INACH

Bringing the Online In Line with Human Rights

Compiled by
Tamás Berecz, Adinde Schoorl &
Naomi Tidball
2025

Monitoring Report 2025

www.inach.net

Table of contents

International Network Against Cyber Hate – INACH

INACH was founded in 2002 to use intervention and other preventive strategies against cyber hate. The member organisations are united in a systematic fight against cyber hate, for example as complaints offices, monitoring offices or online help desks. In their respective countries, they provide important contacts for politicians, internet providers, educational institutions, and users.

Funding for INACH is provided by its members, the European Commission, the BPB and other donors. The International Network Against Cyber Hate (INACH) unites multiple organisations from the EU, North Macedonia, Albania, Israel, Russia, South America, and the United States. While starting as a network of online complaints offices, INACH today pursues a multi-dimensional approach to educational and preventive strategies.

This publication has been produced with the financial support of the Citizens, Equality, Rights and Values (CERV) Programme of the European Union. The contents of this publication are the sole responsibility of the International Network Against Cyber Hate and can in no way be taken to reflect the views of the European Commission.

Supported by the Citizens, Equality, Rights and Values (CERV) Programme of the European Union

1. Background to the Monitoring Exercise

This year the Monitoring Exercise (ME), organized by the European Commission and coordinated by INACH, took place from 3 November until 12 December 2025. The following organisations were part of this ME: Dokustelle, HRHZ, Romea, AKVAH, EHRC, Transfeminines NGO, Point de Contact, GHM, Háttér, Hope & Courage, CESIE, LCHR, LGL, Aditus, Czulent, ILGA, Transparente, DigiQ[1], CERMI, FSG, INACH and SOS Racismo. More than 2000 cases were gathered during these weeks.

A final remark before proceeding to the findings of the ME: due to the purpose of our documentation, X (formerly known as Twitter) will still be referred to as Twitter.

This report contains the quantitative findings of the ME in the first chapter, a chapter dedicated to the types of hate speech and intersectionality and finally the qualitative findings and the observations by NGOs that carried out the ME.

[1] Due to administrative reasons the data from DigiQ and SOS Racismo is not included, so the report will be updated later.

2. Findings of the ME

Overview

The following charts showcase the results of the *Normal User Removal Rates*, the *Trusted Flagger*[2] *Removal Rates*, the *Normal User Feedback Rates* and *Normal Users Feedback Rates*. In the following sections, we provide a description of the data presented in the charts.

Removal Rates

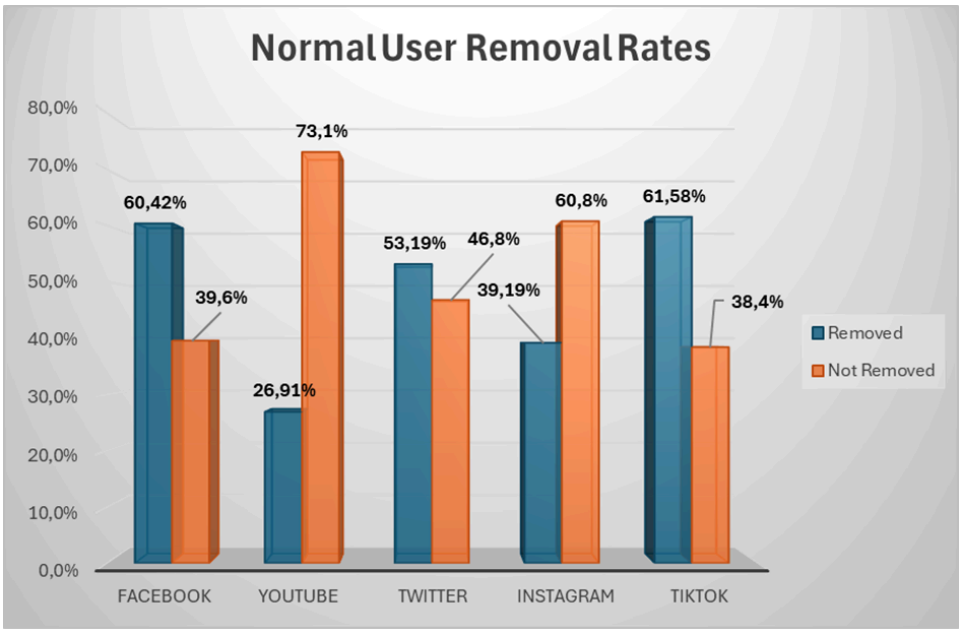

Figure 1: Chart of Normal User Removal Rates

This chart (fig.1) provides an overview of content removal across five platforms: Facebook, YouTube, Twitter (X), Instagram, and TikTok. The findings indicate a nearly balanced outcome between removed and non-removed content on Twitter (X). TikTok shows the highest normal user removal rate, with 61,58% of content removed and

2

The term Trusted Flagger refers here to the organisations that operate as a Trusted Partner and report through the Trusted Partner channels, it does not refer to the Trusted Flaggers under the DSA.

38,4% not removed. Following TikTok, Meta's Facebook records a similarly high removal rate at 60,4%, with 39,6% of content not removed. In contrast, Meta's Instagram demonstrates a higher non-removal rate, with 60,8% of content not removed and only 39,2% removed.

Notably, YouTube exhibits the lowest removal rate among the platforms, with 73,1% of content not removed and only 26,91% removed. However, prior monitoring of YouTube has consistently shown insufficient removal of content on the platform (International Network Against Cyber Hate [INACH], 2021).

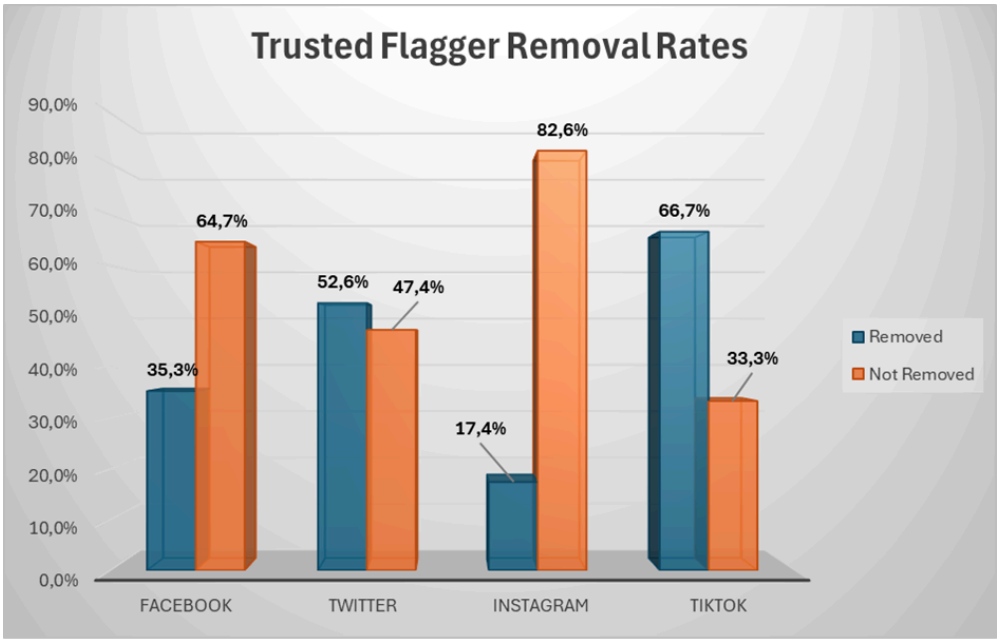

Figure 2: Chart of Trusted Flagger Removal Rates

This chart (fig.2)[3] illustrates content removal rates following reports submitted by organisations operating as Trusted Partners. TikTok removed 66,7% of flagged content, leaving 33,3% not removed. Twitter (X) demonstrated a similar trend, with 52,6% of content removed and 47,4% remaining. In contrast, Meta platforms showed

[3] YouTube results are not presented in this chart, most likely due to the absence of reporting through the Trusted Partner channel of YouTube.

substantially higher non-removal rates: Facebook removed 35,5% of reported content while 64,7% was not removed, and Instagram exhibited the lowest removal rate, with only 17,4% removed and 82,6% not removed.

Overall, the findings reveal some differences between removal rates for reports submitted by normal users and those submitted by Trusted Partners (Fig.1 and Fig.2). Instagram recorded a higher removal rate for normal users (39,19%) than for Trusted Partners. Whereas, TikTok demonstrated an increased responsiveness to Trusted Partner reports, with a removal rate of 66,7%, compared to 61,58% for normal users.

Feedback Rates

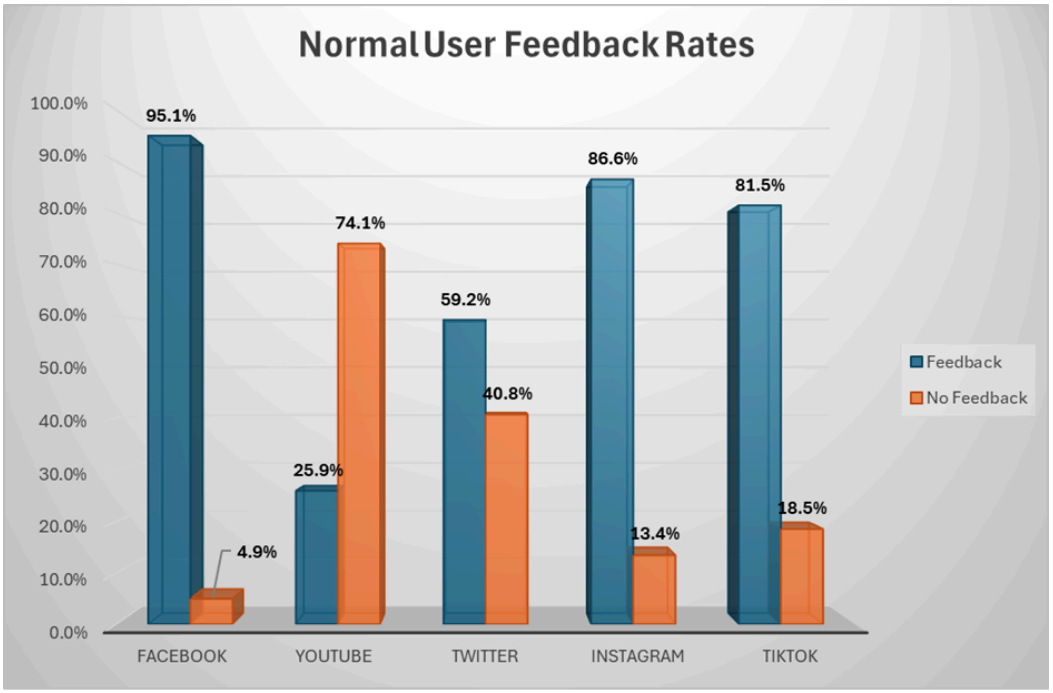

Figure 3: Normal User Feedback Rates

This chart (Fig. 3) evaluates normal user feedback rates across five social media platforms. Overall, the findings indicate relatively high levels of user feedback on most platforms.

Meta's Facebook demonstrates the highest feedback rate, with feedback provided in 95,1% of cases and no feedback in 4,9%. Similarly, Instagram shows strong user engagement, with 86,6% of cases receiving feedback and 13,4% receiving none. TikTok also records a high level of feedback, with 81,5% of cases providing feedback and 18,5% resulting in no feedback.

In contrast, Twitter (X) exhibits more moderate feedback rates, with feedback provided in 59,2% of cases and absent in 40,8%. YouTube shows a substantially lower level of normal user feedback, with only 25,9% of cases receiving feedback, meaning that 74,1% of reports elicited no response.

Assessment Time Ratios

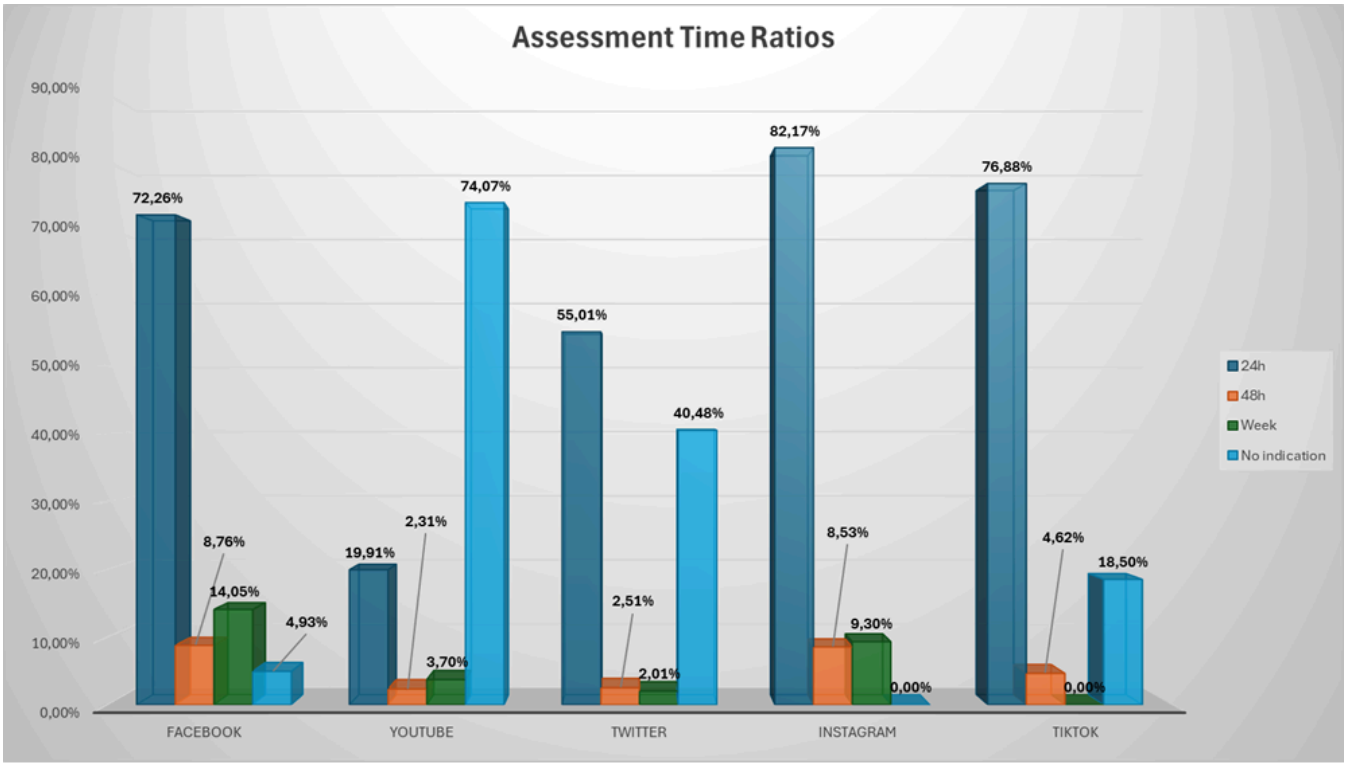

Figure 4: Chart of Assessment Time Ratios

This chart (Fig. 4) presents assessment time results across five social media platforms. The data are based on whether the platforms responded to notifications of reported issues. Assessment times were categorised into four time frames: within 24 hours, within 48 hours, within one week, and no indication of assessment.
Responses within 24 hours were most prevalent on Meta's Instagram (82,17%), followed by TikTok (76,88%) and Meta's Facebook (72,26%). Twitter (X) recorded a moderate response rate within this time frame at 55,01%, while YouTube showed a substantially lower rate of assessment within 24 hours (19,91%).

Within the 48-hour time frame, assessment rates decreased across all platforms. Meta platforms again demonstrated the highest responsiveness, with Facebook at 8,76% and Instagram at 8,53%. TikTok recorded 4,62% of assessments within 48 hours, followed by Twitter (X) at 2,51% and YouTube at 2,31%.

Assessment within one week showed further declines. Meta platforms continued to record higher rates than other platforms, with Facebook at 14,05% and Instagram at 9,30%. YouTube followed with 3,70%, while Twitter (X) recorded 2,01% and TikTok showed no assessments within this time frame (0,0%).

The absence of any indication of assessment was most pronounced on YouTube, accounting for 74,07% of notifications, followed by Twitter (X) at 40,8%. TikTok recorded 18,50% of notifications with no assessment indicated. In contrast, Meta platforms demonstrated significantly lower rates of non-assessment, with Facebook at 4,93% and Instagram at 0,0%.

3. Types of hate speech and intersectionality

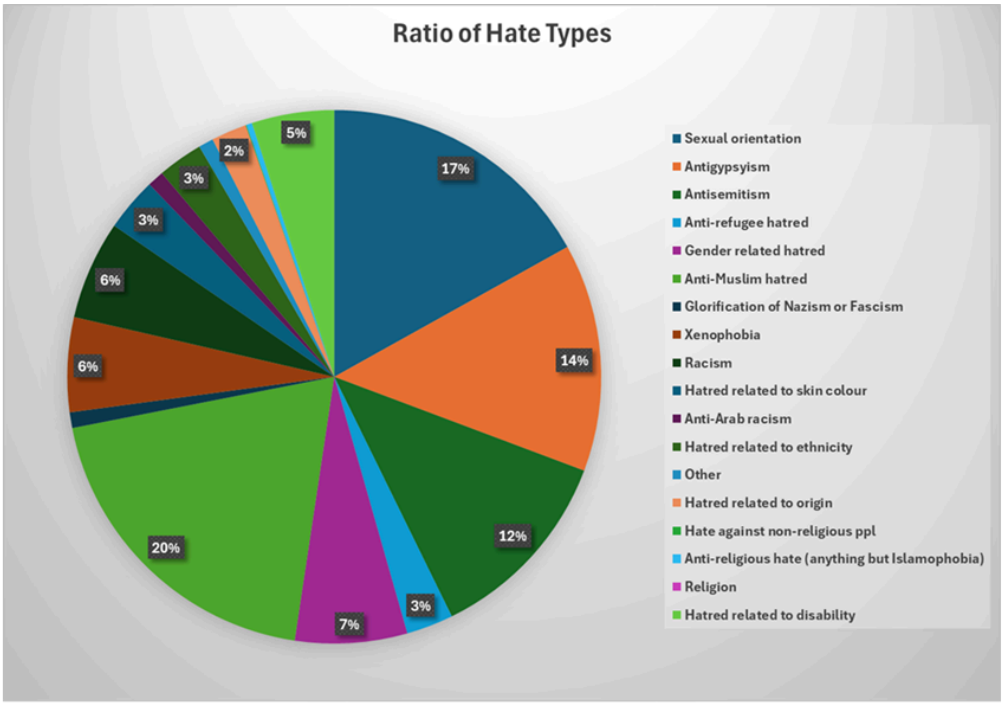

Figure 5: Ratio of Hate Types

Fig.5 demonstrates the different types of hate speech and their percentages: The participation of LGBTQ+ rights NGOs and organizations monitoring anti-Muslim rhetoric and Islamophobia, hate based on anti-Muslim (20%) hatred and sexual orientation (17%) are prominently reflected in the projected data. At the same time, organisations that do not focus on a specific type of hate speech also reported more anti-Muslim hate as well as anti-refugee hate. Other forms of hate such as antisemitism (12%), antigypsyism (14%), and gender-related hatred (7%) are also significantly present in the monitoring results.

Across the analysed countries, online hate speech follows identifiable structural patterns rather than isolated or random dynamics. Three dominant thematic axes emerge: migration and Islamophobia, culture-war narratives targeting LGBTQ+ communities, and historically rooted discrimination against specific minorities such as Roma and Jewish communities. These axes are shaped by geopolitical conflicts, domestic political discourse, and socio-economic tensions.

Western and Southern European countries predominantly report hate narratives centred on migration and religion, often framed as security threats or cultural incompatibility. In contrast, Central and Eastern European countries display stronger culture-war dynamics, with LGBT+ communities as primary targets. Historically marginalised groups, particularly Roma communities, remain focal points of hate in several national contexts, demonstrating the persistence of structural racism in digital environments.

For example, in Estonia an increase was noted of antisemitism during this ME. The increase is linked to the international tensions related to the ongoing conflicts in the Middle East. In Poland it was noted that antisemitism is often embedded in broader conspiratorial and extremist narratives, which gain visibility during periods of heightened geopolitical tension and domestic political debate.

In Hungary it was reported that in terms of the severity of the content, it can be stated that many of them go beyond expressing hostility or violence toward the targeted groups. They explicitly or implicitly encourage killing members of these groups or individuals associated with them. One possible explanation is that Hungarian everyday language contains expressions that are generally violent or inciting (for example, “punch someone” or “shoot someone”), but in this context these carry a far more serious and threatening meaning.

Intersectionality emerges as a key differentiating factor across regions. Western and Nordic countries tend to report higher levels of intersectional hate, where multiple identity markers (e.g. migration status, religion, race, gender, sexuality) overlap within single narratives. By far, most intersectionality appears between anti-migrant sentiments and anti-muslim hate. Migrants are framed as Muslims, and Islamophobia is embedded in migration discourse and security/criminality narratives are used both when talking about muslims and migrants. There is also a clear overlap between anti-migrant sentiments and racism. Migrants are racialised as “others” and cultural

incompatibility narratives are used when talking about migrants. The use of cultural incompatibility narrative is also seen overlapping with anti-Roma narratives.

In Central and Eastern Europe, hate speech is more frequently concentrated on single target groups, reflecting narrower but deeply entrenched patterns of exclusion.

This divergence suggests that online hate in Europe is not uniform but contextually embedded within national political cultures and historical trajectories.

4. IT platform performances and NGO observations

Apart from YouTube and Instagram, the platforms removed more reported content as a normal user than not during this ME. Facebook removed 60 % of the reported content, Twitter 53% and TikTok almost 62%. Interestingly enough the removal rates as Trusted Flaggers are quite different; Meta removed much less (35 % on Facebook and 17 % on Instagram), While Twitter and TikTok removed a similar percentage as with normal user's reported content. But this year's removal rates are in fact higher than the removal rates of last year's ME. An example of a country where removal rates were low is Malta. Due to a lack of data on the Maltese language on many platforms this is often a hindrance to having the comments recognised as hate speech. This is further exacerbated when comments have typos or spelling mistakes which non-native speakers might not immediately understand the context or meaning of. To try and overcome this, the reporters translated the meanings when reporting the comment, however, this remained unsuccessful. No text or video in the Maltese language was taken down.

The feedback rate of this year's ME shows a considerable increase compared to last year's ME. Only Facebook scored the same percentage of 95 % as a feedback rate.
Assessment time ratio varied between platforms and organisations and content. Some content is reviewed within minutes, others take up to a week. But in general the percentage of content that was reviewed within 24 hours, 48 hours and one week was much higher than last year across all platforms. Even the percentage of content that received no indication was much lower compared to last year's ME.

Some organisations also received emails from platforms about reported content, asking for more evidence or writing that they could not find a reason why the content is illegal but still inviting the reporter to send in more evidence. Especially YouTube used emails as a manner of communication. Apart from the fact that it discourages reporting to the

average user, it also increases the assessment time ratio, gives inconclusive answers on whether the platform will remove it or not and lacks user friendliness.

All the organisations implicitly describe platform reporting tools as: complex, opaque and inconsistent. Only TikTok is consistently described as relatively "easy" and fast but not necessarily transparent. Twitter is mostly described as complicated to report and untransparent. Meta and YouTube score in between the other two platforms. NGOs also reported problems with the different reporting mechanisms of the platforms with URLs that change after having reported content and platforms' feedback without sending the URL of the reported case. However, organisations have used different reporting mechanisms during this Monitoring Exercise: as a normal user, as a normal user through the 'illegal under the DSA' button and as a Trusted Partner through the designated reporting channels which makes the experience harder and less uniform to compare. Finally, according to the NGOs, most of the rejected cases do not entail more than automated feedback or minimal feedback, making it hard to understand enforcement decisions.

Three distinct approaches to moderation outcomes emerge: full removal, territorial restriction (geoblocking) or limitation of visibility. X is most frequently associated with geoblocking. Organisations reported content remaining visible globally but inaccessible in specific countries. This was reported mostly in Spain, Portugal and the Netherlands. YouTube uses geoblocking as well, it was reported by organisations in Portugal and Slovakia. Meta has used visibility limitation measures frequently instead of deletion of content. Organisations did express criticism regarding these measures: Geoblocking is perceived as a less transparent and less effective response than removal. Organisations noted uncertainty about whether content was removed globally or merely hidden locally. Several partners highlighted that geoblocking complicates monitoring and accountability, as harmful content may persist in other jurisdictions. Geoblocking and visibility limiting function as intermediary tools that allow platforms to mitigate regulatory risk while preserving content in broader markets. The use of different

measures than complete removal raises concerns about the transparency and effectiveness of platform compliance with European regulatory frameworks, particularly the Digital Services Act.

It is noticeable that there is a shift in hate speech types detected during this year's ME. While last year it was more varied, this year there is a clear focus on anti-LGBTQ+ hate and anti-muslim hate. This year's clear focus on anti-LGBTQ+ hate, however, could be influenced by an increased participation of LGBTQ+ organisations.

Overall, the indicators show that the results of this ME are considerably better than last year when it comes to normal user removal rate, Trusted Flagger removal rate, time assessment ratio and feedback rates. At the same time, reporting organisations have registered quite some difficulties with reporting mechanisms and doubts regarding the use of geoblocking and visibility limitations.

Citations

International Network Against Cyber Hate (2021). *INACH First Shadow Monitoring Report 2021*:**https://www.inach.net/wp-content/uploads/First_FINAL_ShadowME_Report_2021_FINAL.pdf**