Consiglio dell'Unione europea

Bruxelles, 27 settembre 2019
(OR. en)

12522/19

LIMITE

ENFOPOL 422	DROIPEN 149
ANTIDISCRIM 34	DIGIT 145
TELECOM 309	DAPIX 282
SOC 634	CYBER 267
MIGR 155	COSI 201
JAI 993	COPEN 374
FREMP 134	COHOM 108
EDUC 390	AUDIO 102

NOTA INFORMATIVA

Origine:	Commissione europea
Destinatario:	Comitato dei rappresentanti permanenti/Consiglio
Oggetto:	**Valutazione del codice di condotta contro l'incitamento all'odio online** ***Stato dei lavori***

Si allega per le delegazioni una nota informativa sul tema in oggetto, trasmessa dai servizi della Commissione per il Consiglio GAI del 7-8 ottobre 2019.

Consiglio GAI del 7 ottobre

Nota informativa

Progressi nel contrasto dell'incitamento all'odio online grazie al codice di condotta dell'UE

2016-2019

Il codice di condotta per contrastare l'illecito incitamento all'odio online è stato firmato il 31 maggio 2016 dalla Commissione e da Google (YouTube), Facebook, Twitter e i servizi per i consumatori ospitati da Microsoft (es. i servizi di giochi Xbox o LinkedIn). Nel 2018 e nel 2019 hanno aderito Instagram, Google+, Dailymotion, Snap e Jeuxvideo.com. Ciò significa che il codice copre ora il 96% della quota di mercato UE delle piattaforme online che possono essere toccate da contenuti di incitamento all'odio[1]. Il presente documento fornisce una valutazione dei progressi compiuti dal 2016, seguendo la struttura e gli impegni di cui al codice di condotta. Si basa sui dati raccolti dalla Commissione nel corso di esercizi di monitoraggio periodici nonché su informazioni selezionate inviate a intervalli regolari dalle società informatiche.

In sintesi il codice di condotta ha contribuito alla realizzazione di progressi rapidi, anche per quanto riguarda in particolare l'esame e la soppressione tempestivi dei contenuti di incitamento all'odio (nel 2016 è stato rimosso il 28% dei contenuti rispetto al 72% del 2019; nel 2016 il 40% delle segnalazioni è stato esaminato in 24 ore rispetto all'89% nel 2019). Ha rafforzato la fiducia e la cooperazione fra le società informatiche, le organizzazioni della società civile e le autorità degli Stati membri nel quadro di un processo strutturato di apprendimento reciproco e scambio di conoscenze. Tale lavoro completa l'efficace contrasto da parte della legislazione vigente (decisione quadro 2008/913/GAI del Consiglio), che vieta i reati di stampo razzista e xenofobo e l'incitamento all'odio, nonché gli sforzi che devono essere messi in atti dalle competenti autorità nazionali per investigare e perseguire i reati motivati dall'odio, sia offline che online.

Il codice di condotta impone alle società informatiche di:

- **disporre di regole e norme comunitarie che vietino l'incitamento all'odio e introducano sistemi e squadre per esaminare i contenuti che, secondo segnalazioni, violano tali norme.**

Tutte le società informatiche firmatarie del codice dispongono ora di condizioni di servizio, regole o norme comunitarie che sottopongono a continua revisione e che vietano agli utenti di postare

[1] https://gs.statcounter.com/social-media-stats/all/europe

contenuti che incitano alla violenza o all'odio ai danni di gruppi protetti. È interessante osservare che sia Jeuxvideo.com che Dailymotion hanno riesaminato nella sostanza le proprie condizioni di servizio per includere una definizione più precisa dell'incitamento all'odio in quanto contenuto vietato, in considerazione della loro partecipazione al codice. Snap, che ha aderito al codice nella primavera 2018, nel corso dello stesso anno ha ristrutturato completamente il proprio centro di sicurezza, che contiene ora informazioni destinate a cittadini, servizi di contrasto ed educatori in merito ai contenuti vietati, compreso l'incitamento all'odio.

Tutte le piattaforme hanno aumentato considerevolmente il numero di persone che monitorano ed esaminano i contenuti. Facebook riferisce di disporre di una rete globale di circa 15 000 persone che si occupano dell'esame di tutti i tipi di contenuti e presso Google e YouTube più di 10 000 persone si occupano di contenuti che possono violare le politiche della società.

- **esaminare entro 24 ore la maggioranza dei contenuti segnalati e, se necessario, rimuovere i contenuti di incitamento all'odio o bloccarne l'accesso**

In media le società informatiche valutano ora l'**89% dei contenuti segnalati entro 24 ore**, rispetto all'81% di un anno fa. Instagram, che è stato sottoposto a test per la prima volta nel 2018, ha esaminato entro un giorno oltre il 77% delle notifiche. Pochi mesi dopo l'avvio del codice il numero di notifiche esaminate in 24 ore era pari al 40%. Dailymotion e Jeuxvideo.com non sono ancora stati sottoposti a test nel quadro degli esercizi di monitoraggio periodici della Commissione; dichiarano tuttavia che oltre il 90% delle segnalazioni ricevute nel 2019 sono state esaminate entro 24 ore. Snap riferisce che la grande maggioranza dei contenuti segnalati è trattato entro poche ore e in ogni caso tutti i contenuti su Snapchat scompaiono entro 24 ore.

La percentuale di soppressione è ora stabile e supera il 70% in media. Nel 2016, dopo il primo esercizio di monitoraggio dell'attuazione del codice di condotta, veniva soppresso solo il 28% dei contenuti segnalati. La percentuale di soppressione media attuale può essere considerata soddisfacente in un settore come quello dell'incitamento all'odio, considerato che non sempre è facile tracciare una linea di demarcazione tra l'incitamento all'odio e il discorso protetto dal diritto alla libertà di espressione e che tale linea di demarcazione dipende notevolmente dal contesto in cui si inseriscono i contenuti.

Alcune della società informatiche che hanno aderito al codice più di recente riferiscono di avere realizzato una riduzione significativa del numero di segnalazioni di incitamento all'odio (es. per Dailymotion le segnalazioni sono scese da 27 000 nel primo semestre del 2018 a 17 000 nello stesso periodo del 2019) grazie alle strategie messe a punto per conformarsi al codice. I servizi di giochi (es. Xbox o Mixer) hanno attuato misure per promuovere, con un intervento umano, la moderazione

dell'incitamento all'odio sulle chat e sui forum; grazie a tali misure sono stati individuati e bloccati 20 milioni di contenuti nel 2019, anche di incitamento all'odio.

- **fornire regolarmente formazioni al personale**

Tutte le società informatiche riferiscono di tenere formazioni regolari e frequenti e di fornire inquadramento e sostegno ai propri gruppi incaricati di esaminare i contenuti, fra l'altro su aspetti riguardanti specificamente l'incitamento all'odio. Dailymotion organizza per il personale formazioni quindicinali riguardo ai materiali che incitano all'odio. Facebook ha istituito un Product Policy Forum (forum per la politica dei prodotti) che riunisce ogni due settimane tutti i suoi esperti sparsi per il mondo per discutere delle possibili modifiche alle norme comunitarie e individuare nuove questioni, tendenze e sviluppi. I verbali di tali riunioni sono pubblici.

- **partecipare a partenariati e attività di formazione con la società civile per ampliare la propria rete di "segnalatori attendibili"**

Le società informatiche hanno riferito di avere ampliato considerevolmente, dal 2016, la propria rete di "segnalatori attendibili" in Europa, con cui intrattengono contatti regolari per incrementare la comprensione delle specificità nazionali dell'incitamento all'odio. Dalla firma del codice Twitter ha coinvolto 73 nuove organizzazioni di segnalatori attendibili. YouTube dispone oggi di una rete di segnalatori attendibili specializzati nell'individuazione dell'incitamento all'odio quattro volte maggiore rispetto al 2016, passata da 10 a 46 organizzazioni non governative (ONG); Facebook ha incrementato la propria rete dell'82% (dai 9 partner del 2016 ai 51 di oggi).

Dalla firma del codice Facebook/Instagram hanno organizzato in totale 51 sessioni di formazione in merito alle norme comunitarie relative all'incitamento all'odio per un totale di 130 organizzazioni della società civile attive come segnalatori attendibili. Delle 38 sessioni di formazione relative alla sua politica in materia di contenuti e al programma di segnalatori attendibili organizzate da YouTube nel 2018 per le ONG, 18 erano incentrate sull'incitamento all'odio e i contenuti abusivi. Nel 2019 YouTube ha organizzato un ulteriore ciclo di formazioni con 15 ONG in 8 paesi.

YouTube segnala inoltre il considerevole impatto di questa rete estesa sul numero di notifiche da parte di segnalatori attendibili, che è raddoppiato fra il trimestre ottobre-dicembre 2017 e il trimestre aprile-giugno 2019. Per quanto riguarda Facebook, dal quarto trimestre 2017 al primo trimestre 2019 gli interventi effettuati riguardo alle violazioni connesse all'incitamento all'odio sono passati da 1,6 milioni a 4 milioni, con un aumento del 150%.

- **lavorare [con i segnalatori attendibili] alla promozione di contronarrazioni indipendenti e programmi educativi**

Le società informatiche collaborano anche con i loro "segnalatori attendibili" riguardo a campagne di tolleranza e pluralismo online. Tra il 2017 e il 2019 si sono svolti tre workshop presso le sedi di YouTube, Twitter e Facebook al fine di agevolare tali iniziative. Un quarto workshop è previsto per la fine del 2019. Grazie a questi workshop, durante le elezioni europee del 2019, più di 40 ONG hanno lanciato a livello dell'UE una campagna online in 24 lingue, incentrata sulla promozione di conversazioni sane e tolleranti online con l'hashtag # WeDeserveBetter. La campagna ha raggiunto oltre 6 milioni di utenti su Facebook e Twitter e ha ricevuto il sostegno delle società informatiche sotto forma di sovvenzioni pubblicitarie. Un esercizio pilota condotto nel 2018 per testare una campagna ha raggiunto oltre 2 milioni di utenti in vari Stati membri.

Microsoft ha avviato un partenariato con gruppi di riflessione composti di esperti come l'Institute for Strategic Dialogue per combattere l'incitamento all'odio al fine di aiutare le ONG a far emergere e diffondere contenuti incisivi di contronarrativa tramite notifiche su Bing.

- **designare punti di contatto nazionali per ricevere le segnalazioni, in particolare da parte delle autorità nazionali**.

Tutte le società informatiche che hanno aderito al codice di condotta hanno stabilito punti di contatto nazionali per facilitare i contatti con le pertinenti autorità competenti a livello nazionale. È importante sottolineare che i lavori nel quadro del codice di condotta integrano la legislazione in materia di lotta contro il razzismo e la xenofobia (decisione quadro 2008/913/GAI del Consiglio), che impone che gli autori di reati di illecito incitamento all'odio, siano essi online o offline, siano perseguiti in modo efficace. Twitter organizza formazioni di contrasto annuali per le autorità nazionali e i punti di contatto negli Stati membri e ha fornito orientamenti specifici su come segnalare o richiedere informazioni.

- **promuovere la trasparenza nei confronti degli utenti e del pubblico in generale**

Nel 2016 le società informatiche hanno reso disponibili solo le informazioni sul numero di richieste di contrasto e non hanno fornito alcun dettaglio sull'incitamento all'odio online quale motivo specifico per la soppressione. Oggi le soppressioni di contenuti di incitamento all'odio sono chiaramente presentate, su base regolare, in ciascuna delle relazioni sulla trasparenza delle società informatiche, ad esempio vedasi le relazioni sulla trasparenza pubblicate da Facebook, Twitter e YouTube. Nel 2019 sia YouTube che Facebook hanno lanciato nelle loro relazioni sulla trasparenza pagine dedicate all'applicazione delle norme comunitarie concernenti specificamente i contenuti di incitamento all'odio, tra cui la ripartizione dei dati, riguardanti ad esempio la replica alle segnalazioni, e l'individuazione automatica. Il livello di dettaglio resta tuttavia insufficiente: le cifre

pubblicate non forniscono informazioni sul tempo necessario a esaminare le segnalazioni o sulla distribuzione geografica dei contenuti di incitamento all'odio segnalati.

Prima del varo del codice di condotta, raramente gli utenti hanno ricevuto una risposta da parte delle società informatiche quando hanno notificato contenuti di incitamento all'odio. Inoltre, la funzione di comunicazione o segnalazione spesso non era di facile utilizzo. Twitter ha sviluppato vari miglioramenti al sistema di segnalazione degli utenti, anche per consentire segnalazioni multiple di tweet provenienti dallo stesso account. YouTube e Facebook hanno introdotto sistemi di "quadro di controllo" attraverso i quali gli utenti possono monitorare i risultati di ciascuna delle loro segnalazioni. In base ai risultati degli esercizi di controllo, in media circa due terzi delle notifiche ricevono una risposta che descrive i risultati e le misure adottate. Le prestazioni delle piattaforme informatiche differiscono e solo Facebook e Instagram inviano sistematicamente un feedback alle notifiche (più del 95% delle segnalazioni riceve una risposta). Si attendono pertanto ulteriori progressi in questo settore specifico nei prossimi mesi.

La trasparenza e il feedback sono importanti anche per far sì che gli utenti possano ricorrere contro una decisione riguardante i contenuti da loro postati nonché per tutelare il loro diritto alla libertà di parola. Facebook riferisce di aver ricevuto 1,1 milioni di ricorsi concernenti contenuti oggetto di intervento per incitamento all'odio tra il gennaio 2019 e il marzo 2019, e 130.000 contenuti sono stati ripristinati a seguito di una nuova valutazione.

Al di là degli impegni del codice di condotta: il ruolo della tecnologia e degli strumenti di rilevamento automatico

Nell'ambito degli sforzi volti a migliorare il modo in cui i contenuti di incitamento all'odio sono individuati e rimossi, le società informatiche fanno un uso crescente della tecnologia e dei sistemi di rilevamento automatico. Facebook ha riferito che nel primo trimestre del 2019 il 65,4% dei contenuti rimossi è stato segnalato da macchine (con un aumento dal 51,5% rispetto ai mesi precedenti). YouTube riferisce che nel 2017 il 79% dei video rimossi per violazione delle loro politiche è stato inizialmente segnalato da sistemi di segnalazione automatica e nel secondo trimestre del 2019 la percentuale è stata pari all'87%. Un numero considerevole dei video rimossi è ritirato prima di essere visionato anche da un solo utente. Fino ad aprile 2019, grazie all'utilizzo della tecnologia, il 38% dei contenuti illeciti che sono stati oggetto di intervento da parte di Twitter è stato messo in evidenza in modo proattivo in vista di un esame umano, invece di basarsi sulle dichiarazioni degli utenti. Si tratta di un aumento significativo rispetto all'anno precedente, in cui il

20% dei contenuti potenzialmente illeciti è stato segnalato da macchine. Va osservato che tutti i contenuti messi in evidenza tramite un sistema di rilevamento automatico sono valutati dalla squadra di revisori prima di essere oggetto di intervento (supervisione umana).

Che cosa sappiamo riguardo ai volumi di contenuti di incitamento all'odio segnalati alle società informatiche?

Dai dati comunicati da alcune delle società informatiche che partecipano al codice, l'entità delle segnalazioni riguardanti contenuti di incitamento all'odio sembra essere compresa tra il 17 e il 30% del totale[2]. Facebook riferisce di aver rimosso 3,3 milioni di contenuti per violazione delle politiche in materia di incitamento all'odio nell'ultimo trimestre del 2018 e 4 milioni nel primo trimestre del 2019. Nel 2018 oltre 6,2 milioni di account Twitter sono stati segnalati in quanto contenenti comportamenti di incitamento all'odio e la piattaforma è intervenuta in circa 536.000 casi.

Uno studio realizzato da Vox POL per la Commissione nel 2018 ha effettuato un'analisi comparativa dell'attività di un gruppo di circa 175 "hater" in vari Stati membri: mentre nel 2016 il gruppo ha prodotto 60.000 tweet di incitamento all'odio, oggi la sua attività è ridotta a 7.400 tweet.

Gli ecosistemi dell'incitamento all'odio online e l'ampiezza del fenomeno in Europa rimangono un settore in cui sono necessari ulteriori dati e ricerche.

[2] È opportuno osservare che il dato si riferisce alle segnalazioni ricevute e non corrisponde al numero reale di soppressioni per contenuti di incitamento all'odio. Può succedere, ad esempio, che alcuni contenuti siano erroneamente segnalati dagli utenti come incitamento all'odio.