Onorevole Presidente,

Onorevoli Senatrici e Senatori,

Desideriamo in primo luogo rivolgere un sentito ringraziamento alla Commissione per averci concesso l'opportunità di fornire il nostro contributo e condividere alcune riflessioni in merito a una tematica di primaria importanza quale il fenomeno dei discorsi d'odio.

Il tema dell'odio e della violenza online è senza dubbio complesso e porta con sé sfide sempre nuove. Auspichiamo dunque che questo contributo, basato sulla nostra esperienza, possa aiutare a chiarire alcuni aspetti della lotta contro i fenomeni d'odio online e a far luce su ciò che abbiamo fatto e stiamo facendo per contrastarli.

L'esperienza di Meta comincia nel 2004, quando con l'obiettivo di dare alle persone uno strumento tramite il quale potersi connettere ed esprimersi liberamente venne creata Facebook. In pochissimo tempo, la piattaforma ha avuto una crescita esponenziale, arrivando a connettere miliardi di persone in tutto il mondo, in un modo impensabile fino a poco prima. Ciò ha creato nuove opportunità per la libera espressione e la partecipazione politica ma, come sempre accade con la nascita di mezzi così dirompenti, anche nuovi rischi.

In questi anni ci siamo trovati ad affrontare sfide inedite, come quella di proteggere gli utenti dall'odio e dalla violenza online salvaguardando al tempo stesso la loro libertà di espressione. Davanti a una sfida così grande, abbiamo anche commesso degli errori. Il nostro è stato tuttavia un processo di crescita: abbiamo imparato dai nostri sbagli e ci siamo concentrati su come affrontare questi problemi – dall'interno, investendo in persone, processi e tecnologie, e aprendoci all'esterno, collaborando con esperti e legislatori di tutto il mondo.

I contenuti d'odio danneggiano il nostro business

Riteniamo in primo luogo doveroso iniziare con una premessa estremamente importante riguardo a una tesi circolata molto negli ultimi mesi: non è in alcun modo vero che Facebook e Instagram traggono vantaggio dai contenuti dannosi o che le nostre piattaforme permettono la diffusione di messaggi di odio per ricavarne un profitto. Al contrario, il funzionamento stesso del nostro modello di business, basato sulla pubblicità, fa sì che i contenuti d'odio danneggino in primis la nostra azienda.

Nello specifico, Facebook e Instagram guadagnano attraverso la pubblicità che inserzionisti di vario tipo scelgono di promuovere sulle piattaforme. Nessun investitore ha interesse nel pubblicizzare il proprio brand accanto a contenuti che possono turbare utenti e potenziali clienti. Inoltre, gli stessi utenti non vorrebbero utilizzare una

piattaforma in cui non si sentono al sicuro o a loro agio e in cui sanno di non potersi esprimere liberamente.

La tesi secondo cui i nostri algoritmi sono intesi a promuovere contenuti d'odio non rispecchia dunque la realtà dei fatti e non tiene conto del nostro modello di business. Non c'è alcun motivo per cui non dovremmo voler fornire alle persone un'esperienza il più possibile positiva sulle nostre piattaforme, che è ciò che cerchiamo di fare giorno dopo giorno.

Gli algoritmi

Nell'approfondire la tematica degli algoritmi che operano sulle nostre piattaforme, è opportuno sottolineare come la convinzione diffusa che questi siano progettati per favorire contenuti divisivi e sensazionalistici è da considerarsi errata e fuorviante. In un certo senso, parti dei nostri algoritmi sono progettate per fare esattamente l'opposto.

Nel 2018, abbiamo annunciato che ci saremmo concentrati su come aiutare gli utenti ad avere interazioni sociali più significative sulle nostre piattaforme, principalmente promuovendo – attraverso un cambiamento nel nostro approccio alla classificazione dei contenuti – contenuti pubblicati da amici, familiari e Gruppi di cui gli utenti sono parte al posto di quelli provenienti dalle Pagine da loro seguite.

Sebbene fossimo consapevoli che tale cambiamento avrebbe portato le persone a trascorrere meno tempo su Facebook (dal momento che le Pagine pubblicano normalmente contenuti più coinvolgenti anche se meno significativi per gli utenti rispetto, per esempio, ai loro familiari), con un conseguente impatto sui profitti dell'azienda, abbiamo portato avanti con decisione questa scelta dal momento che le ricerche effettuate hanno mostrato come le conversazioni con amici e familiari influiscano positivamente sul benessere delle persone.

La nostra previsione si è rivelata corretta: il cambiamento nell'algoritmo ha portato a una diminuzione nel tempo totale trascorso su Facebook di 50 milioni di ore ogni giorno, migliorando al contempo la qualità del tempo trascorso sulla nostra piattaforma.

Peraltro, uno dei nostri obiettivi è quello di fornire alle persone la possibilità di esercitare maggiore controllo su ciò che visualizzano. A tale riguardo, offriamo già agli utenti la possibilità di "superare" l'algoritmo, permettendo loro di comporre di fatto il proprio Feed su Facebook.

Facebook e l'aumento della polarizzazione

Un ulteriore punto che viene spesso sollevato in relazione alle interazioni e ai contenuti che vengono pubblicati e visualizzati sui social media riguarda il fenomeno della polarizzazione. L'aumento della polarizzazione è stato oggetto di serie ricerche accademiche negli ultimi anni, che non hanno tuttavia portato ad un consenso unanime sulle sue cause. È però opportuno sottolineare che le prove attualmente disponibili non

supportano l'idea che Facebook, o i social media più in generale, siano la causa primaria della polarizzazione. L'aumento della polarizzazione politica precede infatti i social media di diversi decenni.

Se fosse vero che Facebook è la causa principale della polarizzazione, ci aspetteremmo di vedere quest'ultima aumentare laddove Facebook è più popolare – ma la realtà appare ben diversa. A riprova di ciò, è possibile osservare come in alcuni Paesi dove l'utilizzo di internet e di Facebook è aumentato si è assistito ad una diminuzione della polarizzazione.

Ciò premesso, sappiamo di avere un ruolo da svolgere per essere parte della soluzione: ecco perché continueremo a lavorare a miglioramenti che siano coerenti con il nostro obiettivo di rendere l'esperienza online delle persone più significativa. Tuttavia, limitarsi a incolpare Facebook per l'aumento della polarizzazione significa ignorare le cause più profonde di questi fenomeni, nonché ciò che le ricerche svolte effettivamente mostrano.

Investimenti per la protezione degli utenti

Alla luce di quanto premesso, vorremmo ricordare come negli ultimi anni Meta abbia fatto investimenti significativi per far sì che le persone si sentano al sicuro utilizzando i nostri servizi e, al contempo, libere di esprimersi sulle nostre piattaforme. In particolare, abbiamo assunto più di 45.000 persone con l'unico scopo di proteggere i nostri utenti e, solo nel 2021, abbiamo investito circa 5 miliardi di dollari a questo scopo, più di ogni altra *tech company*.

Abbiamo 15.000 esperti che, da oltre 20 location in tutto il mondo, analizzano i contenuti che vengono postati in più di 70 lingue, incluse tutte le lingue ufficiali dell'Unione Europea. Collaboriamo inoltre con più di 80 fact-checkers indipendenti a livello globale, di cui due in Italia (Pagella Politica e Open).

Standard della community

Nell'ambito degli sforzi messi in atto per proteggere i nostri utenti un ruolo chiave è svolto dagli Standard della community e dalle Linee guida della community, vale a dire le regole che tutti gli utenti di Facebook e Instagram devono accettare al momento dell'iscrizione e che definiscono cosa può essere condiviso e cosa no sulle nostre piattaforme.

Queste regole sono valide a livello globale e si applicano a tutti e a ogni tipo di contenuto. Coprono una vasta gamma di fattispecie, tra cui i contenuti che incitano all'odio o alla violenza, e sono definite in collaborazione con esperti in settori quali tecnologia, sicurezza pubblica e diritti umani e sulla base dei feedback ricevuti dalla nostra community. Per garantire che tutti possano esprimersi, prestiamo molta attenzione nel creare standard che includano punti di vista e opinioni diverse, soprattutto quelli di persone e comunità che altrimenti potrebbero essere trascurate o emarginate. Inoltre, è importante sottolineare come queste norme non siano scolpite nella pietra, ma

vengano continuamente aggiornate in base al modo in cui le persone utilizzano i nostri servizi e ai cambiamenti che avvengono nella società.

Per far rispettare queste regole usiamo una combinazione di tre elementi che agiscono in simultanea:

- Le segnalazioni da parte degli utenti;
- La revisione dei contenuti da parte di team di esperti;
- La tecnologia – in particolare basata su intelligenza artificiale – che viene costantemente migliorata per identificare proattivamente e rivedere i contenuti alla luce degli Standard della community.

L'obiettivo dei nostri Standard della community è quello di creare un luogo in cui le persone si sentano libere di esprimersi.

Ci teniamo inoltre a sottolineare che in alcuni casi, allo scopo di sensibilizzare l'opinione pubblica, consentiamo contenuti che sarebbero altrimenti contrari ai nostri standard, qualora si tratti di contenuti rilevanti e di pubblico interesse. Procediamo in questo modo solo dopo aver soppesato i benefici per l'interesse pubblico rispetto ai potenziali danni. Per effettuare queste valutazioni ci basiamo sugli standard internazionali in materia di diritti umani.

Il nostro report sull'applicazione degli Standard della community

Gli sforzi messi in atto in questi anni ci hanno permesso di conseguire importanti risultati. Ogni trimestre condividiamo pubblicamente i dati aggiornati rispetto all'applicazione degli Standard della community attraverso il nostro *Report trimestrale sull'applicazione degli Standard della community*.

Nel nostro ultimo report, pubblicato a febbraio, abbiamo condiviso i dati relativi al periodo che va da ottobre a dicembre dello scorso anno. In questo periodo, sia su Instagram che su Facebook, la diffusione dei contenuti che incitavano all'odio è stata tra lo 0,02 e lo 0,03%. In altre parole, solo 2-3 contenuti ogni 10.000 contenuti visualizzati contenevano incitamento all'odio.

In totale, nell'ultimo trimestre del 2021, su Facebook sono stati rimossi 17,4 milioni di contenuti in quanto violavano le nostre regole in materia di incitamento all'odio, mentre su Instagram ne sono stati rimossi 3,8 milioni.

Di questi contenuti in violazione, il 95,9% su Facebook e il 91,9% su Instagram è stato identificato da noi in modo proattivo, ancora prima che ci venisse segnalato dagli utenti. Per comprendere quanto i nostri sistemi siano migliorati da questo punto di vista, si tenga a mente che a fine 2017, quando abbiamo iniziato a pubblicare questi dati relativamente a Facebook, tale percentuale era del 23,6%.

Gli utenti possono fare ricorso alle nostre decisioni, tranne nei casi che implicano serie preoccupazioni per la sicurezza. Ripristiniamo i contenuti che ci accorgiamo di aver rimosso per errore, o quando le circostanze cambiano e lo richiedono. Ciò può accadere

in seguito a un ricorso da parte degli utenti ma anche se ce ne rendiamo conto autonomamente. Su Facebook, nell'ultimo trimestre del 2021, abbiamo ricevuto ricorsi da parte degli utenti relativamente a 769.000 dei contenuti rimossi perché in violazione delle nostre regole in materia di incitamento all'odio. In totale, circa 293.000 contenuti sono stati ripristinati dopo essere stati inizialmente rimossi per incitamento all'odio – 65.300 dei quali in seguito a dei ricorsi da parte degli utenti, mentre 228.000 sono stati ripristinati da noi in maniera autonoma. Maggiori informazioni sui contenuti oggetto di ricorso si possono trovare a questo link.

Per quanto riguarda i contenuti violenti o che istigano alla violenza, l'ultimo Report ci dice che la loro diffusione tra ottobre e dicembre del 2021 è stata dello 0,03-0,04% su Facebook e dello 0,01-0,02% su Instagram. Anche in questo caso, la stragrande maggioranza dei contenuti rimossi – il 96,6% su Facebook e il 96% su Instagram – sono stati rilevati da noi prima che un utente ce li segnalasse.

Nel Report si possono trovare ulteriori informazioni, relativamente a questi e ad altri Standard della community, sulla prevalenza dei contenuti in violazione, sul numero dei contenuti sui quali siamo intervenuti, sulla percentuale di questi contenuti su cui siamo intervenuti proattivamente (ovvero prima che ci venissero segnalati dagli utenti), sul numero di ricorsi da parte degli utenti e sul numero dei contenuti inizialmente rimossi che sono stati ripristinati in un secondo momento.

Pur consapevoli che ancora molto si può e si deve fare, i dati fin qui condivisi indicano che i nostri sforzi nel rimuovere i contenuti d'odio e in generale i contenuti dannosi stanno dando i loro frutti.

Un approccio olistico alla sicurezza

I miglioramenti che attestiamo sono attribuibili ad un approccio che possiamo definire olistico e nel quale giocano un ruolo i nostri Standard della community, gli algoritmi, l'intelligenza artificiale, la parte operativa relativa alla moderazione globale dei contenuti e un approccio al design dei prodotti che si focalizza sulla sicurezza e sull'integrità.

In particolare, grazie ai progressi compiuti dalle tecnologie di intelligenza artificiale, come il nostro Few Shot Learner – una tecnologia basata su intelligenza artificiale capace di adattarsi più facilmente e agire rapidamente su nuove tipologie di contenuti dannosi – nonché il passaggio all'intelligenza artificiale generalizzata, siamo stati in grado di identificare e intervenire in maniera sempre più corretta e veloce sui contenuti che violano i nostri Standard della community.

In questo lavoro anche i nostri team che si occupano dell'ideazione e progettazione dei prodotti svolgono un ruolo fondamentale, al fine di offrire maggiore sicurezza agli utenti.

Ulteriori informazioni relative all'applicazione dei nostri Standard della community

Lo scorso autunno, su Instagram, abbiamo lanciato lo Stato dell'account, uno strumento che aiuta le persone a capire perché eventuali contenuti sono stati rimossi e quali regole violavano.

Crediamo che strumenti di questo tipo, volti ad aumentare la trasparenza, siano fondamentali in quanto aiutano gli utenti a comprendere meglio i nostri Standard e Linee guida della community, oltre a contribuire al miglioramento dei nostri sistemi fornendoci indicazioni nel caso in cui sia stato commesso un errore. Dopo il lancio di questo strumento, abbiamo infatti assistito a un aumento dei contenuti ripristinati su Instagram relativamente a diverse aree degli Standard della community, tra cui violenza e istigazione alla violenza, incitamento all'odio e bullismo e intimidazioni.

Cerchiamo anche di capire su quali tipi di contenuti, nel tentativo di applicare i nostri Standard, interveniamo in maniera eccessiva o insufficiente, così da poter migliorare. Ad esempio nel 2020, durante il mese per la prevenzione del tumore al seno, su Instagram vi fu un'affluenza di contenuti relativi al cancro al seno, comprese immagini di nudo di natura medica. Sebbene queste non violassero le nostre regole, molte furono erroneamente rimosse dai nostri sistemi. Ciò portò alla rimozione di contenuti pubblicati da importanti account di supporto nella lotta contro il cancro al seno e da account di singoli pazienti sopravvissuti al cancro.

Da allora, abbiamo lavorato costantemente per migliorare l'accuratezza dei nostri interventi sui contenuti relativi a temi che hanno a che fare con la salute, compresi quelli riguardanti il cancro al seno. Abbiamo definito quali sono le nostre eccezioni alle norme sui contenuti di nudo per le immagini volte a sensibilizzare su una causa o a scopo educativo o medico, allenando i nostri sistemi basati su intelligenza artificiale per riconoscere meglio i toraci con cicatrici da mastectomie e rimandando a revisione umana i contenuti di nudo contrassegnati come di natura medica e i contenuti contenenti determinate parole chiave utilizzate di frequente in questi casi, così da migliorare ulteriormente la precisione. A seguito di queste azioni, abbiamo visto ridursi sensibilmente il numero di interventi non giustificati durante il mese per la prevenzione del tumore al seno, cosa che ha permesso alla nostra community di condividere liberamente contenuti legati a questo tema.

Siamo inoltre costantemente alla ricerca di nuovi modi per migliorare la trasparenza e l'integrità delle nostre piattaforme. Da diversi anni pubblichiamo dei report semestrali sulla trasparenza, i quali includono anche il volume delle restrizioni sui contenuti che applichiamo quando un contenuto viene segnalato come in violazione delle leggi locali, ma non è contrario ai nostri Standard della community.

I nostri obiettivi sono ambiziosi ma sostenuti dalla ferma volontà di migliorarci costantemente e di affrontare i problemi che incontriamo.

Comitato per il controllo

Nel 2018, il nostro fondatore Mark Zuckerberg affermò che “Facebook non dovrebbe prendere da sola così tante decisioni che hanno a che fare con la libertà di espressione e la sicurezza degli utenti”. Una delle conseguenze principali della scala alla quale Meta opera è infatti la responsabilità che si trova ad esercitare su questo genere di decisioni, che sono in molti casi difficili da compiere, con valutazioni spesso controverse che hanno un impatto sulla libertà di espressione degli individui. Per questo motivo, negli ultimi anni il dibattito su quali contenuti dovrebbero essere consentiti e quali rimossi, e soprattutto su chi dovrebbe prendere tali decisioni, è diventato sempre più centrale nella società.

Al fine di garantire il rispetto del principio di libera espressione delle persone e gestire in modo ponderato i difficili compromessi che ne derivano, nel 2019 abbiamo promosso la creazione di un Comitato per il controllo (*Oversight Board*), un organo indipendente con il compito di supervisionare le decisioni prese sui contenuti più difficili ed emblematici presenti sulle nostre piattaforme e di formulare raccomandazioni e proposte di modifica alle regole attualmente in vigore.

Le decisioni del Comitato di confermare o annullare le decisioni sui contenuti prese da Meta sono vincolanti per l’azienda, che è obbligata ad implementarle purché tale applicazione non violi la legge.

Nell’ottica di garantire trasparenza sui processi e le decisioni che riguardano il Comitato per il controllo pubblichiamo periodicamente informazioni sui casi sottoposti da Meta al Comitato e un aggiornamento sui nostri progressi nell’attuazione delle raccomandazioni ricevute.

Oltre a fornire agli utenti la possibilità di fare appello al Comitato contro le decisioni prese da Meta sui contenuti, identifichiamo proattivamente alcune delle decisioni più significative e difficili da noi prese e chiediamo al Comitato di esaminarle. È poi il Comitato a decidere su quali di queste effettivamente intervenire. Tali questioni normalmente implicano un impatto nel mondo reale e concernono tematiche gravi, estese e/o importanti per il dibattito pubblico.

Informazioni aggiornate sull’applicazione da parte di Meta delle decisioni prese dal Comitato per il controllo sono disponibili nel più recente *Aggiornamento trimestrale sul Comitato di controllo*, riferito all’ultimo trimestre del 2021.

Regolamentazione dei servizi digitali

Molto di ciò che Meta ha realizzato in questi ambiti è definibile come un tentativo di autoregolamentazione da parte di un’azienda privata che troppo spesso si è trovata ad affrontare dilemmi etici estremamente complessi, bilanciando diritti tra loro contrastanti.

Decidere di volta in volta cosa può restare sulla piattaforma e cosa no o trovare il giusto equilibrio tra libertà di espressione e protezione degli utenti sono questioni molto delicate da un punto di vista etico e sociale. A tale riguardo, siamo sinceramente convinti che

non debbano essere aziende private a compiere scelte di questo genere, che spettano invece a legislatori eletti democraticamente.

Per questo motivo, da tempo auspichiamo lo sviluppo di nuovi quadri normativi che regolino la gestione dei contenuti dannosi online e guardiamo con fiducia alle proposte europee come il Digital Services Act. In questa prospettiva, siamo lieti di vedere che l'Italia è uno dei Paesi che sta aprendo la strada a una maggiore comprensione di questi fenomeni, come dimostrato dal lavoro della Commissione.

Perché ciò avvenga nel modo più efficace, ovvero senza ostacolare l'innovazione o limitare diritti come la libertà di espressione degli utenti, crediamo però siano fondamentali alcuni accorgimenti, tra cui la garanzia di certezze giuridiche e il mantenimento dei principi chiave della Direttiva eCommerce, come il principio di responsabilità limitata, che hanno reso possibile lo sviluppo in Europa del mercato digitale. Bisogna inoltre cercare di evitare qualsiasi tipo di frammentazione tra i vari mercati europei, *conditio sine qua non* per il completamento del mercato unico digitale.

L'importanza dell'educazione digitale

Sebbene le soluzioni tecnologiche siano fondamentali per rendere le piattaforme un luogo sempre più sicuro, parallelamente crediamo sia necessario portare avanti una seria attività di sensibilizzazione sul corretto uso di internet e di formazione sulle competenze digitali: un'attività che sia rivolta a tutti – giovani e adulti – al fine di supportarli nel capire come utilizzare in maniera consapevole e sicura gli strumenti digitali. Siamo convinti che senza uno sforzo educativo e formativo di questo tipo sia impossibile apportare un cambiamento profondo, significativo e duraturo nella società.

Con tale obiettivo in mente, nel 2018 abbiamo inaugurato Binario F, uno spazio fisico, situato nel cuore di Roma, dedicato alla formazione sulle competenze digitali di persone, imprese, associazioni e istituzioni. A causa della pandemia e della conseguente chiusura temporanea dello spazio, abbiamo riconvertito tutte le nostre attività online, riuscendo in tal modo a raggiungere un numero sempre maggiore di utenti.

In questi anni, le attività promosse da Binario F in collaborazione con numerosi partner hanno raggiunto oltre 160.000 persone in modo completamente gratuito, contribuendo a formarle su materie che spaziano dalla sicurezza online all'imprenditoria digitale, dal contrasto alla disinformazione ai discorsi d'odio.