Rat der
Europäischen Union

Brüssel, den 27. September 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**

INFORMATORISCHER VERMERK

Absender:	Europäische Kommission
Empfänger:	Ausschuss der Ständigen Vertreter/Rat
Betr.:	**Bewertung des Verhaltenskodex zu Hetze im Internet**
	Sachstand

Die Delegationen erhalten in der Anlage einen informatorischen Vermerk zum eingangs genannten Thema, den die Dienststellen der Kommission für die Tagung des Rates (Justiz und Inneres) am 7./8. Oktober 2019 vorgelegt haben.

Tagung des Rates (Justiz und Inneres) am 7. Oktober

Informatorischer Vermerk

Fortschritte bei der Bekämpfung von Hetze im Internet durch den EU-Verhaltenskodex 2016-2019

Der Verhaltenskodex zur Bekämpfung illegaler Hetze im Internet wurde am 31. Mai 2016 von der Kommission und Google (YouTube), Facebook, Twitter und den von Microsoft gehosteten Verbraucherdiensten (z. B. Xbox-Spieledienste oder LinkedIn) unterzeichnet. In den Jahren 2018 und 2019 sind Instagram, Google+, Dailymotion, Snap und Jeuxvideo.com beigetreten. Damit erfasst der Verhaltenskodex nunmehr 96 % des EU-Marktanteils an Online-Plattformen, die potenziell von hetzerischen Inhalten betroffen sind.[1] Dieser Vermerk enthält eine Bewertung der Fortschritte, die seit 2016 erzielt wurden, und orientiert sich dabei am Aufbau des Verhaltenskodex und den darin festgelegten Verpflichtungen. Er stützt sich auf die von der Kommission bei ihren regelmäßigen Bewertungsrunden erhobenen Daten und auf ausgewählte Informationen, die von den IT-Unternehmen in regelmäßigen Abständen übermittelt werden.

Zusammenfassend lässt sich sagen, dass der Verhaltenskodex zu zügigen Fortschritten beigetragen hat, darunter insbesondere bei der raschen Prüfung und Entfernung hetzerischer Inhalte (2016 wurden 28 % der Inhalte entfernt gegenüber 72 % im Jahr 2019; 2016 wurden 40 % der gemeldeten Inhalte binnen 24 Stunden geprüft, 2019 bereits 89 %). Dadurch haben Vertrauen und Zusammenarbeit zwischen den IT-Unternehmen, den Organisationen der Zivilgesellschaft und den Behörden der Mitgliedstaaten in Form eines strukturierten Prozesses des Voneinander-Lernens und des Wissensaustauschs deutlich zugenommen. Diese Arbeit ergänzt die wirksame Durchsetzung der geltenden Rechtsvorschriften (Rahmenbeschluss 2008/913/JI des Rates) zum Verbot rassistischer und fremdenfeindlicher Hassverbrechen und Hetze und die notwendigen Bemühungen vonseiten der zuständigen Behörden der Mitgliedstaaten um die Ermittlung und Verfolgung hassmotivierter Straftaten, sowohl offline als auch online.

Mit dem Verhaltenskodex werden die IT-Unternehmen verpflichtet,

- **Vorschriften und Gemeinschaftsstandards zum Verbot der Hetze einzuführen sowie Systeme und Teams einzurichten, die als Verstöße gegen diese Standards gemeldete Inhalte prüfen;**

Sämtliche IT-Unternehmen, die den Verhaltenskodex unterzeichnet haben, verfügen inzwischen über Nutzungsbedingungen, Vorschriften oder Gemeinschaftsstandards, die es Nutzern verbieten, Inhalte zu posten, die zu Gewalt oder Hass gegen geschützte Gruppen aufrufen, und sie überarbeiten diese Regelungen fortlaufend.

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

Interessanterweise haben sowohl Jeuxvideo.com als auch Dailymotion ihre Nutzungsbedingungen grundlegend überarbeitet und mit Blick auf ihren Beitritt zu dem Verhaltenskodex eine präzisere Begriffsbestimmung für Hetze als verbotenen Inhalt aufgenommen. Snap, das dem Verhaltenskodex im Frühjahr 2018 beigetreten ist, hat noch im selben Jahr sein Sicherheitszentrum vollständig umgestaltet; es enthält inzwischen Informationen für Einzelpersonen, Strafverfolgungsbehörden und Pädagogen zu verbotenen Inhalten, einschließlich Hetze.

Alle Plattformen haben zudem ihr Personal zur Überwachung und Prüfung von Inhalten erheblich aufgestockt. So verfügt Facebook nach eigenen Angaben über ein weltweites Netzwerk von rund 15 000 Personen, die mit allen Arten der Überprüfung von Inhalten befasst sind, und bei Google und YouTube sind über 10 000 Personen damit beschäftigt, gegen Inhalte vorzugehen, die gegen die Unternehmenspolitik verstoßen könnten.

- **die Mehrheit der gemeldeten Inhalte binnen 24 Stunden zu prüfen und – sofern erforderlich – hetzerische Inhalte zu entfernen oder den Zugang zu diesen zu blockieren;**

Die IT-Unternehmen prüfen mittlerweile im Durchschnitt **89 % der gemeldeten Inhalte innerhalb von 24 Stunden**, gegenüber nur 81 % vor einem Jahr. Instagram, das erstmals 2018 beurteilt wurde, prüft über 77 % der Meldungen innerhalb eines Tages. Einige Monate nach der Einführung des Verhaltenskodex lag die Anzahl der innerhalb von 24 Stunden geprüften Meldungen noch bei 40 %. Dailymotion und Jeuxvideo.com waren noch nicht Gegenstand der regelmäßigen Bewertungsrunden der Kommission, allerdings wurden nach ihren eigenen Angaben über 90 % der 2019 eingegangenen Meldungen innerhalb von 24 Stunden geprüft. Snap gibt an, dass die große Mehrheit der gemeldeten Inhalte innerhalb weniger Stunden entsprechend behandelt und dass ohnehin sämtliche Inhalte auf Snapchat binnen 24 Stunden entfernt würden.

Die **Entfernungsquote** ist inzwischen stabil und liegt im Durchschnitt bei über 70 %. Die erste Bewertungsrunde bezüglich der Umsetzung des Verhaltenskodex im Jahr 2016 hatte ergeben, dass lediglich 28 % der gemeldeten Inhalte entfernt wurden. Die derzeitige durchschnittliche Entfernungsquote kann in einem Bereich wie Hetze als zufriedenstellend betrachtet werden, zumal eine Abgrenzung gegenüber Äußerungen, die durch das Recht auf freie Meinungsäußerung geschützt sind, nicht immer einfach ist und sehr stark von dem Kontext abhängt, in dem ein Inhalt gepostet wird.

Einige der IT-Unternehmen, die dem Kodex erst vor Kurzem beigetreten sind, geben an, dass sie dank der zur Einhaltung des Verhaltenskodex eingeführten Strategien eine deutliche Abnahme der Meldungen von Hetze erreicht hätten (z. B. Dailymotion: Reduzierung von 27 000 Meldungen im ersten Halbjahr 2018 auf 17 000 im gleichen Zeitraum 2019). Spieledienste (wie Xbox oder Mixer) haben Maßnahmen umgesetzt,

um eine menschliche Moderation von Hetze in Chats oder Foren zu fördern, was dazu geführt hat, dass im Jahr 2019 20 Millionen Inhalte, darunter hetzerische Inhalte, ermittelt und blockiert werden konnten.

- **regelmäßige Schulungen für ihr Personal anzubieten;**

Alle IT-Unternehmen geben an, dass sie regelmäßig und häufig Schulungen durchführen und den für die Prüfung der Inhalte zuständigen Teams entsprechendes Coaching und Unterstützung anbieten, unter anderem in Bezug auf die besonderen Merkmale hetzerischer Inhalte. Dailymotion führt alle vierzehn Tage Personalschulungen zu hetzerischem Material durch. Facebook hat ein Forum für Produktpolitik eingerichtet, in dem sämtliche weltweit für Facebook tätigen Experten alle zwei Wochen zusammentreffen, um mögliche Änderungen an den Gemeinschaftsstandards zu besprechen und neue Fragen, Trends und Entwicklungen aufzugreifen. Die Protokolle dieser Treffen sind öffentlich zugänglich.

- **Partnerschaften mit der Zivilgesellschaft einzugehen und Schulungsmaßnahmen mit ihr durchzuführen, um ihr Netzwerk "vertrauenswürdiger Hinweisgeber" auszubauen;**

Die IT-Unternehmen geben an, dass sich ihr Netzwerk vertrauenswürdiger Hinweisgeber in Europa seit 2016 erheblich erweitert hat. Sie tauschen sich regelmäßig mit ihnen aus, um das Verständnis für die nationalen Besonderheiten von Hetze zu verbessern. Twitter hat seit Unterzeichnung des Kodex 73 neue Organisationen vertrauenswürdiger Hinweisgeber angeworben. Das Netzwerk vertrauenswürdiger auf Hetze spezialisierter Hinweisgeber von YouTube ist gegenüber 2016 auf das Vierfache angewachsen, wobei die Anzahl der beteiligten Nichtregierungsorganisationen (NRO) von 10 auf 46 gestiegen ist; Facebook konnte sein Netzwerk um 82 % vergrößern (von 9 Partnern 2016 auf derzeit 51 Partner).

Seit Unterzeichnung des Verhaltenskodex hat Facebook/Instagram insgesamt 51 Schulungen zu seinen Gemeinschaftsstandards in Bezug auf Hetze für rund 130 Organisationen der Zivilgesellschaft, die als vertrauenswürdige Hinweisgeber tätig sind, durchgeführt. Von den 38 Schulungen, die YouTube 2018 für NRO zu seiner Politik zu Inhalten und seinen Programmen für vertrauenswürdige Hinweisgeber angeboten hat, waren 18 schwerpunktmäßig hetzerischen und missbräuchlichen Inhalten gewidmet. YouTube hat 2019 einen weiteren Schulungszyklus mit 15 NRO in 8 Ländern durchgeführt.

YouTube gibt ferner an, dass sich dieses erweiterte Netzwerk erheblich auf die Zahl der Meldungen durch vertrauenswürdige Hinweisgeber auswirke: Vom vierten Quartal (Oktober bis Dezember) 2017 bis zum zweiten Quartal (April bis Juni) 2019 hätten sich diese verdoppelt. Bei Facebook haben sich die Maßnahmen gegen illegale Hetze von 1,6 Millionen im vierten Quartal 2017 auf 4 Millionen Maßnahmen im ersten Quartal 2019 erhöht (was einem Anstieg um 150 % entspricht).

- **[mit vertrauenswürdigen Hinweisgebern] an der Förderung unabhängiger Gegennarrative und Bildungsprogramme zu arbeiten;**

Die IT-Unternehmen arbeiten mit ihren vertrauenswürdigen Hinweisgebern ferner bei Kampagnen zugunsten von Toleranz und Pluralismus im Internet zusammen. Zwischen 2017 und 2019 fanden an den Hauptsitzen von YouTube, Twitter und Facebook drei Workshops statt, um solche Initiativen zu erleichtern. Ein vierter ist für Ende 2019 geplant. Diese Workshops haben dazu geführt, dass über 40 NRO während der Europawahl 2019 eine europaweite Internet-Kampagne in 24 Sprachen eingeleitet haben, die unter dem Hashtag #WeDeserveBetter (wir verdienen Besseres) insbesondere der Förderung eines respektvollen und toleranten Umgangstons im Internet gewidmet war. Diese Kampagne hat über sechs Millionen Nutzer auf Facebook und Twitter erreicht und wurde von den IT-Unternehmen in Form von Werbezuschüssen unterstützt. Ein 2018 durchgeführtes Pilotprojekt zur Erprobung einer Kampagne hat über zwei Millionen Nutzer in verschiedenen Mitgliedstaaten erreicht.

Microsoft ist eine Partnerschaft für Gegenrede mit Thinktanks von Experten, wie etwa dem Institute for Strategic Dialogue (Institut für Strategischen Dialog) eingegangen, um NRO dabei zu unterstützen, wirkungsvolle Gegennarrative über Werbeanzeigen auf Bing zu präsentieren und zu verbreiten.

- **nationale Kontaktstellen zu benennen, die Meldungen entgegennehmen, insbesondere durch die nationalen Behörden;**

Alle IT-Unternehmen, die sich dem Verhaltenskodex angeschlossen haben, haben nationale Kontaktstellen eingerichtet, um den Kontakt zu den jeweils auf nationaler Ebene zuständigen Behörden zu erleichtern. Es sei darauf hingewiesen, dass die Arbeit im Rahmen des Verhaltenskodex die Rechtsvorschriften zur Bekämpfung von Rassismus und Fremdenfeindlichkeit (Rahmenbeschluss 2008/913/JI des Rates) ergänzt, die eine wirksame strafrechtliche Verfolgung der Urheber illegaler Hetze – sei es online oder offline – vorschreiben. Twitter veranstaltet jährliche Schulungen im Bereich Strafverfolgung für die nationalen Behörden und die Kontaktstellen in den Mitgliedstaaten und hat spezifische Richtlinien für Auskunftsanträge oder Meldungen vorgelegt.

- **die Transparenz gegenüber Nutzern und der breiten Öffentlichkeit zu fördern.**

2016 legten die IT-Unternehmen lediglich Informationen über die Zahl der Ersuchen von Strafverfolgungsbehörden vor, jedoch keinerlei detaillierte Angaben zu Hetze im Internet als spezifischer Begründung für die Entfernung von Inhalten. Inzwischen enthalten die jeweiligen Transparenzberichte der IT-Unternehmen regelmäßig klare Angaben zur Entfernung hetzerischer Inhalte, so etwa die von Facebook, Twitter und YouTube veröffentlichten Transparenzberichte. Sowohl YouTube als auch Facebook haben 2019 spezielle Internetseiten mit ihren Transparenzberichten über die Durchsetzung von Gemeinschaftsstandards insbesondere in Bezug auf hetzerische Inhalte – einschließlich einer Aufschlüsselung der Angaben, z. B. zu Gegendarstellungen – und die automatische Erkennung eingerichtet.

Allerdings mangelt es immer noch an Detailtiefe, da die veröffentlichten Zahlen keinerlei Auskunft über den Zeitpunkt der Prüfung der Meldungen oder die geografische Verteilung der gemeldeten hetzerischen Inhalte geben.

Vor Einführung des Verhaltenskodex erhielten die Nutzer selten eine Antwort der IT-Unternehmen auf ihre Meldungen hetzerischer Inhalte. Darüber hinaus war die Meldefunktion häufig nicht sehr benutzerfreundlich. Twitter hat das Meldesystem für Nutzer weiterentwickelt und einige Verbesserungen vorgenommen, unter anderem indem die Mehrfachmeldung von Tweets desselben Absenderkontos ermöglicht wurde. YouTube und Facebook bieten inzwischen ein "Dashboard"-System an, mit dem die Nutzer das Resultat jeder ihrer Meldungen verfolgen können. Den Ergebnissen der Bewertungsrunden zufolge wird auf durchschnittlich rund zwei Drittel der Meldungen eine Antwort erteilt, aus der das Resultat und die jeweils ergriffenen Maßnahmen hervorgehen. Die IT-Plattformen schneiden im Leistungsvergleich unterschiedlich ab; lediglich Facebook und Instagram geben systematisch Rückmeldungen auf eingegangene Meldungen (über 95 % der Meldungen werden beantwortet). Daher werden in diesem spezifischen Bereich in den kommenden Monaten weitere Fortschritte erwartet.

Transparenz und Rückmeldung sind ebenfalls wichtig, um sicherzustellen, dass die Nutzer eine Entscheidung bezüglich eines von ihnen geposteten Inhalts anfechten können, und dienen zugleich als Absicherung, damit ihr Recht auf freie Meinungsäußerung geschützt ist. Facebook gibt an, zwischen Januar und März 2019 1,1 Millionen Beschwerden in Bezug auf als Hetze behandelte Inhalte erhalten zu haben, und 130 000 entfernte Inhalte wurden nach neuerlicher Bewertung wiederhergestellt.

Über die Verpflichtungen im Rahmen des Verhaltenskodex hinaus: die Rolle von Technologie und Tools zur automatischen Erkennung

Im Rahmen ihrer Bemühungen, die Verfahren zur Erkennung und Entfernung hetzerischer Inhalte zu verbessern, nutzen die IT-Unternehmen in zunehmenden Maße Technologien und Systeme zur automatischen Erkennung. Facebook gibt an, dass im ersten Quartal 2019 65,4 % der entfernten Inhalte von Maschinen gemeldet wurden (was eine Zunahme gegenüber den vorangegangenen Monaten (51,5 %) bedeutet). YouTube gibt an, dass 2017 79 % der aus Gründen eines Verstoßes gegen ihre Unternehmenspolitik entfernten Videos ursprünglich von automatischen Meldesystemen gemeldet wurden, während es im zweiten Quartal 2019 bereits 87 % waren. Eine Großteil dieser Videos wird bereits entfernt, bevor auch nur ein einziger Nutzer sie zu sehen bekommt. Bis April 2019 konnten bereits 38 % der von Twitter entsprechend behandelten missbräuchlichen Inhalte mithilfe entsprechender Technologie proaktiv für die anschließende Überprüfung durch Menschen ermittelt werden, anstatt auf Meldungen durch Nutzer zu vertrauen. Dies ist eine deutliche Steigerung gegenüber dem Vorjahr,

in dem lediglich 20 % der potenziell missbräuchlichen Inhalte von Maschinen gemeldet wurden. Es sei darauf hingewiesen, dass sämtliche von automatischen Erkennungssystemen ermittelten Inhalte durch ein Team von Prüfern bewertet werden, ehe entsprechende Maßnahmen eingeleitet werden (Human-in-the-Loop).

Was ist über den Umfang der den IT-Unternehmen gemeldeten hetzerischen Inhalte bekannt?

Den Daten zufolge, die einige der an dem Verhaltenskodex teilnehmenden IT-Unternehmen übermittelt haben, dürfte der Anteil der Meldungen hetzerischer Inhalte zwischen 17 und 30 % der Gesamtmeldungen ausmachen.[2] Facebook gibt an, im letzten Quartal 2018 3,3 Millionen Inhalte aufgrund eines Verstoßes gegen die in Bezug auf Hetze verfolgte Politik entfernt zu haben, im ersten Quartal 2019 hingegen 4 Millionen Inhalte. Im Jahr 2018 wurden über 6,2 Millionen Twitter-Konten wegen hasserfüllten Verhaltens gemeldet, und die Plattform ist gegen rund 536 000 Konten vorgegangen.

In einer 2018 von Vox POL im Auftrag der Kommission durchgeführten Studie wurde eine vergleichende Analyse der Aktivität einer Gruppe von etwa 175 "Hetzern" in mehreren Mitgliedstaaten durchgeführt, mit folgendem Ergebnis: 2016 verfasste diese Gruppe 60 000 Hass-Tweets, während ihre Aktivität inzwischen auf 7 400 Tweets zurückgegangen ist.

Die Ökosysteme der Hetze im Internet und das Ausmaß dieses Phänomens in Europa erfordern nach wie vor mehr Forschung und eine bessere Datenlage.

---

[2] Es sei allerdings angemerkt, dass sich diese Angaben auf die eingegangenen Meldungen beziehen und nicht der Anzahl der tatsächlich entfernten Hassinhalte entsprechen. Es kann beispielsweise vorkommen, dass Inhalte von Nutzern fälschlicherweise als Hetze gemeldet werden.