Nexa Center for Internet & Society
Politecnico di Torino

Grandi piattaforme e moderazione dei contenuti: dati ed evidenze sulla limitazione della libertà d'espressione online

Prof. Maurizio Borghi

Dott. Beatrice Balzola
Dott. Giacomo Conti

Dicembre 2024

Studying the Internet, exploring its potential & experimenting new ideas

Via Pier Carlo Boggio 65/A, 10129 Torino, Italia

(http://nexa.polito.it/contacts-en)

+39 011 090 7217 (Telephone)

+39 011 090 7216 (Fax)

info@nexa.polito.it

Mailing address:

Nexa Center for Internet & Society

Politecnico di Torino - DAUIN

Corso Duca degli Abruzzi, 24

10129 TORINO

ITALY

The Nexa Center for Internet & Society is a research center affiliated to the Department of Control and Computer Engineering of Politecnico di Torino (http://dauin.polito.it).

Grandi piattaforme e moderazione dei contenuti: dati ed evidenze sulla limitazione della libertà d'espressione online

Sommario

- Negli ultimi anni la digitalizzazione e la sempre più massiccia diffusione della comunicazione online ha comportato una ridefinizione empirica del concetto di libertà di espressione. Le piattaforme digitali, nate come meri intermediari per la comunicazione tra terze parti, svolgono ora un ruolo attivo nel controllo dei contenuti, acquisendo un crescente potere di incidere sul pluralismo e sul libero scambio delle informazioni.

- L'attuale quadro normativo, segnato dal Regolamento 2022/2065 (Digital Services Act), ha inciso soprattutto sulle grandi piattaforme, introducendo specifici obblighi di diligenza, trasparenza e moderazione dei contenuti. Questo assetto avvicina la funzione di tali piattaforme a quella di "editori" piuttosto che a quella di intermediari passivi e neutrali. In particolare, il Regolamento subordina l'esenzione di responsabilità alla rimozione *ex ante* di contenuti illeciti o contrari ai termini d'uso.

- Per raccogliere evidenza empirica sulle pratiche concrete di rimozione dei contenuti, è stato predisposto un questionario anonimo rivolto agli utenti italiani e volto a raccogliere testimonianze dirette sulle esperienze di "censura online" da loro esperite. I provvedimenti oggetto di attenzione consistono nella sospensione dell'account (o di sue funzionalità essenziali), nella rimozione del contenuto, o nello *shadow ban* del contenuto stesso (ossia riduzione della visibilità senza preavviso).

- Dall'analisi emerge una frequenza significativa di contenuti oggetto di rimozione che non sono configurabili né come illeciti né come contravvenzioni ai termini d'uso. Questo evidenzia come le grandi piattaforme esercitino un controllo discrezionale e di fatto editoriale, che va oltre gli obblighi del Digital Services Act. Tale pratica può, in particolare, configurare una limitazione della libertà di espressione e una forma di censura preventiva incompatibili con l'ordinamento costituzionale italiano.

- L'analisi qualitativa dei contenuti rimossi evidenzia inoltre una frequente illogicità nei provvedimenti, configurando un modello di controllo caratterizzato da imprevedibilità e contraddittorietà, prevalentemente orientato alla repressione di opinioni minoritarie o controverse. Tale modello appare determinato dagli interessi mutevoli delle piattaforme (e dei regolatori di riferimento) e privo di adeguata base giuridica. Ad aggravare la situazione contribuisce l'inefficacia degli strumenti di ricorso, i quali, per gli utenti colpiti da provvedimenti, consistono spesso in una mera procedura automatizzata sprovvista di un controllo umano significativo.

- In linea con il dettato normativo del Digital Services Act, la necessità di una maggiore trasparenza algoritmica si deve necessariamente accompagnare a procedure di ricorso realmente accessibili e a definizioni più precise e meno interpretabili dei concetti di contenuti "pericolosi" o "dannosi", che, nella mancanza di una formulazione adeguata, lasciano eccessivo arbitrio censorio alle piattaforme digitali e ai regolatori pubblici a cui fanno riferimento.

Indice

Contents

Sommario

Il quadro normativo

L'articolo 21 della Costituzione Italiana rappresenta un pilastro per la libertà d'espressione, garantendola come diritto fondamentale e sottolineandone il ruolo cruciale nella democrazia[1]. Questo diritto si estende oltre il semplice atto del parlare o dello scrivere, abbracciando una vasta gamma di forme di espressione, inclusa la libertà di stampa e, per interpretazione estensiva, la libertà di informazione. È peraltro utile, per facilità di comprensione e per attinenza al tema in esame, scindere il diritto in un duplice significato: da una parte, la libertà di manifestare il proprio pensiero senza limitazioni; dall'altra, la possibilità di carpire la manifestazione del pensiero altrui, cioè quella di essere in qualche misura partecipi del libero scambio di pensieri, opinioni e informazioni[2].

La digitalizzazione e l'evoluzione tecnologica hanno radicalmente trasformato i modi in cui la libertà d'espressione viene esercitata, portando alla luce nuove sfide non sempre efficacemente gestibili con i pregressi strumenti del diritto. Le piattaforme digitali, pur offrendo opportunità senza precedenti per la condivisione di idee e informazioni, sollevano infatti numerose questioni riguardanti il loro ruolo nel controllo e nella distribuzione dei contenuti[3]. I filtri automatici e gli algoritmi possono influenzare la visibilità delle informazioni, censurandole *tout court* oppure silenziandole di fatto attraverso meccanismi che ne rendono difficoltoso il reperimento, perciò mettendo in discussione il pluralismo e la diversità delle opinioni: elementi che usualmente si ritengono fondamentali per un efficace esercizio dei diritti fondamentali in una moderna democrazia.

Il secondo comma dell'articolo 21 della Costituzione Italiana vieta la censura sulla stampa, intesa come controllo preventivo delle informazioni prima della loro diffusione. Da qui l'idea che siano generalmente problematici, rispetto all'art. 21, tutti i sistemi che, online, filtrano o bloccano ex ante contenuti, specie se in modo automatizzato o pseudo-automatizzato, impedendone la pubblicazione o rendendone difficile la reperibilità.

Questa lettura è rafforzata dall'art. 10 CEDU sulla libertà di espressione e dalla giurisprudenza della Corte di giustizia UE sui blocchi di siti e contenuti decisi dalle autorità pubbliche[4].

Per quanto inerisce all'esercizio dei poteri pubblici, la giurisprudenza europea[5] e la Carta dei Diritti Fondamentali (art. 52), vincolano la limitazione della libertà di espressione a tre condizioni:

1. Essa deve essere prevista dalla legge: serve una base normativa dotata di forza di legge, sufficientemente chiara, accessibile e prevedibile da parte dei destinatari;

[1] C. Esposito, *La libertà di manifestazione del pensiero nell'ordinamento italiano*, Milano, Giuffrè, 1958

[2] Il tema è espresso in particolare all'art. 11 CDFUE, laddove si dice che il diritto alla libertà di espressione "include la libertà d'opinione e la libertà di ricevere o di comunicare informazioni senza che vi sia ingerenza da parte delle autorità pubbliche". Nonostante la Costituzione italiana parli di "manifestazione del pensiero" e le fonti europee citino normalmente la libertà di "espressione", i due concetti sono sostanzialmente sinonimi, anche se il secondo è forse più vasto, e ricomprende il primo – mentre potrebbe non valere l'opposto.

[3] Un tema, questo, di cui si discute da molti anni: si veda ad es. Y. Benkler, *The Wealth of Networks: How social production transforms markets and freedom,* Yale University Press, 2006.

[4] Seppure in più casi si faccia espresso riferimento al blocco di contenuti da parte di autorità pubbliche, e non da parte di soggetti privati come invece sempre più spesso avviene nell'uso quotidiano delle piattaforme digitali, si rilevano alcune sentenze della CGUE interessanti sul tema come la C-70/10 del 24 novembre 2011 (*Scarlet Extended*) o la C-314/12 del 27 marzo 2014 (*UPC Telekabel Wien*).

[5] Sentenza *WebMindLicenses* (C-419/14, punto 81); *Parere 1/15 dell'Accordo PNR UE-Canada*, 26 luglio 2017 (EU:C:2017:592, punto 146)

2. Deve rispettare il contenuto essenziale del diritto: anche quando si perseguono altri interessi legittimi (sicurezza, tutela di diritti altrui, ordine pubblico), non si può svuotare del tutto la libertà di espressione;

3. Deve risultare necessaria e proporzionata rispetto all'obiettivo perseguito: la misura deve essere idonea, indispensabile e non eccessiva.

Applicato a Internet, ciò significa che misure molto invasive, come l'oscuramento di un intero sito o l'imposizione di filtri generalizzati sul traffico, sono ammissibili solo se rispettano questi criteri, e sono sottoposte al controllo dei giudici nazionali e, in ultima istanza, della Corte di giustizia.

La questione si complica quando non è più solo lo Stato a limitare la circolazione delle informazioni, ma lo fanno attori privati, in particolare le grandi piattaforme di condivisione di contenuti (social network, servizi di video sharing, ecc.).

In questo quadro assume rilevanza centrale la Direttiva 2000/31/CE sul commercio elettronico, trasposta nell'ordinamento italiano dal D.Lgs. 9 aprile 2003 n. 70. La direttiva stabilisce uno sgravio di responsabilità per i prestatori di servizi della società dell'informazione, quali gli hosting providers, per i contenuti illeciti caricati dagli utenti. Tale beneficio è subordinato al sussistere di tre condizioni cumulative: che il fornitore di servizi svolga un ruolo meramente tecnico e "passivo" (limitandosi a ospitare ciò che altri creano); che non abbia conoscenza del contenuto illecito ospitato; e che, una volta venutone a conoscenza (ad esempio tramite segnalazioni), agisca tempestivamente per rimuovere il contenuto o per renderlo inaccessibile. Allo stesso tempo, la Direttiva vieta esplicitamente agli Stati membri di imporre ai prestatori di servizi della società dell'informazione obblighi generali di sorveglianza sui contenuti che ospitano, pur consentendo obblighi "specifici" riferiti a singoli casi e singole violazioni. In tal modo, la Direttiva ha configurato una responsabilità di natura principalmente "reattiva" per gli intermediari Internet, liberandoli da oneri di controllo preventivo sui contenuti presenti sulle loro infrastrutture.

L'impianto originario della Direttiva E-Commerce ha tuttavia subito due significative revisioni.

La prima, di settore, ha riguardato specificamente il campo del diritto d'autore online con la Direttiva 2019/790 (Direttiva sul copyright nel mercato unico digitale). Questo intervento ha introdotto obblighi specifici per le piattaforme che rientrano nella definizione di "prestatore di servizi di condivisione di contenuti online",[6] in particolare per quanto concerne l'ottenimento di una licenza da parte dei titolari dei diritti e la prevenzione delle violazioni diffuse del diritto d'autore.[7] Per effetto di questi obblighi, la Direttiva ha reso di fatto necessaria l'adozione di sistemi di filtraggio e rilevazione automatica dei contenuti illeciti violanti il copyright[8], configurando un obbligo di controllo preventivo del tutto nuovo e potenzialmente limitativo della libertà d'espressione.

[6] Art. 2(6), definito come "un prestatore di servizi della società dell'informazione il cui scopo principale o uno dei principali scopi è quello di memorizzare e dare accesso al pubblico a grandi quantità di opere protette dal diritto d'autore o altri materiali protetti caricati dai suoi utenti, che il servizio organizza e promuove a scopo di lucro." Dalla definizione sono esclusi "I prestatori di servizi quali le enciclopedie online senza scopo di lucro, i repertori didattici o scientifici senza scopo di lucro, le piattaforme di sviluppo di e condivisione di software open source, i fornitori di servizi di comunicazione elettronica ai sensi della direttiva (UE) 2018/1972, i mercati online, i servizi cloud da impresa a impresa e i servizi cloud che consentono agli utenti di caricare contenuti per uso personale".

[7] Art. 17 "Utilizzo di contenuti protetti da parte di prestatori di servizi di condivisione di contenuti online.

[8] A. Scenna, *Il bilanciamento tra i diritti d'autore e i diritti fondamentali alla luce della direttiva UE 2019/790. Note alla sentenza della Corte di giustizia europea,* DPCE Online, 3/2022, P.1765, https://www.dpceonline.it/index.php/dpceonline/article/view/1678/1685

La Polonia ha impugnato l'art. 17 davanti alla Corte di giustizia, sostenendo che il dispositivo violi l'art. 11 CDFUE (libertà di espressione), poiché l'attuale tecnologia non è in grado di distinguere tra contenuti illeciti e leciti (ad esempio l'uso di opere protette per parodia o citazione).[9]

Ma la Corte, pur riconoscendo il rischio di limitazioni eccessive, ha respinto il ricorso, ritenendo che non si introduca un obbligo generale di sorveglianza, grazie al combinato disposto dell'art. 17, par. 8, e del requisito che i titolari forniscano indicazioni specifiche sui metodi di controllo, e ha fondato la compatibilità sull'osservanza dei principi di necessità e proporzionalità ex art. 52 CDFUE, e sulle garanzie procedurali di reclamo e ricorso, che però, come si vedrà, sono nel concreto spesso inefficaci.

Resta tuttavia, nella pratica, il problema dei falsi positivi e dell'*overblocking*: filtri che rimuovono o bloccano contenuti legittimi, specie se critici, satirici o dai toni "alti".

L'uso crescente di algoritmi di moderazione automatica (utilizzati oltre il copyright) pone generalmente tre tipi di problemi:

1) Un rischio di eccesso: rimozione o deindicizzazione di contenuti leciti, con impoverimento del dibattito e rafforzamento delle echo chambers
2) Un problema di trasparenza e accountability: spesso gli utenti non capiscono perché un contenuto è stato rimosso, né dispongono di effettivi strumenti di ricorso.
3) La mancata comprensione del contesto: l'algoritmo fatica con ironia, linguaggi minoritari, dialetti, riusi creativi, e tende a bloccare contenuti provocatori ma leciti.

La seconda revisione, di portata ben più ampia e sistemica, è stata operata dal Regolamento (UE) 2022/2065, noto come Digital Services Act (DSA). Questo provvedimento ha radicalmente aggiornato e sostituito le regole per tutti i servizi digitali, introducendo un quadro di responsabilità articolato e proattivo, obblighi di diligenza rafforzati e meccanismi di controllo per le piattaforme online, specialmente quelle di grandi dimensioni.

In particolare, il DSA individua un'ulteriore categoria di soggetti, le "Very Large Online Platforms" (VLOPs), che a causa della loro dimensione e del loro percepito impatto sul discorso pubblico sono destinatari di vincoli più rigorosi. Se gli obblighi introdotti dall'articolo 17 della Direttiva 790/2019 sono specificamente indirizzati alla protezione del diritto d'autore e alla prevenzione della condivisione di contenuti protetti, il DSA si distingue per l'introduzione di regole più estese volte a contrastare un ampio spettro di comportamenti illeciti. Di particolare rilevanza per questo studio sono le norme che disciplinano la "moderazione di contenuti", definita come l'insieme di attività volte a contrastare la diffusione di *contenuti illegali* o *incompatibili con le condizioni generali* (cioè i *Terms of Service*) della piattaforma.[10]

[9] Causa C-401/19 *Repubblica di Polonia contro Parlamento europeo e Consiglio dell'Unione europea*.

[10] La definizione all'Articolo 3(t) recita: "le attività, automatizzate o meno, svolte dai prestatori di servizi intermediari con il fine, in particolare, di individuare, identificare e contrastare contenuti illegali e informazioni incompatibili con le condizioni generali, forniti dai destinatari del servizio, comprese le misure adottate che incidono sulla disponibilità, sulla visibilità e sull'accessibilità di tali contenuti illegali o informazioni, quali la loro retrocessione, demonetizzazione o rimozione o la disabilitazione dell'accesso agli stessi, o che incidono sulla capacità dei destinatari del servizio di fornire tali informazioni, quali la cessazione o la sospensione dell'account di un destinatario del servizio".

Attraverso il ricorso a una definizione ampia di "contenuto illegale",[11] il DSA si propone di creare un quadro normativo che affronti una varietà di problematiche online, che vanno dalla propaganda terroristica alla pornografia infantile, dalla violazione del copyright all'informazione ingannevole al consumatore, fino ad estendersi a contenuti non illegali "*ma comunque dannosi*"[12] come i discorsi d'odio ("hate speech"[13]) e la c.d. "disinformazione"[14], con l'obiettivo dichiarato di garantire un ambiente online sicuro, prevedibile e affidabile.

I principali obblighi introdotti dal DSA riguardo alla "moderazione dei contenuti" si possono riassumere nella tabella seguente:

Condizioni contrattuali (*Terms and conditions*)	Le piattaforme devono predisporre **condizioni contrattuali chiare e di agevole consultazione** in materia di moderazione dei contenuti, incluse le politiche relative ai contenuti soggetti a restrizioni, all'utilizzo di strumenti automatizzati e ai motivi di sospensione dell'account (Art. 14).
Segnalazione e azione	Le piattaforme devono mettere a disposizione un meccanismo che consenta agli utenti (e ai "segnalatori attendibili") di notificare la presenza di contenuti illeciti. Tali notifiche devono essere trattate in maniera solerte, diligente, non arbitraria e obiettiva (Art. 16).
Motivazione	Quando rimuove o limita l'accesso a contenuti di un utente (o sospende un account), la piattaforma deve fornire all'utente stesso una **motivazione chiara e specifica**, salvo che l'illecito sia manifestamente evidente (ad esempio, contenuti riguardanti gravi reati) (Art. 17).
Gestione dei reclami	Le piattaforme devono offrire un sistema interno **gratuito e di facile utilizzo** per consentire agli utenti di impugnare le decisioni di moderazione dei contenuti (ad esempio, rimozioni, sospensioni). I reclami devono essere trattati con celerità e le decisioni devono essere adottate in modo tempestivo, non arbitrario e obiettivo (Art. 20).

[11] Definito all'Articolo 3(h) come "qualsiasi informazione che, di per sé o in relazione a un'attività, tra cui la vendita di prodotti o la prestazione di servizi, non è conforme al diritto dell'Unione o di qualunque Stato membro conforme con il diritto dell'Unione, indipendentemente dalla natura o dall'oggetto specifico di tale diritto".

[12] Come ribadito ai Considerando 5 e 68.

[13] Sulla definizione di "hate speech" vale la pena osservare quanto sostenuto da G. Vasino in *Censura "privata" e contrasto all'hate speech nell'era delle Internet Platforms,* Op. Cit., laddove viene mostrato come il concetto e la disciplina che identifica *l'hate speech* sia virata verso un concetto di "incitement to hatred" e non come "incitement to violence, discrimination e hostility" quale era in origine. A onor del vero, l'autrice reputa la definizione "vaga, fuorviante e giuridicamente non adeguata in quanto collegabile ad uno stato emotivo piuttosto che al verificarsi di un pericolo certo", trovandosi peraltro in contrasto con i principi di proporzionalità e gradualità.

[14] Per quanto riguarda il concetto di "disinformazione", occorre rilevare l'assenza di una sua definizione giuridica formale in testi normativi quali il DSA. Il regime attuale si fonda su linee guida di soft law elaborate nell'ambito dell'Unione Europea (tra cui i lavori dell'High Level Expert Group e la Comunicazione della Commissione relativa al Piano d'Azione per la Democrazia Europea) poi recepite a livello pratico nei Codici di Condotta del 2018 e del 2022. Per una ricostruzione critica si vedano Vincenzo Zeno-Zencovich, The EU regulation of speech. A critical view, Medialaws, https://www.medialaws.eu/wp-content/uploads/2023/06/1-23-Zeno-Zencovich.pdf e S. Foà "Pubblici poteri e contrasto alle fake news. Verso l'effettività dei diritti aletici?" *federalismi.it* 11/2020.

Risoluzione extragiudiziale delle controversie	Gli utenti devono poter accedere a organismi di risoluzione extragiudiziale delle controversie certificati per risolvere le controversie in materia di decisioni di moderazione dei contenuti, **senza alcun costo** per l'utente (Art. 21).

A questi si sommano poi obblighi aggiuntivi per le piattaforme designate come VLOP che riguardano la valutazione e l'attenuazione dei "rischi sistemici", l'adozione di sistemi di risposta alle crisi, la produzione di audit e rapporti di trasparenza, e altri ancora.[15]

Una prima questione attiene a come il DSA tenda a raggruppare sotto la stessa etichetta di "contenuti illegali (o comunque dannosi)" fenomeni estremamente diversi dal punto di vista giuridico, etico e sociale. Del resto, è evidente come esistano nel quotidiano comportamenti di odio o di falsità non sanzionabili dall'ordinamento a meno che non integrino gli estremi di un reato o violino norme specifiche. L'approccio semplificatorio-riduttivo del DSA, sebbene miri a razionalizzare la gestione dei contenuti online, solleva legittimi dubbi sull'opportunità di trattare condotte così disparate in maniera uniforme, semplicemente perché veicolate attraverso Internet[16]. Anche in assenza di una definizione euro-unitaria precisa di *online hate speech*[17], si sostiene che la portata di internet sia tale da poter influenzare la vita comune al di là del mondo digitale, ponendo rischi per la comunità e la tenuta democratica che imporrebbero la necessità di una "seconda generazione di regole", più stringenti di quelle precedenti, che miravano innanzi tutto a non ostacolare uno sviluppo repentino ed economicamente irrinunciabile del medium[18].

Estendendo la sua portata anche e soprattutto ai contenuti "dannosi", il DSA potenzia gli obblighi di moderazione e controllo per i fornitori di servizi Internet e per le piattaforme online. E' proprio questo potenziamento di responsabilità che solleva le maggiori preoccupazioni in termini di censura preventiva e di impatto sulla libertà di espressione: non si tratta qui semplicemente di contrastare la violazione di diritti ed evitare la diffusione di contenuti penalmente rilevanti, ma di intervenire su forme di espressione che, pur astrattamente considerabili come "dannose", non costituiscono di per sé un illecito[19].

Le VLOPs, identificabili nei servizi offerti da giganti del web come Google, Meta, Twitter, Instagram, AliExpress, TikTok,[20] si trovano infatti a dover implementare sistemi algoritmici per prevenire attivamente la diffusione di contenuti illegali o dannosi attraverso i loro servizi. A questo, ovviamente, si deve aggiungere il rischio di una censura algoritmica (quindi preventiva e potenzialmente in conflitto con l'articolo 21 cost.) che potrebbe limitare indebitamente la libertà di espressione, rimuovendo contenuti che, pur non essendo illeciti, possono essere comunque classificati come dannosi secondo criteri spesso opachi e arbitrari. Lo stesso problema di "chi controllerà i controllori" si pone anche per

[15] Definiti agli Articoli 33-43.

[16] Vincenzo Zeno-Zencovich, *The EU regulation of speech. A critical view,* Medialaws, https://www.medialaws.eu/wp-content/uploads/2023/06/1-23-Zeno-Zencovich.pdf

[17] I. Anrò, *Online hate speech: la prospettiva dell'unione europea tra regolamentazione della condotta dei prestatori di servizi intermediari e ricorso al diritto penale,* Osservatorio sule fonti, Vol. 16, p.15, https://air.unimi.it/bitstream/2434/967934/2/OSF%201%202023%20Anr%C3%B2.pdf

[18] M. Husovec, *Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation of Global Internet Rules,* Berkeley Technology Law Journal, Vol. 38, 10 ottobre 2023, https://tinyurl.com/smyamvvx

[19] R. Ó Fathaigh, N. Helberger, N. Appelman, *The Perils of Legally Defining Disinformation*, University of Amsterdam, Institute for Information Law Research Paper No. 2022-05, dicembre 2021

[20] L'elenco dei servizi designati dalla Commissione UE come VLOPs ("Very Large Online Platforms") e VLOSE ("Very Large Online Search Engines") è disponibile all'indirizzo https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses

la soluzione introdotta dal DSA di individuare delle figure di riferimento per la segnalazione di contenuti illegali (i "segnalatori attendibili" o *trusted flaggers*),[21] che non sono oggetto però di questo scritto.

Non solo: alcuni autori hanno anche messo in risalto la natura di *lex specialis* della Direttiva 790/2019 rispetto al DSA, dalla quale discende che le cautele imposte alle piattaforme digitali per la protezione del copyright (imposte nella Direttiva 790) permangono anche sotto il DSA, potenzialmente configurando una stessa piattaforma *sia* come *Online Content Service Provider* (OCSSP), sia come *(Very Large) Online Platform* (VLOP). Del resto, l'articolo 17, secondo comma, della Direttiva 790/2019 afferma che la limitazione di responsabilità per l'hosting dell'art. 14 della Direttiva e-Commerce si applica ancora agli OCSSP "per scopi che esulano dal campo di applicazione di questa Direttiva". Si consideri dunque una piattaforma che venga qualificata dal DSA come VLOP. Se essa organizza e promuove a scopo di lucro contenuti protetti caricati dagli utenti, sarà altresì qualificata come OCSSP, e la sua responsabilità sarà determinata dalla natura del contenuto da regolare di volta in volta: violazione del diritto d'autore o diffusione di *harmful speech*[22].

L'ultimo, ma non meno importante, problema dell'impianto normativo così descritto sta nella sempre più sottile differenza tra *editori* e piattaforme digitali.

Come si è avuto modo di vedere, soltanto le *piattaforme digitali* godono di uno sgravio di responsabilità rispetto ai contenuti pubblicati da terze parti sui loro servizi, seppure alle condizioni indicate (tra cui in particolare gli obblighi relativi alla "moderazione dei contenuti"). Al contrario, gli *editori* scelgono attivamente cosa pubblicare, indipendentemente dal fatto che il contenuto pubblicato sia creato da loro stessi o da soggetti terzi. Svolgono, a tutti gli effetti, una "scelta editoriale" e pertanto sono responsabili anche penalmente di ciò che pubblicano sui loro media.

A seguito del Digital Services Act, le VLOPs sono messe in una posizione di equilibrio intermediale piuttosto delicata. Da un lato, infatti, tali piattaforme sono di fatto obbligate a implementare sistemi di moderazione algoritmica e procedure di controllo efficaci per prevenire la diffusione di contenuti illegali, pena la violazione degli obblighi di diligenza imposti dal regolamento. Dall'altro, un eccesso di intervento rischia di trasformarle in soggetti che esercitano un controllo editoriale sui contenuti, avvicinandole alla figura dell'editore e compromettendo così il regime di responsabilità limitata che caratterizza gli intermediari online.

Questo equilibrio impone alle piattaforme di moderare abbastanza da garantire un "ambiente sicuro", ma non tanto da assumere scelte discrezionali e preventive sulla liceità o "nocività" dei contenuti. In caso contrario, la moderazione (algoritmica o meno) potrebbe degenerare in una forma di censura preventiva privata, potenzialmente in contrasto con l'art. 21 della Costituzione e con i principi europei sulla libertà d'espressione.

In estrema sintesi, l'azione delle piattaforme nell'adempiere agli obblighi del DSA è vincolata dal diritto primario UE e dai principi costituzionali. Tale cornice impone tre precisi limiti per il mantenimento del ruolo di "piattaforma" (e dunque soggetto esentato da responsabilità per i contenuti pubblicati dagli utenti): 1) il divieto di rimuovere contenuti che non siano illeciti o contrari ai propri termini di servizio; 2) l'obbligo di motivare in modo chiaro e trasparente ogni decisione di rimozione o limitazione; 3) il divieto di discriminare i contenuti in base alla loro espressione ideologica o politica.

Questo le distingue dagli editori: un giornale o una casa editrice hanno un *controllo editoriale* pieno. Decidono che cosa pubblicare, come presentarlo, e con quale taglio, senza dover rendere conto al

[21] Articolo 22 DSA.

[22] J.P. Quintais, S. F. Schewemer, *The Interplay between the Digital Services Act and Sector Regulation: How Special is Copyright?,* European Journal of Risk Regulation, 2022, https://tinyurl.com/2wjdfkmu

pubblico sul perché un tale pezzo è uscito, o sul perché sia stato ritenuto, o no, accettabile. Al contrario, una piattaforma dovrebbe essere un intermediario: un mero facilitatore della comunicazione tra terzi, con limitate funzioni di controllo sulla diffusione di contenuti illegali.

Tuttavia, proprio a seguito dei nuovi obblighi imposti dal DSA, una piattaforma può essere spinta ad assumere un ruolo di quasi-editore quando esercita un controllo attivo sui contenuti (seleziona, promuove, modifica, "oscura" sistematicamente certi temi, magari mediante algoritmi e sistemi indiretti di contenuti correlati e consigliati); quando usa algoritmi che, al di là della tecnica, riflettono scelte di valore su ciò che è più "rilevante", attraente, o semplicemente monetizzabile; quando applica linee guida vaghe per rimuovere contenuti ingannevoli, nocivi, sensibili, stabilendo dunque nella pratica che cosa sia accettabile o meno nel discorso pubblico.

Alla luce del quadro normativo e concettuale descritto, è parso necessario acquisire dati empirici per ottenere una rappresentazione concreta delle pratiche di moderazione adottate dalle piattaforme, analizzandole specificamente attraverso la lente della percezione degli utenti.

Attraverso un questionario anonimo rivolto agli utenti delle piattaforme, si è voluto dare evidenza qualitativa e quantitativa al rischio di una "censura privata". Un rischio che il quadro normativo del DSA, con il suo mix di doveri di diligenza, controllo dei rischi sistemici e moderazione proattiva, potrebbe finire per alimentare, pur senza averne l'intenzione esplicita.

Le risposte raccolte ambiscono a offrire uno spaccato empirico delle forme di censura effettivamente sperimentate dagli utenti e delle pratiche di controllo dei contenuti attuate dalle piattaforme, colto nel momento critico di attuazione dei nuovi obblighi imposti dal Digital Services Act.

I dati raccolti confermano che le pratiche censorie si concentrano soprattutto nei social network generalisti, in particolare su Facebook e Instagram, e che si manifestano attraverso una pluralità di provvedimenti: dalla rimozione diretta dei contenuti alla sospensione temporanea o permanente degli account, fino a misure più opache come lo shadowbanning, che riduce silenziosamente la visibilità dei post senza notificare formalmente alcuna sanzione all'utente interessato.

Oltre a segnalare la diffusione di questi fenomeni, il questionario stimola una riflessione più ampia sui criteri di trasparenza, proporzionalità e prevedibilità delle decisioni di moderazione. In molti casi, infatti, gli utenti non ricevono spiegazioni chiare sui motivi della rimozione o sulla natura del contenuto contestato, né dispongono di reali strumenti di ricorso, configurando una condizione di asimmetria tra il soggetto pubblicante e la piattaforma che decide la sorte del suo contenuto pubblicato.

Evidenze empiriche sulla censura online

Il presente capitolo si fonda sui dati raccolti attraverso un questionario anonimo progettato per indagare esperienze concrete di censura online così come percepite dagli utenti. La metodologia adottata ha privilegiato un approccio aperto, mirato a raggiungere un campione eterogeneo di rispondenti e a cogliere la varietà delle situazioni vissute. Il questionario ha utilizzato la piattaforma Limesurvey per essere creato e somministrato, e ha raccolto risposte a partire da marzo 2023 sino a marzo 2024. Alcune risposte riferivano eventi cronologicamente precedenti alla data di somministrazione del questionario.

In una prima fase, il questionario è stato diffuso tramite passaparola all'interno di cerchie professionali e accademiche vicine ai ricercatori, al fine di testarne la chiarezza e la comprensibilità, nonché di raccogliere le prime risposte esplorative. Successivamente, la raccolta è stata aperta al pubblico online, consentendo la partecipazione spontanea di chiunque volesse contribuire in modo anonimo. In un'ultima fase, il questionario è stato attivamente pubblicizzato attraverso i canali digitali a

disposizione del gruppo di ricerca: siti web, social network e community tematiche, per ampliare il bacino di partecipanti e ottenere un campione più rappresentativo. Sono state raccolte in tutto 147 risposte. Le domande poste, e le risposte previste in caso di domande chiuse, sono raccolte nell'Appendice del documento.

La scelta di un formato anonimo è stata dettata dalla volontà di ridurre al minimo le inibizioni e favorire una risposta sincera anche su temi potenzialmente sensibili, come la percezione di censura, le opinioni politiche o l'uso di linguaggi ritenuti inappropriati dalle piattaforme. L'assenza di vincoli identificativi, unita alla semplicità della compilazione online, ha permesso di ottenere una pluralità di prospettive difficilmente raggiungibile con metodi più strutturati o controllati.

Dal punto di vista analitico, il questionario ha combinato domande chiuse, utili a quantificare la frequenza e la distribuzione dei diversi tipi di provvedimenti e ad avere un numero finito di categorie nelle quali far rientrare gli episodi di censura testimoniati, con domande aperte, pensate per raccogliere testimonianze e percezioni personali dei partecipanti. Tale combinazione ha consentito di coniugare un approccio quantitativo e qualitativo, fornendo al tempo stesso dati misurabili e spunti interpretativi.

L'obiettivo complessivo perseguito è stato quello di offrire uno spaccato empirico delle esperienze individuali legate alla moderazione dei contenuti online. I risultati, pertanto, vanno interpretati come indicatori di tendenze e percezioni emergenti, capaci di illuminare criticità e ambiguità dei meccanismi di censura nelle piattaforme digitali.

I dati provenienti dal questionario somministrato anonimamente mirano ad offrire uno sguardo sui tipi di censura che gli utenti hanno concretamente sperimentato. Tali casi non solo sollevano interrogativi sulla trasparenza e l'equità delle politiche di moderazione delle piattaforme, ma anche sulla soggettività della percezione di ciò che è considerato inaccettabile o dannoso.

La stragrande maggioranza delle censure individuate prendono le mosse dai social network tradizionali, con Facebook nettamente in testa, seguito da Instagram.

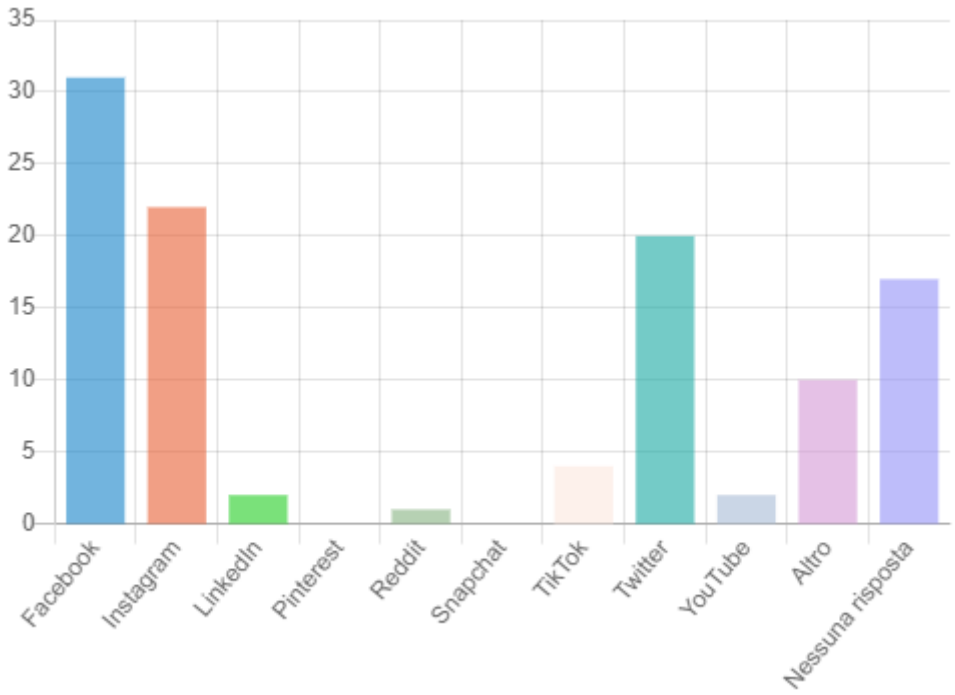


Fig. 1: Ripartizione delle piattaforme digitali segnalate come aventi messo in atto una o più operazioni di censura online.

Si segnalano una varietà di provvedimenti presi come risposta alla pubblicazione di contenuti ritenuti illeciti. Il più frequente attiene alla semplice rimozione del contenuto, ma figurano anche numerose limitazioni all'accesso alla piattaforma digitale: più spesso il cosiddetto *ban* è temporaneo, ma in un numero cospicuo di casi è stato segnalato come permanente. In altri casi, l'utente sospetta che il suo contenuto sia stato limitato in visibilità, ma non può averne la certezza: si tratta dello *shadowbanning*, evidenziato nella seguente immagine dalla lettera E.: un provvedimento per definizione non segnalato che rimuove in modo silenzioso la fruibilità del contenuto da parte di altri ma non da parte dell'account dell'utente-pubblicatore, che in questo modo non si rende conto del provvedimento di censura a suo carico, ma crede semplicemente di essere stato ignorato dalla comunità online.

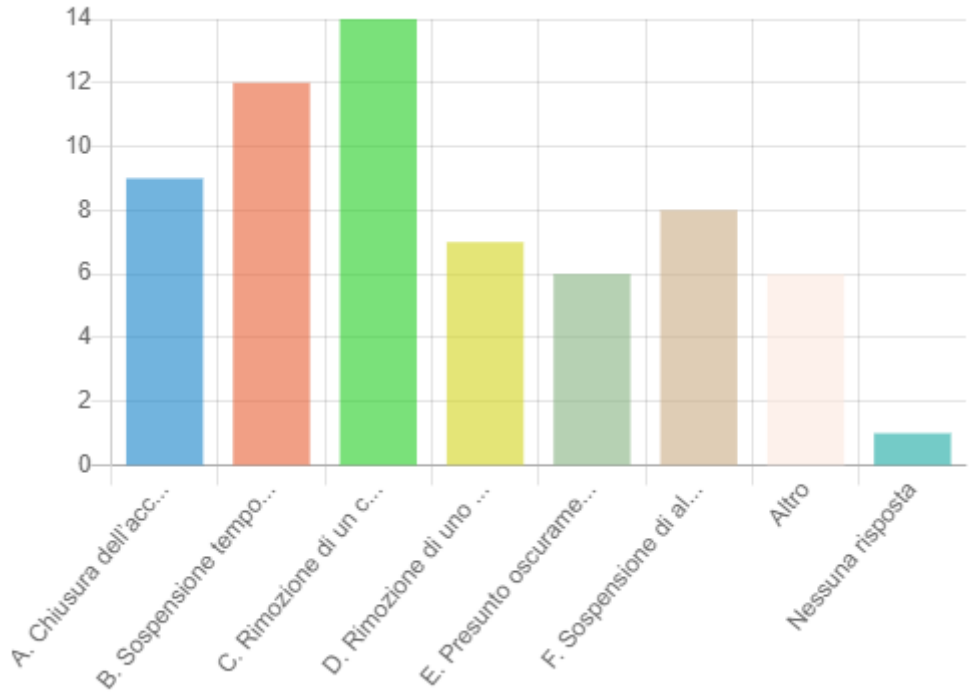


Fig. 2: I tipi di provvedimenti subiti a seguito della pubblicazione di contenuti inaccettabili da parte delle piattaforme.

Secondo i dati raccolti pare inoltre emergere una certa propensione alla recidiva da parte degli utenti interessati ai provvedimenti delle piattaforme. Più della metà dei rispondenti, infatti, ha ammesso di avere avuto altre sospensioni o limitazioni dell'account al di là del caso specifico descritto in sede di questionario.

Dunque, una proporzione significativa degli utenti che hanno subito sanzioni da parte delle piattaforme online tende a incorrere in ulteriori provvedimenti disciplinari, suggerendo che le attuali politiche di moderazione potrebbero non essere sufficientemente efficaci nel promuovere un cambiamento duraturo nel comportamento degli utenti. Questo dato apre alcune riflessioni: intanto, le sanzioni attuali potrebbero non essere abbastanza mirate o efficaci per indurre una riflessione o un cambiamento nel comportamento degli utenti. Spesso, sono esageratamente severe o non permettono

adeguatamente di distinguere tra comportamenti diversi in termini di gravità della sanzione o presenza di possibili condizioni scriminanti.

Può anche essere auspicabile un migliore metodo di pubblicizzazione delle regole per la comunità nei vari servizi online: un tema, peraltro, che come si è visto viene passato ampiamente in rassegna dal DSA, che acuisce gli obblighi di chiarezza e trasparenza per le piattaforme online.

Esattamente nella metà dei casi segnalati (e quando l'account dell'interessato non è stato temporaneamente o definitivamente limitato), gli utenti hanno fatto ricorso contro la sanzione della piattaforma, chiedendo per lo meno delle spiegazioni sul provvedimento da loro subito. Di nuovo, soltanto nel 50% dei casi è stata offerta una risposta a chi la chiedeva.

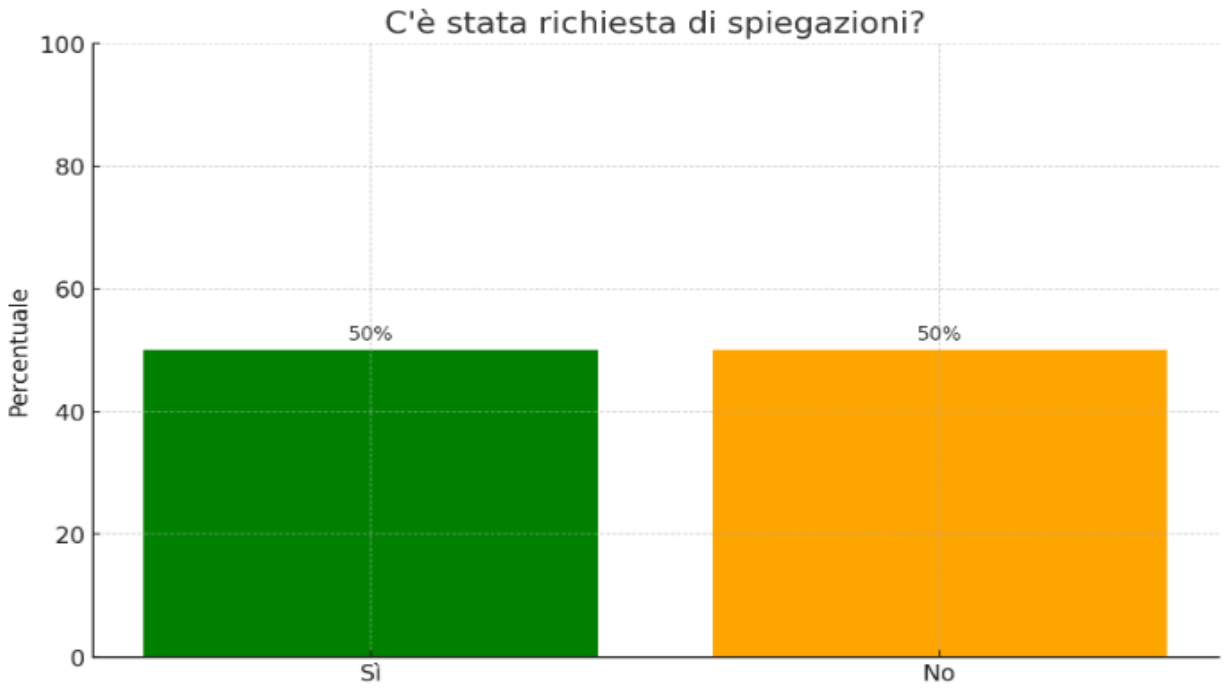


Fig. 3: Se è stata avanzata alla piattaforma una richiesta di spiegazioni a seguito del provvedimento.

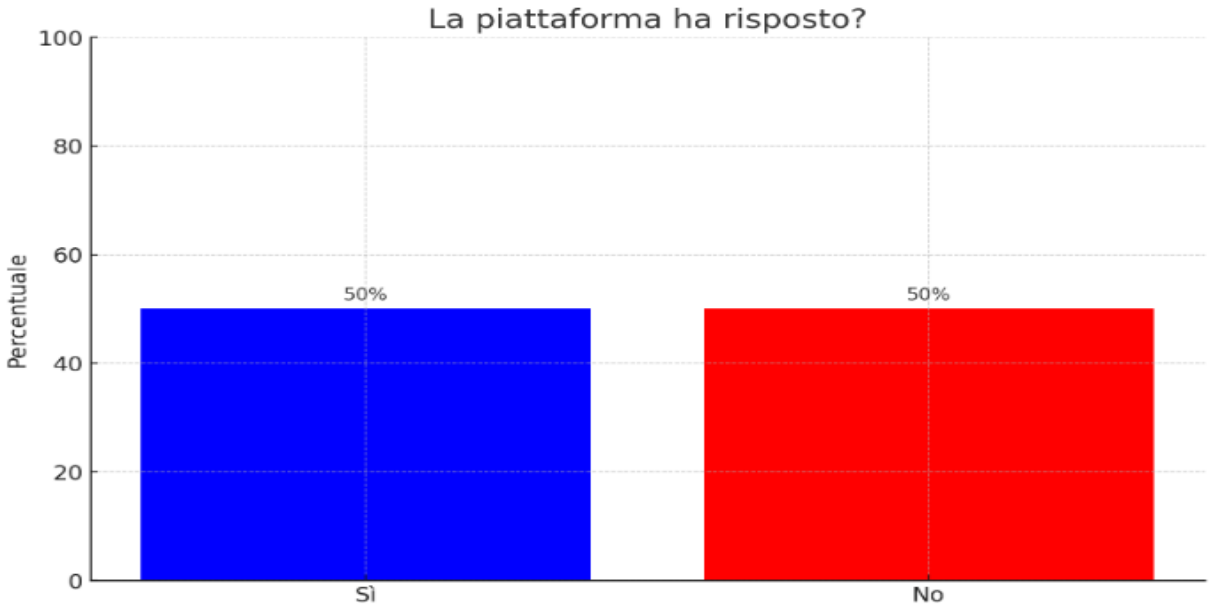


Fig. 4: Se la piattaforma ha risposto alla richiesta di spiegazioni.

Oltre alla censura sono stati segnalati anche casi più sottili, nei quali, ad esempio, venivano posti dei *disclaimer* o comunque degli annunci sotto certi tipi di contenuti con l'obiettivo, da parte della piattaforma digitale, di influenzare i fruitori inducendoli a non dare loro conto. In un caso, un utente ha cercato su Google un paper scientifico pubblicato su rivista internazionale inerente a taluni rischi vaccinali durante il periodo del COVID[23], ed ha lamentato di trovarsi dinnanzi non ad uno ma a ben due avvisi da parte di Google stesso che lo invitavano a diffidare dalle informazioni che avrebbe reperito.

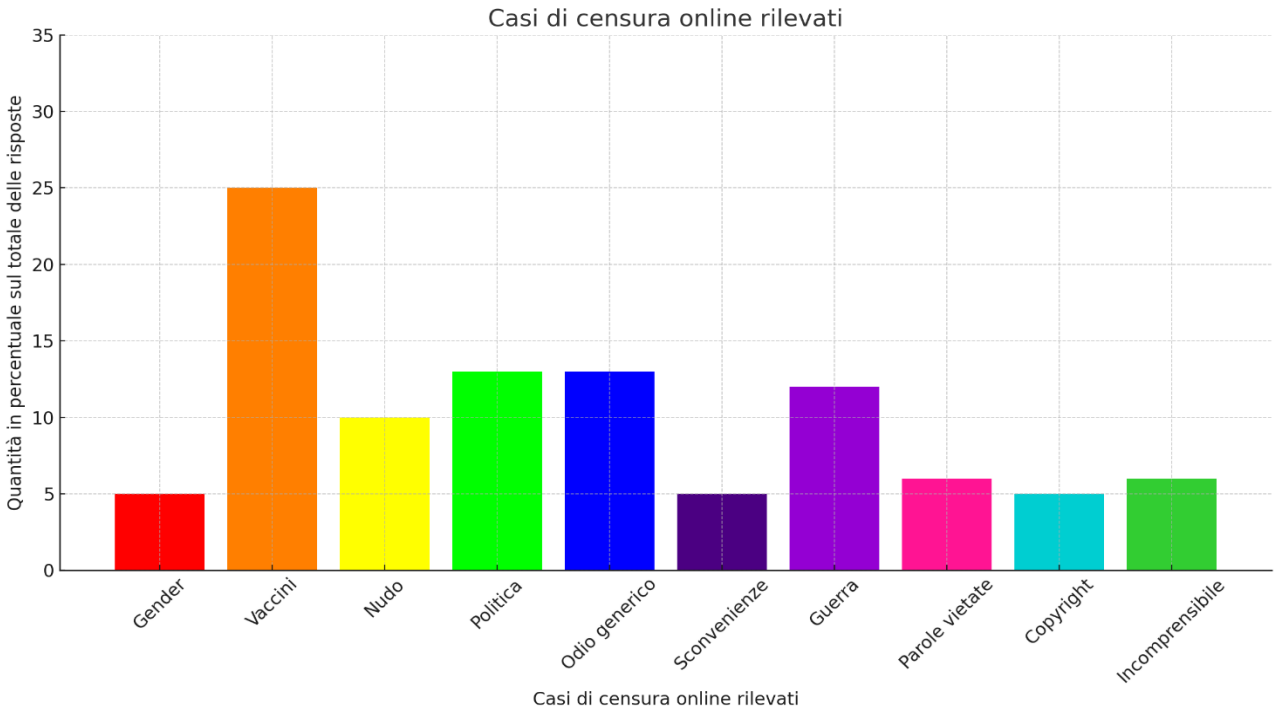


Fig. 4: Argomenti dei contenuti avverso i quali le piattaforme hanno preso provvedimenti censorei

Il tema dei vaccini è quello citato più spesso nei dati raccolti: molti utenti lamentano di aver subito limitazioni del proprio account anche solo per aver "postato foto uscite sui giornali italiani un anno prima" che sono risultate poi solo parzialmente corrispondenti al vero, e pertanto ritenute *tout court* false dalla piattaforma con conseguenti limitazioni sull'account di chi le aveva caricate.

Decisamente sotto al tema dei vaccini figurano numerosi casi di censura per la generica espressione di opinioni politiche, insieme a messaggi definibili come di "odio generico" nella misura in cui non figurano in alcuna categoria particolare ma comprendono quegli interventi scomposti, insultanti o comunque volgari su qualsiasi argomento.

La categoria delle opinioni politiche censurate pone certamente problemi particolari: nel valutare tali casi di censura, emerge in primo luogo la possibilità di una reticenza da parte dell'utente di auto-riportare in modo oggettivo il caso concreto, insieme alla necessità di interrogarsi su quanto questo si potesse configurare come effettivamente lesivo o pericoloso, indipendentemente dal suo contenuto. Nei casi segnalati come "politici", si rileva inoltre un discreto successo per ciò che concerne le procedure di ricorso rispetto ad altri temi: più volte a utenti diversi è stata rimossa la sanzione precedentemente comminata loro, o è stato ripristinato il contenuto rimosso.

[23] Si trattava di "Safety of COVID-19 Vaccines in Patients with Autoimmune Diseases, in Patients with Cardiac Issues, and in the Healthy Population https://www.mdpi.com/2076-0817/12/2/233", di L. Frasca, G. Ocone e R. Palazzo.

Affine ma separata dalla categoria delle opinioni politiche propriamente dette vi è quella riguardante i contenuti inerenti alla guerra in Ucraina, negli ultimi anni, comprensibilmente, particolarmente preponderante. Il tema è separato da quello delle opinioni politiche dal momento che contiene due diversi sottoinsiemi, soltanto uno dei quali compatibile con il tema delle opinioni politiche. Il primo attiene all'espressione di contenuti ritenuti dalla piattaforma discutibili o mal posti, in modo del tutto simile a quanto accade per le opinioni politiche. Il secondo, invece, attiene alla semplice informazione (o presunta tale) di notizie sulla guerra – al netto ovviamente di potenziali valutazioni personali che farebbero rientrare i casi nel primo insieme. In questo senso si rilevano procedure di de-indicizzazione o de-monetizzazione (laddove la piattaforma preveda sistemi per guadagnare sui contenuti pubblicati, come YouTube o Twitch.tv). Il secondo insieme appare quindi contrastato con procedure più sottili: non rimosso di per sé, ma disincentivato nella sua ideazione.

I temi tradizionali della censura online, come la rimozione dei contenuti in violazione di copyright o della nudità online o la proibizione all'utilizzo di espressioni particolarmente volgari o vietate (si pensi alla bestemmia) sembrano invece essere meno frequenti, o meno percepite come censoree. In questo senso pare registrarsi maggiormente una sorta di "consenso diffuso" su quanto sia accettabile in determinati spazi della rete. Si segnalano comunque un numero rilevante di lamentele circa la rimozione di contenuti di nudo ritenuto artisticamente valido o accettabile per gli standard della piattaforma.

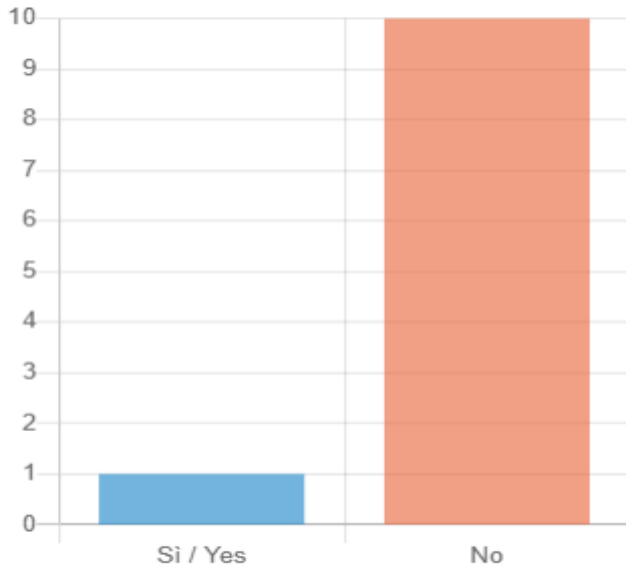


Fig. 5: Il grafico mostra se l'account interessato dal provvedimento era ancora sospeso al tempo della risposta sul questionario.

Per ciò che concerne la longevità dei provvedimenti, almeno per ciò che concerne le piattaforme digitali più diffuse, sembra esistere una notevole predilezione verso gli allontanamenti temporanei, anziché permanenti. La domanda è stata tuttavia per lo più ignorata dai partecipanti, e pertanto i dati sono pochi e parzialmente contraddittori con quelli di cui alla figura 2, dove era stato rilevato un buon numero di "chiusure dell'account", intendendo con ciò evidentemente un provvedimento irreversibile che costringesse l'utente, se avesse voluto tornare a partecipare alla comunità online a seguito dell'allontanamento, a crearne uno completamente nuovo.

La predominante temporaneità di questi provvedimenti, o "ban", in gergo, è tuttavia interessante perché si distingue completamente dallo standard utilizzato invece nelle comunità online propriamente "virtuali", o per meglio dire "digitali", come quelle dei giochi online o del giuoco di ruolo, laddove invece di un account nominativo, su modello di quello dei social network, se ne ha

tendenzialmente uno anonimo o pseudonimizzato, con il quale interagire, in un mondo virtuale, con altri giocatori o membri della comunità, anch'essi dotati di un account anonimo o pseudonimizzato.

In questi casi, la stragrande maggioranza dei ban sono permanenti già dalla prima offesa[24], e spesso comminati in modo piuttosto discutibile ed arbitrario[25] e, anche se in alcuni casi non può venire del tutto negato l'accesso al servizio da parte del contravventore, questi continuerà a partecipare con un account "marchiato" (o "flaggato", in gergo) da un precedente ban[26]. È superfluo dire che queste punizioni esagerate e perduranti non hanno nel corso del tempo affatto limitato né il problema degli *haters* né dei *cheaters,* intendendo con ciò gli utenti dal comportamento nocivo e di quelli che, con l'ausilio di software di terze parti, ottengono trucchi e vantaggi illeciti.

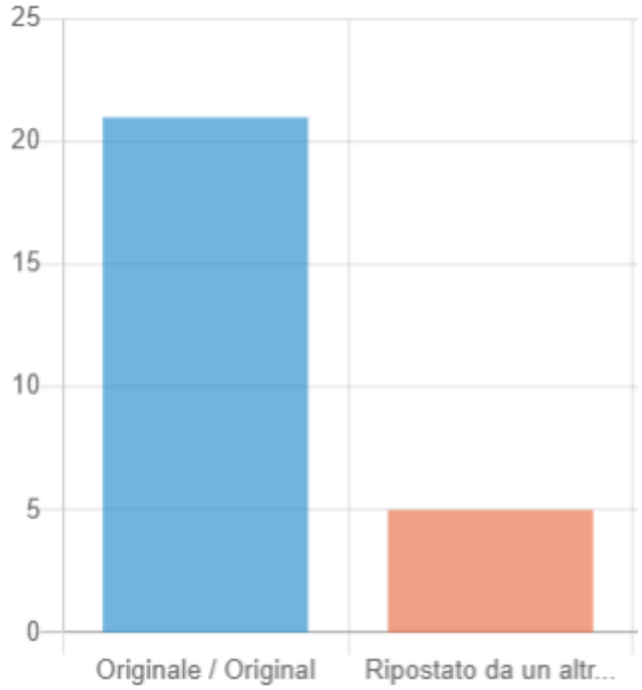


Fig. 6: Distribuzione del contenuto in base alla sua origine: se partorito dalla mente dell'utente che ha subito poi il provvedimento censorio, o se semplicemente ripreso altrove e ripostato da quest'ultimo sulla piattaforma interessata.

Secondo i dati raccolti, emerge una predominanza di contenuti originali tra quelli che portano a sanzioni o provvedimenti di allontanamento dagli spazi digitali. La prevalenza di questi ban può indicare una

[24] Mantenendo l'esempio di cui alla nota 10, si noti come nei *Call of Duty Security and Enforcement Policy* questo concetto sia espresso chiaramente: "[*even at] first offense, user may be permanently suspended from playing the game online"*. https://support.activision.com/articles/call-of-duty-security-and-enforcement-policy

[25] Di norma sono utilizzati volutamente termini generici: ancora sull'esempio di Call of Duty, si legge: "at any time, the account may be returned to normal state, placed in a limited state again, or it may receive a temporary or permanent ban."

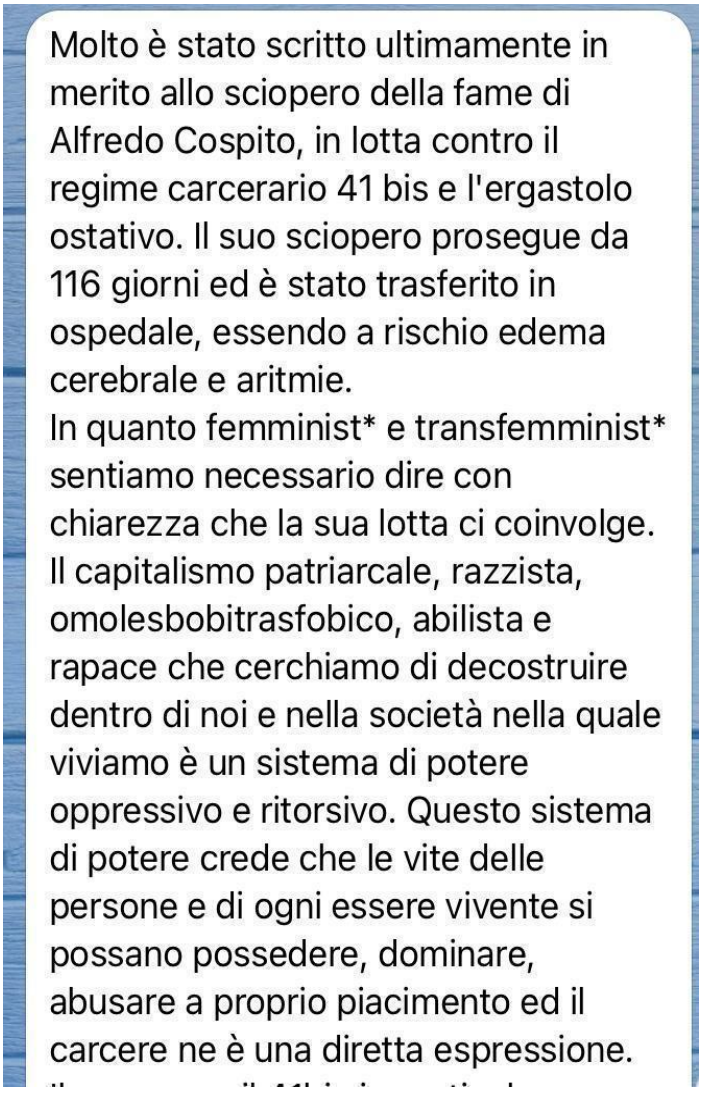
[26] E' il caso del discutibilissimo sistema VAC, o Valve Anti Cheat, il sistema più in voga per la protezione del gioco online onesto e non truccato nel sistema di *digital delivery* più diffuso attualmente nel mondo dei videogiochi: Steam, di Valve Corporation. Tramite Steam è possibile acquistare ed accedere ad un catalogo di decine di migliaia di videogiochi. Se in uno di tali giochi, protetti da VAC, è rilevata una scorrettezza da parte di un utente, questi è allontanato dal gioco in questione ed il suo account è "marchiato" anche al di fuori del gioco stesso, portandosi dietro l'infamia del ban per il resto dei suoi giorni: https://help.steampowered.com/it/faqs/view/571A-97DA-70E9-FF74

maggiore visibilità o un impatto emotivo più forte di questi contenuti, che li rende più soggetti a segnalazioni e revisioni da parte delle piattaforme. È possibile che i contenuti originali, spesso espressioni dirette di opinioni personali, idee politiche, o manifestazioni artistiche, siano più propensi a incitare reazioni, sia positive che negative, rispetto ai contenuti ripostati che possono essere percepiti come meno personali. Inoltre, a livello pratico, il contenuto ripostato è spesso tratto da un link terzo, o, nel caso dei social network nello specifico, da una pagina diversa da quella personale dell'utente che ripubblica. Perciò, è spesso sufficiente oscurare il contenuto alla fonte per raggiungere l'obiettivo di renderlo non visibile, senza dover intaccare direttamente l'account dell'utente ripostatore.

Alcuni esempi concreti

Si riportano qui di seguito alcuni esempi concreti di censura, reale o percepita, subìta da alcuni partecipanti al questionario, che per interesse, rilevanza o generale magnitudine del provvedimento sono stati ritenuti particolarmente interessanti, e ai quali è dedicata questa sezione specifica dello studio.

Il caso del post su Cospito, transfermminismo e capitalismo patriarcale



Il contenuto censurato era un post originale pubblicato ed ideato dall'utente su Instagram, in cui si esprimeva solidarietà ad Alfredo Cospito, collegando il suo sciopero della fame contro il regime del 41-bis a una critica transfemminista del capitalismo patriarcale e dei sistemi di potere oppressivo.

Alfredo Cospito, militante anarchico, divenne oggetto di un intenso dibattito pubblico tra il 2022 e il 2023 per il suo sciopero della fame contro l'applicazione del 41-bis, misura normalmente riservata a detenuti mafiosi o terroristi. La vicenda ha assunto un valore simbolico nelle discussioni su diritti carcerari, dissenso politico e gestione della sicurezza.

Instagram ha rimosso il post per presunta violazione delle linee guida su "violenza e organizzazioni pericolose", verosimilmente a causa del riferimento a un soggetto associato a movimenti anarchici considerati sensibili dai filtri automatici della piattaforma. L'utente non ha presentato ricorso.

Le stesse linee guida citate come matrice del provvedimento esplicitamente indicano "Siamo consapevoli del fatto che gli utenti potrebbero condividere contenuti che includono riferimenti a organizzazioni e persone pericolose designate per segnalare, condannare o discutere in modo neutrale delle stesse o delle loro attività. Le nostre normative sono state concepite per lasciare spazio a questi tipi di discussioni limitando al contempo il rischio di possibili casi di violenza offline. [...] Le nostre normative consentono discussioni su diritti umani di persone designate o membri di entità pericolose designate, a meno che i contenuti presentino altre forme di elogio, sostegno sostanziale o rappresentazioni di entità designate o altre violazioni delle normative, come l'incitamento alla violenza".

Il caso mostra un "overblocking" tipico della moderazione algoritmica: un contenuto politico e critico viene assimilato a supporto di soggetti considerati pericolosi. Ad un'analisi oggettiva, il testo esprimeva solidarietà politica nei confronti di Cospito e interpretava il regime carcerario, incluso il 41-bis, come una manifestazione di un più ampio sistema di dominio e controllo sui corpi e sulle vite. Il post in questione sembrerebbe rientrare all'interno della cornice – tollerata dalla piattaforma – in tema di discussioni su diritti umani di persone designate o membri di entità pericolose designate. Il testo non presenta forme di elogio o sostegno sostanziale alla figura di Alfredo Cospito, né contiene espressioni di incitamento alla violenza.

Non vi erano immagini di violenza, né inviti a compiere atti ostili o incitamenti espliciti, ma il linguaggio era fortemente critico verso istituzioni e strutture di potere.

L'assenza di ricorso riflette la scarsa fiducia degli utenti nei meccanismi di contestazione e la percezione di scarsa trasparenza delle decisioni.

Contenuti legati al raduno elettorale della "Lega – Salvini Premier" a Milano

Caso riguardante la pubblicazione di numerosi tweet, alcuni dei quali oggi non più disponibili.

https://twitter.com/sherpa810/status/1129771885049389057
https://twitter.com/sherpa810/status/1129760533131988992
https://twitter.com/Politphilo/status/1129751829582163976
https://twitter.com/sherpa810/status/1129753992421154817
https://twitter.com/sherpa810/status/1129748677231300611
https://twitter.com/sherpa810/status/1129736162640367616
https://twitter.com/Simo_Pirro/status/1129735492050870274
https://twitter.com/AlbertoBagnai/status/1129733015469527045
https://twitter.com/sherpa810/status/1129734090570248193

L’utente aveva pubblicato su Twitter alcuni tweet e retweet relativi al raduno elettorale della “Lega – Salvini Premier” in Piazza Duomo a Milano (maggio 2019). I contenuti, sotto forma di fotografie e commenti neutrali, documentavano la manifestazione senza toni offensivi o incitamenti.

Nonostante ciò, l’account è stato improvvisamente sospeso: Twitter ha richiesto il numero di telefono per procedere allo sblocco, richiesta che l’utente ha rifiutato per principio. Dopo alcune settimane, l’account è stato riattivato senza spiegazioni, ma risultava soggetto a shadowban (limitazione occulta della visibilità).

La piattaforma non ha fornito alcuna motivazione né notifica formale.

Si ipotizza un’applicazione automatica delle norme sulla moderazione di "entità violente e cariche di odio" o di "condotta che incita all’odio". Tuttavia, può trattarsi anche di un errore algoritmico legato al contesto politico polarizzato o a precedenti interazioni dell’utente su temi sensibili, in particolare legate al tema del COVID e dei vaccini.

Nessuna opzione di ricorso è stata resa disponibile.

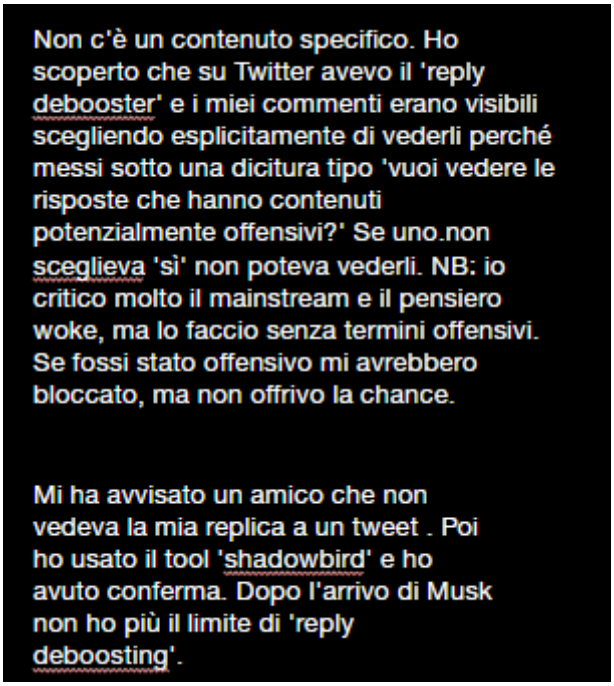
Durante campagne elettorali e temi politici divisivi, le piattaforme digitali adottano filtri più severi per prevenire contenuti di odio o manipolazione politica. In questo scenario, automatismi e controlli preventivi possono colpire anche contenuti neutrali e documentaristici, soprattutto quando associati a figure politiche e movimenti percepiti come polarizzanti.

Questo caso mostra diversi elementi critici ricorrenti: una assenza totale di trasparenza sulle ragioni del provvedimento; la moderazione opaca attuata tramite shadowban, senza avvisi all’utente, e perfino un potenziale errore algoritmico di classificazione, con sovrapposizione fra contenuto politico legittimo e contenuti sensibili.

Permane inoltre la mancanza di garanzie per il diritto di ricorso, fenomeno diffuso nei dati empirici raccolti.

Si tratta di un tipico esempio di “censura silenziosa” su contenuti politici non solo leciti ma parte integrante del dibattito democratico.

Commenti politici e critica al sistema “woke”



L’utente riferisce di non aver pubblicato un singolo contenuto specifico, ma di aver subito una limitazione sistematica della visibilità dei propri commenti su Twitter (oggi X). Si trattava di contenuti originali, principalmente critiche politiche e osservazioni sul cosiddetto “pensiero woke”, formulate, almeno secondo quanto riportato dall’utente, senza toni offensivi.

Il profilo risultava soggetto a *reply deboosting*: i commenti venivano nascosti automaticamente sotto un avviso: “Vuoi vedere le risposte che potrebbero contenere contenuti potenzialmente offensivi?”, rendendoli visibili solo se gli altri utenti cliccavano manualmente sull’opzione “sì”.

Un amico dell’interessato ha segnalato di non vedere una sua replica, confermando la limitazione. L’utente ha inoltre utilizzato un tool di verifica dei casi di *shadowban*, che ha confermato il declassamento del profilo nelle risposte.

Dopo il cambio di proprietà della piattaforma e l’arrivo di Elon Musk, l’utente riferisce che il proprio profilo non risultava più soggetto ai limiti di visibilità.

Non è stato presentato alcun ricorso formale alla piattaforma.

Il caso esemplifica una forma di censura algoritmica silenziosa: nessuna rimozione esplicita, nessuna notifica, ma una riduzione della visibilità delle interazioni tramite filtri automatici. Questa dinamica conferma due tendenze emerse nel questionario: da una parte l’incertezza degli utenti nel comprendere i motivi della limitazione; dall’altra la difficoltà di distinguere tra moderazione legittima e meccanismi opachi di penalizzazione automatica.

Il ripristino della visibilità senza alcuna spiegazione evidenzia inoltre la natura altamente discrezionale e imprevedibile degli algoritmi di ranking e moderazione.

Ripubblicazione di una notizia ANSA sui potenziali effetti collaterali dei vaccini Covid-19

L'utente ha ripostato su Facebook un articolo dell'ANSA riguardante uno studio statunitense sulle possibili cause delle miocarditi associate alla vaccinazione anti-Covid. Si trattava quindi di contenuto ripreso da una fonte giornalistica nazionale, non modificato e non accompagnato da commenti aggiuntivi.

Facebook ha rimosso il post per assenza di fact-checking. Nonostante la fonte fosse un'agenzia di stampa mainstream e verificata, l'algoritmo ha probabilmente classificato il tema come "alto rischio" nel contesto del Covid-19, applicando filtri preventivi pensati per limitare la diffusione di contenuti sanitari sensibili. L'utente non ha presentato ricorso.

Il caso illustra un tipico esempio di moderazione severa, immediata e decontestualizzata: anche contenuti provenienti da fonti giornalistiche riconosciute possono essere rimossi quando toccano temi sanitari ad alta sensibilità algoritmica. La rimozione di materiale informativo mainstream indica una difficoltà delle piattaforme nel distinguere tra disinformazione e semplice condivisione di notizie legittime, soprattutto in ambiti dove i sistemi di fact-checking vengono applicati in modo rigido e generalizzato. L'assenza di ricorso conferma la sfiducia degli utenti verso la trasparenza e l'efficacia dei meccanismi di contestazione.

Ripubblicazione di un video sulla guerra in Ucraina

L'utente aveva condiviso su Facebook un contenuto proveniente da terze parti: un video proveniente da YouTube, che offriva una lettura alternativa della guerra in Ucraina. Il post è stato oscurato automaticamente dalla piattaforma, che ha applicato l'etichetta di "informazioni false" avendo ritenuto assente il fact-checking sulla notizia.

Si tratta di un caso di moderazione potenziata su temi particolarmente sensibili e oggetto di campagne, reali o percepite, di disinformazione, per i quali Facebook applica protocolli più restrittivi.

L'utente non ha potuto presentare ricorso, riferendo che la piattaforma non offriva alcuna opzione per contestare il provvedimento.

Dall'inizio dell'invasione russa dell'Ucraina (24 febbraio 2022), i social network hanno introdotto meccanismi particolarmente severi di fact-checking e limitazione dei contenuti legati al conflitto, soprattutto quando riconducibili a fonti ritenute parziali o a narrazioni alternative rispetto al quadro informativo prevalente. Ciò ha prodotto numerosi casi di segnalazioni o rimozioni basate su automatismi.

Il caso evidenzia un'applicazione particolarmente rigida dei sistemi di verifica su temi geopolitici delicati, nonché un sistema di bloccaggio e rimozione, con conseguenze sull'utente ripostante, anche per contenuti non originali, replicati cioè da altre fonti ma non necessariamente accompagnati da commenti manipolativi. Emerge una difficoltà strutturale per gli utenti di contestare la decisione, soprattutto quando la piattaforma non prevede un canale di ricorso per contenuti classificati come informazioni false.

L'episodio si inserisce nella tendenza emersa dal questionario: i contenuti ripostati su temi sensibili (vaccini, guerra, salute pubblica) vengono frequentemente limitati senza possibilità di revisione.

Post eterogenei su vaccini, cambiamento climatico e guerra in Ucraina

L'utente segnala di aver condiviso su Facebook tre contenuti ripostati:
1. La traduzione di un articolo del Wall Street Journal critico verso l'obbligo vaccinale anti-Covid;
2. Un manifesto firmato da centinaia di scienziati che contestava l'allarmismo climatico;
3. Un'immagine del battaglione Azov, legata al conflitto ucraino.

Il mix di temi trattati, tutti inerenti a temi sensibili: vaccini, cambiamento climatico e guerra, rifletteva un insieme eterogeneo di posizioni politiche e scientifiche.

Facebook ha sospeso alcune funzioni dell'account per presunta violazione della *community policy*. È probabile che i sistemi automatici abbiano classificato i contenuti come potenzialmente fuorvianti, sensibili o associati a gruppi pericolosi (probabilmente facendo leva sulla natura neo-nazista del Battaglione Azov), attivando varie forme di limitazione algoritmica.

Non è stato presentato alcun ricorso.

Questo caso mostra un esempio che può essere ritenuto di *bundle censorship*: contenuti diversi per tema, fonte e tono vengono trattati come un'unica categoria rischiosa dai filtri algoritmici, che reagiscono alla ricorrenza di parole chiave sensibili più che al contesto.
La sospensione parziale delle funzioni, anziché la rimozione diretta, evidenzia un uso crescente di

misure intermedie che penalizzano l'utente senza una motivazione chiaramente articolata. Anche qui l'assenza di ricorso indica sfiducia nei meccanismi di contestazione o impossibilità pratica di attivarli.

Post a contenuto potenzialmente sessuale

L'utente ha pubblicato su Facebook una foto personale, in una risoluzione estremamente bassa, raffigurante un uomo nudo di schiena mentre una donna sembra massaggiarlo. A seguito della pubblicazione, la piattaforma ha limitato alcune funzionalità dell'account (es. dirette streaming, pagamenti). Non è stato presentato ricorso.

Facebook ha indicato soltanto una generica violazione degli "Standard della community".

Si presume un riferimento alla norma su immagini di nudo o contenuti a sfondo sessuale, anche se il contenuto — in base alla descrizione — non appare rientrare nelle ipotesi vietate dalle linee guida (assenza di genitali visibili, assenza di atti sessuali).

La moderazione su nudo e sessualità su Facebook è notoriamente ampia e ambigua, con elevata dipendenza da sistemi automatici che tendono a rimuovere contenuti che possano anche solo vagamente sembrare sessuali.

Il caso segnala un'applicazione eccessiva e non contestualizzata dei filtri per il nudo, peraltro subordinati a una normazione estremamente vaga ("presenza di derivati di atti sessuali"), usata come criterio elastico per giustificare la rimozione.

"Al rogo" il dizionario Zanichelli

L'utente è stato sospeso per 3 giorni da Facebook a causa di un commento originale rivolto ad un altro utente che recitava: «al rogo subito. Tanto oramai lo usiamo in pochissimi». Il significato reale, chiarito successivamente dall'autore, era riferito ironicamente ai dizionari Zanichelli citati nella discussione, e non a persone fisiche.

Ciononostante, la piattaforma ha classificato il commento come violenza o istigazione alla violenza, ritenendolo un invito a danneggiare altri utenti. Pur avendo l'utente presentato richiesta di riesame, la decisione è stata confermata.

Se da una parte è ovvio che il commento presentasse una struttura linguistica ambigua, senza punteggiatura e con un riferimento non immediatamente riconoscibile, dall'altra è altrettanto evidente come il sistema di moderazione automatica abbia interpretato la frase come una minaccia credibile: una conclusione che nessun essere umano avrebbe potuto ragionevolmente raggiungere. La mancanza di considerazione del contesto ha portato quindi evidentemente ad una sovraestensione della sanzione.

Il caso evidenzia pertanto ovvie carenze della moderazione algoritmica nel comprendere il contesto semantico, un rischio elevato di sanzioni errate in presenza di ironia o linguaggio colloquiale e pure l'inefficacia degli strumenti di ricorso, che non ha corretto l'errore iniziale neppure in un caso così palese. Siamo dunque in presenza di overblocking su espressioni che, pur con toni forti, non avevano alcun intento violento e, visti gli oggetti cui erano riferiti, non avrebbero materialmente potuto averlo.

"Uccidi il maschio bianco cisetero in te"

L'utente riferisce di aver pubblicato su Instagram una foto raffigurante una scritta su un muro: «uccidi l'uomo bianco cisetero che c'è in te».

Si trattava di un contenuto originale, rimosso automaticamente dalla piattaforma con la motivazione di incitazione all'odio. L'utente non ha richiesto il riesame.

Instagram ha interpretato il messaggio come un attacco basato su caratteristiche protette (razza, sesso, identità di genere e orientamento sessuale). La piattaforma, in assenza di una chiara contestualizzazione da parte dell'utente, ha applicato la normativa che vieta i contenuti percepiti come violenti o discriminatori verso gruppi specifici.

Lo slogan rientra nel repertorio comunicativo di frange transfemministe radicali e mira alla decostruzione simbolica del modello sociale del "maschio bianco cis-eterosessuale", identificato come beneficiario di privilegi patriarcali. Sebbene il linguaggio adottato sia volutamente provocatorio e iperbolico, esso va letto sul piano culturale, politico e soprattutto metaforico.

Il caso mostra i limiti della moderazione automatica nel distinguere la critica politico-sociale da attacco a una categoria protetta, nonché di cogliere l'uso metaforico del verbo "uccidere", riferito (si dovrebbe supporre) alla decostruzione di un modello di maschilità e non alla violenza fisica.

L'assenza di ricorso elimina ogni possibilità di verifica umana del contesto, contribuendo a casi di censura su contenuti certamente provocatori ma non finalizzati a incitare odio reale.

Conclusioni

Dall'analisi dei casi raccolti tramite il questionario, nonché dalle premesse riportate nel capitolo introduttivo, emerge un quadro concreto nel quale la rimozione o la limitazione dei contenuti non riguarda soltanto materiale effettivamente illecito o manifestamente in violazione dei termini d'uso

delle piattaforme. Al contrario, una parte significativa degli episodi riportati coinvolge contenuti leciti, spesso di natura politica, informativa o generalmente opinabile, ma che non risultano chiaramente incompatibili con le policy dichiarate dalle piattaforme, né figurano illeciti, e meno che mai integrano alcuna fattispecie di reato. Ciò evidenzia un problema sistemico di moderazione eccessiva (*over-moderation*), generata in larghissima e forse esclusiva parte da filtri automatici e da interpretazioni cangianti delle regole interne alle piattaforme.

A fronte di questi interventi, gli obblighi introdotti dal Digital Services Act, che impongono alle piattaforme di fornire all'utente motivazioni puntuali e comprensibili per ogni provvedimento, risultano spesso inosservati o adempiuti in modo minimale. In numerosi casi analizzati, infatti, gli utenti non hanno ricevuto alcuna spiegazione specifica circa i provvedimenti comminati avverso di loro, né sono stati istruiti sulle ragioni concrete della rimozione, con l'effetto di lasciare l'interessato nell'impossibilità pratica di comprendere la natura dell'infrazione contestata. Ciò rende peraltro inattuabile qualsiasi progetto di "rieducazione" di un utente che ha sbagliato: anche in eventuali casi gravi ed espliciti di violazione di norme, non potendo comprendere a fondo il suo eventuale errore, qualsiasi soluzione è per lui preclusa da un deficit informativo.

Parallelamente, anche un secondo requisito del DSA, quello relativo agli strumenti effettivi di ricorso, mostra gravi criticità: solo in una minoranza di casi i partecipanti hanno tentato di contestare la decisione e, laddove avrebbero voluto, spesso la piattaforma non offriva un meccanismo di reclamo facilmente accessibile o chiaramente visibile. Anche questa mancanza procedurale contribuisce ad alimentare un clima di sfiducia, rendendo di fatto inefficace uno dei più importanti presidi di garanzia pensati dal legislatore europeo, oltre ad acuire il problema di incertezza generale che aleggia su questi provvedimenti.

Successive opacità si manifestano anche nelle forme più subdole di sanzioni, come lo shadowbanning, pratica per cui la visibilità dei contenuti viene limitata senza notifica all'utente. Questa tipologia di intervento sottrae all'interessato la possibilità stessa di rendersi conto della sanzione subita e, conseguentemente, di esercitare qualsiasi diritto di contestazione o richiesta di chiarimenti. Si tratta di un punto particolarmente critico, poiché, come si è visto ampiamente, nonostante la trasparenza sulle misure sanzionatorie delle piattaforme digitali rientri a pieno titolo tra gli obblighi posti dal DSA, risulta ad oggi scarsamente implementata.

È importante sottolineare, infine, che il questionario è stato somministrato in una fase molto iniziale dell'applicazione del Digital Services Act. Durante questo periodo, alcuni obblighi erano solo formalmente in vigore, mentre le piattaforme non avevano ancora sviluppato, implementato o consolidato i meccanismi necessari per garantirne l'effettiva applicazione. Per questa ragione, pur essendo significativi, i dati raccolti devono essere interpretati come una fotografia preliminare e non per forza conclusiva.

Sarà quindi necessaria ulteriore ricerca, condotta in un momento più avanzato dell'attuazione del DSA, per verificare se le garanzie procedurali introdotte dal regolamento europeo riusciranno concretamente a superare l'attuale contestabile situazione di fatto, e mitigare i rischi di censura privata, migliorando la trasparenza e rendendo più equilibrato il rapporto tra utenti e piattaforme digitali.

Glossario

Account: profilo utente che incarna un soggetto reale in un ambiente digitale. Può riportare l'identità reale dell'utente o mascherarla tramite uno pseudonimo.

Account "flaggato" (marchiato): account contrassegnato da un precedente provvedimento (es. ban), con effetti reputazionali o di visibilità.

Ban (temporaneo / permanente): esclusione dall'accesso a una piattaforma

De-indicizzazione: rimozione o abbassamento della visibilità di un contenuto nei risultati di ricerca o nelle raccomandazioni.

De-monetizzazione: esclusione di un contenuto dai meccanismi di guadagno (pubblicità, abbonamenti), senza rimuoverlo.

Disclaimer: avviso posto sotto/accanto a un contenuto per guidare la fruizione (es. "informazioni controverse"); non rimuove, ma orienta.

Haters: utenti che esprimono comportamenti ostili o molesti, spesso ripetuti, verso persone o gruppi.

Piattaforma digitale: servizio online che ospita e distribuisce contenuti e interazioni tra utenti (es. social network).

Reclamo: richiesta formale di spiegazioni o revisione di una sanzione/moderazione decisa dalla piattaforma.

Rimozione del contenuto: cancellazione di un post/media ritenuto illecito o contrario alle regole.

Shadowbanning: limitazione occulta della visibilità di un contenuto/account (l'utente vede il proprio contenuto, altri no o molto meno).

Appendice: domande del questionario

Di seguito sono riportate le domande poste durante il questionario, insieme con la lista delle loro risposte possibili qualora ve ne fossero di selezionabili. Le domande sono precedute da lettere ad indicare la loro priorità gerarchica (le domande B seguono le domande A, le domande C seguono le B e così via). Nel caso in cui alcune domande fossero state poste condizionalmente, soltanto in seguito a determinate risposte, tali risposte sono indicate tra parentesi quadre. Tra parentesi è indicato il tipo di risposta consentita durante la compilazione del questionario.

A1 | **Su quale piattaforma hai aperto il tuo account? (Solo una risposta possibile)**

- ☐ Facebook
- ☐ Instagram
- ☐ LinkedIn
- ☐ Pinterest
- ☐ Reddit
- ☐ Snapchat
- ☐ TikTok
- ☐ Twitter
- ☐ YouTube
- ☐ Altro (Testo aperto)

A2 | **Quanto tempo fa hai aperto il tuo account (Testo aperto)**

- ☐ meno di un mese
- ☐ tra 1 e 6 mesi
- ☐ tra 6 mesi e 1 anno
- ☐ 1-2 anni
- ☐ più di 2 anni

aa2 | In quale paese ti trovavi nel momento in cui hai aperto il tuo profilo?

- ☐ Italia
- ☐ Altro

A3 | **Nome dell'account e link (se disponibile) (Opzionale)**

A4 | **Come descriveresti il tuo account (al momento del provvedimento di censura)? (Più scelte possibili)**

- ☐ Account personale
  - privato
  - pubblico
- ☐ Account professionale, aziendale o istituzionale

A5 | **Di cosa si occupa il tuo account? (Massimo 2 risposte possibili)**

- ☐ Contenuti personali
- ☐ Informazione
  - (Opzionale) Specificare
- ☐ Istruzione e cultura
- ☐ Attività politica o sindacale

- (Opzionale) Specificare
- ☐ Attività non profit e volontariato
  - (Opzionale) Specificare
- ☐ Promozione e/ ovendita di prodotti o servizi
  - (Opzionale) Specificare
- ☐ Arte, fotografia, video, musica
- ☐ Sport
- ☐ Case editrici
- ☐ Cucina
- ☐ Salute e fitness
- ☐ Altro
  - Specificare

A6 | **Che tipo di provvedimento hai subito da parte della piattaforma? (Più risposte possibili)**

- ☐ A. Chiusura dell'account
- ☐ B. Sospensione temporanea dell'account
- ☐ C. Rimozione di un contenuto creato da te stesso/a/
- ☐ D. Rimozione di uno o più links a un contenuto creato da altri
- ☐ E. Presunto oscuramento dell'account o di uno o più contenuti pubblicati sul tuo account ("shadowbanning")
- ☐ F. Sospensione di alcune funzionalità dell'account (impossibilità di fare dirette streaming, impossibilità di postare, ...)
  - Specificare (testo aperto)
- ☐ G. Altro/Other
  - Specificare (testo aperto)

B1 | [In caso A] **Hai subito altre volte la chiusura di un account su questa piattaforma? (Sì/no)**

BB1 | La chiusura/sospensione dell'account era avvenuta per le stesse ragioni della volta in oggetto?

- Sì
- No

BB2 | L'account è stato riaperto?

- Sì
- No

BB3 | Per quanto tempo l'account è stato sospeso?

- Settimane
- Mesi
- Anni

BB4 | Al momento, l'account è ancora sospeso?

- Sì
- No

B2 | [In caso A,B**] Per quanto tempo l'account è stato chiuso/sospeso? (Specificare se lo è ancora) (Testo aperto)**

B3 | [In caso C, D, E] **Quale contenuto è stato oggetto del provvedimento? (Più opzioni possibili)**

- ☐ Testo
- ☐ Foto
- ☐ Video
- ☐ Altro
  - Testo aperto

BB6 |Descrivi brevemente il contenuto oggetto del provvedimento (testo aperto)

B4 | [In caso C o D] **La piattaforma ti ha segnalato la rimozione del contenuto? (Sì/no e possibilità di allegare screenshot)**

- ☐ No
- ☐ Sì
  - C1: Se sì, che cosa ti ha riferito?
  - C2: Se possibile, allega il messaggio da parte della piattaforma
  - C3: La piattaforma ha specificato se il provvedimento fosse conseguente ad una segnalazione da parte di un altro utente?

B5 | [In caso A, B, E, F e G] La piattaforma ti ha comunicato la ragione del provvedimento? (Sì/no)

- ☐ Sì
- ☐ No

D1: Se no, hai chiesto spiegazioni alla piattaforma riguardo al provvedimento da te subito?

- No
- Sì

D2: Se sì, la piattaforma ha risposto?

- No
- Sì
  - **E1: Se sì, cosa ha risposto?** (Testo aperto)
  - **E2: Dopo quanto tempo** (Testo aperto)
  - **E3: Se possibile, allega il messaggio da parte della piattaforma**

B6 | [In caso E] Da quali elementi hai dedotto di aver subito shadowbanning (risposta aperta)

B7 | [In caso E] Hai chiesto chiarimenti alla piattaforma riguardo al presunto shadowban? (sì/no)

- ☐ **H1**: Se sì, la piattaforma ha risposto?
  - Sì
  - No

- ☐ **H2**: Che cosa ha risposto? (Se possibile, copiare e incollare il messaggio della piattaforma) (testo aperto)
- ☐ **H3**: Se sì, dopo quanto tempo?

- Settimane
- Mesi
- Anni

A7 | Il contenuto censurato era: (Una sola risposta)

- originale
- ripostato da un altro account/reposted from another account

A8 | Quando hai subito il provvedimento da parte della piattaforma? (Testo aperto)

A9 | Hai chiesto il riesame del provvedimento? (Testo aperto)

A10 | Ritieni di aver subito una forma di censura ingiustificata? (Sì/no)

- **G1 Perché ritieni di aver subito una forma di censura ingiustificata? (Sì/no con testo aperto opzionale)**