INACH

Bringing the Online In Line with Human Rights

Overview of the latest transparency reports under the DSA

Executive Summary

This brief compares the latest transparency reports from the VLOPs: Meta (Feb 2025), TikTok (Dec 2024), X (April 2025), YouTube (Feb 2025). Across platforms, the transparency reports vary considerably in completeness and usefulness. While all reports provide tables detailing the number of human moderators per language, it is immediately apparent how limited the moderation capacity is for smaller or less widely spoken languages. The clarity of the data also differs significantly. For instance, X’s Transparency report does not provide a meaningful link between the volume of illegal hate-speech reports received and the number of items removed either globally or nationally. Similarly, several platforms do not specifically report on illegal hate speech: Meta, for example, aggregates hate speech within a broader “illegal content” category that also includes other forms of violations. TikTok has almost no data available specifically on hate speech as a separate category. Some platforms, such as YouTube, do offer figures related to submissions by Trusted Flaggers, but this remains the exception rather than the norm. None of the platforms give any information on how removal decisions are made or how they demote content. Importantly, none of the reviewed reports provide a breakdown of hate-speech enforcement by country or by protected characteristic or type of hate speech, leaving critical blind spots in understanding enforcement disparities across the EU.

Meta

Human moderation data

These data apply to both Facebook and Instagram. Some of the languages (e.g. English, Spanish and Portuguese) also are used globally outside the EU which explains the higher number of reviewers in this table.

EU Languages	**Number of Reviewers**

Bulgarian	55
Croatian	56
Czech	63
Danish	39
Dutch	154
English	212
Estonian	6
Finnish	24
French	630
German	470
Greek	37
Hungarian	44
Italian	427
Latvian	4
Lithuanian	11
Maltese	1
Polish	112
Portuguese	2088
Romanian	74
Slovak	49
Slovenian	8
Spanish	3110
Swedish	78

Facebook

Meta reports 538,826 removals for hate-speech violations, of which 386,892 were carried out automatically by AI, without human review. In addition to removals, Meta applies “organic content demotions,” meaning that content is not deleted but its visibility is intentionally reduced across feeds, search results, recommendations, and other ranking surfaces. For hate speech specifically, 65,115 items were demoted, with 65,105 of these demotions applied automatically by AI.

The platform does not provide a distinct category for “illegal hate speech.” Instead, it reports notices under a broad “other illegal content” category, which explicitly excludes intellectual property, defamation, and privacy violations but does not isolate hate speech as its own legal category. Meta received a total of 113,638 notices regarding alleged illegal content. Of these, 16,052 notices resulted in the removal of content for policy violations, and an additional 1,852 notices resulted in content being restricted due to alleged illegality.

Instagram

Hate speech removals totalled 716,246, of which 639,852 were carried out automatically by AI systems without human review. In addition, Instagram applied 53,526 demotions to hate-speech content, with 53,522 of these demotions performed through automated means.

The report does not provide a dedicated category for *illegal* hate speech; instead, it groups such content under the broad label of “other illegal content,” which excludes intellectual property, defamation, and privacy violations. Across the reporting period, Instagram received 61,888 notices. These resulted in 12,418 instances where content was removed on the basis of policy violations, and 655 cases where access to content was restricted due to alleged illegality under applicable law.

TikTok

Human moderation data

Number of people dedicated to content moderation	**Official Member State language**
Bulgarian	38
Croatian	29
Czech	53
Danish	15
Dutch	99
English	1,524
Estonian	17
Finnish	31
French	509
German	532
Greek	50

Hungarian	51
Italian	290
Latvian	22
Lithuanian	19
Polish	146
Portuguese	160
Romanian	99
Slovak	33
Slovenian	37
Spanish	531
Swedish	72

During the reporting period, the platform received 22,429 user reports concerning illegal hate speech. In addition to general user reporting, Trusted Flaggers submitted 5 specialised reports on illegal hate speech. However, beyond this data there is no information to be found in this report on hate speech specifically.

YouTube

Human moderation data

Number of people dedicated to content moderation	**Official Member State language**
Bulgarian	22
Croatian	37
Czech	26
Danish	19
Dutch	53
English	4,187
Estonian	10
Finnish	22

French	276
German	250
Greek	48
Hungarian	31
Italian	175
Latvian	11
Lithuanian	21
Polish	230
Portuguese	264
Romanian	56
Slovak	11
Slovenian	16
Spanish	543
Swedish	29
Agnostic	5747

Content moderators who review non-language content (e.g., an image) are included in the ‘Agnostic’ category. Agnostic reviews are primarily done when no language is needed to conduct the review (e.g., adult content) or in specific cases when YouTube cannot identify the language.

YouTube received 7,869 notices related to *Hate and Harassment*, covering abusive insults directed at individuals or groups. Trusted Flaggers submitted an additional nine notices in this category, bringing the total number of hate-related notifications to 7,878. While YouTube reports 374,059 actions taken on content deemed illegal and 6,009 actions taken for violations of its own policies, these figures encompass *all* types of illegal or policy-violating material and do not provide a breakdown specific to illegal hate speech. However, YouTube’s own-initiative enforcement offers clearer insight into the scale of harmful content removal: 3,103,427 items were removed for *Hateful or Abusive* content.

Why civil society monitoring remains crucial

Despite the progress made under the DSA, the current transparency reports still leave major blind spots that limit meaningful oversight. Key metrics remain inconsistent across platforms, with no harmonised definitions, no breakdown of illegal hate speech by protected characteristic, and no clarity on moderation outcomes per country or per language. Several platforms provide only partial insight into how they detect, assess, and act on hate speech reports, while others such as X publish data that is too aggregated to meaningfully evaluate enforcement effectiveness. These gaps make independent NGO monitoring not just valuable, but essential. Civil society organisations provide the contextual, qualitative understanding that platforms do not disclose: they identify emerging narratives, track real-world harms, assess the adequacy of platform responses, and highlight systemic under-enforcement affecting vulnerable groups. NGO monitoring is therefore indispensable for ensuring accountability, informing policymakers, strengthening user protections, and ensuring that platforms' DSA reporting translates into real-world safety for the communities most affected by online hate.

Find the full latest transparency reports here:

X
Facebook:
Instagram:
YouTube:
TikTok: