. Topidi (2021). The Regulation and Governance of Online Hate Speech in the Post-truth Era: A European Comparative Perspective in The Indian Yearbook of Comparative Law 2021
84. Treaty on the Functioning of the European Union (2009). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12012E%2FTXT
85. Tsesis (2002). Destructive messages: How hate speech paves the way for harmful social movements. New York University Press, New York
86. TWG (2020). Freedom and Accountability: A Transatlantic Framework for Moderating Speech Online. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/07/Freedom_and_Accountability_TWG_Final_Report.pdf
87. Twitter (2022). Twitter Rules and Policies. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
88. UK Draft Online Safety Bill (2021). https://www.gov.uk/government/publications/draft-online-safety-bill
89. UK Online Harms White Paper (2019). https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper
90. UN General Assembly, International Covenant on Civil and Political Rights (1966). https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights
91. UN's Office on Genocide Prevention and the Responsibility to Protect (2021). Addressing hate speech on social media: Contemporary challenges. https://unesdoc.unesco.org/ark:/48223/pf0000379177/PDF/379177eng.pdf.multi
92. UN's Office on Genocide Prevention and the Responsibility to Protect (2021), Addressing hate speech on social media: Contemporary challenges. https://unesdoc.unesco.org/ark:/48223/pf0000379177/PDF/379177eng.pdf.multi
93. Uyheng & Carley (2020). Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines. Journal of computational social science, 3(2), 445-468
94. Uyheng & Carley (2021). Characterising network dynamics of online hate communities around the COVID-19 pandemic. Applied Network Science, 6(1), 1-21
95. Vegetti & Littyay (2022). Belief in conspiracy theories and attitudes toward political violence. Italian Political Science Review/Rivista Italiana di Scienza Politica, 52(1), 18-32

96. Velasquez et al. (2021). Online hate networks spread malicious COVID-19 content outside the control of individual social media platforms. Scientific reports, 11(1), 1-8. https://doi.org/10.1038/s41598-021-89467-y
97. Wachs & Wright (2019). The moderation of online disinhibition and sex on the relationship between online hate victimization and perpetration. Cyberpsychology, Behavior and Social Networking, 22, 300–306. https://doi.org/10.1089/Cyber.2018.0551
98. Waqas et al. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. PloS one, 14(9), e0222194
99. Watanabe et al. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access, 6, 13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394
100. Weimann & Masri (2020). Research note: spreading hate on TikTok. Studies in conflict & terrorism, 1-14
101. Windisch et al. (2021). PROTOCOL: Online interventions for reducing hate speech and cyberhate: A systematic review. Campbell Systematic Reviews 17(1), e1133
102. Youtube (2022). Community Guidelines. https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436