INACH

Bringing the Online In Line with Human Rights

licra

INACH Monitoring Report 2021

Compiled by
Maia Feijoo and Tamás Berecz
2021

TABLE OF CONTENTS

1) BASIC INFORMATION ABOUT THE MONITORING EXERCISE 1

2) FINDINGS OF THE MONITORING EXERCISE 1

3) INTERSECTIONALITY IN THE FINDINGS 3

4) DISCREPANCIES IN THE PERFORMANCE OF THE PLATFORMS 3

5) INTERNAL EVALUATION AND HOW TO IMPROVE THE MONITORING PROCESS 3

6) CONCLUSION 4

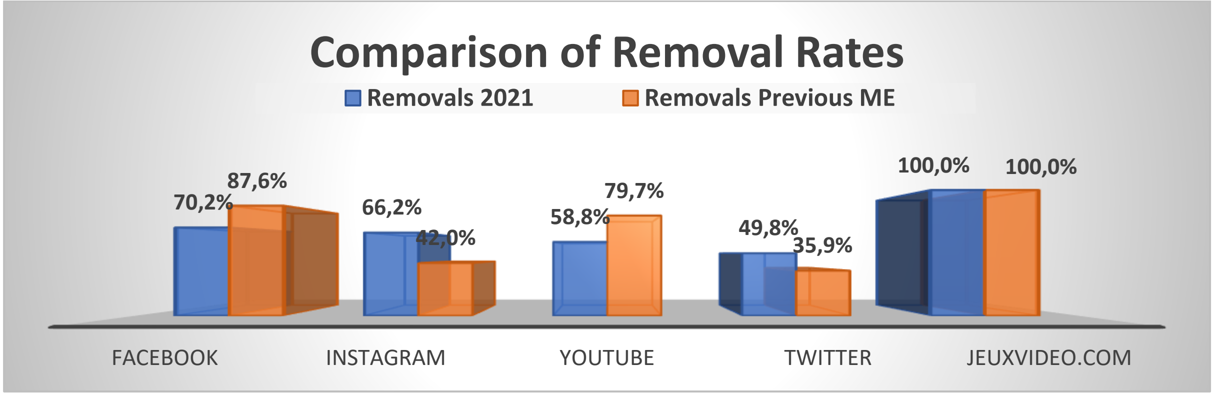
Therefore, INACH, Licra and our partners have three main policy suggestions to the platforms: 5

1) Basic Information about the Monitoring Exercise

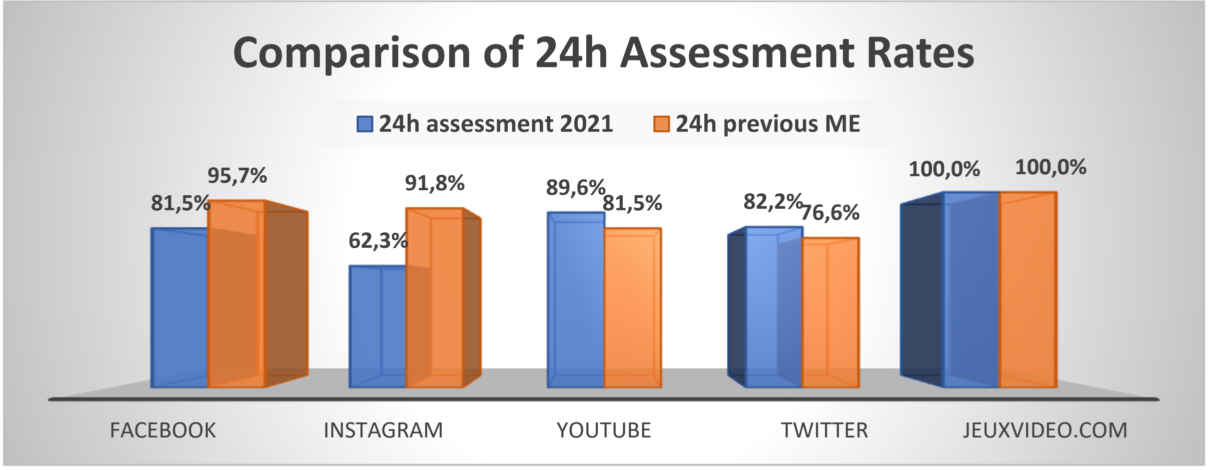
The sixth monitoring exercise has been coordinated by INACH and LICRA and has gathered information from 19 NGO's coming from different European countries. The following NGOs took part in the exercise: INACH, LICRA, Fundacion Secretariado Gitano, Estonian Human Rights Center, ILGA, Jugendschutz.net, Latvian Center for Human Rights, Institutet for juridic och internet, CESIE, ZARA, Greek Helsinki Monitor, LGL, Human Rights House of Zagreb, Integro Association, Never Again Association, ROMEA, ActiveWatch, Digitalna Inteligencia and Háttér Society.
The exercise begun on the 1st of March and ended on the 9th of April 2021, a period of 6 weeks to monitor IT companies and verify if they are respecting the Code of Conduct on Countering Illegal Hate Speech Online. A total of 6 social media has been monitored by the NGOs: Facebook, Twitter, YouTube, Instagram, Tik Tok and Jeuxvideo.com. In total, the NGOs reported more than 3500 cases of hate speech.

2) Findings of the Monitoring Exercise

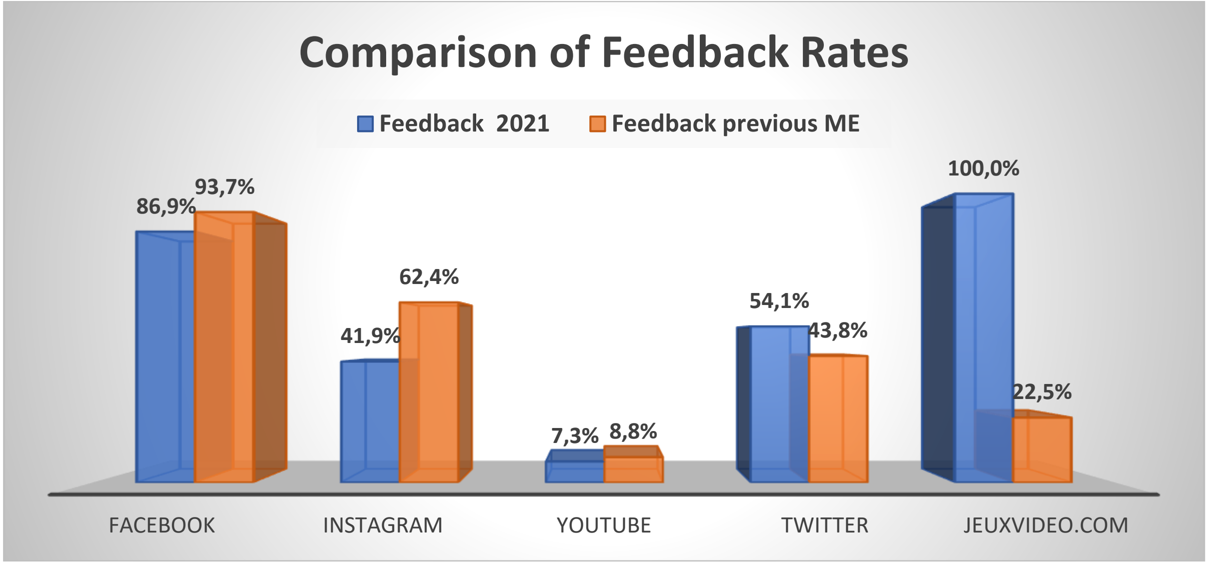
We observe that the results of this monitoring exercise highlight that the IT companies have worse results than the previous year. The removal rate has reduced from 71% to 62,5%.



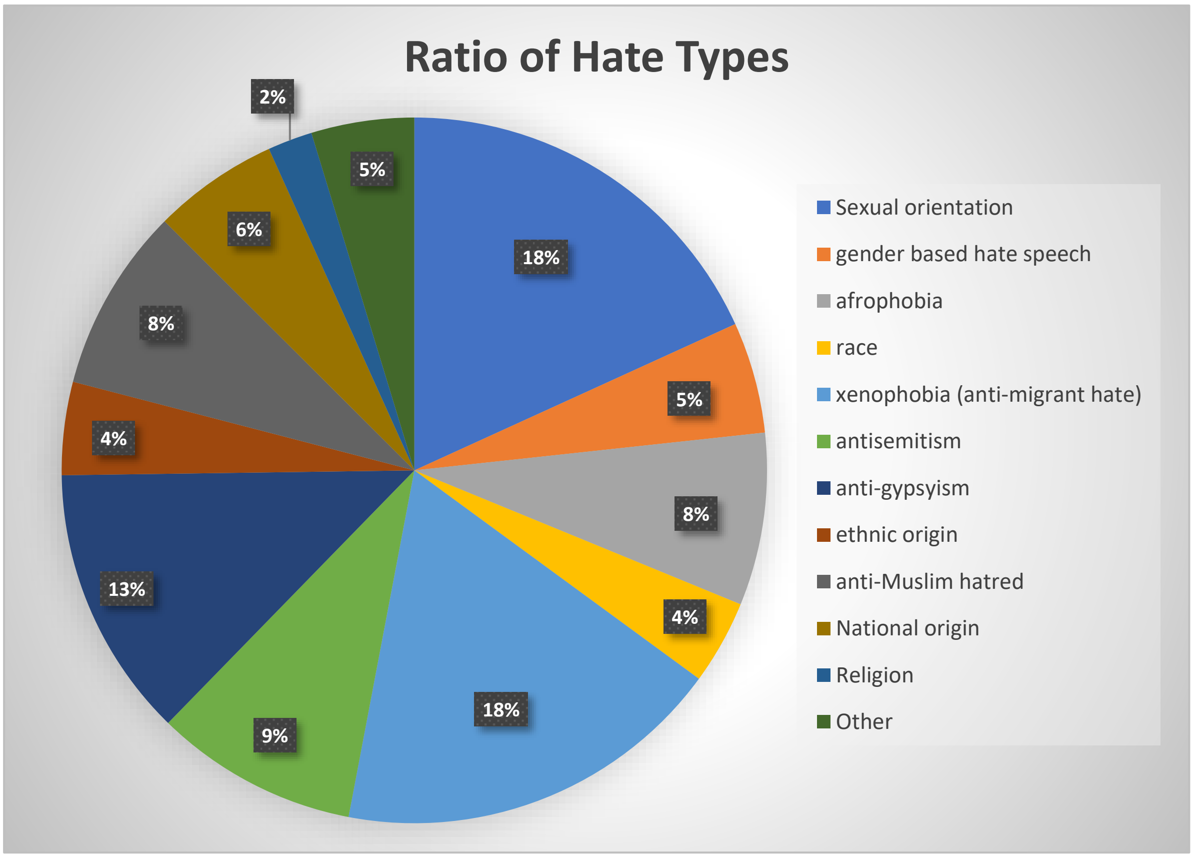
The respect of the 24 hours assessment has reduced, from 90,4% to 81%.



The percentage of feedbacks from the IT companies has also reduced, from 67,1% to 59,9% this year. There is a serious lack of consideration from them according to the results of our exercise.



Regarding the type of hate speech, during this year's monitoring, NGOs found that the most prevalent hate types were hate based on sexual orientation, anti-refugee hate and anti-Gypsyism.



3) Intersectionality in the Findings

For the last past years, the European Commission has been focusing on research about intersectionality, which is a notion used in sociology and political reflexion, referring to the situation of a person suffering different forms of discrimination at the same time. According to the European Institute for Gender Equality, intersectionality is an "analytical tool for studying, understanding and responding to the ways in which sex and gender intersect with other personal characteristics / identities, and how these intersections contribute to unique experiences of discrimination". During this monitoring exercise, we took into consideration the notion of intersectionality in our research and analysis of online hate speech. We noticed that intersectionality has been found by the NGOs, especially in gender-based hate speech, Afrophobia, hate speech against racism and origins.

4) Discrepancies in the Performance of the Platforms

During this monitoring exercise, we noticed a difference of notifications, assessment, and removal rate between the IT companies. In general, Facebook has been the most efficient IT company, since it has been responding to all the reports[1], it has sent notifications and removed most of the reported hate speech. Twitter had also quiet good rates of removal and notifications. Instagram and YouTube are following with fewer rates of notifications and removal.

We must highlight the fact that there are big differences between countries when it comes to notifications, 24 hours assessment and removal rate. The example of Facebook is revealing, it has been the "best" IT company in most countries, but in Greece, the results in terms of notifications, assessment, and removal of hate speech are poor.

We also must underline the fact that there is a huge difference of treatment between a normal user report and a trusted flagger one. All the feedbacks from the 18 NGOs that took part in the monitoring exercise have the same conclusion: the IT companies are favouring the reports of hate speech made by trusted flagger users. This is showing a lack of means from the platforms. It also disadvantages the consideration of normal user reports in their moderation of hate speech.

5) Internal Evaluation and How to Improve the Monitoring Process

INACH did not only collect data on the reported content, removal rates/times and the prevalence of certain hate types. We also collected information from our participating members and expert NGOs about the preparation processes the preceded the main monitoring exercise and on their expert opinion about how the ME methodology could be enhanced, made more representative or include other foci points besides removal and the timeliness of removal.

INACH Secretariat and Licra, the two coordinating partners of the monitoring efforts held two trainings before the main monitoring exercise. These trainings focused on the methodology of the ME, i.e., how long the monitoring period would be, how to record the data in our internal Excel sheets and the Commission's system, how to check the reported content and in what time increments these checks should be done. With

[1] Disclaimer: this report is talking about the general findings of this monitoring exercise, some particularities exist, and they are summed up in the Commission's report.

these trainings we aimed to help NGOs that were new to this type of monitoring and also to make the monitoring process as uniform as possible, so the collected data would be as comparable as it could be.
The responses to these preparatory trainings were very positive. Participants said that they received useful and sufficient information on all topics that were presented, and they did not have a lot of trouble carrying out their monitoring efforts based on the input they received beforehand. However, we also received some criticisms. These criticisms were mainly rooted in the fact that the circumstances of this ME were a little bit different than before. This was the first major ME that was carried out within a service contract with the EC, which slightly changed what we needed from the NGOs. Due to this, some expectations were communicated late towards the participants, which led to some misunderstandings and frustration. The Secretariat will learn from these criticisms, and we will put a much bigger emphasis on communicating the expectations towards our partners in a more timely and precise manner in the future.

Our participating members and partners also had valuable input on what could be changed in the monitoring process and/or methodology to make the findings more representative or just even more in-depth.
Some suggested that either the timeframe of the ME should be lengthened or the number of collected cases should be lowered. They argued that the lengthened timeframe would produce a more representative outcome or lowering the number of collected cases, which is right now a hundred per country, could allow for more time to discuss cases with the platforms and produce more fruitful discussions about the illegality of the reported content. These disputes then could help the platforms to train their moderators better and enhance their algorithms so there would be less over-, and under-removal than there is right now.

Multiple partners mentioned the communication by the companies towards the reporters and their feedback given to complaints, or rather the lack thereof. We discuss this issue elsewhere in this report, but we have to stress it here too. The companies are still doing an abysmal job in providing proper, or sometimes any, feedback to normal user reports, and quite often even to trusted flagger reports. INACH and our partners would like to stress how important proper feedback is and how pivotal it is for companies to give clear and concise information to complainants on why the content they reported was not removed. The lack of clear feedback discourages people from reporting content that might be illegal, since they do not understand why something is removed while other content stays online after multiple complaints, even though the cases seem very similar to one another. Thus, many of our partners suggested monitorings that are less focused on removal times and rates and more on the feedback and feedback times provided by the platforms. This would be even more important in cases of non-removal. Proper clear and concise feedback in cases that are not removed could be used to understand how the platforms and their algorithms work and in return, our experts could provide the platforms with advice on how they could improve their algorithms and the work of their moderators. It would also encourage members of the public to report content that they think is hate speech, since they would receive information that clearly explains why the content they reported was not removed.

6) Conclusion

In conclusion to the 2021 main monitoring exercise, we have to – sadly – make the observation that almost all major platforms did worse than in the previous main exercise. The removed less cases, they provided timely assessment in less cases and they provided feedback in less cases. There were a few exceptions in all these indicators. There were platforms that removed more reported content than before (Twitter and Instagram), platforms that provided 24-hour assessment in more cases than before (YouTube and Twitter) and there were platforms that provided feedback in more cases than before (Twitter and Jeuxvideo.com). A caveat has to be noted here: Twitter had a lot of room for improvement and their numbers, even though they are better than they used to be, are still far from good.
The platforms will most likely blame the issues and hardships caused by the pandemic for their slipping performance. We think that this explanation is a little bit lacking, and we hope that the platforms will reach their previous standards during the next monitoring round. We also hope that the findings of this main

monitoring and the shadow exercise that followed it will nudge the companies to do better the next time and in general.

To close this report, we want to focus on two main issues that can be seen from the numbers and the qualitative feedback we received from our experts carrying out the monitoring efforts. The first is insufficient feedback or the complete lack of it and the second is the difference between the assessment and removal rates of normal user and trusted flagger reports.

Hardly any platform provides clear and concise feedback to reporters, especially when it comes to non-removal. They usually just let the complainant know that the content did not breach their community guidelines, but there is no explanation as to why that is the case and most platforms' appellate systems are just as confusing. This clearly discourages reporting illegal hate speech to the platforms. Furthermore, it still clearly shows that the platforms rely too heavily on their own guidelines and not the local laws of the countries where the complaints are received from. The companies must do better in this field to maximise the encouragement of their users to report content to them. It is their paramount interest too to keep their platforms as clean of hate speech and disinformation as possible. They cannot do that without the reports and complaints of their users and expert NGOs.

The second highlighted issue is also an old one. From the very first monitoring in 2016, NGOs observed that there are major differences in removal rates when a certain piece of content is reported via the normal user channel or via the trusted flagger channels that are only open to hate speech experts. At first glance this seems logical and probably even the platforms would argue that this is a good thing. They utilise this trusted flagger system to harness the expertise of NGOs and thus, they take the reports coming through these channels more seriously. However, quite often content that does not get removed after a normal user report will get removed after it gets reported via a trusted flagger channel. This is simply unacceptable. Content should be removed because it is illegal, not because it was reported by an expert instead of an everyday user. A piece of content is either illegal hate speech or not, it is either in breach of the community guidelines or not. Hence, there should be basically no or very little difference between removal of normal user and trusted flagger reports.

Lastly, it has to be underscored that there are still major differences between removal and assessment rates in different countries on the same platforms. This also needs be remedied. A platform should adhere to the Code of Conduct to the same extent in all EU countries, i.e., its removal and assessment rates should be very similar everywhere within the Union.

Therefore, INACH, Licra and our partners have three main policy suggestions to the platforms:

1) Improved feedback to complaints.

2) Less differentiation between normal user and trusted flagger reports.

3) Harmonising their efforts in adhering to the Code of Conduct in all EU countries.