INACH

Bringing the Online In Line with Human Rights

The state of policy on Cyber Hate in the EU

2021

Compiled by Adinde Schoorl

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 organizations 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 of educational and preventive strategies.

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.

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

bpb: Bundeszentrale für politische Bildung

Funded by

Federal Ministry for Family Affairs, Senior Citizens, Women and Youth

as part of the federal programme

Demokratie leben!

Introduction

Cyber hate does not disappear from the news for one moment. For example, when former President of the United States, Donald Trump, was banned from Twitter, we quickly learned there is a whole world out there with new alternative social media platforms that care a lot less about fighting hate speech. These platforms are not part of the Code of Conduct or a policy to moderate the hatred that is disseminated through their channels. On the contrary, they actually seem to be created for that purpose; to receive the extreme opinions that have been banned from mainstream platforms. Or when the Covid-19 pandemic started, it almost immediately led to a sharp increase of hate against Asians everywhere in the world, both online and offline. Eventually, the different mutations of the virus had to be given a name following the order of the Greek alphabet, to avoid any stigmatizing effects for people from Brazil, China or South Africa. All of these places had originally named a Covid-19 mutation after them.

Even more recently, black football players of the British National Team received an enormous amount of online racism after losing against Italy in the final of the Eurocup in June 2021. These players had failed to score the penalties and therefore became the scapegoats for losing the event, even though the British had not made it to the final in decades. Where one would have expected pride, they were greeted with hatred.

These examples show that the purpose of our organization, bringing the online in line with human rights, is still very much up-to-date and it is inevitable to keep highlighting the importance of fighting cyber hate. This paper analyzes where we are regarding that; what is being done to stop online hate speech and what steps can still be taken? Even though a lot of developments have taken place, and even though there is certainly more sympathy for the fight against online hate, there is still a long road ahead of us to ban hatred from the internet.

Our last policy paper was written two years ago. We will summarize and quickly revise what has been written back in 2019 before moving on. After that, we will take a look at six legal documents and/or proposals, the first being the Digital Security Act (DSA) of the

EU. The DSA was announced in late 2020 and should be a step forward in protecting human rights online in Europe when finally enacted. Second, we will revise the Network Enforcement Act (NEA) in Germany again. Since we took a look at it in our paper in 2019, now is an opportunity to give an update on the situation. The Austrian lawmaker set up a similar legal framework, which will be discussed after. The fourth and fifth topics are article 612 of the Italian Criminal Code and a proposal of the Polish prime minister to penalize social media companies that ban hate from their platform. Article 612 is a huge step forward to protect online human rights while the Polish proposal is actually a worrying setback. Our last example is of the Penal Code in Estonia, where the EU has started infringement proceedings since the Estonian laws do not comply with the EU framework decision.

Then, we will analyze the developments in the relationship between social media companies and NGO´s, the moderation of content by AI, and finally take a closer look at the phenomenon of fake news. We will close off this paper with revising INACH´s policy recommendations of 2019 and list our new recommendations.

1. Short summary of our previous policy paper

Before we dive deeper into the issues of this policy paper, it is important to summarize what the focus of our last policy paper in 2019 was. The topics analyzed were firstly two legal documents of the European Union (EU): the Code of Conduct (CoC) and the European Commission's Communication, and thirdly a new German law called the Network Enforcement Act (NEA).

In May 2016, the Commission made the agreement with Facebook, Microsoft, Twitter and YouTube that they should abide by a Code of Conduct. In 2018, Instagram, Google+, Snapchat and Dailymotion joined the CoC. More recently, Jeuxvideo.com also joined in January 2019, which makes the coverage of the EU market share of online platforms possibly affected by hateful content to be 96%. The EC´s Communication embodies the fact that the Commission felt it was necessary to give policy suggestions on an EU level regarding tackling cyber hate. Even though it is not a law per se and only recommendations, it was still a big step in the right direction and showed the issue is a top priority. This Communication became a part of the proposal for a regulation preventing the dissemination of terrorist content online from 2018 and it played a role in setting up the Digital Services Act, which will actually be reviewed in this policy paper.

Lastly, the new German law NEA was examined. What was so exciting about this law was that it finally, like never before, forced companies to adhere to the law, with, amongst other things, the existence of fines, whilst at the same time offering access to a clear set of rules concerning the handling of reports.

The policy paper also took a look at the development of the Monitoring Exercises, run by the EC. It concluded that the decision to make the start of the Monitoring Exercises confidential was a big step forward, giving a much more representative look at how social media companies deal with flagged content.

Finally, the paper focused on the social media companies themselves and concluded a lot can still be improved regarding the transparency in decision making and lack thereof. The reason why content is removed, or is not, is often completely unclear and not communicated towards the flaggers. Besides that, there is a clear difference between how normal users and trusted flaggers are treated, which obviously should not be the case. The reporting of normal users should be taken just as seriously as those of trusted flaggers. There are much more normal users than trusted flaggers, 98% of the reported content is done by normal users. Besides that, they are the ones who are actually mostly confronted with online hate speech. The status of trusted flaggers is good to build a relationship between NGOs and social media platforms but often content is not removed when reported by a normal user but is actually removed when reported again by a trusted flagger. This discourages reporting done by normal users.

There were three main areas where policy recommendations were needed; social media, the EU and the national level. Social media companies should find a solution to the problem of the discrepancies between what is being removed and what is not, by working on harmonizing, detailing and clarifying their content guidelines. Social media companies should ask NGOs to train their moderators on hate speech and on the laws that regulate illegal speech in different EU countries. On an EU level, work should be done to attain a more harmonized definition of hate speech, changes should be made to make the monitoring exercise less biased, and the Code of Conduct could be developed further. Social media's adherence to the Code of Conduct should be kept in check through continuous monitoring exercises. The methodology of these exercises should be fine-tuned to mitigate bias. The Communication published by the EC should be the minimum standard in the fight against cyber hate on an EU level. The EU should consider tougher approaches to policing illegal online content if the CoC and the Communication do not reach the intended goals in the coming years. On a national level, the German law should be taken as an example in general terms, including the necessary development regarding its missing regulations on the deletion of legal content. More should be done in educating the public (hence the potential complainants), with a focus on younger people, the elderly and authorities in charge of helping those complainants, such as the police. Besides that, NGOs should move away from the cyber nanny approach and gear their work more towards education, counter speech and prevention. The organizations should put a larger emphasis on building a relationship with the public, become better known and build an image that is easier to digest.

You can read the policy paper of 2019 here: Policy paper 2019

2. The legal environment

The six main legal documents and/or proposals mentioned in the introduction will now be summarized and their impacts in each department will be analysed. The first one is the Digital Services Act (DSA), which was introduced by the EU in late 2020. The EU announced that it plans on enacting the DSA in the coming years. So, even though it has not been put in practice yet, it is a good moment to see whether it is a step forward in protecting human rights online. The DSA aims to create a safe and more open digital space in which the fundamental rights of all users of digital services are protected. The Digital Services here range from simple websites to social media platforms.

The DSA is a timely initiative that aims to fill the gaps and shortcomings of existing legal frameworks, and to ensure greater transparency, protection and safety of users and accountability and harmonization at the EU level. This ambitious incentive is an important step towards protection of human rights online. It introduces for example notice and action procedures to protect fundamental rights and to effectively remedy breaches (including transparent decision making and a complaint handling system). It adds liability obligations to intermediary services to also fight against illegal content and aims to strengthen coordination between different authorities. The DSA also includes the establishment of more user-friendly mechanisms to report content to social media companies. It is important to have easy-to-use mechanisms that are accessible for children and young people, as difficult reporting mechanisms discourage reporting. The DSA mandates these organizations to provide content creators with information and reasoning when their content is removed. However, it doesn't provide for more transparency towards users reporting content to the platforms. It is important for users to receive information on the decisions regarding reported content, especially if the users are very young or directly affected by the reported content. The DSA recognises a procedure that has been well-established for several years to utilise the experience of civil society organisations by giving them the status of Trusted Flagger. However, it remains to be seen who will be certified as trusted flaggers. It is important to have input from thematically diverse organisations and not only recognise those organisations with a large lobbying power. Furthermore, reports from children and young people should also be treated with priority, especially if they concern online risks to minors like cyberbullying or grooming. Social media platforms have been using technologies for years to voluntarily detect and remove certain forms of content (e.g. Child Sexual Abuse Material (CSAM) or terrorist materials). This could be applied to more subjects. Known illegal content could be identified using hash values and AI could be used to flag potentially illegal content for human review.

It is advisable as well to put more emphasis on the fact that social media platforms are actually not just acting as neutral vessels of content or information that is created by others and not at all controlled - but that they are actually selling attention to information. They have developed many forms and elaborate methods of gathering and connecting information about their users and the content they are spreading. Therefore, it seems appropriate to pay more attention to the fact that these companies are actually creating revenue and profit (also) from spreading illegal content. Any regulation on EU-level should contain clearly negative incentives for making money with illegal content. As there are, for sure, precise monitoring mechanisms in place that allow the exact allocation of revenues, these should be a starting point for intervention. As a principle, money gained by spreading illegal content should be forfeit or set aside and invested in counter-measures regarding the illegal content it owes its existence to.

The negotiations on the Digital Service Act are still ongoing between the Commission, the Council and the Parliament. Therefore, we will not be sure what the DSA will look like until its official adoption. Yet, it is a good thing that users will be empowered to report illegal content in an easier and more effective way. But there remain a lot of doubts about how social media platforms will react to the DSA. From previous Monitoring Exercises we know that the rules are not always followed by social media platforms. For example, when a case of hate speech is reported by one user it is often not removed but when it is reported again by 100 users, it is. That discourages normal users to report because the process of the removal of content is very unclear and intangible. Therefore, even when there are rules in place, it doesn't always mean these are followed by companies. How effective can the supervision then be? Is the definition of illegal content clearer in the DSA or even more ambiguous than it was before? Will platforms minimize risk of penalty by taking down more content that is legal? These are all questions that remain to be answered. However, the DSA is clearly a step in the right direction.

The EU also has the responsibility to look further than the mainstream social media companies. There are many new alternative platforms that have a Terms Of Service that are not strictly against hate speech and sometimes even encourage it. These companies obviously have not joined the Code of Conduct of the EU. Some of the members of INACH believe that the EU should establish contact with new alternative online platforms and encourage them to join the CoC. However, other members of INACH believe that it would be useful if the CoC were made mandatory for each and every platform that has a number of subscribers/account holders that exceeds an agreed minimum threshold. Whether it should be made mandatory by the EU or not, it is clear that the EU has to find a way to have these new social media companies sign the CoC. At the same time, when the DSA will be put in practice, it should cover all the platforms regardless of whether or not they have signed the CoC, so there will be the same rules for all of them. There are some warnings about the risk that the DSA will create new access barriers to platforms and that differentiation between large and very large platforms could be disingenuous. So there are still loopholes where content can escape by moving to smaller platforms and we need to have a look at how to prevent that.

The second legal document that we will explain about is the Network Enforcement Act (NEA) in Germany. The NEA entered into force in October 2017. It includes obligations for social media companies to process reports and delete illegal content, including illegal hate speech, in a timely manner. It also includes obligations for transparent reporting, the establishment of effective reporting mechanisms and the appointment of a representative of the platform in Germany. Finally, it includes fines for non-compliance with the provisions of the NEA. In 2021, the NEA was amended in order to comply with the EU's Audiovisual Media Services Directive (AVMSD) and to strengthen the rights of social media users. The changes to the NEA include:

- Users can appeal the decision of social media platforms to (not) remove content reported to them. The social media company will have to review the content and provide reasoning for their decision.
- Reporting mechanisms must be easy to find and use. (After the implementation of the NEA in 2017, some platforms established reporting mechanisms that were difficult to find and very complicated, thus discouraging reporting.)

- Transparency reports have to outline changes in relation to the past two transparency reports, include information on the social media organizations' handling of appeals to their decisions to (not) remove content reported to them and provide basic information on automated tools used to detect potentially illegal content on the platform.

Since these changes to the NEA only entered into force on 28 June 2021, we have not been able to evaluate them yet. Furthermore, it remains to be seen what implications the adoption of the DSA might have on the NEA, due to the primacy of EU law.

Austria actually took the NEA as an example and aimed at overcoming some of the shortcomings of it by strengthening access to justice for victims of certain offenses of online hatred and maybe making it easier to comply with the right to freedom of expression. From January 2021, a comprehensive legal framework is in place aiming at facilitating access to justice for victims of online hatred. The new legislative package against online hatred is an important step towards extending and strengthening the protection of victims. One important change is that the scope of application of the offense of incitement to hatred has been extended. Certain insults by individuals are now covered if made on the basis of an (ascribed) group membership (art 283 para 1 sub-para 1 of the Austrian Criminal Code). Furthermore, in the context of cyberbullying, the one-time publication of, for example, nude photos can also be qualified as incitement. Nevertheless, the corresponding post must be retrievable for a longer period of time to be considered unlawful under this provision (art 107c of the Austrian Criminal Code). Several provisions make it easier for victims of online hatred to gain access to justice, like the new mandate procedure, allowing for seeking injunctive relief for serious violations of personality rights by using an online form and at low cost (art 549 of the Austrian Civil Procedure Code). Victims of defamation, insults, or an accusation of a criminal offense on the Internet that has already been dismissed can now file an application for investigation of the suspect with the regional court, whereby the court can order investigative steps by the authorities (art 71 para 1 of the Austrian Criminal Procedure Code). Persons who experience persistent persecution, cyberbullying, or incitement are now entitled to legal and psychosocial process support (art 66b para 1 lit c of the Austrian Criminal Procedure Code). Victims of defamation, insults, and libel may also receive legal and psychosocial process support if, based on certain indications, it can be assumed that the offense was committed online (art 66b para 1 lit d of the Austrian Criminal Procedure Code).

Another interesting example of a step forward when it comes to protecting online human rights, was given by our Italian member CESIE. They consider the creation of article 612 in the Criminal Code of Italy as an improvement regarding the protection of women against misogynistic hate speech. This law punishes so called 'revenge porn' or the dissemination of unauthorized content, also by third parties. The law prohibits *´…the unlawful dissemination, sale, or publication of sexually explicit images or videos of a person without the person's consent, in order to harm the person.´*[1]
The law even states increased penalties when the crime is committed by the spouse, even if divorced or separated, or by those who have been sentimentally connected with the offended person, or when the offense is committed through computer or electronic means.[2] Especially the fact that the dissemination through electronic means will be penalized higher, makes it a worthwhile article to mention here in relation to online rights. While the victims of this crime could be anyone, it is important to underline that this is a highly gendered crime. In Italy over 80% of the victims of revenge porn are women, and the perpetrators are mostly their male (ex)partners.[3] In 2020 it was reported that the Italian authorities were investigating 1000 alleged cases of revenge porn. This spike in reporting was probably taking place due to the Covid-19 lockdown.[4]
Italy is one of the few European countries that introduced a law regarding revenge porn, next to the United Kingdom, France and Germany. However, due to the transnational nature of the internet, a more unified international approach is needed. Therefore, the

[1] https://www.internetjustsociety.org/fighting-against-revenge-porn
[2] https://www.loc.gov/item/global-legal-monitor/2019-09-09/italy-new-law-enters-into-force-to-protect-victims-of-domestic-and-gender-violence/
[3] https://www.internetjustsociety.org/fighting-against-revenge-porn
[4] https://cde.news/case-of-revenge-porn-spike-in-italy/

DSA is an opportunity to offer that exact unified approach.

Fighting hate speech is often confused with censoring freedom of speech. Even though these are two very separate concepts they are regularly, and sometimes on purpose, treated as synonyms, especially in politics. In Poland for example, the Polish Prime Minister Mateusz Morawiecki has promoted his government's plans for a new law to prevent social media companies from censoring what he calls ´free speech´. According to Morawiecki, these companies do not have the right to decide what views are correct or incorrect.[5] The Polish Minister of Justice, Zbigniew Ziobro, proposed that social media companies should receive a fine when they delete content or ban users who are not posting something illegal. Social media users who have been blocked or had content deleted will be able to complain with the platform and the platform would have to respond within 24 hours. If a social media company refuses to comply, they could receive a fine from a Freedom Of Speech Council, which will be created by the Polish government. The measures could come into effect by January 2022.[6] The proposal of the Polish government is worrying, particularly because there is in Poland more a situation of under-removal than over-removal by social media platforms. Especially LGBTQIA+ groups are often very much unprotected against online attacks. Therefore, the proposal of the Polish government means a huge incentive for the companies not to bother removing hate speech in general.

The claim here, and in other cases, that freedom of speech is protected by making sure that every opinion can be disseminated on the internet, is a false one. Because the opposite is true; if a safe and open digital space where human rights are respected can't be guaranteed, freedom of speech can never take place.

Our last example is that of Estonia. At first glance, Estonia has a quite complete law regarding hate speech. The Estonian Penal Code includes provisions on prohibition of incitement to hatred. Activities which publicly incite to hatred, violence or discrimination

[5] https://www.vice.com/en/article/v7mpkj/poland-wants-to-ban-social-media-companies-from-banning-hate-speech

[6] https://www.bbc.com/news/technology-55678502

against listed groups are punishable by a fine of up to three hundred fine units or by detention (misdemeanour). However, at the same time the law also includes a condition restricting the scope of the article making incitement illegal only in cases where the victim's health, life or property are at stake. That would be hard to prove and therefore makes the Estonian Penal Code a lot less resolute.

Also, Estonia fails to transpose the EU framework decision on combating certain forms of expressions of racism and xenophobia by means of criminal law. The EC has initiated infringement proceedings against Estonia. According to the EC, ´Estonia has failed to transpose criminalization of the specific forms of hate speech, namely the public condoning, denying or gross trivialization of international crimes and the Holocaust when such conduct aims at inciting violence or hatred.'[7] And additionally: ´Estonia has not correctly criminalized hate speech, by omitting the criminalization of public incitement to violence or hatred when directed at groups and has not provided for adequate penalties.´[8] Finally, according the EC the Estonian criminal code does not ensure that the racist and xenophobic motivation of crimes are taken into account as aggravating circumstances so that such crimes are effectively and adequately prosecuted. The infringement process started in October 2020.[9] According to the Estonian Human Rights Centre, there is no public information about the development of the process. The case of Estonia shows the importance of designing laws that are complete and do not offer escape routes which means they are just empty laws, in place to satisfy the public but not actually protecting human rights.

[7] https://news.err.ee/1153405/european-commission-launches-infringement-proceedings-against-estonia

[8] https://news.err.ee/1153405/european-commission-launches-infringement-proceedings-against-estonia

[9] https://news.err.ee/1153405/european-commission-launches-infringement-proceedings-against-estonia

3. Social media (and AI)

In the past two years much has changed regarding the interaction between NGOs and social media companies. There clearly is an improvement in the relationship between them, even though a lot still needs to be done. Social media companies organise more calls and events to ask NGOs for their expertise and feedback and in return they sometimes participate in events and conferences organized by NGOs. Furthermore, all popular platforms have implemented Trusted Flagger programmes. Several of our members have noticed an improvement in the approachability of the social media platforms. Needless to say this mostly happens during the official Monitoring Exercises. In particular, better cooperation has been established with YouTube. Also, some social networks have shown improvement in the transparency regarding decisions whether or not to remove content – for example, Facebook does give feedback within hours. Transparency is usually only achieved through public reports of the platforms or through discussion in calls. However, most monitored platforms have improved their feedback to reports, and also to user reports. An exception is Pinterest – they do remove content that is reported to them via mail, but no personal contact exists with them. With TikTok there is no transparency at all and removal of the reported content is very rare.

Of course, there is also a lot of room for improvement. Transparency in general is still missing. Even though social media companies communicate better and faster, they could do more in explaining their decisions about (non)removal. Platforms could also still try to include NGOs more and use their expertise before rolling-out new features. On top of that, there is a clear regional difference in how social media companies deal with flagged content. For example, hateful content in Polish is less likely to be removed than in English or German. This is worrying because it means hate speech is ignored more in some regions than in others. One of INACH´s members from Slovenia, Spletno Oko, has noticed a growing number of hateful comments and hate speech on TikTok lately. They believe language plays an important role in companies' decisions connected to interaction with the NGOs, so smaller countries like Slovenia are neglected in this area. However, our member of Estonia, the Estonian Human Rights Centre, reported that due to the fact that Estonia is a small country they are actually able to communicate quicker and more efficiently with the offices of social media platforms than would be the case in bigger countries. To the very least, it shows that social media platforms act very differently in every country and region of the world. Tighter future collaboration between the platforms and NGOs (funding, monitoring, participation in joint webinars, etc.) would surely benefit both parties and advance the goal of removing hate speech from social media.

The moderation of social media is obviously not just done by people, the AI technology that detects hate speech and other illegal content becomes smarter every day. On one hand, this is a good development. Moderating online hate speech can be very taxing on moderators, especially if there is a very high workload and limited staff capacities. And by now, the stories about the psychological damaging effects on moderators are widespread. Ethically speaking, it is not responsible to keep using humans as moderators even when their labour circumstances are improved greatly. AI could help alleviate the mentioned burdens. Since it is impossible to manually monitor social media posts on a large scale, the introduction of relevant AI is not only helpful but necessary and inevitable. However, in its current state, it is not yet the answer. AI is definitely better than classical recognition techniques as it is more flexible and its ability to adapt to new content is better and quicker. But AI still lacks accuracy (as any automated solution). There have been many protests about automatic (non)removal of content on different platforms. Sometimes, educational or awareness-raising content has been blocked, while hate content was not removed. In order to detect hate speech in text-based content, AI would have to be able to understand context, satire, humour and dialects in the monitored languages, as well as combinations of text and images and online codes as for example leetspeak (replacing letters by numbers). This can be a problem when somebody uses offensive language to make a point about hate speech or when sarcasm is used in referral to illegal content. That is a huge challenge for a technological tool that will never become an independent human being. AI will never manage to replace human beings, but it might be useful to spot fake accounts, bots, and algorithms created to bombard social media accounts with fake news and propaganda content. Of course it needs to be developed further and at some point, there will be a higher accuracy. - but the accuracy of an AI solution highly depends on the amount and quality of the material the AI is trained with beforehand. Furthermore, training material would also have to be free from discriminatory biases (such as racism, antisemitism, antigypsyism etc.). AI cannot be a stand alone solution, but can be used as a technical support. While we know that the use of AI is inevitable in content moderation, it has to be stated that it is still in a fetal state and lacks the above-mentioned development. AI cannot understand context or use it to analyse content. It also makes a lot of mistakes and over-removes certain content (especially when it is about ´nudity´). And it completely lacks the understanding of irony or sarcasm, which leads to the removal of actual anti-racist content that mocks racist topes or other hateful ideas. A hybrid human-AI monitoring tandem is the key to efficient monitoring. Continuous negotiation of definitions (what constitutes hateful content?) and adjusting the AI accordingly, can only be done by humans. In addition, continuous checks are necessary in order to make sure that 'hardcore' hateful content isn't allowed to be posted, to begin with, or is quickly removed when flagged by users. AI can be used to filter content on which it decides that the content is true-positive above a certain threshold (e.g. 95% plus) and it can relay content with a specific uncertainty (below a certain threshold) to a human moderator for review. Platforms should provide information if content has been blocked based on AI. AI should not be able to block or delete content permanently without the possibility of human review.

Of course AI is one of the answers but it cannot, and should not, be the only one. We believe the combination of AI and human moderation is the only way, but transparency in the process is needed.

4. NGOs

NGO´s play a paramount role in fighting online hate speech. However, they too have to keep up with all the developments that happen online. New social media platforms are often not as well known as Facebook, Twitter and others. In order to stay up-to-date, international cooperation and exchange of expertise on these newer platforms is important. Extremists usually advertise newer/ fringe social media platforms or "safe havens" on more popular platforms, in order to mobilise their followers to join them on these new platforms. For NGOs monitoring extremist groups and hate speech on popular social media, it is therefore possible to be aware of these new platforms. Here lies a great opportunity for the network of INACH to arrange the exchange of knowledge on these new developments, and maybe even help to divide the tasks in following these changes. With developed specialised expertise on certain platforms, different NGOs can share their knowledge with the rest of the network. NGOs such as ours must keep up with the ever-evolving online social sphere. To do that, it is also important to engage teenagers, young adults, and students. They could be taught about the basics of monitoring hateful content online and they could implement them across social media networks, including new and upcoming ones. NGOs can indeed train them on what hate speech is, on how fake news works and what can everyone do to challenge them, while young students can teach NGOs staff how to use and navigate new platforms which are little known by adults.

Finally, NGOs play a crucial role in monitoring the social media platforms´ adherence to the CoC. They are the gatekeepers of checking whether these social media companies are doing their job in moderating hateful content. Of course NGOs such as ours should continue doing that. INACH and the participating members have the responsibility to continue with the Monitoring Exercises and the shadow exercises. These Monitoring Exercises have led to an improvement of cooperation between social media companies and NGOs and led to these platforms taking hate speech seriously. Therefore, that work needs to be continued in order to keep everybody on their toes in respecting the CoC and in improving the Monitoring Exercise as well by adding new platforms to it to monitor and improving the process itself.

5. Fake news

Our policy paper of 2019 started the introduction with describing the issue of fake news. Only two years later fake news has become an even hotter topic. Ultimately, fake news and hatred are intertwined, therefore this is a very important subject to take into account when talking about online hate speech.

When we say that fake news and hatred are intertwined, we mean that there are groups who spread disinformation online in order to attract new members. The Covid-19 pandemic is only the latest example of it. There are clear connections between antisemitism and conspiracy theories on one hand and anti-lockdown and anti-vax movements on the other hand. White supremacist groups lure in people who are part of the anti-lockdown and anti-vax movements and convince them of their antisemitic ideas. For example, one of the fake news stories they use is that Covid-19 is a hoax made up by jews in order to make money. Or another one is that Covid-19 is not a hoax and was actually created by jews in order to serve their Zionist agenda. Bill Gates and George Soros, who are used as a stand-in for ´jews´, are often mentioned as the people who have created the virus for their own financial gain by forcing people to get vaccinated. The danger here obviously lies in the fact that a whole new group of people is being dragged into antisemitic ideas of extreme right wing groups who would probably not have been reached in pré-Covid-19 times. In this manner these dangerous ideas keep spreading, especially on the internet. This is not a new phenomenon and is not exclusively used against Jews. When Barack Obama was elected president of the US, there were persistent theories that Obama was not born in the US and therefore should not have been the President, the theory was called ´birtherism´ and was widespread. Or when referring to the aftermath of the killing of George Floyd in the US, it was assumed by many that the numbers of police brutality were false and an excuse for letting black people act upon their aggressive nature.[10] Even the European news stations covering the ´refugee crisis´ in recent years, have often been dominated by a conspiracy theory that has almost become mainstream: the immigrants are coming to Europe to dominate the continent and islamization is around the corner.

Do you want to read more about this? You can read our report on it here: Covid-19 a plague of racism and conspiracy theories

These examples show the importance of combating fake news, because fake news intensifies and spreads hatred. In Germany an effort is being made to put rules on the distribution of fake news. The new Interstate Media Treaty, which entered into force in November 2020, imposes due diligence obligations on "commercial, journalistic-editorial telemedia offers, which regularly contain news or political information", stating: "Before they are disseminated, news stories are to be checked by the provider for content, origin, and truth with the due diligence required for the circumstances." (Article 19). Of course, this also necessitates due considerations of freedom of expression.[11]

Next to setting up rules, awareness about fake news is equally essential. In order to combat disinformation and fake news, it is important to strengthen the critical media competence of young users, not only on content but also on media design and formats. Fake news often use specific designs, cues or codes in order to pass as regular news articles. Furthermore, participatory media services should be established, which offer de-escalating counter arguments to fake news and enable critical discussions.

However, if fake news are exclusively or primarily used to disseminate hate speech and incite hatred and violence, they have to be clearly labelled as such and countered accordingly. Therefore, we believe that fake news should be monitored similarly to hate

[10] https://www.pbs.org/wgbh/frontline/article/how-conspiracy-theorists-have-tapped-into-race-and-racism-to-further-their-message/

[11] https://www.die-medienanstalten.de/fileadmin/user_upload/Rechtsgrundlagen/Gesetze_Staatsvertraege/Interstate_Media_Treaty_en.pdf

speech and other types of prohibited content. A clear definition of fake news should be set and taught. NGOs should collaborate with local school systems and involve community leaders in this regard. The online training of INACH about hate speech is a great example of that, particularly since it includes a game in which the participant gets to design fake news himself. The underlying goal should be to join forces in the fight against fake news and disinformation. The effect of fake news and disinformation spread online, on offline events (such as a rise in covid-19 cases among people who chose not to be vaccinated due to fake news) should be emphasized, when discussing the importance of taking action against the spread of fake news online, and when seeking funding to sponsor and support such actions.

It is also necessary to continue with the fact-checking practices. We know that fake news tends to spread much faster than real news. Warnings or links with accurate information are useful tools, if people come across them before they encounter misinformation (the so-called pre-bunking). Finally, social media companies should be encouraged to use algorithmic solutions that downrank unreliable content.

All of our efforts should be focused upon reducing polarisation in society that exists because everybody's social media feed is unique and everyone gets served the information they ask for. We all need to step out of our bubbles and connect again. In order to do that, INACH has some final policy recommendations.

6. INACH´s policy recommendations

Before we jump to our recommendations, we have to examine the suggestions we made in 2019 and whether they were implemented, at least to a certain extent, or not. Two years ago, we made the following policy recommendations to the EC and social media:

1. **Social media companies should find a solution to the problem of the discrepancies between what is being removed and what is not, by working on harmonizing, detailing and clarifying their content guidelines.**

This is still an issue. After this year´s Monitoring Exercise one of the main problems again was the lack of transparency. And the discrepancy between the treatment of normal users and trusted flaggers remains very present.

2. **On an EU level, work should be done to attain a more harmonized definition of hate speech, changes should be made to make the Monitoring Exercise less biased, and the Code of Conduct could be developed further.**

INACH maintains this recommendation. There is still work to be done to make the Monitoring Exercise less biased and find a more harmonized definition of hate speech. And to add to that, in order to generate more impact more attention should be given to the results of these MEs as well.

3. **Social media's adherence to the Code of Conduct should be kept in check through continuous Monitoring Exercises. The methodology of these exercises should be fine-tuned to mitigate bias.**

This recommendation remains very important. INACH and its partners keep working on this continuously together with the EU.

4. **The Communication published by the EC should be the minimum standard in the fight against cyber hate on an EU level.**

Since the EC´s Communication was introduced five years ago and has been maintained ever since, we can conclude this recommendation has been implemented and can be removed from our list of recommendations.

5. **The EU should consider tougher approaches to policing illegal online content if the CoC and the Communication do not reach the intended goals in the coming years.**

The DSA is a step forward regarding this. However, while we await the final version of the DSA to see if it will fulfill its promises, we maintain this recommendation.

6. **On a National level, the German law should be taken as an example in general terms, including the necessary development regarding its missing regulations on the deletion of legal content.**

We maintain this recommendation since this German law remains one of the few that actually involves a state-regulated system with fines. We can see that there are still countries that take a step back instead of forward and therefore a law like the German one still serves as an example of what the protection against online hate speech should look like. And again it is important to underline that the DSA would be a chance to eliminate the dependency on national politics for respecting human rights everywhere in the same manner.

7. **More should be done in educating the public (hence the potential complainants), with a focus on younger people, the elderly and authorities in charge of helping those complainants, such as the police.**

There remains a lot to do in this aspect. Too often hate speech is confused with offensive speech, free speech or bullying while these are all clearly separate concepts. We should add to this recommendation that an effort should be made to cooperate with groups like the youth to have an exchange of knowledge take place about new platforms and teach about hate speech. If NGOs want to keep up with the developments online, the youth is a chance to exchange knowledge with.

8. **Social media companies should ask NGOs to train their moderators on hate speech and on the laws that regulate illegal speech in different EU countries.**

We still maintain this recommendation. If social media companies would take advantage of the expertise of NGOs to teach about hate speech, it would be a great step forward in combating it online.

9. **NGOs should move away from the cyber nanny approach and gear their work more towards education, counter-speech and prevention.**

Since this has largely been implemented, we do not have to maintain this recommendation. NGOs in the area of hate speech are not merely complaints bureaus anymore. They are investing a lot more time and effort in education and countering hate speech.

10. **NGOs should put a larger emphasis on building a relationship with the public, become better known and build an image that is easier to digest.**

This recommendation has been largely implemented as well. NGOs have been able to raise their profile and reach out to the public by using social media and many are still working on this to develop that further. Therefore, we feel we do not have to maintain this recommendation.

A few policy recommendations should be added to this list:

1. **The EU should find a way to have the new social media platforms sign a CoC. The DSA in general is a real opportunity to offer a unified approach to respecting online human rights.**
2. **AI cannot be the only tool in place to handle the monitoring of hate speech. It needs to be done in close cooperation with humans. There is a strong need however to keep developing AI regarding hate speech and keep the context that is used to teach AI free of discrimination, in order to make the technology smarter and therefore more useful in the future.**

3. **In order to stay up-to-date on the developments, NGOs should make an effort to exchange knowledge regarding the new social media platforms. INACH should use its network to organize the time and place to make that exchange possible through webinars.**
4. **Since disinformation and conspiracy theories are closely intertwined with hate speech, more efforts should be made to counter fake news by NGOs such as ours. Fake news should be monitored in the same manner as hate speech.**

So, to summarise, these are INACH´s policy recommendations:

1. **The EU should find a way to have the new social media platforms sign a CoC. The DSA in general is a real opportunity to offer a unified approach to respecting online human rights.**
2. **AI cannot be the only tool in place to handle the monitoring of hate speech. It needs to be done in close cooperation with humans. There is a strong need however to keep developing AI regarding hate speech and keep the context that is used to teach AI free of discrimination, in order to make the technology smarter and therefore more useful in the future.**
3. **In order to stay up-to-date on the developments, NGOs should make an effort to exchange knowledge regarding the new social media platforms. INACH should use its network to organize the time and place to make that exchange possible through webinars.**
4. **Since disinformation and conspiracy theories are closely intertwined with hate speech, more efforts should be made to counter fake news by NGOs such as ours. It should be monitored in the same manner as hate speech.**
5. **Social media companies should find a solution to the problem of the discrepancies between what is being removed and what is not, by working on harmonizing, detailing and clarifying their content guidelines.**

6. **On an EU level, work should be done to attain a more harmonized definition of hate speech, changes should be made to make the monitoring exercise less biased, and the code of conduct could be developed further.**
7. **Social media's adherence to the Code of Conduct should be kept in check through continuous monitoring exercises. The methodology of these exercises should be fine-tuned to mitigate bias.**
8. **The EU should consider tougher approaches to policing illegal online content if the CoC and the Communication do not reach the intended goals in the coming years.**
9. **On a National level, the German law should be taken as an example in general terms, including the necessary development regarding its missing regulations on the deletion of legal content.**
10. **More should be done in educating the public (hence the potential complainants), with a focus on younger people, the elderly and authorities in charge of helping those complainants, such as the police.**
11. **Social media companies should ask NGOs to train their moderators on hate speech and on the laws that regulate illegal speech in different EU countries.**