European Commission

Countering illegal hate speech online

6th evaluation of the Code of Conduct

Factsheet | **7 October 2021**

Didier Reynders
Commissioner for Justice

Directorate-General for Justice and Consumers

The sixth evaluation on the Code of Conduct on Countering Illegal Hate Speech Online shows that while the average of notifications reviewed within 24 hours remains high (81%), it has decreased compared to 2020 (90.4%). At 62.5% the average removal rate was also lower than in 2019 and 2020. However, broken down by IT company the progress of Instagram (66.2% removals in 2021, 42% in 2020) and Twitter (49.8% versus 35.9%) is noteworthy. TikTok was included in the evaluation for the first time and performed well (80.1% removals).

Key figures

1. Notifications of illegal hate speech

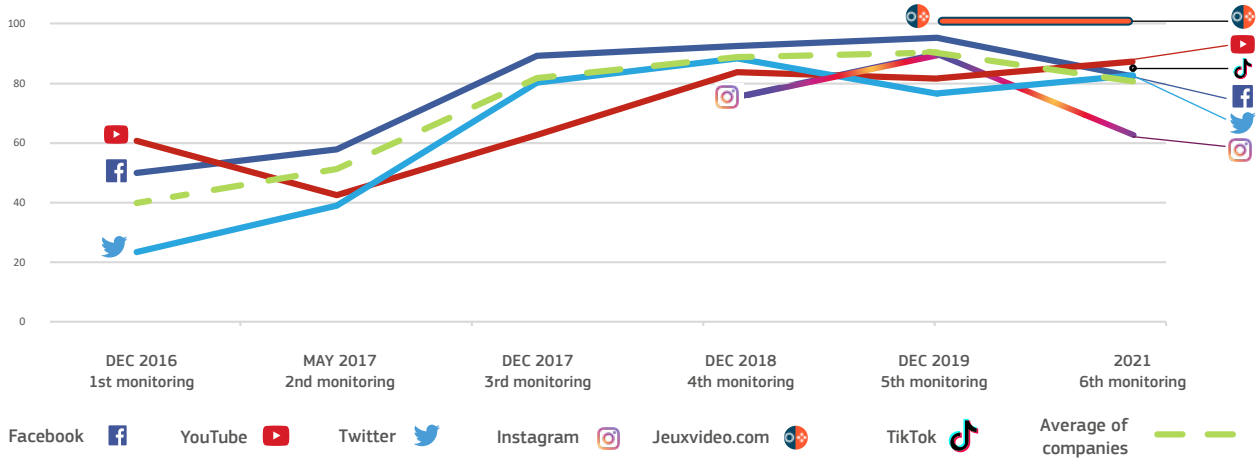
- **35 organisations** from 22 Member States **sent notifications relating to hate speech deemed illegal to the IT companies** during a period of approximately 6 weeks (1 March to 14 April 2021). In order to establish trends, this exercise used the same methodology as the previous monitoring rounds (see Annex).
- A total of **4543 notifications** were submitted to the IT companies taking part in the Code of Conduct.
- **3237 notifications** were submitted through the reporting **channels available to general users**, while **1306** were submitted through **specific channels available only to trusted flaggers/reporters**.
- **Facebook received the largest amount of notifications (1799),** followed by Twitter **(1595)**, YouTube **(519)**, Instagram **(401)** and Jeuxvideo.com **(30)**. Snapchat, Dailymotion and Microsoft did not receive any notification in the course of the monitoring exercise. TikTok, which joined the Code in September 2020, received **199 notifications**.
- In addition to flagging the content to IT companies, the organisations taking part in the monitoring exercise submitted **315 cases of hate speech** to the police, public prosecutor's bodies or other national authorities.

2. Time of assessment of notifications

- In **81% of the cases** the IT companies assessed the notifications **in less than 24 hours,** an additional **10.1%** in less than 48 hours, **8.1%** in less than a week and in **0.8%** of cases it took more than a week.
- **The Code of conduct prescribes that the majority of notifications is assessed within 24h.** All IT companies are therefore on target, yet, the average results are lower than in 2020 (**90.4%**).

Facebook assessed notifications in less than 24 hours in **81.5%** of the cases and an additional **10.6%** in less than 48 hours. The corresponding figures for YouTube are **88.8%** and **6.7%** and for Twitter **81.8%** and **8.9%**, respectively. Instagram had **62.4%** and **17.6%**, TikTok **82.5%** and **9.7%**. Jeuxvideo.com assessed all notifications in less than 24h. Twitter and YouTube improved their performance with respect to 2020 while the other platforms have a slight decrease.

Percentage of notifications assessed within 24 hours - Trend over time

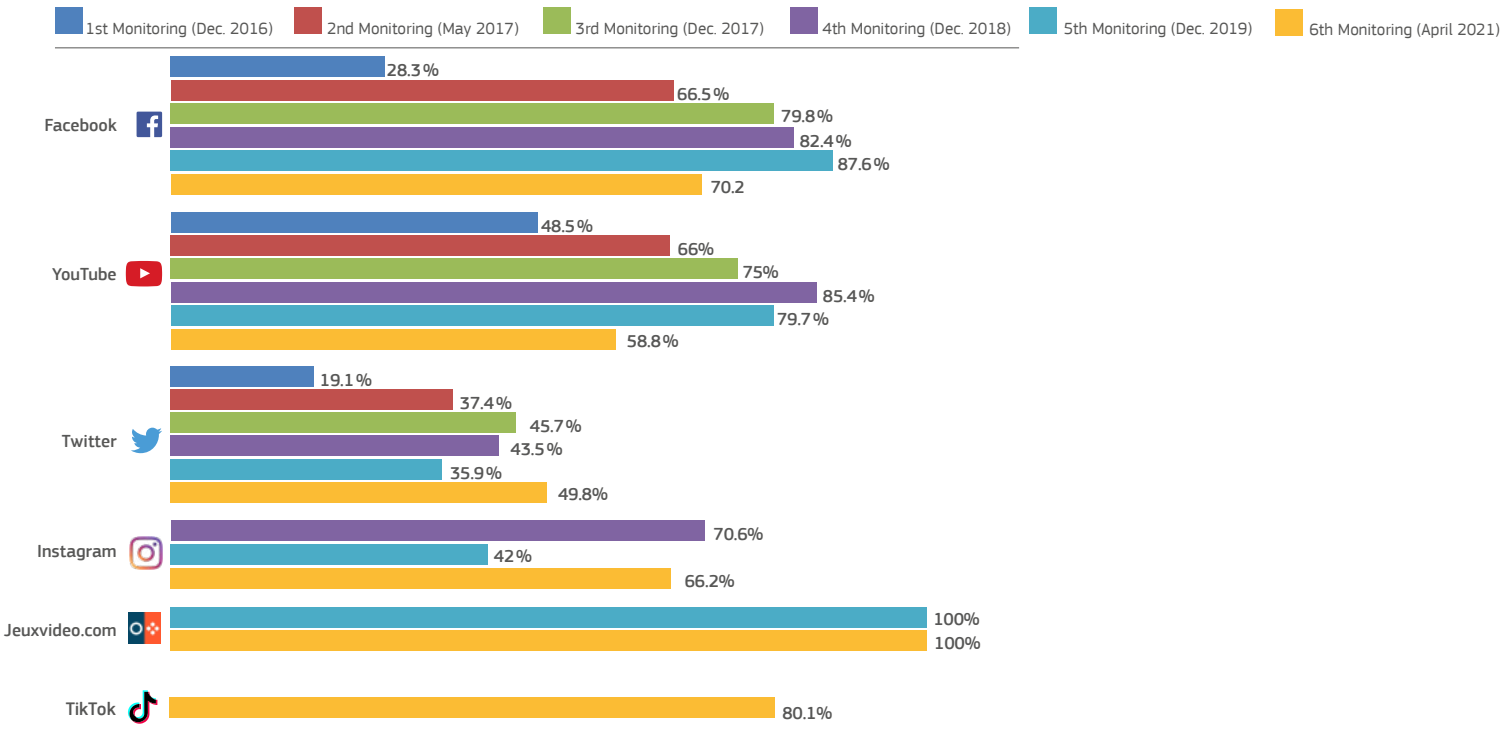


3. Removal rates

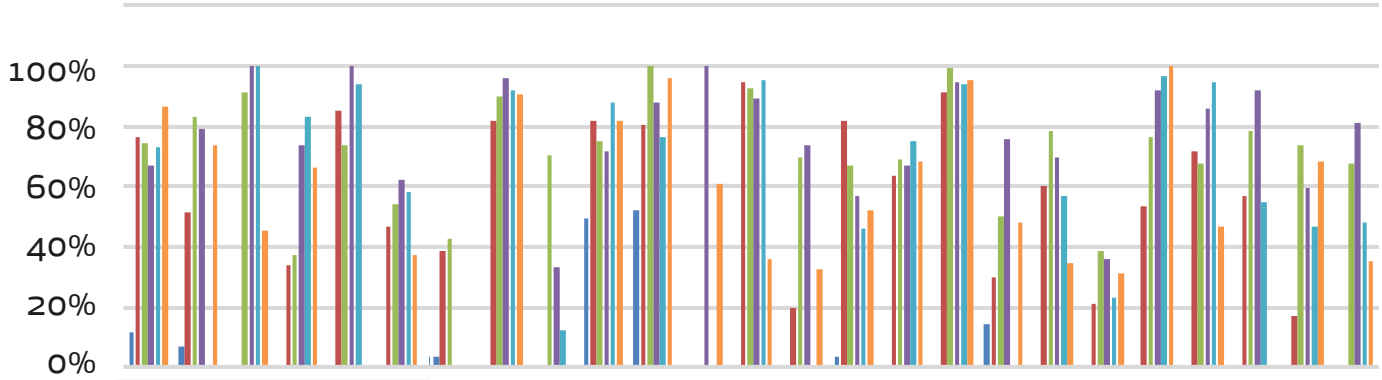
- Overall, IT companies removed **62.5%** of the content notified to them, while **37.5%** remained online. This result is lower than the average of **71%** recorded in 2019 and 2020.
- **Removal rates varied depending on the severity of hateful content.** On average, **69% of content calling for murder or violence against specific groups was removed,** while content using defamatory words or pictures to name certain groups was removed in **55%** of the cases.
- The divergence in removal rates of content reported using trusted reporting channels as compared to channels available to all users **was 13.5** percentage points. **This difference is similar to the one observed in 2020 (16.2%)**. This seems to suggest that **notifications from general users continue to be often treated differently** than those sent through special channels for "trusted flaggers".
- IT companies were invited to make a self-assessment on the results of the exercise. They reported cases in which they disagreed with the notifying organisations, i.e. where according to their assessment the content notified was not in violation of terms of services and/or local laws. This resulted in Facebook disagreeing on **12%** of cases flagged to them, Instagram on **11.9%**, and YouTube on **10%**. This shows the complexity of making assessments on hate speech content and calls for enhanced exchanges between trusted flaggers, civil society organisations and the content moderation teams in the IT companies.

Facebook removed **70.2%** of the content, YouTube **58.8%**, Instagram **66.2%** and Twitter **49.8%**. Twitter and Instagram made progress compared to 2020, while Facebook and YouTube had higher removal rates during the previous monitoring exercise in 2020. TikTok had a good first test, with **80.1%**. Jeuxvideo.com removed all flagged content.

Removals per IT Company



Rate of removals per EU country (in %)[1]



	AT	BE	BG	HR	CY	CZ	DK	EE	FI	FR	DE	EL	HU	IE	IT	LV	LT	NL	PL	PT	RO	SK	SI	ES	SE
1st monitoring (Dec. 2016)	11.4	6.9					3.4			49.5	52				3.6			14.3					0%	0%	
2nd monitoring (May 2017)	76.1	51.2		33.6	84.8	46.5	38.9	81.8		82	80.1		94.5	20	81.7	63.3	90.9	29.5	60.3	21	53.3	71.4	56.8	17.2	
3rd monitoring (Dec. 2017)	74	83	91.3	37.1	73.8	54.5	42.5	89.8	70.6	75	100		92.6	69.6	66.9	69	99	50	78.4	38.6	76.2	67.7	78.3	73.8	67.3
4th monitoring (Dec. 2018)	66.7	78.9	100	73.4	100	62.5		95.8	39.1	71.8	87.6	100	89.2	73.5	56.6	66.9	94.9	75.8	69.7	35.7	92.2	85.8	91.7	59.7	81.3
5th monitoring (Dec. 2019)	73.3		100	83.2	94.1	57.9		92	12.1	87.8	76.6		95		46.2	74.7	94.1		56.8	23.2	96.5	94.3	54.8	46.9	47.9
6th monitoring (2021)	86.8	73.5	45.2	66.3		37.4		90.6		81.5	95.9	60.6	36.1	32.3	52	68.4	95.4	47.9	34.5	31.2	100	46.5		68.2	35.4

[1] The table does not reflect the prevalence on illegal hate speech online in a specific country and it is based on the number of notifications sent by each individual organisation., Malta, is not included given the too low number of notifications made to companies (<20). For Slovenia, Cyprus, Finland, Luxembourg, and Denmark the organisations did not submit cases for this exercise. Three organisations from the United Kingdom took part to the monitoring exercise: CST (15 notifications), Galop (48) and Media Diversity Institute (78) with a total number of 151 notifications sent. Their work resulted on an average removal rate of 43%.

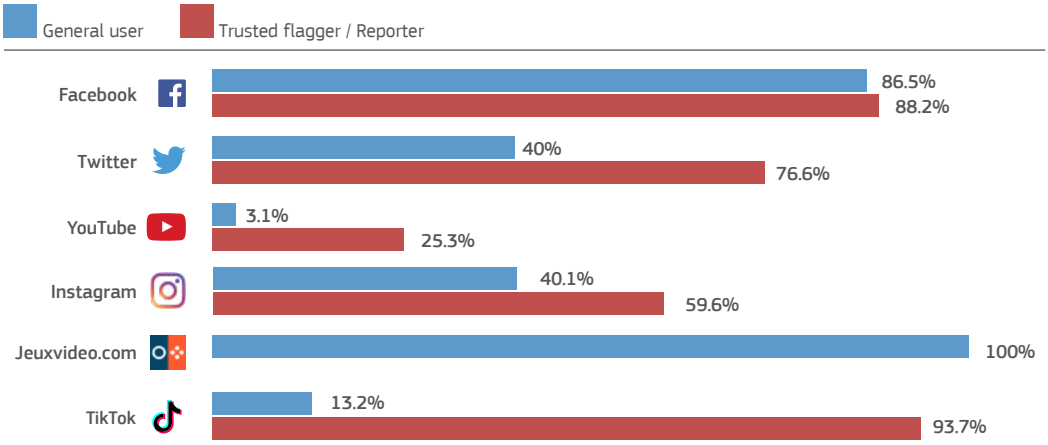
4. Feedback to users and transparency

- On average, the **IT companies responded with a feedback to 60.3%** of the notifications received. **This is lower than in the previous monitoring exercise (67.1%).**
- The Digital Service Act proposal adopted in December 2020 highlights the importance of clearer 'notice-and-action' procedures including transparency and feedback to users' notifications.

Facebook is informing users most systematically (**86.9%** of notifications received feedback). Twitter gave feedback to **54.1%** of the notifications, Instagram to **41.9%** and YouTube only to **7.3%**. Jeuxvideo.com sent feedback to all the notifications and TikTok to **28.7%**.

While Facebook is the only company informing consistently both trusted flaggers and general users, Twitter, YouTube, TikTok and Instagram provide feedback more frequently when notifications come from trusted flaggers. Jeuxvideo has increased its performance on feedback to users (it was 22.5% in 2020).

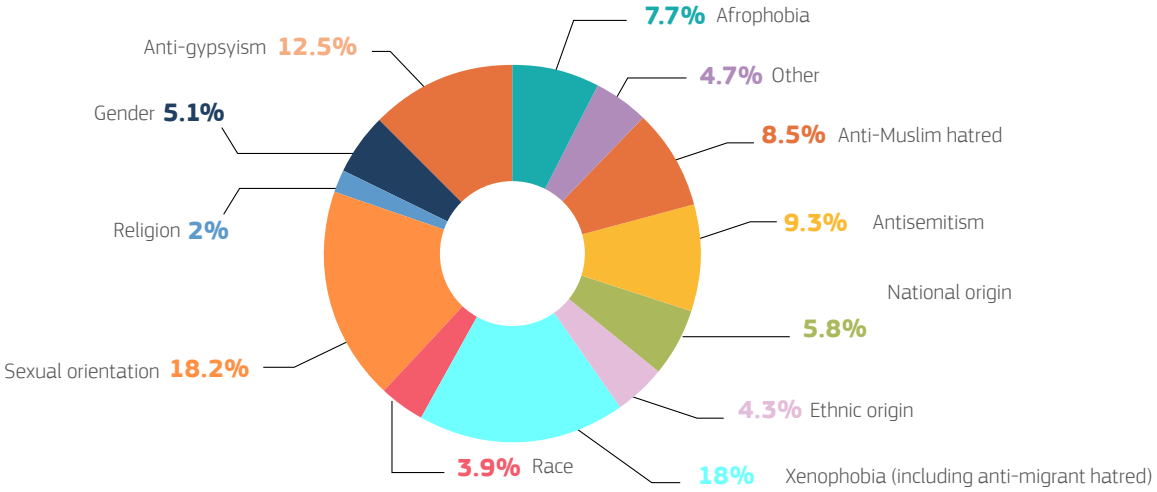
Feedback provided to different types of user



5. Grounds for reporting hatred

- In this monitoring exercise, sexual orientation and xenophobia (including anti-migrant hatred) are the most commonly reported grounds of hate speech (18.2% and 18% respectively) followed by anti-gypsyism (12.5%).
- The data on grounds of hatred are only an indication and are influenced by the number of notifications sent by each organisation as well as their field of work.

Grounds of hatred 2021



ANNEX

Methodology of the exercise

- The sixth exercise was carried out for a period of approximately 6 weeks, from 1 March to 14 April 2021, using the same methodology as the previous monitoring exercises.
- 35 organisations and 4 public bodies (in Belgium, France, Spain) reported on the outcomes of a total sample of 4543 notifications from 22 Member States.
- The figures do not intend to be statistically representative of the prevalence and types of illegal hate speech in absolute terms, and are based on the total number of notifications sent by the organisations.
- The organisations only notified the IT companies about content deemed to be "illegal hate speech" under national laws transposing the EU Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law.
- Notifications were submitted either through reporting channels available to all users, or via dedicated channels only accessible to trusted flaggers/reporters.
- The organisations having the status of trusted flagger/reporter often used the dedicated channels to report cases which they previously notified anonymously (using the channels for all users) to check if the outcomes could diverge. Typically, this happened in cases when the IT companies did not send feedback to a first notification and content was kept online.
- The organisations participating in the sixth monitoring exercise are the following:

COUNTRY	N° OF CASES
BELGIUM (BE)	
CEJI - A Jewish contribution to an inclusive Europe	19
Centre interfédéral pour l'égalité des chances (UNIA)	12
BULGARIA (BG)	
Integro association	105
CZECH REPUBLIC (CZ)	
In Iustitia	104
Romea	99
GERMANY (DE)	
Jugendschutz.net	98
ESTONIA (EE)	
Estonian Human Rights Centre	96
IRELAND (IE)	
ENAR Ireland	31
GREECE (EL)	
Greek Helsinki Monitor	104
SPAIN (ES)	
Fundación Secretariado Gitano	177
Federación Estatal de Lesbianas, Gais, Transexuales y Bisexuales (FELGTB)	85
Spanish Observatory on Racism and Xenophobia (OBERAXE)	290
Spanish Ministry of Interior	150
Khetane Platform	37
FRANCE (FR)	
Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA)	210
Platforme PHAROS	82
CROATIA (HR)	
Centre for Peace Studies / Human Rights House Zagreb	104
ITALY (IT)	
Ufficio Nazionale Antidiscriminazioni Razziali (UNAR)	83
CESIE	98
Centro Studi Regis	124
Amnesty International Italia	76
Associazione Carta di Roma	103

COUNTRY	N° OF CASES
LATVIA (LV)	
Mozaika	101
Latvian Centre for Human Rights	107
LITHUANIA (LT)	
National LGBT Rights Oganisation (LGL)	260
HUNGARY (HU)	
Háttér Society	108
AUSTRIA (AT)	
Zivilcourage und Anti-Rassismus-Arbeit (ZARA)	75
POLAND (PL)	
HejtStop / Projekt: Polska	93
Never Again Association	104
PORTUGAL (PT)	
Associação ILGA Portugal	93
ROMANIA (RO)	
Active Watch	56
SLOVAKIA (SK)	
digiQ	141
SWEDEN (SE)	
Institutet för Juridik och Internet	96
NETHERLANDS	
INACH/Magenta	69
MALTA	
MGRM	1