INACH

Bringing the Online In Line with Human Rights

Compiled by
Tamás Berecz &
Cosima A. Hofacker

2024

Monitoring Report 2024

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, 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.

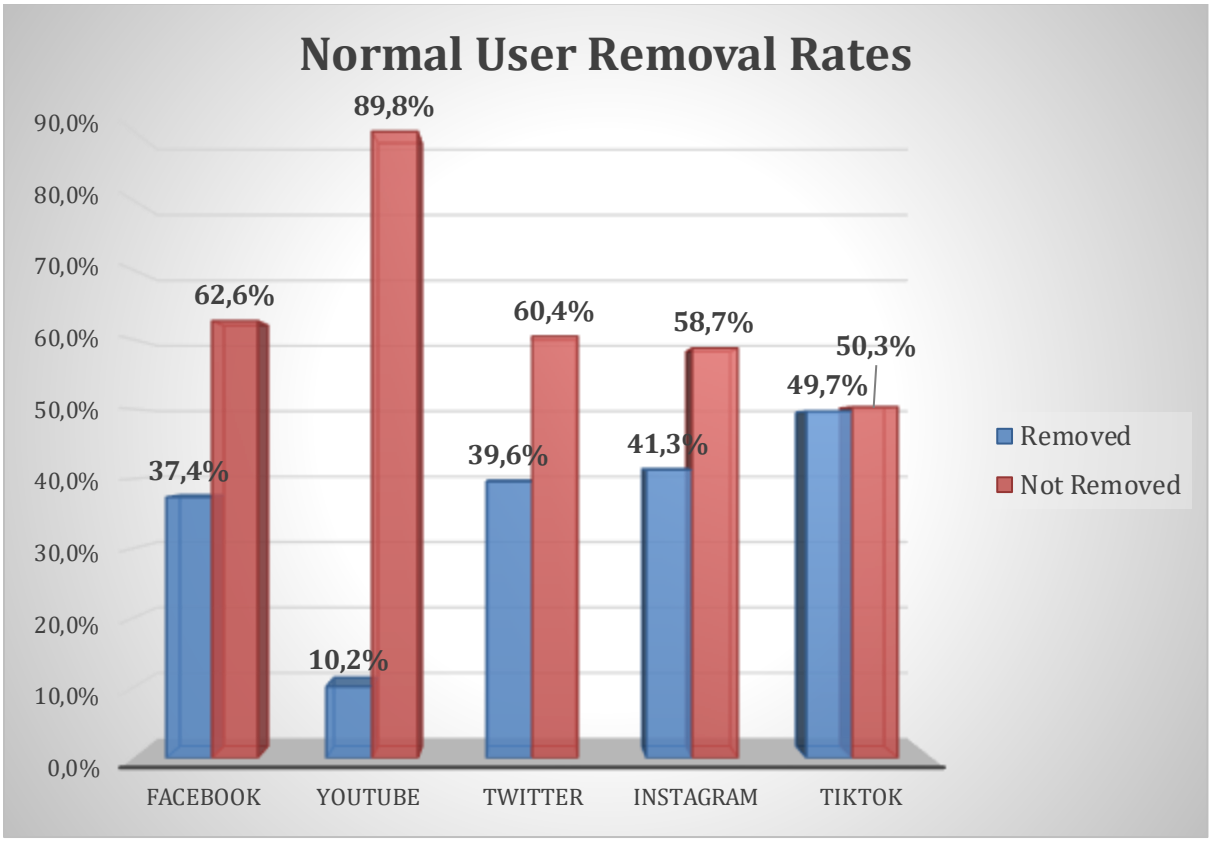
Supported by the Citizens, Equality, Rights
and Values (CERV) Programme of the
European Union

1. Basic information on the Monitoring Exercise

This year the normal annual Monitoring Exercise (ME) organized by the European Commission was cancelled. However, INACH and other partners organized the Shadow Monitoring Exercise with its partners from the 9th of September until the 18th of October 2024. The following organisations were part of this ME: CEJI, CESIE, DigiQ, Dokustelle, Estonian Human Rights Center, Fighting Online Antisemitism, Fundación Secretariado Gitano, Greek Helsinki Monitor, Háttér Society, Human Rights House Zagreb, INACH, Integro, Latvian Centre for Human Rights, LGL, LICRA, Movimiento contra la Intolerancia, MIIJI, Never Again Association, ROMEA and ZARA. These twenty organisations participated in the Shadow ME, covering Austria, Belgium, Bulgaria, Croatia, Czech Republic, France, Germany, Greece, Hungary, Italy, Latvia, Lithuania, the Netherlands, Poland, Slovakia, Slovenia, Spain, and Sweden. More than 1900 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.

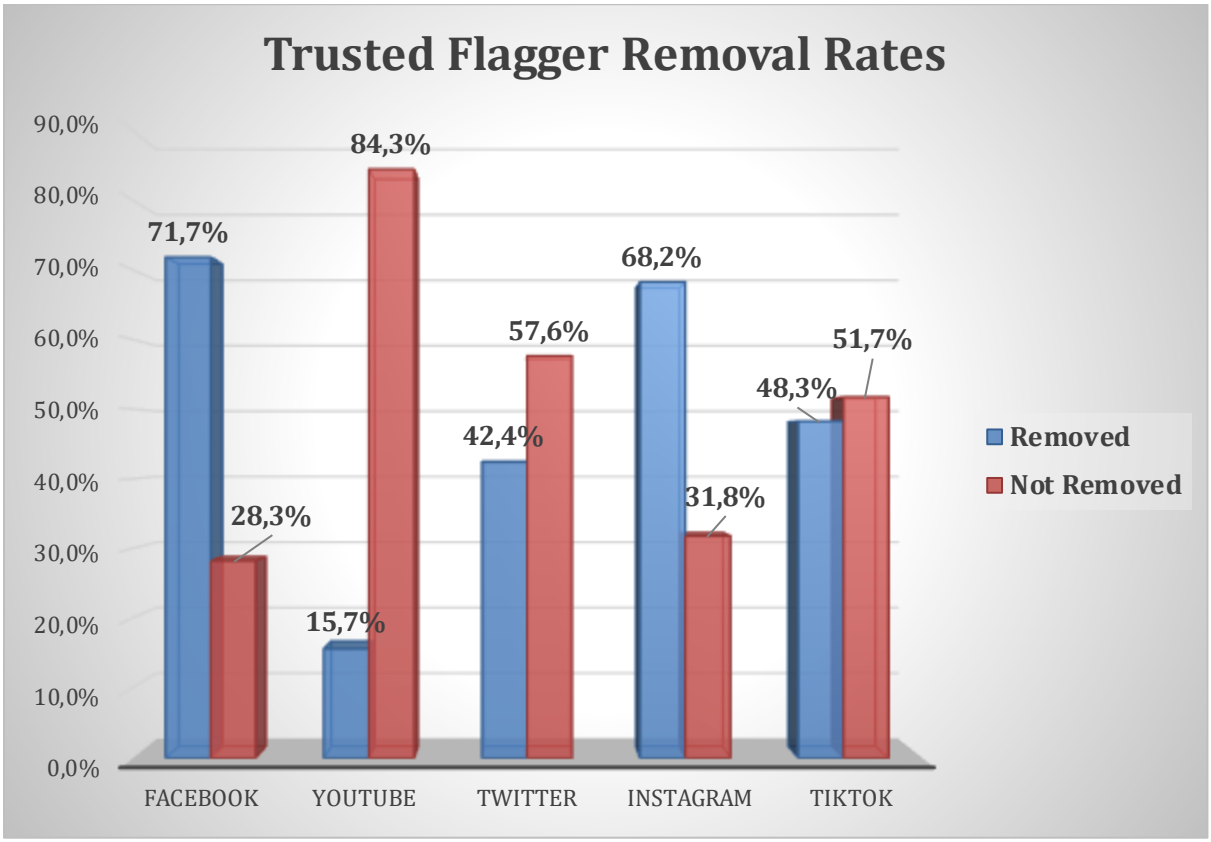
2. Findings of the ME

The following charts showcase the results of the *Normal User Removal Rates*, the *Trusted Flagger Removal Rates*, the *Normal User Feedback Rates* and *Assessment Time Ratios*.



The *Normal User Removal Rates* show the existing gap between the reported content that was removed and the content which was not removed on all five monitored platforms. YouTube has the lowest percentage of content removed, with only 10.2% of reported content removed. The results for the four other platforms are very similar: Facebook removed 37.4%, Twitter removed 39.6%, Instagram removed 41.3% and TikTok removed 49.7%.

Therefore, the most striking result is that of YouTube, which did not remove the reported content in almost ninety per cent of cases.

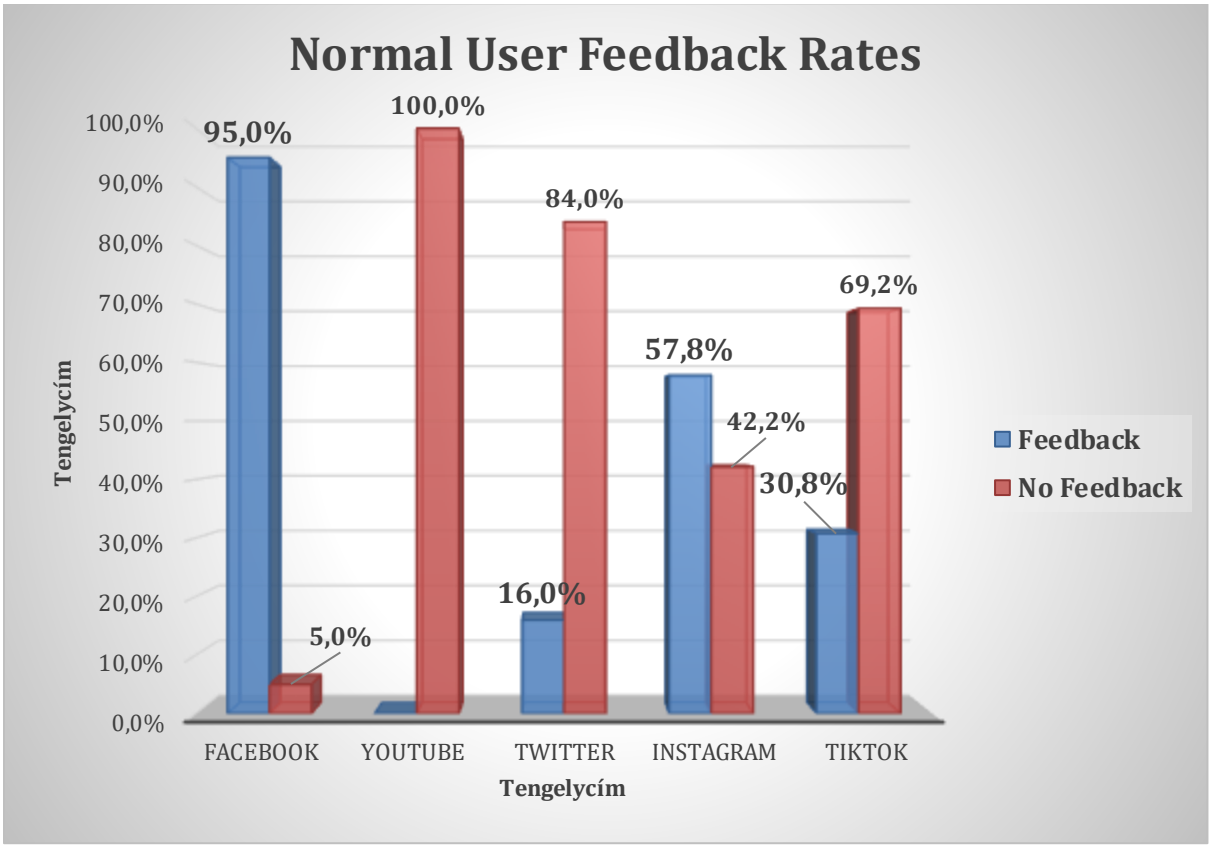


The *Trusted Flagger Removal Rate* reveals different results. All platforms' see increased removal rates via the Trusted Flagger System compared to the normal user removal rates. Facebook removed 71.7% of the reported content through the Trusted Flagger channels and Instagram removed 68.2%. Fairly similar are TikTok, which removed 48.3% of the reported content by Trusted Flaggers, and Twitter which removed 42.4%.

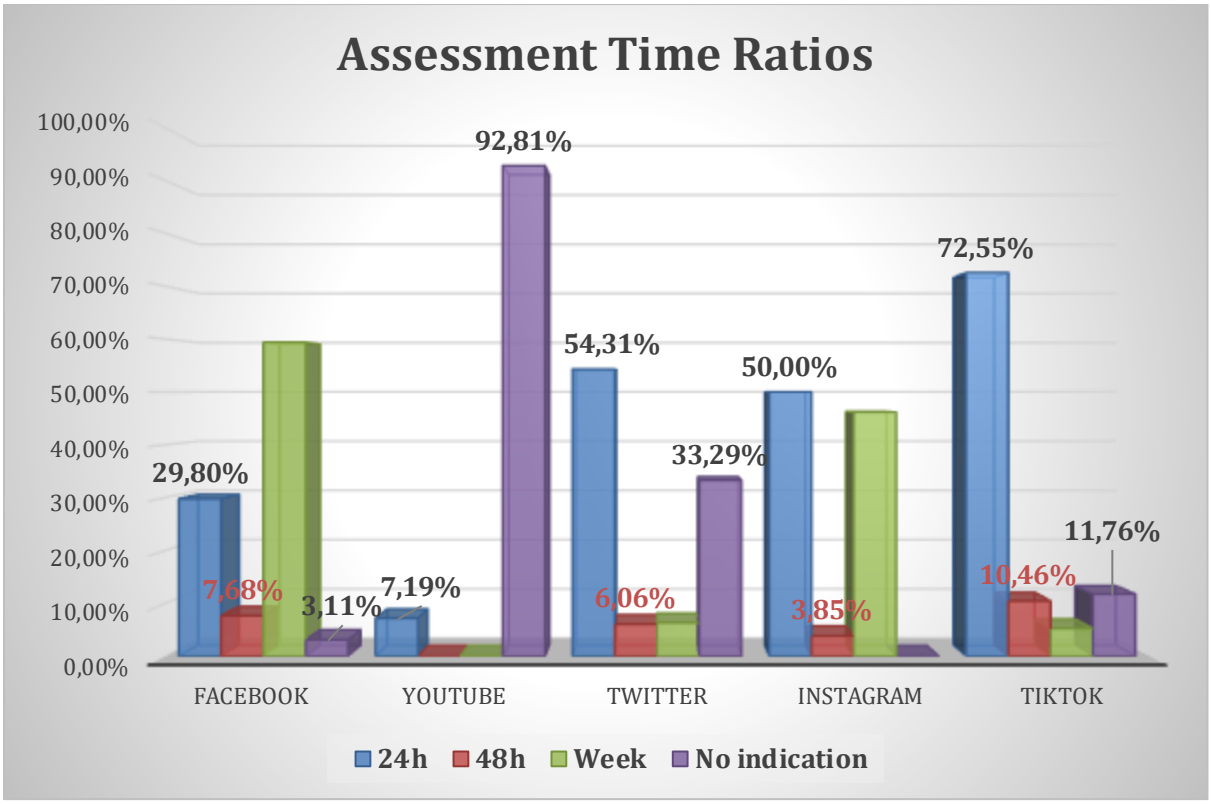
YouTube stands out again with a removal rate of only 15.7%. TikTok, Twitter and YouTube, three of the five monitored platforms, still have a lower removal rate than the 'non-removal rate,' despite the Trusted Flagger channels.

The following findings reflect the *Normal User Feedback Rates*. The results show generally lower feedback rates for reported content. Solely Instagram with a feedback rate of 57.8% and Facebook with a feedback rate of 95% exceed the 'non-feedback rates'.

Conversely, TikTok, Twitter and particularly YouTube have low to no 'feedback rates', with TikTok having a feedback rate of 30,8%, Twitter 16% and YouTube a strikingly low feedback rate of 0% and, thus, contrasting with Facebook, which has a comparatively high and the best feedback rate in this monitoring.



Lastly, the findings below illustrate the results of the *Assessment Time Ratios*.



The results differ greatly between the individual platforms. Receiving assessment within 24h was most prevalent on TikTok with 72.55%, followed by 54.31% on Twitter, 50% on Instagram, 29.80% on Facebook and 7.19% on YouTube.

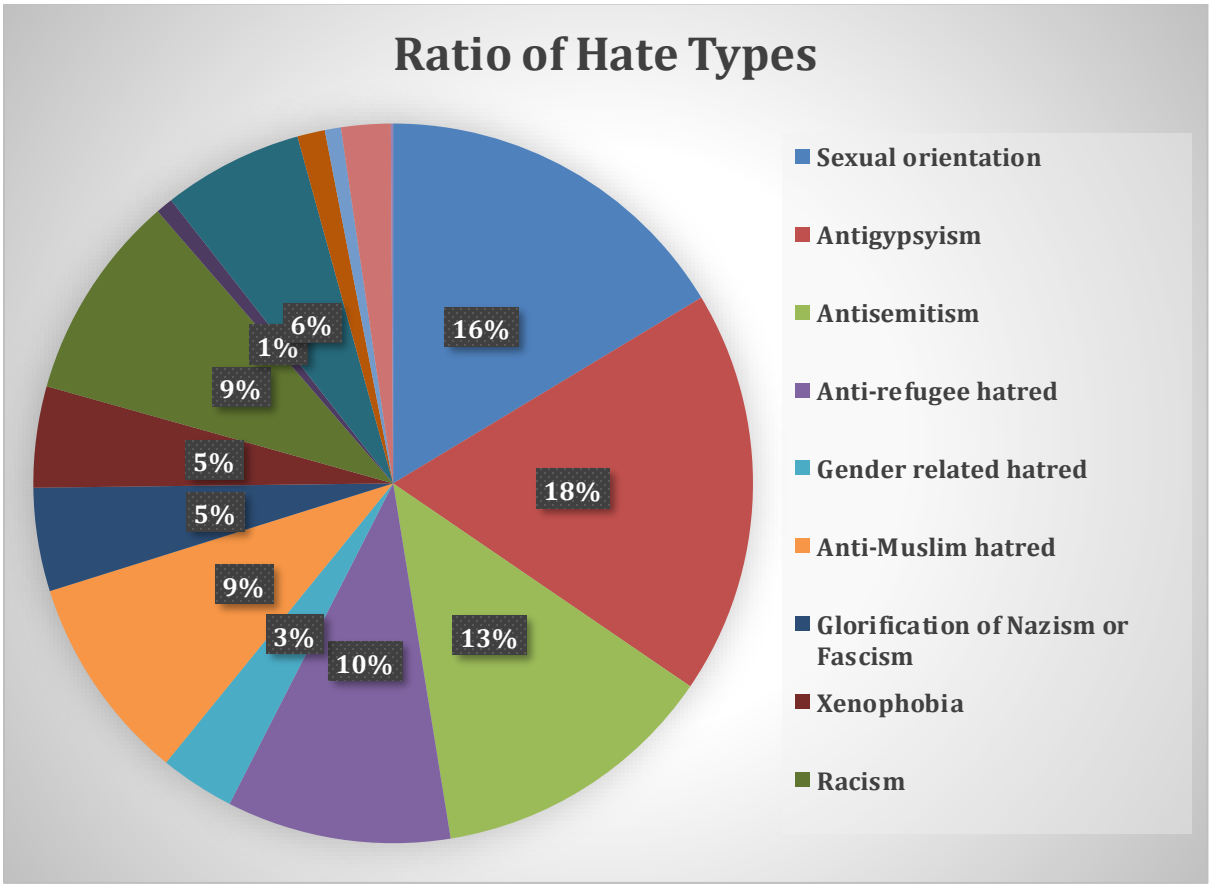
The data shows the generally lowest rates for the 48-hour assessment time. It is again most prevalent on TikTok with 10.46%, followed by Facebook with 7.68%, Twitter with 6.06%, Instagram with 3.85% and YouTube with 0%.

The Assessment Time Ratio within a week is 59.41% on Facebook, 46.15% on Instagram, 6.33% on Twitter, 5.23% on TikTok, and 0% on YouTube.

The rate that the reported content was never assessed ('no indication') was significantly highest on YouTube at 92.81%, followed by Twitter at 33.29%, TikTok at 11.76%, Facebook at 3.11% and Instagram at 0%.

Overall, the predominant assessment time for Facebook is one week, for TikTok, Twitter and Instagram it is 24h and for YouTube, it is most likely that there is no indication of an assessment. For Facebook and Instagram, 'no indication' was monitored the least.

3. Types of hate speech and intersectionality



The most prevalent *Types of Hate* during this Shadow ME were: anti gypsyism at 18.2%, hatred related to sexual orientation at 16.4%, and antisemitism at 12.9%. Anti-refugee hatred is 10.1%, while anti-Muslim hatred and racism are at 9.3%. This is followed by slightly less commonly found types: Anti-Arab racism at 6.3%, glorification of Nazism and fascism at 4.6% and xenophobia at 4.5% and 3.4% gender-related hatred.

It is important to note that the types of hate speech that organisations focus on depend on the objectives of the organisation in question, resulting in slightly skewed data.

The aspect of intersectionality must be taken into account here. Discrimination is not always clearly delineated and limited to one type within a content piece. While there may be a predominant form of discrimination or target, it can (subtly) overlap with other forms of hate and discrimination. For example, anti-Muslim hatred intersects with anti-migrant and anti-Arab hate or antigypsyism with women (sexism) and refugees.

4. IT platform performances and NGO observations

Generally, the removal rates for all 20 participating organisations are in the medium to low percentage range. The removal rates when reporting as a normal user are particularly disappointing. While three of the five platforms are still in the lower forty per cent removal rate range (the exception is YouTube with a removal rate of less than eleven per cent), no platform removed at least half of the reported content. Solemnly TikTok almost managed to do so. This improves slightly when the reporting is done via the Trusted Flagger channels. Sometimes, cases were only removed via the Trusted Flagger channels after having been rejected through normal user reporting, leading the organisations to highlight the increased removal rate via the Trusted Flagger channels. Despite rejections also via these channels and the subsequent need to contact the platforms directly, the monitoring findings show that what is reported via the Trusted Flagger channels works better regarding response time and higher removal chances, emphasising the importance and increasing positive impact of this system.

When looking at the communication with platforms there seems to be a lack of consistency in terms of the time needed to respond. There is a general tendency to assess reportings either quickly (within 24h) or slowly (one week). However, this differs per platform. Considering, for instance, Facebook's high *Normal User Feedback Rate* (95%) – despite its *Normal User Removal Rate* of 37.4% removed content – the platform needed up to a week for the assessment almost sixty per cent of the time or just 24h (29.80%). YouTube in general did not give feedback to any organization and took either 24h (7.19%)

for the assessment of the content (only 10.2% was removed), but mostly gave no indication about the assessment at all (92.81%). This makes it impossible to assess whether the reported content is reviewed or once the content is taken down, to assess whether it was due to the issued reporting, other reporting, or other reasons.

It was noted that the IT companies most often did not explain why the reported content was not considered illegal and therefore not removed. In other cases, the content was removed but no notification was given, and the content was not reported as a breach of security after it was reported. Sometimes reminders including screenshots had to be sent to the IT company contact. Issues occurring also included appeals against decisions that were lodged, but the content was still not removed on the same grounds, or there was no possibility to report content as illegal under specific legislation directly from the website, only via a specific form.

Generally, the communication experiences differ between platforms and monitoring organisations (countries), but in sum, it remains challenging, and not enough, including conflicting feedback and differing issues. The decision to remove content appears random and potentially automated at times (suspicion of increasing use of algorithms/AI to make initial assessments of reported content, as there have been cases where a negative response was received within seconds). Also, there seems to be no clear pattern as some content is removed, some is not – despite severe hate expressions or clearly illegal content – or only after increased effort, by escalating or reaching out to a contact person. Thus, the communication by the IT companies must improve regarding the assessment and removal time but also concerning explanations as to why some reported content was not considered illegal and removed and general feedback such as receipt confirmation.

Sometimes it was communicated that the content is restricted in one country without any explanation on why this particular case is excluded from being removed everywhere.

Platforms can restrict content only in the country where it is reported – 'geoblocking.' Some reported content was geo-blocked, however, there is no recognisable pattern as to why certain content is geo-blocked and why other content is not. Those organisations who noticed geoblocking mostly found it to be occurring on Twitter (and when reported as illegal under EU law), or also on TikTok. Other platforms might not have mentioned that option when the content was confirmed to be removed. Geo-blocked content also remained visible within the respective country in some cases with it only being labelled as limited visibility, and in another instance, the content was geoblocked within the country although the content was in English.

Among the most prevalent types of hate speech observed by the NGOs were hatred related to sexual orientation, antisemitism, anti-refugee hatred, anti-Muslim hatred, racism and anti-gypsyism.

Hate speech can be rooted in historically-based discrimination and stereotypes, influenced by current events and the general increasing polarisation and tensions in society. 2024 was a super-election year with many national elections in European countries and the European Parliament's elections, where right-wing parties gained ground, amplifying the existing and persistent hatred against certain groups in society. The rise of right-wing populist governments and parties in Europe, e.g. in Italy, Austria, the Netherlands, Germany, the Czech Republic, Hungary and Slovakia, which are often prominent on social media and spread and normalise hatred and fear against immigrants and refugees or other sexual orientations and gender concepts, is reflected in the hatred observed online and can also be seen transferring to the streets. For example, the riots in Amsterdam around a soccer match between an Israeli and a Dutch team, where a cocktail of antisemitism and anger over the war in Palestine, Israel and other countries in the Middle East was taken to the streets, or in Poland, where the so-called civic patrols against migrants were organised by far-right supporters. The matter of migration (and 'remigration') as a hot topic for right-wing populist parties stirs hate against, refugees,

and the Muslims and Arab communities, spreads false and misleading information, instigates fear and hostility and generates sensationalistic content, likely amplifying hate against these groups. Moreover, conflicts such as the one in Gaza or the war in Ukraine reinforce hate against marginalised social groups, causing increased anti-Muslim and antisemitism hatred as well as ethnic hate speech against Ukrainians living in Europe. But also, the 2024 Olympics, with its opening ceremony that included a drag performance, and competitive sports in general were seen as a source of hatred against members of the LGBTQIA+ community, particularly (presumably) trans athletes/people.

All of these societal issues and tensions seem to shift what is sayable. Reported content called for limiting rights, bullying, harassment, violent acts, physical harm, murder, annihilation, arson attacks, extermination, genocide, and expulsion from the respective country of certain groups or individuals. Degrading and defamatory expressions or words and overgeneralisations ultimately aimed at insulting and dehumanising (e.g. comparisons to animals) the targeted group or individual through language and (moving) image were found to be used. Furthermore, glorification denying or grossly trivialising historical events, such as the Holocaust were found.

Among the hate types overlapping were anti-Muslim and antisemitism, antisemitism and xenophobia, Anti-Arab, and Anti-Migrant/refugee, also combined with Anti-Muslim hatred (most prevalent). Gender-related hatred and Anti-Muslim hatred, Antisemitism (including Holocaust denial or revisionism) and Glorification of Nazism or Fascism, and Antigypsyism and Racism, or Anti-Gypsyism and sexism, specifically targeting Roma women, were observed to intersect, as well as racist hate and sexual orientation. In some cases, sexual orientation and gender-related hatred intersected, while in others hate speech was expressed against several ethnic groups within one content piece/comment - such as Ukrainians and Jews, or Ukrainians and Roma.

This highlights the aforementioned importance of considering intersectionality, as many cases do not only target one group or individuals but merge different types of hate.

Compared to last year's ME, it can be observed that the removal rates for the *Normal User Removal Rates* have improved for Facebook, Twitter, and Instagram, but went down a few percentage points for YouTube and TikTok. In contrast, the amount of content removed for *Trusted Flagger Removal Rates* only improved on Twitter, while it decreased by a few percentage points on YouTube and TikTok and fell slightly on Facebook and Instagram. Thus, while the *Trusted Flagger Removal Rates* 2024 appear significantly better than the *Normal User Removal Rates* 2024, they have not significantly improved compared to last year's *Trusted Flagger Removal Rates*.
As far as *Normal User Feedback Rates* are concerned, giving feedback has improved dramatically on Facebook in particular, but also on Instagram. Twitter and TikTok have also seen an improvement, while YouTube's feedback rate has dropped from an already extremely low percentage rate last year to zero per cent this year.

Comparing the *Assessment Time Ratios*, Facebook's assessment time has shifted from predominantly 48h in 2023 to either 24h or (in most cases) a week. Instagram's assessment time has shifted from predominantly 'no indication' (most cases) or 48h in 2023, to mostly assessing within 24h or a week and 0% 'no indication' rate – and has thus improved significantly. While TikTok and Twitter have only seen minor changes, – both have fortunately increased their assessment time within 24h. Strikingly, YouTube has unfortunately seen a further increase in the 'no indication' rate this year, while the assessment time frames of 48h and one week have disappeared completely and the 24h assessment time has only increased slightly (less than one per cent).

Again, this year, the three most commonly reported types of hate were antigypsyism, antisemitism and hate related to sexual orientation, with the only difference being that this year antigypsyism overtook hate related to sexual orientation as the most commonly

found type of hate. While hatred related to sexual orientation decreased more noticeably, antigypsyism and antisemitism subsequently increased. In addition, the numbers for anti-refugee hatred, anti-Muslim hatred and racism rose compared to last year.

Thus, this year, participating organizations noticed little to medium removal rates, and mixed communication with IT companies and, similar to last year, some uncertainty about knowing whether content was removed or not remains. Overall, YouTube performed very poorly in all analysing factors. Interestingly, last year's ME could report that in 2022 YouTube was one of the best-performing platforms with a very high removal rate.