INACH

Bringing the Online In Line with Human Rights

licra

INACH Final Shadow Monitoring Report 2022

Compiled by Adinde Schoorl & Maia Feijoo
Edited by Tamás Berecz

2022

Table of Contents

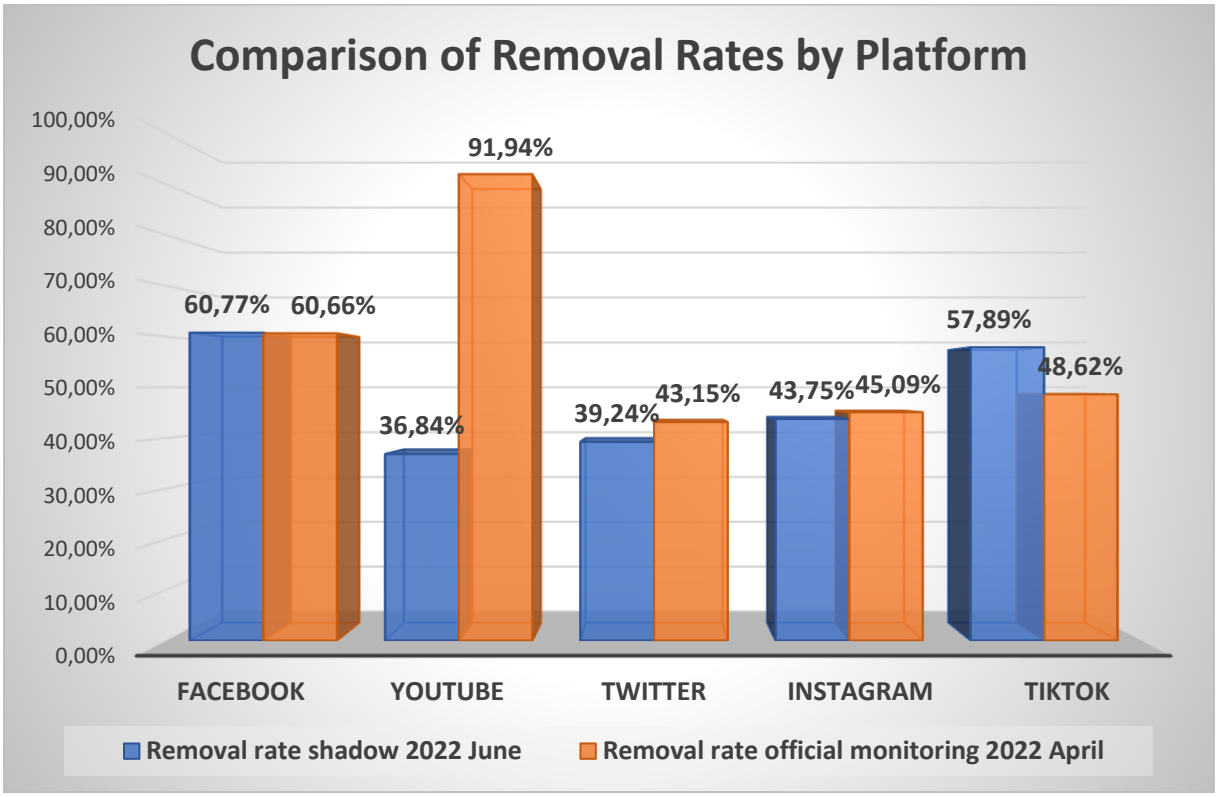
1) Introduction / Basic information about the Shadow Monitoring Exercise

The 2022 Shadow Monitoring Exercise, following the seventh Official Monitoring Exercise, has begun on the 23rd of May, and ended on the 10th of June, a period of three weeks where five European NGOs monitored online illegal hate speech and analysed how the IT platforms moderated their reports. The objective was to analyse how the platforms act when they don't know they are monitored, contrary to the official ME.

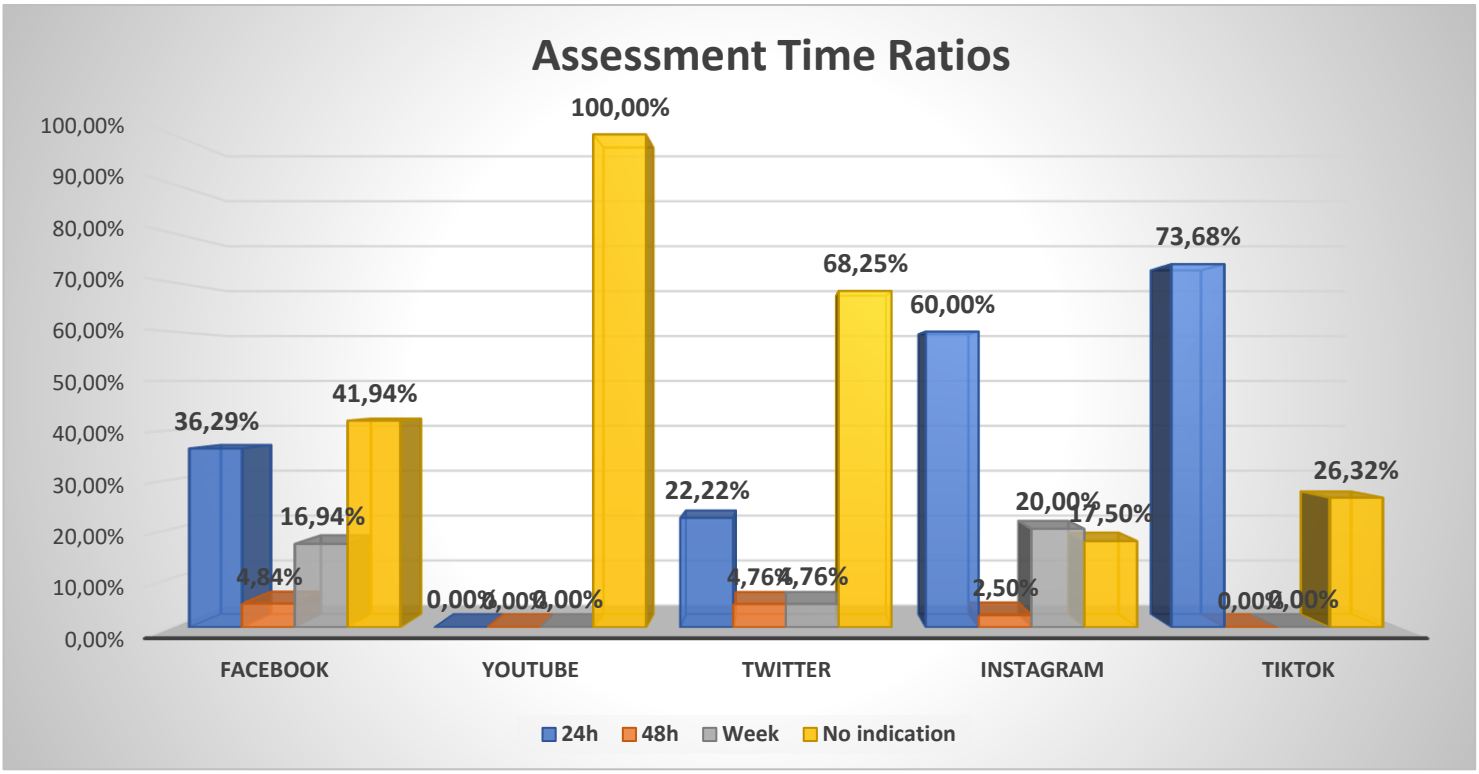
Active Watch (Romania), LICRA (France), CESIE (Italy), LCHR (Latvia) and INACH (Netherlands) took part in this exercise. Together, they monitored five social companies: Twitter, Facebook, Instagram, TikTok and YouTube and reported 269 cases.

2) Findings of the Shadow Monitoring Exercise

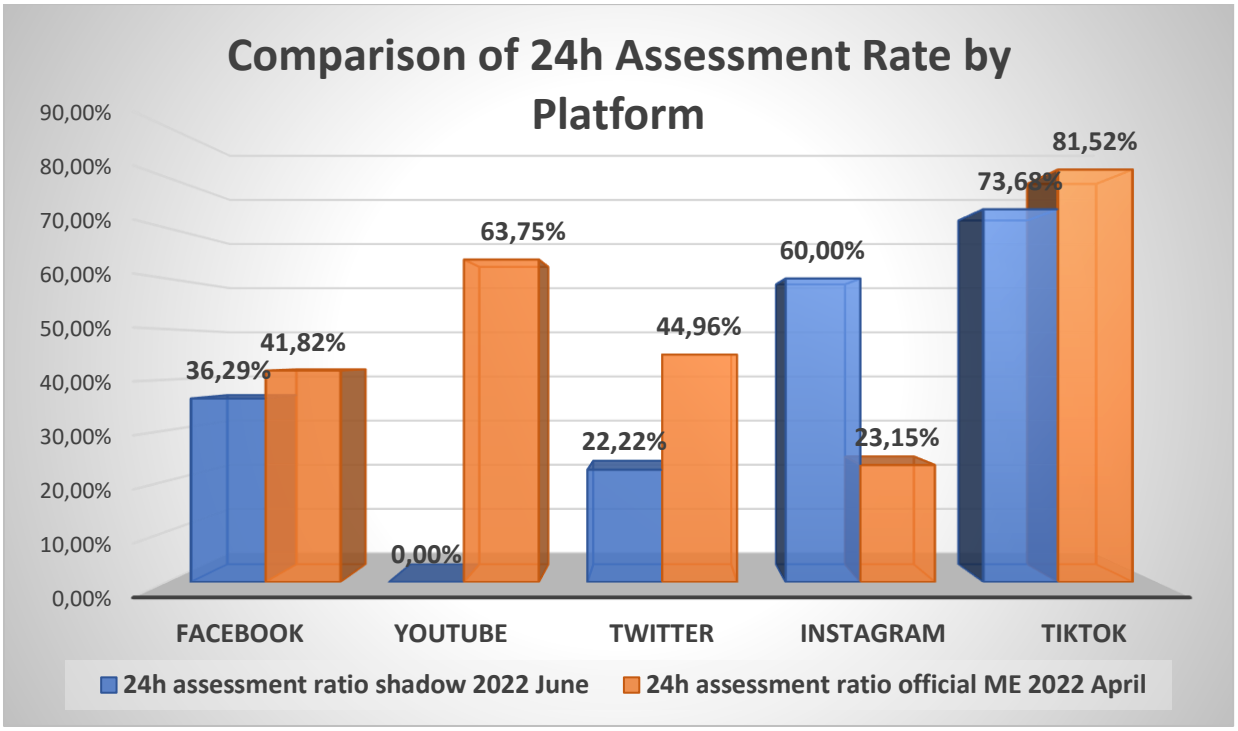
The results of this shadow monitoring exercise in terms of removal rate are quite similar to the official monitoring exercise that took place in April-May 2022. Facebook has almost exactly the same removal rate while the differences for Twitter, Instagram and TikTok are negligible between the shadow exercise and official exercise. The only outlier is YouTube. It had a removal rate of over 90% during the official monitoring exercise, while during the shadow exercise it only reached 36%.



The assessment time ratio shows very different results between all the platforms. TikTok has answered to more than 70% of the reported content within 24 hours. Interestingly enough, the remaining 26% content never received an indication, so it seems to be an 'all or nothing' approach for them. YouTube had a 100% no indication ratio. It means that not one piece of content with illegal hate speech reported to YouTube received an answer during this Shadow Monitoring round. Facebook had very different results. Even though almost half of the reported cases did not receive an answer at all, at least over one third of the reported cases received an assessment within 24 hours. Instagram has a 60% within 24 hours assessment ratio and only 17% did not receive an answer at all. With Twitter, almost 70% of the reported cases did not receive any indication at all, while 22% did receive an assessment within 24 hours.

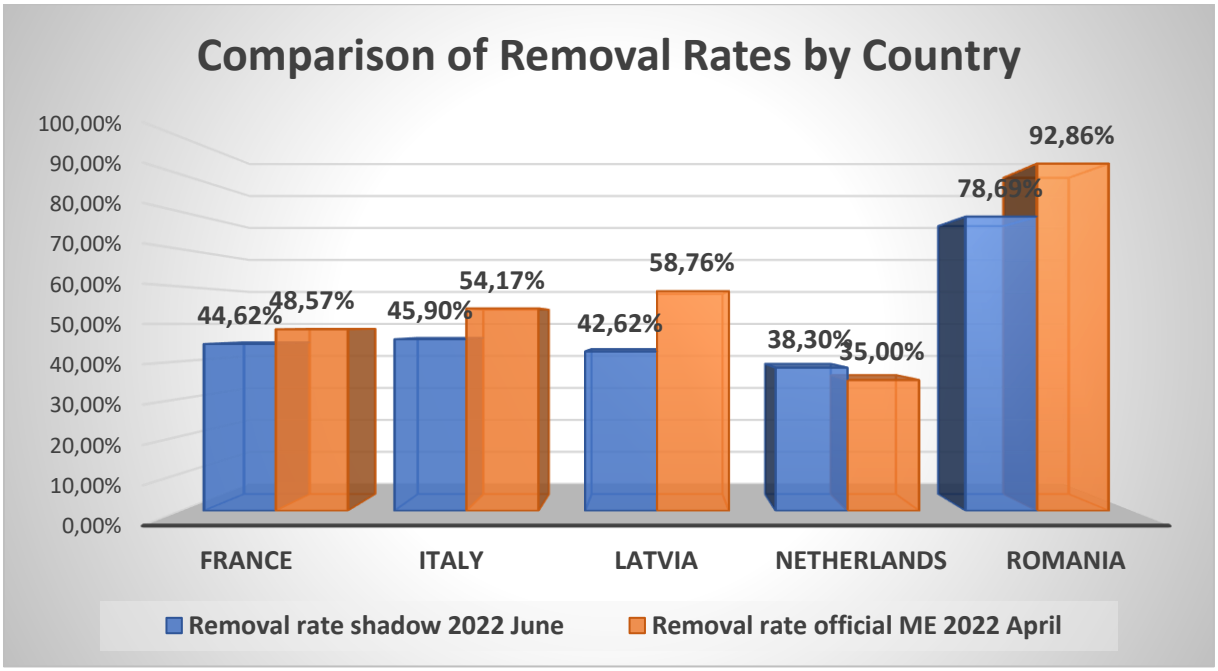


The differences between the official ME and the shadow ME are bigger when it comes to 24 hours assessment rates. Again, YouTube shows a huge difference: over 60% received an assessment within 24 hours during the Official Exercise while 0% received an indication within 24 hours during the Shadow Exercise. Facebook and TikTok show small differences between both rounds, but both have better results during the official ME than in the shadow. Instagram actually performed better during the shadow ME: 60% received an answer within 24 hours while that was much lower during the Monitoring Exercise. Twitter did perform better during the official Monitoring Exercise, with almost 45%, while only 22% received an answer within 24 hours during the Shadow Exercise.

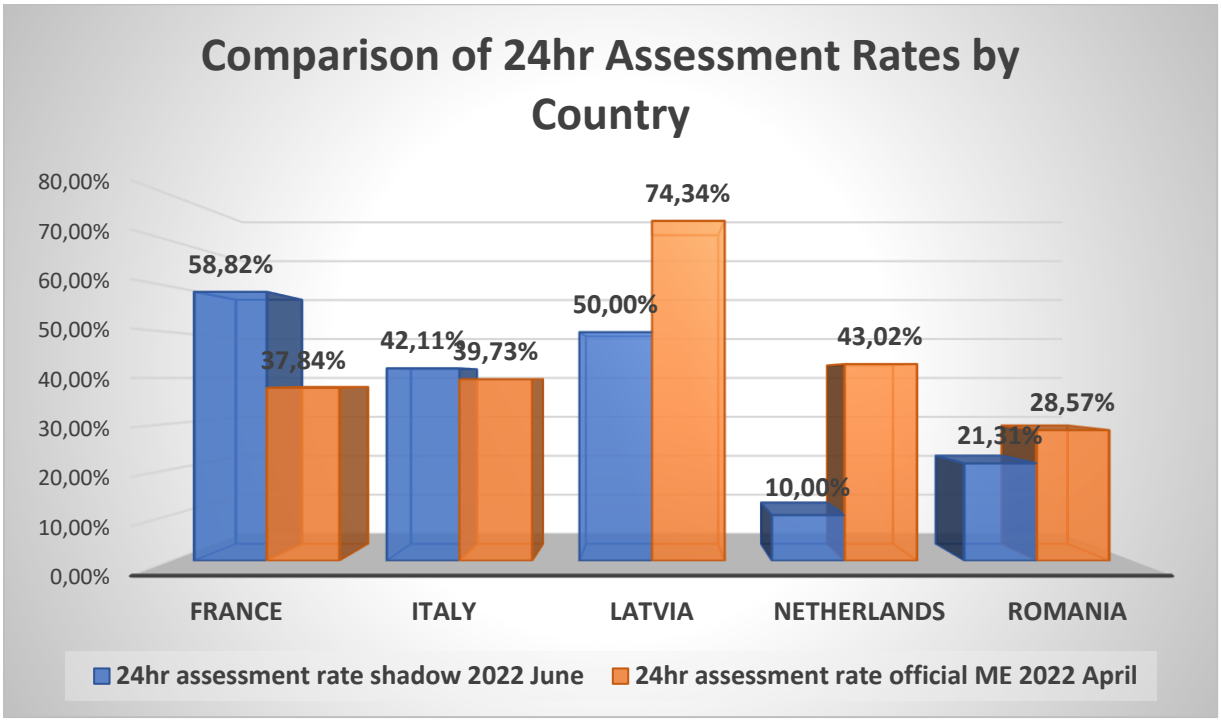


Below one can see the comparison of removal rates by country. One can compare between the official ME and the Shadow ME, but also between the different countries that participated in both exercises. For instance,

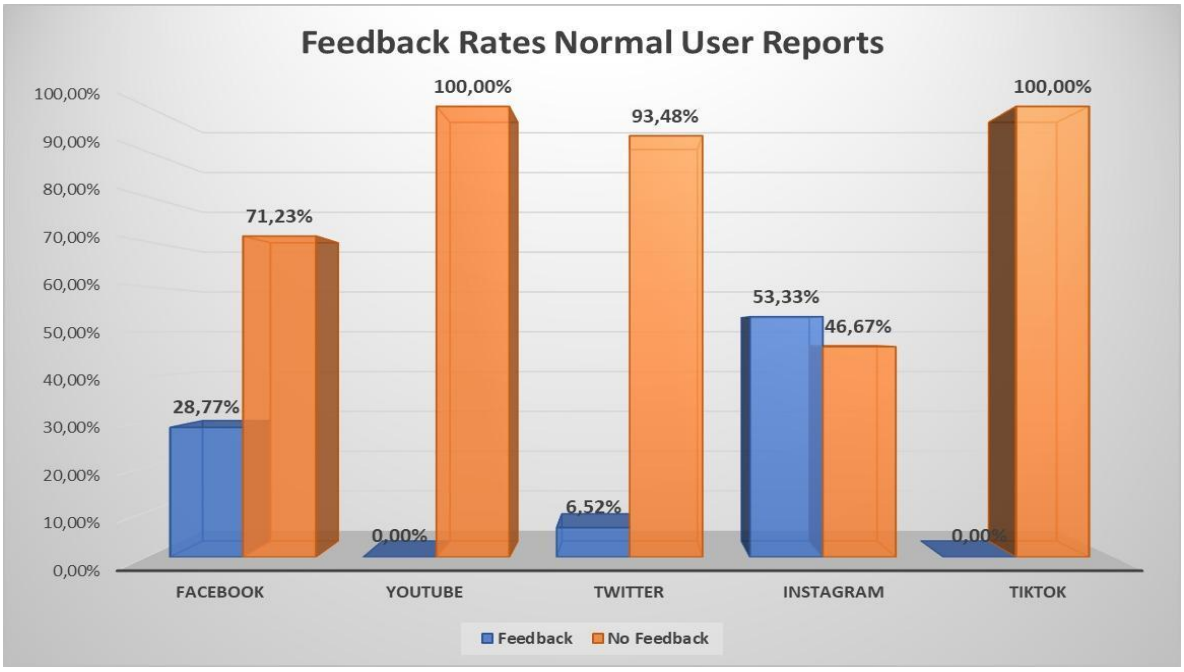
Romania has by far the best results of removal rates in both the official- and shadow round. The other countries show similar percentages although the Netherlands shows the lowest removal rate of all. All countries show quite similar results between both rounds with negligible differences.



Below one can see a comparison of the 24 hours assessment ratios between countries. Here, Latvia performs the best while, surprisingly, Romania performs quite bad. Thus, despite the fact that they have such a high removal rate, they do experience obstacles with assessing the content within 24 hours. Italy shows similar percentages and very small differences between the two rounds. The Netherlands has low results here as well: only 10% of the reported content received an answer within 24 hours during the shadow exercise while during the official ME they performed on a similar level as France and Italy. France is the only outlier insofar as the 24-hour assessment rate was higher by 20% than during the official ME. All other countries did worse during the shadow round, except for Italy, but the difference there is marginal.

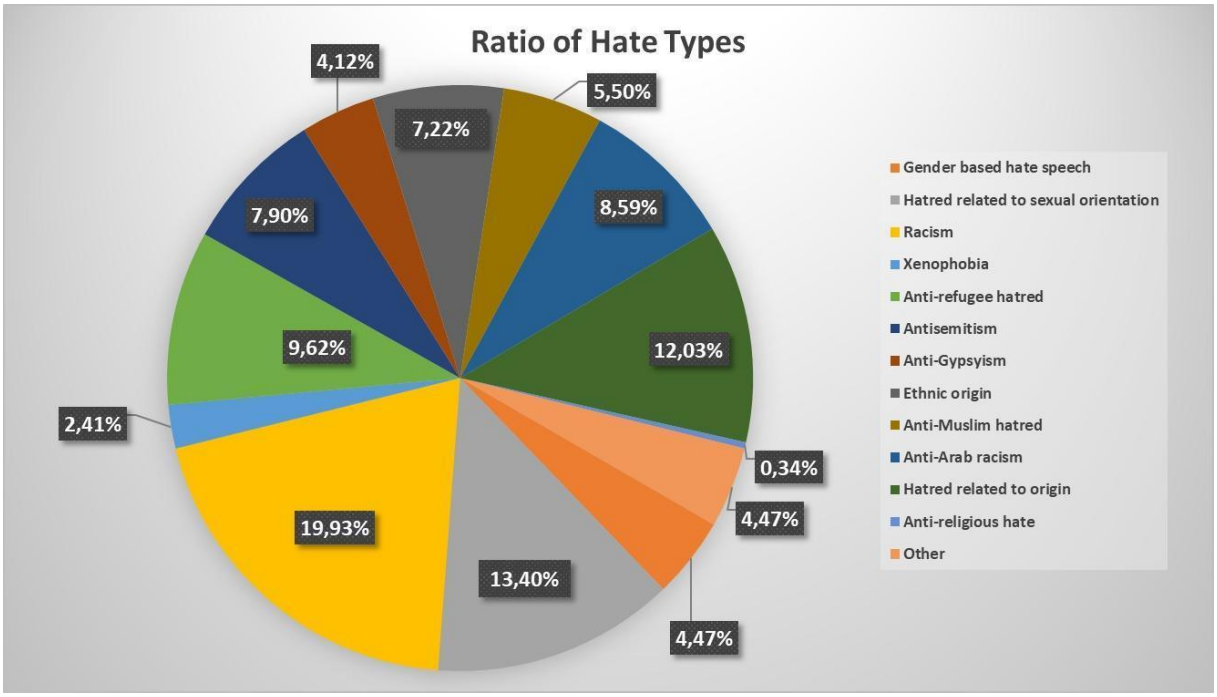


Finally, one can find below the Feedback Rates. One can see that both YouTube and TikTok score the worst here. Also, Twitter has a very low Feedback Rate of a bit over 6%. Facebook follows with almost 30% of a Feedback Rate. Instagram has the highest Feedback Rate 53%.



3) Types of hate and intersectionality

The most prevalent types of hate speech during this shadow monitoring exercise were racism, hatred related to sexual orientation, hatred related to origin and anti-refugee hatred. Cases of intersectionality between anti-refugee and anti-African hate speech has been noticed in France.



4) IT platforms performances and observations from NGOs

The five participating NGOs highlighted the fact that there are big differences between the removal and feedback rates between the countries. The feedback rate is low in Romania and in the Netherlands. In France, the country with the higher 24 hours assessment rate, it is still perceived as not enough, with a total of 58,82%. The NGOs had the feeling that the moderation is still unclear and that if the assessment rate is fast, the non-removal rate is higher.

The NGOs also noted some technical issues. They noticed the fact that when they copied the URL of a TikTok comment, it didn't take them back to the reported comment and they had to check every comment on a post. This made it almost impossible for them to check if the content reported was still online or not, especially if they did not receive any feedback from the platform. In addition, the NGOs could only copy the URL of a comment on the web browser version, and not on TikTok's application. It is important to highlight the fact that the same issue has been reported to the platform during the official monitoring exercise, and the contact of the platform said this incident was fixed. It seems that it wasn't the case several weeks after, during the shadow exercise.

INACH and LICRA also faced a technical issue with Instagram. In some cases, when they made a report as a normal user, they received an automatic response explaining that the reported content could not be monitored by the platform, due to a high number of reports they were receiving. This means that Instagram didn't review any content if it was not reported with a trusted flagger status.

5) Conclusion

This 2022 Shadow Monitoring Exercise was relevant because it has expressed that there are still some differences in the results when it is compared to the official monitoring exercise. The more notable is YouTube, that did not send any feedback during the shadow exercise. The NGOs also noticed that in some cases, IT platforms do not respect their engagement to solve an issue that has been raised during the official monitoring exercise. It has been the case with the URL comments with TikTok and with the lack of moderation on a content reported as normal user on Instagram.

It also needs to be noted that in every country monitored during the shadow round, the feedback rates plummeted on all platforms, except for Instagram. However, Instagram had the lowest feedback rates among the platforms during the official ME, so their improvement must be taken with a pinch of salt. The 24-hour assessment timeframe is also hardly adhered to by the companies. This is one of the most important caveats of the Code of Conduct, yet all platforms struggled to maintain a high level in this indicator. Instagram was the only platform that did mostly better in almost all indicators during the shadow ME. This suggests that most likely it has more capacity in the countries that participated in the shadow exercise than globally within the whole of the EU.