THE OUTLIERS: Addendum to INACH's joint monitoring report with the sCAN Project

By Tamás Berecz

Bringing the Online in Line with Human Rights

INACH

2019

Supported by the Rights,
Equality and Citizenship (REC) Programme
of the European Union

bpb: Bundeszentrale für politische Bildung

Legal Disclaimer

This publication has been produced with the financial support of the Rights, Equality and Citizenship (REC) 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.

The INACH Secretariat would like to thank the sCAN Project partners for their tireless work during the joint monitoring exercise and the writing of the monitoring report.

We would also like to thank the following INACH members that contributed to this addendum by participating in the monitoring without receiving any funding or other financial help:

Without their participation, INACH would never have been able to shine a light on the weaknesses of the official monitoring exercises and we would not have been able to highlight the major discrepancies in the attitudes of social media companies to the removal of illegal hate speech in different countries.

Table of Contents

1. Introduction

During the summer of 2019 INACH and the sCAN Project (a project that is almost solely run by INACH members) ran an unannounced, silent monitoring exercise to check the adherence of social media companies to the Code of Conduct (CoC) that they had signed with the European Commission (EC) in 2016. In this CoC, Facebook, Twitter, Google and Microsoft signed that they would provide substantial and speedy feedback to complaints about illegal content, namely online hate speech, and they would remove said content within 24 hours if they indeed found it illegal or against their community standards.

The EC initiated the first monitoring exercise (ME) to check the adherence of the companies to the CoC in late 2016. Since then, 4 MEs have been held and the 5th is underway right now in November-December 2019. These official EC-run exercises have proven very useful in keeping the companies' promises in check and they do show improvements in the field of hate speech removal on the major platforms, such as Facebook. However, the NGOs that actually carry out the ME – collect cases, report them to the companies, record their responses and the removal or non-removal of reported cases, etc. – have voiced many concerns and criticisms about the methodology of these MEs. Some of these criticisms and concerns can be found in our first and second policy papers, which were published in 2017 and 2019 respectively. That is why INACH Secretariat and several INACH members thought it important to run MEs that are independent of the EC-run ones and run them silently, i.e. without announcing them to the social media companies.

The findings of our independent ME were fascinating and, even though they were not completely different to the findings of the official MEs, they did unearth some issues that the official ones could not.

INACH Secretariat, within its Framework Partnership Agreement (FPA) with the EC, and the sCAN project jointly authored a report and published the findings of this unannounced ME in September 2019, which we summarised the following way:

"*Although the overall removal rate of 70,6 % turned out to be only slightly lower compared to the last monitoring (-1,1 percentage points), this result is mostly owed to Facebook's consistently high removal rate of 84,5% (+0,9 percentage points) and Instagram's improvement to 77,2% (+6,6 percentage points). Twitter's performance remained low at 44,9% (+1,4 percentage points), and YouTube only removed 67,8% of illegal hate content, a major drop of 17,6 percent points compared to its last checked performance.*

This monitoring exercise, which has been coordinated and conducted by the International Network Against Cyber Hate (INACH) and its partners of the project sCAN (Platforms, Experts, Tools: Specialised Cyber Activists Network) to check the compliance of social media platforms with the European Commission's Code of Conduct on Countering Illegal Hate Speech, was the first one in which the platforms have not been aware of the monitoring.

Between May, 6th and June, 21th , 12 organizations which are specialized in dealing with online hate have reported 432 cases to the platforms through their public reporting channels, out of which 90 have been re-reported through reporting channels available to organisations recognized by the IT companies as "trusted flaggers" after having been rejected by the platforms.

With the EU Code of Conduct, the companies have agreed to assess and remove illegal hate speech online that is against national law or their Terms of Services within 24 hours. Yet, only Facebook managed to reach a tolerable level in removing reported hate speech within that timeframe (64%). Instagram, Twitter and YouTube remained below 50%.

In addition, the companies' performance in providing feedback was poor: to 42% of reports the companies provided absolutely no feedback, reactions within the required 24 hours came to not even half of reports (46%). Again, only Facebook provided timely feedback to 70% of reports while YouTube remained silent to 97% of reports.

Providing no feedback, late feedback or meaningless feedback is a major issue that needs to be addressed by the companies as soon as possible. If people report online content that is hateful, discriminatory or inciting violence, it is not enough for platforms to send an automated reply stating that they received the report, or not even that. "Users need to know that their efforts in making the internet a friendlier place are taken seriously so they feel encouraged and valued" emphasises Ronald Eissens, General Director of INACH.

Hence, INACH, together with the partners of the sCAN project and other member organizations that have participated in the monitoring, urges the social media companies, especially YouTube, Instagram and Twitter, to improve their removal practices further and to react and respond meaningfully to all users reporting hateful content."[1]

Nine INACH members participated in this ME within the framework of the sCAN project. However, thanks to the coordinating efforts of the Secretariat, three more INACH members joined the effort that were not funded by sCAN. These were the following: digiQ from Slovakia, the Greek Helsinki Monitor (one of the newest members of INACH) and the Never Again Association from Poland. Their work and contribution were thanked in our joint report that was published in September, but their findings were not included in that paper. The reason for that was twofold. One: the ME report was cowritten by the Secretariat and the sCAN Project, thus

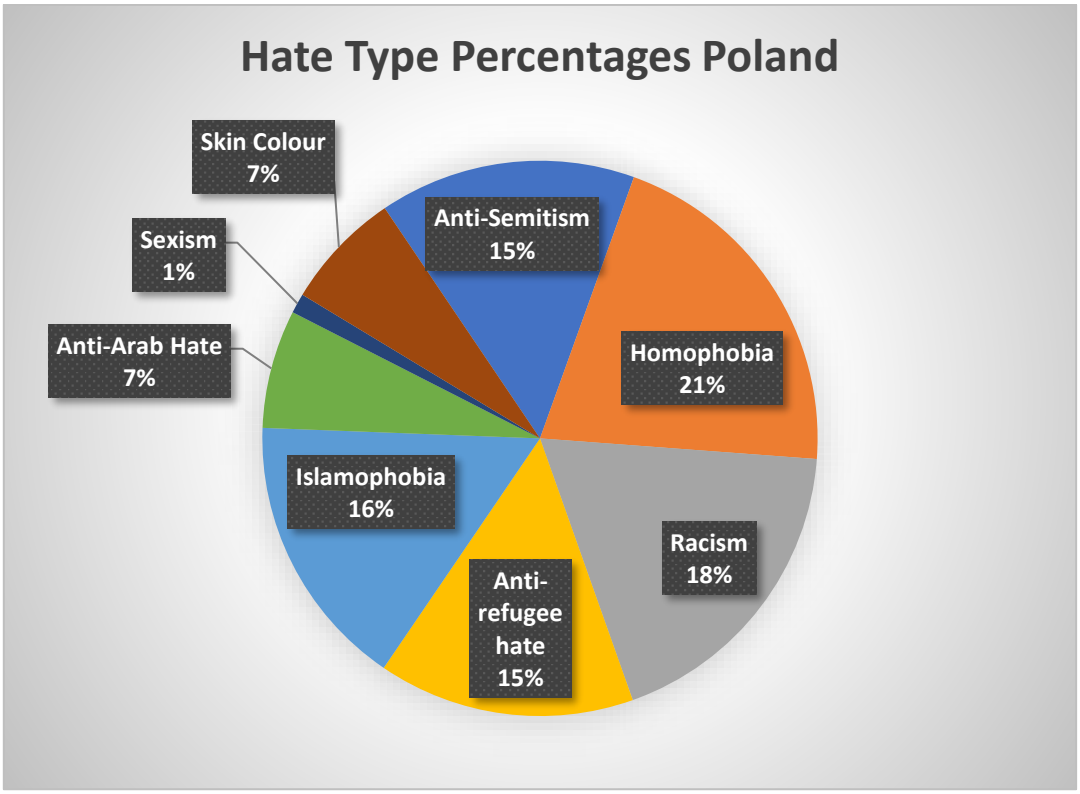
[1] http://www.inach.net/wp-content/uploads/INACH_sCAN_ME_press_release_12092019_fin-002.pdf

its production was funded within the framework of the FPA and the sCAN Project. Hence, it was decided not to include data from INACH members that are not funded through these channels. Two: the findings of these NGOs, especially in Greece and Poland were so interesting and troubling that we did not want them to be lost by mixing them in with the findings from the other countries (France, Italy, Belgium, Austria, Germany, Croatia, Czech Republic, Slovenia and Latvia). So, let us take a look at these troubling findings from Greece and Poland.

2. The Outliers

2.1 Poland

The findings of the Never Again Association in Poland are probably the most troubling that the Secretariat has ever seen since the MEs were started by the EC. Our member reported 52 cases during our unannounced exercise. They decided to only report to Facebook, a company that has shown the most improvement in the field of cyber hate removal since 2016 and produced the best results in all other monitored countries. The Never Again Association paid attention to report cases that were a good presentation of the state of cyber hate within the country. That is why they reported to Facebook cases that fell within eight different hate types.



21 per cent of cases they reported were homophobic, 18 per cent were racist, 16 per cent were Islamophobic that was followed by antisemitism and anti-refugee hate both with a 15 per cent share of the cases. **According our Polish colleague, who worked on the ME:**

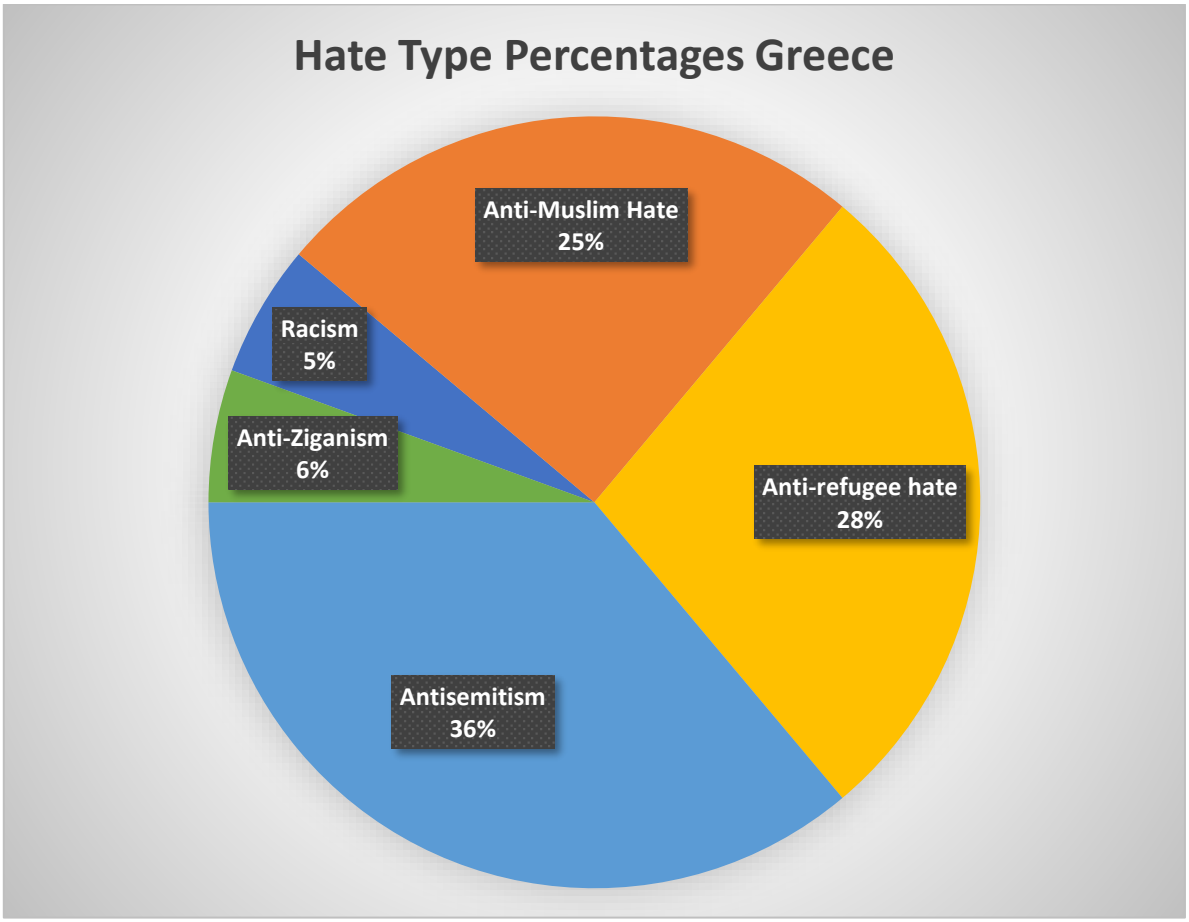
"*I personally reported to Facebook around 50 hateful comments. I tried to focus on different kinds of hate (antisemitism, anti-refugee and anti-Muslim, homophobic, racism). Many of those comments or materials encouraged directly to violence, incited hate and contained glorification of war crimes or murders. None of the comments was removed. I never received any reply from Facebook following my reporting. They completely ignored the reports, all of those hateful comments and materials are still available publicly on different Facebook pages. This is how situation looks in Poland when it comes to reporting hate speech to Facebook. I do hope that some mechanisms are implemented that will make it more effective, because hate and direct incitement to violence is dangerous and has an impact on what happens on our streets*."

This is such a major discrepancy and difference compared to our findings in all other countries that the sCAN Project members monitored, that it is almost hard to believe. However, according Rafal Pankowski, co-founder of Never Again, states that – for them – the outcome of the ME was not a huge surprise. In his opinion, the Polish government has been actively pressuring Facebook not to remove extremist content from its platform, so much so that the Polish government established an extra option to appeal on a ministerial level where people can turn if the company blocks or deletes content they posted.[2] In such a political climate, it is indeed unsurprising that Facebook did not delete anything that our reported during our unannounced ME.

[2] https://www.wprost.pl/kraj/10186022/andruszkiewicz-chwali-sie-sukcesem-resortu-polska-jest-pierwszym-krajem.html

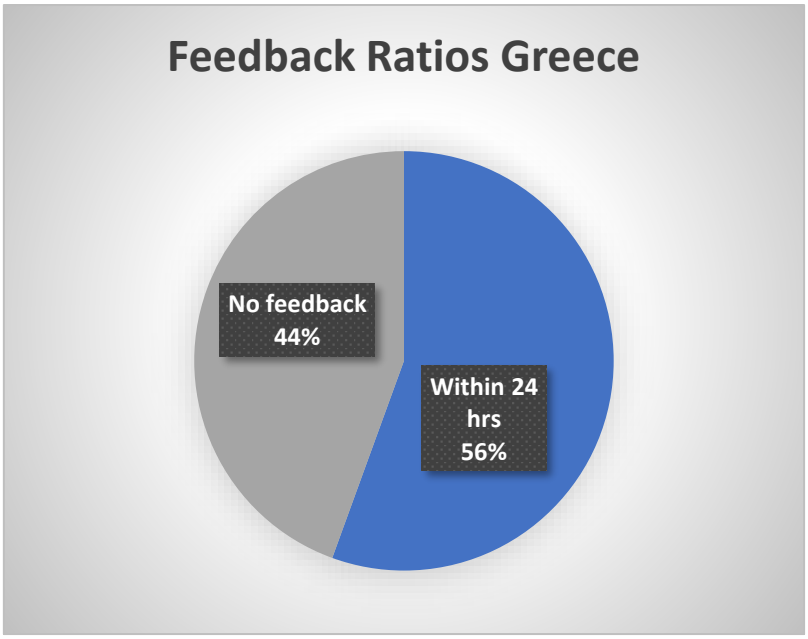
2.2 Greece

The Greek Helsinki Monitor (GHM) reported cases to all three major social media companies: Facebook, Twitter and YouTube. Altogether they sent in complaints about 36 instances of cyber hate to these companies. 20 to Facebook, 8 to Twitter and 8 to YouTube. Just like our Polish colleagues, they reported cases that fell into multiple hate types, namely five.



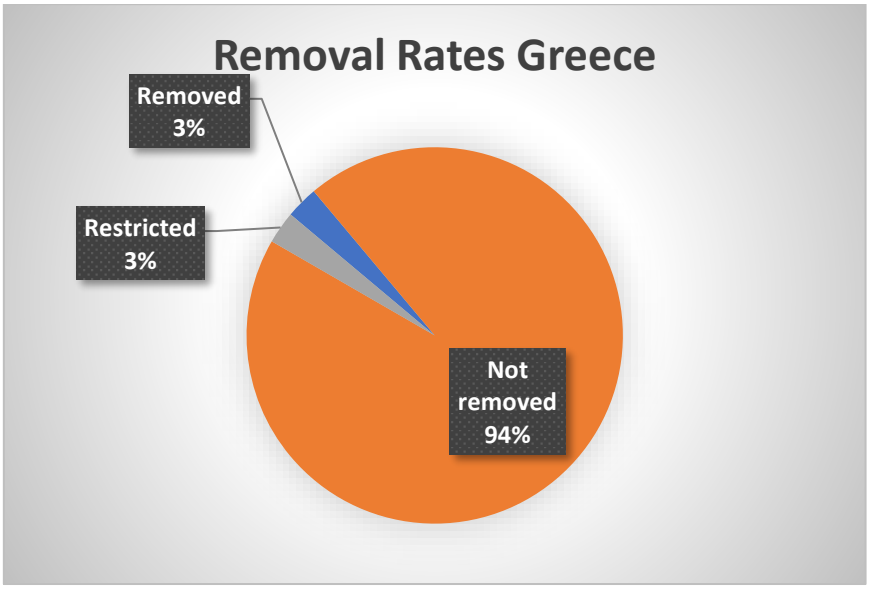
Based on this small sample, it is clear that antisemitism (36%), anti-refugee hate (28%) and Islamophobia (25%) are by far the biggest issues that plague social media in the country. However, we should not gloss over anti-Roma hate (6%) that our member found, since antigypsyism is one of the major drivers of cyber hate within Europe, especially in the eastern member states.

We can provide a bit more information about the findings of the Greek Helsinki Monitor than the Never Again Association, because they faired a little bit better in getting hateful content removed from social media platforms, but their numbers are just as worrying and abysmal as our Polish colleagues'.

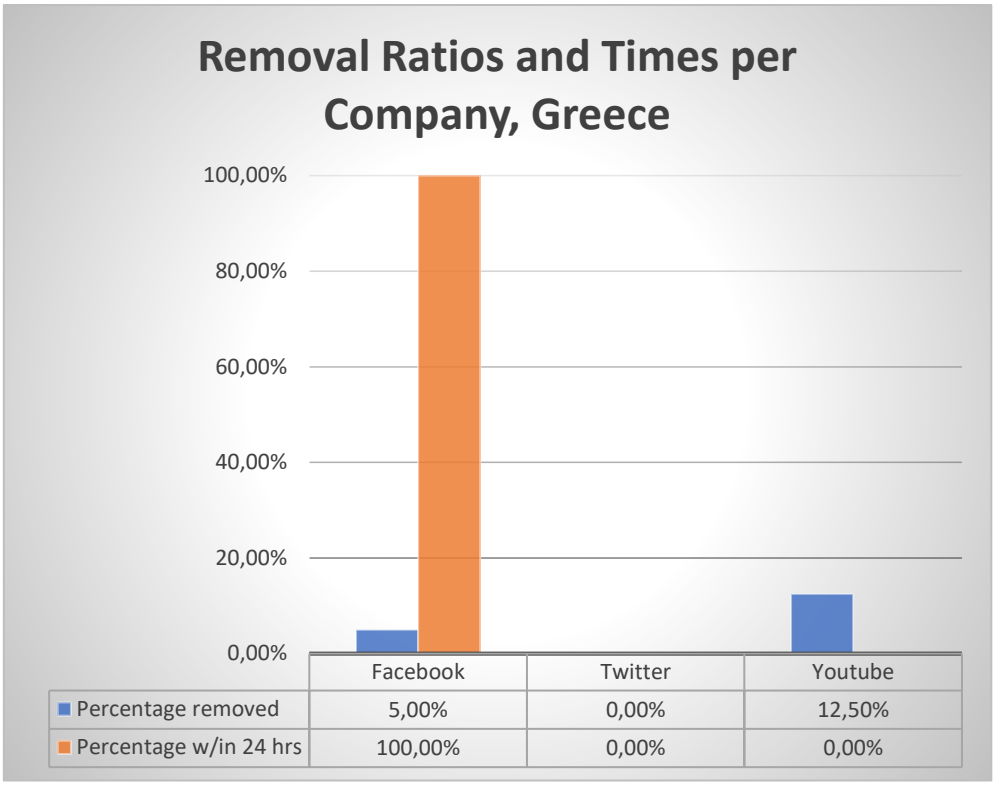


As one can see, the GHM either received feedback from the companies within the first 24 hours after reporting (56% of cases) or they received no feedback at all. This falls close to the findings of the sCAN Project partners, although it is a somewhat below it.

Besides this, however, GHM's findings are magnitudes worse than what the sCAN NGOs found.



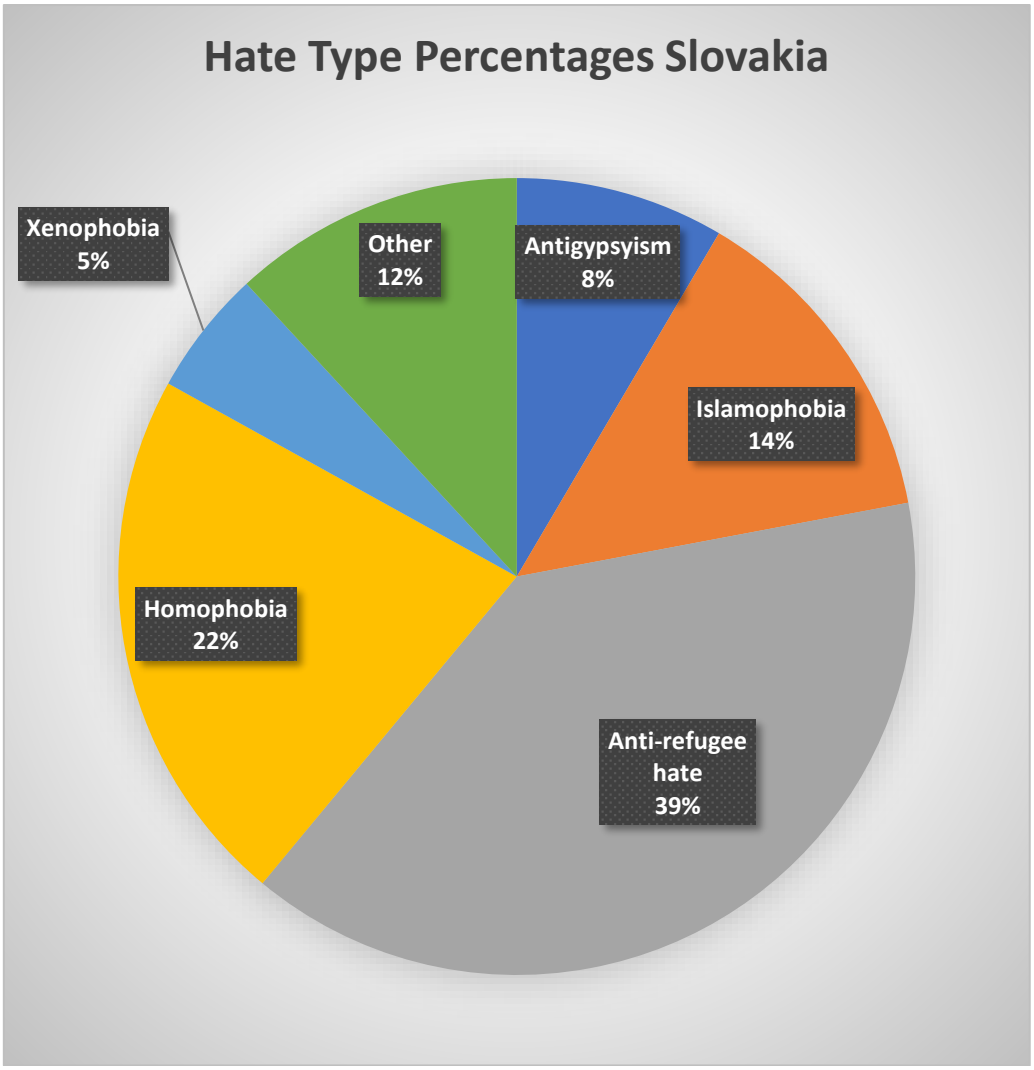
Out of the 36 cases they reported to the different companies one was removed by Facebook, one was restricted (e.g. geo-blocked) by YouTube and a whopping 34, 94 per cent of all reported cases, were not removed. A removal rate that is so deplorable that INACH and its members have never seen such low numbers in any country during any MEs that we have participated in since 2016.



After these appalling numbers and findings, the only positive thing we can add is that – at least – Facebook removed that one case within 24 hours and YouTube restricted their one case within 48 hours. As one can see, Twitter did not remove any of the 8 cases reported to them by GHM.

3. Closer to the Average: Slovakia

Our member from Slovakia, digiQ reported 50 cases during the monitoring period. Just like our Polish and Greek members, they focused on multiple hate types to provide a varied and representative sample of cyber hate prevalent within the country. And just like the Never Again Association in Poland, digiQ solely reported instances of cyber hate to Facebook. Additionally, they reported content via multiple Facebook profiles that were unknown to the company and they did not use their trusted flagger channel during the unannounced ME.

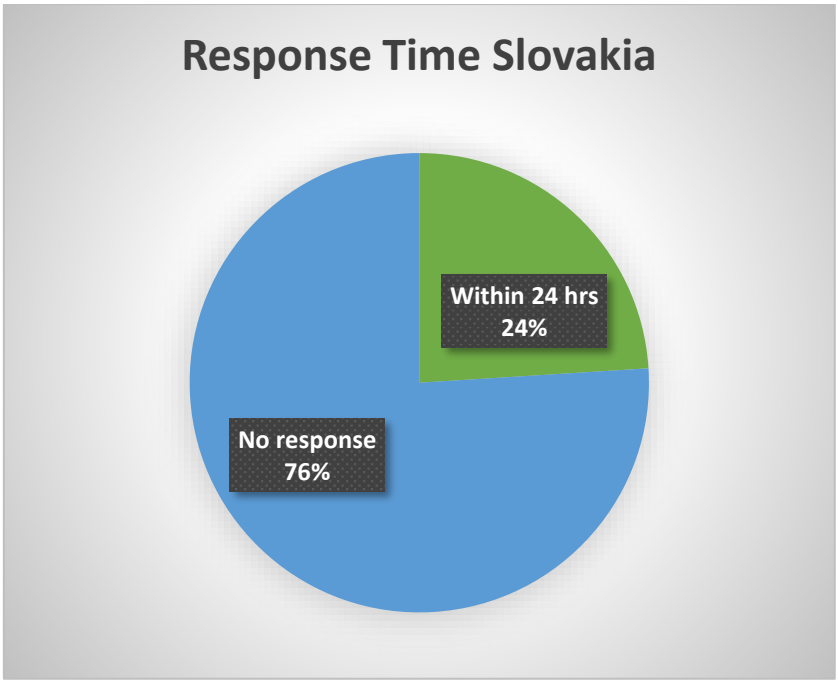


It is very clear that anti-refugee hatred (39%) and homophobia (22%) are the dominant types of hate right now in Slovakia, followed by Islamophobia (14%) and antigypsyism (8%).

The removal rate that digiQ reported is far closer to the average of the sCAN Project partners' findings.

Facebook removed 62 per cent of the 50 cases reported to them by our Slovak member. A much better performance than in Greece of Poland, yet it is still very low and below the 80 – 90 per cent range that would be acceptable.

The company did not produce much better numbers when it comes to responding to complaints or removing them within the desired 24-hour timeframe.



The company provided zero feedback or response to 76 per cent of digiQ's complaints. This is an abysmal showing from a company that managed to provide – on average – timely responses to 70 per cent of complaints sent in by the sCAN Project partners.

Finally, out of the 31 cases that Facebook removed, the company only deleted 39 per cent within 24 hours.

The removal time for the rest of the cases were recorded as unknown, because without receiving feedback, digiQ was only able to go back and manually check whether a certain content was still online or not. If it was not, they recorded the removal, but they did not add a removal time, because they had no way to know when the content was actually deleted by Facebook.

In summary, even though the findings of our member in Slovakia are closer to the numbers recorded by the sCAN Project partners and they seem less of an outlier than Poland or Greece, their records still show that Facebook did appallingly, even compared to itself in other countries and they fell far below the acceptable ranges in responding to complaints, removing hateful content and removing said content within the 24-hour timeframe that they signed up for when they signed the CoC.

4. Conclusions and Recommendations

INACH's joint monitoring exercise with the sCAN project and its other participating members showed that there is definitely methodological value in not announcing an ME to the social media companies and keep the exercise a secret as much as possible.

Furthermore, our ME unearthed some deeply troubling findings that cannot be ignored neither by the EC, nor by the NGOs that help the EC to carry out these exercises. It is absolutely flabbergasting that the companies, and especially Facebook, can produce such different results in different countries. Moreover, it is completely preposterous for Facebook to produce results, such as our member's findings in Poland. It is unacceptable for a platform to provide no feedback and remove no hateful content when 52 instances of cyber hate are being reported to its moderators.

The findings presented in this addendum clearly show that, even though there is a sliver of hope and there has been progress in certain countries within hate speech removal from social media platforms, there are still countries within the EU where illegal hate speech is mainly ignored by the companies. Letting extreme views and extremists in general flourish on their sites.

There is one more issue that is clearly highlighted by the numbers presented above. The way the EC presents the findings of the official MEs is insufficient to show the whole picture. The commission only publishes averages that are aggregates of all numbers collected from 26 different EU countries. The only data that they publish broken down per country is the aggregate

removal rate by all social media companies together. This is clearly insufficient and probably masks the reality to a certain extent.

Therefore, INACH recommends to the Commission to – henceforth – publish all of the findings of the future ME's broken down per EU member state. Not just removal rates, but response rates, removal rates within 24 hours, and feedback rates within 24 hours. Hopefully, this way, the curve will not be skewed by countries where social media companies pay more attention to online hate speech due to governmental pressure or stricter anti-hate speech laws.

One more thing also has to be reiterated here. The social media companies are too involved in the development and organisation of the official monitoring exercises. This too deep involvement skews the outcome of these exercises and provides an environment that is too biased towards the interests of the companies. Thus, we repeat it here again, INACH recommends to the EC to keep social media companies in the dark, not just about the starting date of the exercises, but everything else related to the ME's (starting date, running time, participating NGOs, number of cases recommended to report, etc.). This is the only way to gain a representative and full picture of the efficacy of the CoC as a self-regulating measure for the social media companies in the fight against illegal online hate speech.