INACH

Bringing the Online In Line with Human Rights

The state of policy on cyber hate in the EU

Compiled by
Adinde Schoorl

2023

www.inach.net

TABLE OF CONTENTS

International Network Against Cyber Hate – INACH

INACH was founded in 2002 to use intervention and other preventive strategies against cyber hate. The member organisations are united in a systematic fight against cyber hate, for example as complaints offices, monitoring offices or online help desks. In their respective countries, they provide important contacts for politicians, internet providers, educational institutions, and users.

Funding for INACH is provided by its members, the European Commission, the BPB, and other donors. The International Network Against Cyber Hate (INACH) unites multiple 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 Citizens, Equality, Rights and Values (CERV) Programme of the European Union. The contents of this publication are the sole responsibility of the International Network Against Cyber Hate and can in no way be taken to reflect the views of the European Commission.

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

Introduction

Online hate speech is an ever evolving and developing issue. It changes form on a continuous basis, along with the events that shape the international order and local contexts. On top of that, the development of recent technologies facilitates new forms of spreading hate and political events are usually followed by different types of hate. Societies go through waves of xenophobia, gender-based hate and LGBT+ hate that particularly hit public figures like politicians, journalists, and activists. Also, society contributes to the normalization of hate. Politicians use it as a political strategy to attract voters. For example, in some European societies it is not allowed anymore to even talk or read about LGBT+ issues and many accept the normalcy of targeting female politicians with hate campaigns.

Hatred is on the front pages of newspapers when reporting about the conflict in Israel and Palestinian Territories. It has become almost impossible to talk about the events without igniting hatred against either of the communities involved. The hatred also happens online. Amnesty International urged social media companies to address the online hate happening on their platforms. The organization found an alarming rise in advocacy of hatred that constitutes incitement to violence, hostility, and discrimination on these platforms. A significant number of posts glorify Israel's attacks on civilians in Gaza, support the destruction of Gaza and violence against Palestinians. Moreover, they received reports on censorship by platforms shadow banning content from Palestinian accounts. The organization also documented antisemitic posts, advocating hatred and violence against Jewish people (Amnesty International 2023).

Hate speech is increasingly intertwined with disinformation. By using false statements or half-truths that confirm existing stereotypes of minority groups, hatred is ignited and spread repeatedly. It also shapes European politics. For example, the attitude of the European governments regarding refugees is hardening. There seems to be an attitude of acceptance that it is allowed to ignore international laws, human rights and basic needs of people who are fleeing war and/or persecution. There is less and less public outcry about the push backs of boats on the mediterranean sea, of the circumstances in refugee camps and the manner in which the issue is addressed rhetorically.

All in all it shows 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 online hate speech. We aim to prevent and counter online hate speech, raise awareness about hate speech and connect and cooperate with all our partners in the field of hate speech. This policy paper aims to draw a picture of the current state of affairs regarding online hate speech by describing different issues that concern us and suggesting policy changes to contribute to solutions.

The first chapter will summarize the previous policy paper which we published in 2021. After that, this paper is entirely focused on four topics that INACH deems important to explain and suggest policy changes on.

First, the newly introduced Digital Services Act (DSA) remains a concern since so much remains unclear on what the implementation of the DSA will look like nationally.

Second, Image Based Sexual Violence (IBSV). Gender based hate, and more specifically IBSV, is an issue that is highly influenced by technology and online hate that deserves all of our attention.

Third, ChatGPT and more generally Artificial Intelligence (AI) has been the centre of attention lately in any news medium. However, we have to keep an eye on the consequences that AI has for online hate and what policies should protect against it.

Finally, we will dive deeper into the topic of disinformation. This issue has become more and more of a problem that is intermingled with hate speech. Disinformation focuses on increasing distrust between different groups in society and between citizens and governments. Once citizens have decided they do not trust mainstream media for their information and turn to alternative media sources that use a lot of disinformation, the truth becomes subjective. When disinformation contains hate to target minority groups, hate speech is on the rise.

This policy paper will conclude with a review of our policy suggestions from 2021 and will give new policy suggestions regarding the discussed trends.

Summary of our previous policy paper

Before we delve deeper into the issues of this policy paper, it is important to summarize what the focus of our last policy paper was. Our last policy paper focused first on a few different legal developments regarding online hate speech:

- The Digital Services Act: In 2021, the negotiations on the Digital Services Act were still ongoing between the Commission, the Council, and the Parliament. Therefore, we were not sure what the DSA was going to look like until its official adoption. Yet, we viewed it as a good thing that users will be empowered to report illegal content in an easier and more effective way. But there remained a lot of doubts about how social media platforms would react to the DSA. From previous Monitoring Exercises we know that the rules are not always followed by social media platforms. We will come back to the current status of the DSA in this policy paper.
- The NEA in Germany: 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.
- 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.´
- The Polish initiative to fine social media companies if they delete or ban content that is not illegal according to the government. 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. 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.
- In Estonia, the criminal code includes incitement to hatred. 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.

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.

The use of Artificial Intelligence while moderating hate on social media platforms 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). 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. We will come back to the developing role of AI in this policy paper.

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. We will come back to the developing role of misinformation in this policy paper.

You can read our previous policy paper here.

Digital Services Act

In this section we will review what we know about the Digital Services Act (DSA) so far and our concerns connected to online hate speech. The DSA is a new European law, introduced in November 2022. The DSA is a large set of rules that apply to everything online related, including illegal online hate speech. Since November 2022, online platforms have had three months to report the number of active users on their website. Based on that, the European Commission decided whether a platform could be viewed as a very large online platform (VLOP) or very large online search engine (VLOSE). Different rules apply to these platforms as to what they should be doing in terms of moderation and removal of online hate ('The DSA: ensuring a safe and accountable environment' N.D.). In April 2023, the first designation decisions were made. The big mainstream platforms that fall within the category VLOP are Alibaba, Amazon, Apple AppStore, Booking.com, Google Play, Google Shopping, Wikipedia, Facebook, Instagram, LinkedIn, Pinterest, TikTok, Twitter, Zalando, Snapchat, Google Maps and YouTube ('The DSA: Commission designates first set of VLOPs and VLOSEs' 2023). Since then, these platforms have had four months to comply with the obligations under the DSA. EU Member States will need to set up their Digital Service Coordinators by February 2024, which is when the DSA is fully applicable for all entities. Bing and Google Search fall within the category VLOSE ('The DSA: ensuring a safe and accountable environment' N.D.).

The DSA is a great initiative to have a common framework for all member states on a European level. It requires member states to harmonize their national hate speech laws as much as possible and to set up common processes for addressing illegal hate speech. It is an opportunity to prevent and counter online hate speech in a more uniform manner and it means the approaches of it will depend less on the question whether national governments are willing to give attention to online hate speech. It does not mean however that all governments in the EU need to apply the same hate speech laws, the DSA only requires harmonization. More than anything, the DSA puts compliance and process rules on the social media platforms dealing with online hate speech. And it gives users legal tools and protections (Keller 2022).

As we just described, the DSA is still in process of being implemented and therefore there are still quite a few uncertainties. Based on what we know so far, there are a few points to be made.

First of all, the new Trusted Flagger system, as it seems it will be implemented until now, remains very unclear. The national Digital Service Coordinators (DSC) will appoint the organizations that will fulfil the role of Trusted Flagger. Within this role the reporting to social media platforms by these organizations will have priority and be dealt with by the platforms as soon as possible and taken very seriously. However, this runs the risk of

abuse by governments that might be hostile towards these organizations. The European Union includes a lot of different national political contexts where organizations that promote LGBT+ rights or refugee rights or women's rights are seen as an obstacle by their national governments. To give the DSC's the power to decide who will be the Trusted Flagger within their country means that a huge amount of power is in the hands of governments as opposed to NGOs.

Next to that, NGOs are concerned that a DSC will not ensure that the included Trusted Flagger organizations are diverse enough. For example, according to INACH member #StopFisha, in France no NGO that focuses on gender-based hate is included. This means that this type of hate will have a high chance to be overlooked. A second concern of #StopFisha is that while the DSA requires complete independence, in France negotiations are already taking place with organizations that receive support from social media platforms.

Also, the requirements to comply with the Trusted Flagger status in terms of reporting are quite strict and there is a lot of doubt whether small NGOs will be able to do so. Until now there is no clarity whether there will be resources available for NGOs from the EU to work within this system. That would be the case in Estonia for example, where INACH member Estonian Human Rights Centre is based.

Finally, in many countries it is not clear at all what organization will be the DSC in the first place, and what the process of appointing Trusted Flaggers will look like. This means NGOs are in the dark about the role they will play because it completely lacks transparency.

Policy suggestions

Our policy suggestions focus on the doubts raised concerning the Trusted Flagger system.

- First of all, funding is needed for those NGOs that decide to be a Trusted Flagger since it implies a lot of extra work for them. Most organizations are small and understaffed and need funds to be able to comply with the new regulations.
- Second, more transparency is needed as soon as possible on the Trusted Flagger system. It is pivotal for NGOs to receive more clarity on the application process in order to have time to be able to work with it.
- Third, there needs to be oversight as to how organizations are chosen on the national level, in order to ensure that they are chosen in a fair manner, that diversity is accomplished and to guarantee that there is no abuse of power taking place by governments. Democracies should enforce principles of transparency

and oversight which in turn are needed to ensure that the impact of online platforms do not go against what they stand for.

Concluding, although the DSA applies to platforms inside the EU, its impact will be global, which is why we must ensure the most successful outcome possible and set an example and global standard in the regulation of this technology that can affect so many around the world.

Image Based Sexual Violence

The second trend that we are reviewing in this policy paper is connected to our previous policy paper when we described Article 612 in Italy to criminalize 'revenge porn'. However, we do not use this term anymore as it does not give an accurate description of the phenomenon. 'Revenge porn' is not only restricted to ex-partners wanting to take revenge. It is a much more complicated and multi-layered issue that requires a review.

Image Based Sexual Violence (IBSV) is an umbrella term that includes different forms of digital violence but always includes the non-consensual making and/or sharing of intimate images, by publishing it online on social media platforms, on porn websites or sharing them in WhatsApp or Telegram groups. But it can also involve using someone's profile picture for the use of deep fake porn, where someone is doing something in videos and/or pictures that he or she is not doing in reality.

IBSV is part of gender-based violence or gender-based hate speech. It entails the creation, theft, extortion, threatened or actual distribution, or any use of sexually explicit or sexualized materials without the meaningful consent of the person/s depicted and/or for purposes of sexual exploitation. It also includes sexual violence or harassment committed towards a person's representation (e.g., a person's avatar) in virtual reality or online gaming (National Center on Sexual Exploitation 2023). The hate that follows on the sharing and exposing of the material usually targets women, women of colour and LGBT+ people.

It is important to realize that there are different methods of obtaining the content, for example by hacking, consensual sharing, coercion, and hidden cameras. Regarding consensual sharing, the lines often blur as well. Pressure and/or alcohol/drugs are often involved here. Many victims are often as young as 13 or 14 when the consensual taking of pictures or videos takes place (Huber 2022).

Research shows increasing levels of misogyny online. This has become so widespread that it leads to a normalisation of online abuse, with threats of rape and violence against women becoming more common. There also seems a persistent assumption that online sexual violence is less significant compared to sexual violence offline. In reality, the consequences and traumas caused by IBSV are grave and more far reaching than one might assume. When images are distributed, victims often become facially recognisable and personal identifying information, including names, links to social media profiles, telephone numbers and locations are disclosed. In some cases, images are also sent to employers, co-workers, families, and friends. This public identification is inevitably increasing victims' chances of further abuse. Research concludes that 49% of victims had been stalked and harassed online by those who had seen the material. Anonymity is unequally distributed here; the perpetrators can hide their identity easily online while

they are explicitly exposing the victims with their identity (Huber 2022).

With IBSV more specifically, it is argued that the impact of victimisation may become amplified due to the permanence of online material. Something said in public can potentially be forgotten but the sharing of material on the internet is a permanent feature. Research has identified a range of emotional, physical, and financial consequences for victims: suicide, shame, humiliation, reputational damage, forced changes or loss in occupation, fear, anxiety, and an increased vulnerability to further harassment and/or abuse. The ease and normality of using technology plays a fundamental role in obtaining images, but it also increases the ease with which images can be obtained non-consensually (Huber 2022).

Victim blaming with IBSV is very common. Generally, parents and educators advise youngsters not to take any nudes and send them to someone. If they still do it and they get abused, victims feel that it was their own fault because they were warned not to do so in the first place. This type of victim blaming is comparable to asking a girl what she wore when she was being raped. But sexting is a modern form of flirtation, and it should be possible to take place in a safe manner where everyone's rights are protected. The boundaries between the offline and online world are becoming increasingly non-existent and therefore, to understand victimisation in today's world, it is crucial we pay more attention to this interwoven relationship (Huber 2022). Not only is this realization essential, but also laws are needed that recognize the digital reality.

The issue of IBSV proves there is no division between the online and offline reality. There is a need to recognize and further research how our society is a digital one, in which the interrelationship between society and technology requires us to not restrict our understanding of technology as something that can be separated from our everyday lives. On 8 March 2022, the European Commission adopted a proposal for a directive on combating violence against women and domestic violence. According to the EU:

'The proposal sets out targeted rules for the protection of this group of crime victims in order to strengthen the actions taken by the Member States. It aims to ensure a minimum level of protection across the EU against such violence, regardless of whether it takes place online or offline' (European Commission).

The directive will also ensure that victims have:

- access to justice
- the right to claim compensation
- access to free of charge helplines and rape crisis centres (European Commission)

This directive will also include IBSV and gender-based hate and we are looking forward to seeing the final version of it.

Policy suggestions

It is obvious that IBSV is a complicated, widespread, and multilayered issue. Therefore, interventions on different levels are also necessary.

- First of all, awareness is needed. Education on what IBSV is and everyone's role in it. For example, not everyone who receives and shares pictures in WhatsApp or Telegram groups realizes that they are part of the problem when they do so. It should be addressed that sharing intimate pictures has grave consequences. Also, as a society it is important to be educated on the dangers of victim blaming. In general, we need to accept that sexting is part of modern flirtation culture, but the sharing of it and the hate that follows, is not.
- Moreover, laws are necessary to ask for age and identity verification of the person depicted when creating, uploading, and distributing pornography (National Center on Sexual Exploitation).
- Social media platforms play an essential role here and they could do a lot more. For instance, they should be pressured to have algorithms in place that can identify content that has already been removed once and not let it be published again. Platforms could also have harsher punishments for users who keep regularly sharing illegal content.
- With IBSV the porn platforms are a huge problem. So far, despite being some of the most visited websites in the world, the DSA has not included porn platforms as VLOPs, so the rules do not apply to them. Of course, they should be.
- Platforms like Telegram are not based in the EU or do not have a contact person to reach out to. We recommend them to have a contact person so that it will become easier to discuss matters with them.

We would like to conclude this section with a few initiatives that could contribute to the solution of the problem:

- Here is an interesting initiative asking Google to prevent the re-uploading by quickly removing and preventing IBSA from Google search results, and to stop surfacing pornography sites for criminal, violent, racist, and incest themed pornography searches: click here
- Panorama Global facilitated the creation of the Reclaim Coalition. This is an international network of survivors, organizations, experts, and hotlines / helplines that cooperate on this relatively new issue. The network includes a survivor-centred approach by working with them as experts on what is needed to confront this issue. Read more about it here.

Artificial Intelligence

The third issue to be discussed in this policy paper is Artificial Intelligence (AI). The new AI programme Chat GPT and the current technological revolution appear daily on the front pages of newspapers. The tone in the media varies from enthusiastic and in awe of the new possibilities for our society, to the fear that AI will become stronger than humans are. We would like to highlight the ways in which AI can help with the detection, countering, and removal of online hate speech. But we would also like to review the concerns that we have regarding AI, especially the biases that it contains.

ChatGPT was created by the company OpenAI. It has shown great potential in performing several tasks, including hate speech detection (Nextias N.D.). Detection of hate speech by human moderation requires a huge workforce but even then, it is still impossible to screen the great amount of content that appears every second on social media platforms. Besides that, the personal impact for people watching the extreme content is immense. One can only imagine what it must be like to see the shocking variety of illegal content; many moderators struggle with PTSD. Therefore, the assistance of AI with detection is of immense help. The speed and the amount of content that can be reviewed with AI is unprecedented. Social media platforms already use it, usually in combination with human moderation.

Research and experiments show how AI can not only help to detect, but also to counter hate speech. The organization Samurai Labs developed the first 'Cyber Guardian.' This is a chatbot that shows empathy and gives out friendly warnings to hate posters online (Samurai Labs N.D.). This is effective and leads to haters to change their tone once they feel seen by someone. It also adds motivation to other users to do the same because it encourages others to stand up against haters if they see it happening.

Another research showed the potential of 'hope-' or 'help speech'. AI was developed that would detect positive comments regarding the Rohingya refugees in Myanmar. Instead of focusing on detecting and removing hate, the researchers argued that there is a lot to say for finding the positive comments and amplifying their reach instead. The hope is that, especially social media platforms, will take this into consideration. If hope speech becomes more visible instead of hate speech, it will change the online environment drastically (Waters 2020).

INACH is working on its very own 'Cyber Hate Neutralization Hub' which includes an algorithm that will be able to show the origins and patterns of online hate. It will be able to show the trends of online hate, the bots and accounts and the events that trigger it. This will be of essential help to countering and preventing online hate.

However, a huge drawback of AI when it comes to detection of hate speech is there are plenty of ways to circumvent detection. AI is able to spot the patterns of hate speech

based on word vectors and the positions of words with certain connotations. Thus, it is easier to spot emerging hate speech that went undetected earlier, as current politics or social events may trigger new forms of online aggression. Unfortunately, people spreading hate have shown serious determination to overcome automated systems of spotting hate speech by combining common ways of fooling machines (like using acronyms and euphemisms) and perpetuating hate (Budek 2019). OpenAI added rules or guardrails to help ChatGPT avoid problematic answers from users asking the chatbot to, for example, commit crimes or offer any Nazi ideology. However, users found it extremely easy to get around this by rephrasing their questions or simply asking the program to ignore its guardrails, which prompted responses with questionable, and discriminatory, language (Piantadosi 2022) (Getahun 2023). This certainly means that not enough effort has been made yet to intensively test and improve the guardrails that are in place so far.

Also, it remains unclear how AI will perform for low-resource languages. Most of what we know about the software concerns the English language. Additionally, the model's ability to distinguish between protected and non-protected target groups is more effective in English compared to non-English languages. This leads to the misclassification of abusive content towards non-protected groups as hate speech for non-English languages (Das, Pandey and Mukherjee). The model's ability to classify posts targeting specific communities varies based on the language. Hence, further research is needed to achieve adequate performance across all target communities (Das, Pandey & Mukherjee). In addition, hate speech detection can be tricky when complex emotions, actions, and intentions appear in the post. While the ChatGPT model's performance is excellent for detecting hateful posts, it fails to identify non-hateful counter speeches and often misclassifies them as hate speech. Counter speech plays a crucial role in mitigating the spread of hate speech, and mislabelling counter speeches as hate speech would unjustly impact users engaging in counter speech activities (Das, Pandey, Mukherjee). Also, hate speech detection with AI will remain challenging in the future because AI needs to be continuously trained on hate speech due to the fact that hate speech changes all the time over time as well (Eliot 2023).

As a language model, ChatGPT does not have the capacity for ethical reasoning or decision-making. Thus, it can be used for a variety of purposes, some of which may raise ethical concerns. This could be a problem regarding disinformation and conspiracy theories. ChatGPT may generate text that is factually incorrect or misleading, especially when it is used to generate news articles, social media posts, or other forms of content that can spread rapidly online. Hate speech is often mixed with disinformation and conspiracy theories and therefore this is very concerning.

Finally, AI is often biased due to the data it is trained with. ChatGPT may perpetuate and even amplify biases present in the data it was trained on. This can lead to unfair and inaccurate predictions or generated text **(**Nextias N.D.**)**. It is a crucial obstacle because if

the foundation of the developed software is unethical, we are welcoming a broken product into the world without worrying about the consequences of its functioning. Especially with online hate speech it is of utmost importance to trust that the AI is as unbiased as possible. Of course, that is a high bar to set but where would we end up if we do not have high expectations of these new technologies? If AI will be used in all kinds of aspects of our society - administrative, law enforcement, educational, medical - we run the risk of introducing discriminatory biases there as well. Progress in our societies will be extremely difficult if we allow biased AI into our systems. Not everyone working in AI seems to be conscious of their responsibility to fix the obstacle of biased training data. Some AI experts reason that the biased training data for AI is a reflection of our society, it is a reality check of how we are, and therefore inevitable. Sean McGregor, the founder of the Responsible AI collaborative, told Insider that biased data is inevitable and OpenAI's release of ChatGPT allows people to help make the "guardrails" that filter biased data more robust: "You can do your best to filter an instrument and make a better dataset, and you can improve that. But the problem is, it's still a reflection of the world we live in, and the world we live in is very biased and the data that is produced for these systems is also biased." (Getahun 2023).

However, other AI experts do advise to prioritize these ethical issues over everything else. And indeed, it seems ludicrous as an industry to excuse one's responsibility for the effects of the products that it produces, as if they are not able to fix it. Because if that is the case, should we not argue to stop using AI in the first place? Even leaders in AI development have questioned the same. At the beginning of 2023, a letter was published asking to hold off the development of AI due to concerns of ethical issues. More than a thousand tech leaders and researchers - among them, Elon Musk - signed the open letter calling for a six-month pause in developing the most powerful AI systems. They noticed that AI technology is getting so good that technology experts are worrying about the profound negative impacts it could have on society (Mohammed, Jarenwattananon & Summers 2023). So far, it does not seem that this letter had any impact on pausing the developments.

There are numerous factors to take into account when it comes to the use of AI on detection and moderation of content. How algorithmic systems shape the visibility and promotion of content on social media platforms, and its societal and ethical impact, is an area of growing concern. For example, in 2019 engineers at Meta discovered that a moderation algorithm at Instagram was 50% more likely to ban black users than white users (Solon 2020). Other research showed that the hate speech detection systems Meta used on Facebook aggressively detected comments denigrating white people more than attacks on other demographic groups (Dwoskin, Tiku & Timberg 2021) (Wiggers 2021). Measures adopted under the DSA call for algorithmic accountability and transparency audits. This means a lot for the hate speech field. We know that content that generates a lot of emotions are amplified, among them are hate speech messages. That means

that platforms make profit off of hate and of course that is an obstacle to removing content. Recently, the European Centre for Algorithmic Transparency (ECAT) was created. The purpose of this organization is to contribute to a safer, more predictable, and trusted online environment for people and business. Up until now, it has been a huge obstacle that social media platforms have not been transparent on how their algorithms work. There are assumptions but there is no data available to show their functioning due to the competitiveness between companies. Social media platforms consider their algorithms business secrets, and they do not want the competition to know how they tweak it. With ECAT and the DSA there is the opportunity to obtain more insights into this. Ultimately, the solution to hate speech lies in achieving higher standards of transparency and accountability on these platforms. Because so far, data mining processes and the algorithms that promote content are untransparent, content moderation decisions are often inconsistent, and the ability to audit data is severely circumscribed (Havlicek 2023).

Before we proceed to the policy suggestions regarding AI, we would like to highlight an interesting development that is in process. The European Union is working on a law that targets AI: the EU AI Act. It will be the first law on AI by a major regulator. The plan is that the law will assign applications of AI to three risk categories. First, applications and systems that create an unacceptable risk, such as government-run social scoring of the type used in China, are banned. Second, high risk applications, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated (Future of Life Institute N.D.).

Policy suggestions

We would like to end this section with a few policy suggestions regarding AI.

- First of all, it is unrealistic to expect businesses to self-regulate. Their focus lies on generating profit. Therefore, regulations from governments and international organizations are very much needed. These regulations should focus on putting in place ethical standards that AI would need to live up to. This to ensure that AI functions without any biases before going public. The European AI Act is an opportunity for this. We underline the need for ethical standards that ensure the training data that is used does not contain any biases and that AI is tested on possible biases and effort has been made to remove any biases before using it in any form. To say that biased AI is just a reflection of our biased society is irresponsible, unfair and a failing attempt in simplifying a very new and complex technology that needs to function in an equally complex society. Attention, thought, and time needs to be given to developing AI. It is of essential importance to ensure that minority groups are protected from any bias. If technology is to

help, develop and enlighten societies, then it can only do so if the safety and rights of all groups of people are guaranteed.

- Second, we have high expectations of the EU Data Act and ECAT and the transparency it will hopefully cause. This is at last an opportunity to gain the much-needed transparency organizations have asked for for a long time. Social media platforms have always been reluctant to be open about how their algorithms work. We know that content with strong emotions gets pushed to the top of one's feed and that algorithms pick up one's preferences to what type of content they look at. However, much more insights into the functioning of the algorithms are needed. Especially, in order to understand how online hate works and in that way being able to prevent- and counter it.
- Finally, much more research and development are needed on how AI tools can help us to identify hate patterns in order to be able to counter it. But also, in order to be able to identify the silent majority groups who need to be convinced of stepping up against online hate speech. We have given a few examples of what is possible with AI to work in the hate speech field, but more knowledge and experiments is necessary to use AI for good. Of course, NGOs can carry out the work, but funds provided by governments and international organizations are needed in order to help out with that.

Disinformation

In our previous policy paper, we wrote about fake news. The COVID-19 pandemic showed how incorrect information leads to distrust, conspiracy theories, protests, and disobedience. However, we would not like to discuss the issue anymore by labelling it as 'fake news'. Since extreme right politicians dismiss facts as 'fake news' it has become an empty word. Therefore, we prefer to talk about disinformation. There remains a lot of confusion on the difference between the concepts misinformation, disinformation and malinformation. All three are quite different concepts, but in this section, we will focus on disinformation: the deliberate and purposeful distribution of false information. Usually, disinformation is used as a campaign to disrupt and influence public opinion. It is especially used by political movements that try to attract voters.

Disinformation has always existed and since the invention of the internet, the medium has made sure it gets spread at an unprecedented speed throughout the world. But since the COVID-19 pandemic, disinformation has increased steeply. The distrust between citizens and governments led to a huge distrust towards facts about COVID-19 as well. A new disease of which very little was known was a great opportunity for the spread of disinformation. In other words, the truth has become a subjective phenomenon and even science has become an area to be questioned and doubted. Today, and in the near future, we will not know anymore what information is real and what is not. Due to technological development, this does not only go for written information but also for images and videos. It increases the distrust between citizens and governments even more and therefore undermines the institute of democracy.

Disinformation shows up in almost any political issue ranging from climate change to refugees and hate speech is almost always intertwined with it. For example, the amount of hate that Greta Thunberg has received as a consequence of her protesting to demand measures against climate change is overwhelming. Especially extreme right groups like to question the accuracy of climate change caused by humans and they target Greta Thunberg with a huge amount of disinformation trying to prove that climate change is just extreme weather. Greta has been called a leftist pawn and a Nazi. Conspiracy theories have also been used, suggesting that she is a puppet 'exploited by sinister forces'. Narratives have focused on a link between her and George Soros. Other narratives focus on her mental state or are clearly misogynist (Serhan 2021).

The secretary-general of the United Nations, Antonio Guterres, recently recognized the danger of disinformation. He said his intention is to set up: '…guidelines and guardrails for governments to promote facts, while exposing conspiracies and lies, and safeguarding freedom of expression and information' (Fowler 2023). He also referred to the fact that disinformation is: '…' weaponised information, in sowing confusion, feeding hate, inciting violence and can actually prolong conflict.'(Fowler 2023)

There is a clear connection between disinformation and hate speech. Specifically in hate speech, a division between the in-group and out-groups is cultivated:

'These out-group descriptions are based on stereotypes and are founded on negative associations rather than empirical evidence or expert knowledge. In addition, hostile and damaging terms are used to describe the other, which lack an empirical basis. Communicators of hate speech may deliberately disseminate incorrect information related to out-groups to cultivate polarized divides in society and to create support for their radical right-wing issue positions' (Hameleers, van der Meer & Vliegenthart 2021).

The above quote becomes very visible in practice when it comes to the refugees coming into the European Union. Disinformation, half-truths and flat out lies about refugees confirm the stereotype of refugees in the EU. It feeds the idea that they cause trouble, that Europe is too full, that they steal our jobs and rape our women. It feeds into the hate against refugees and changes the attitude of politicians. Politics across Europe are shifting to the right when it comes to refugees (and other issues) to attract voters. Politicians use this discourse and feed that type of disinformation to win elections (Robinson 2023). However, it means that anti-refugee hate is a persistent phenomenon which INACH and other NGOs of our network notice in the daily monitoring work. Xenophobia and anti-refugee hate is always one of the main hate speech types that appear on social media platforms (INACH 2022).

According to a report written by INACH member Jugendschutz.net, Telegram serves as a switch board for the extreme right, especially for distributing disinformation and their so called 'information war':

'Telegram has become a key platform for the distribution of disinformation. This was true before the Corona pandemic and has intensified since it began. On this messenger service, enormous numbers of articles and videos are disseminated by self-designated "alternative media." In addition, there is supposedly "revelatory" information posted by right-wing extremists along with quotes from mainstream media that seem to corroborate their ideological position. Anything that corresponds with their worldview bears the promise of achieving wide resonance and being seen as plausible. Disagreement or other points of view are seldom to be found. Right-wing extremists use Telegram to initiate and coordinate hate campaigns and online attacks. Political opponents and other persons considered undesirable are designated as targets. Corresponding propaganda material, for example in the form of vilifying memes, are made available as files. Thus, Telegram serves as a "switchboard" for managing the self-declared "information war." At the same time, Telegram is a significant, international networking node for right-wing extremists of various persuasions. Sharing content or direct links to other groups and channels gives rise to an extensive network of postings and contacts (Jugendschutz.net 2021).'

In 2018 the European Union set up the EU Code of Practice on Disinformation and then set up the strengthened Code of Practice on Disinformation in 2022. The Code set up commitments between the EU and a range of different stakeholders. These commitments include demonetising the dissemination of disinformation, transparency of political advertising and cooperation with fact checkers (European Union 2022). This Code is completely different from the CoC on Illegal Hate Speech Online since it seems that civil society is not included in monitoring the commitment of the social media platforms to this Code. It is more up to the platforms to show what they have done to prevent amplification of disinformation.

Finally, we would like to highlight a recent research project INACH participated in with the VU University of Amsterdam. INACH provided a research question to the students to research: 'What to do against **legal harmful speech** if it cannot be removed based on laws and rules? From the perspective of NGOs, international organizations, governments, and social media platforms, what can be done to prevent and/or counter harmful hate speech?'. They provided the following policy recommendations:

1. More education and awareness are needed about harmful hate speech. Not only targeting citizens but especially to train authorities, law enforcement and social media platforms.
2. More research is needed on the trends of harmful hate speech. Not just broad research but also on local trends and 'smaller topics.' Hate speech is usually in line with current national political debates or social events and therefore it is essential to research that in order to map out the hate and gain a deeper understanding of it.
3. A change of social norms is needed. Hate speech should not necessarily be divided in illegal and legal sections. Hate speech should be considered as harmful, regardless of its illegality or legality. This change of social norms can only be reached with a holistic approach.
4. Advocacy by civil society on setting up broad soft law is needed. With that we mean: soft law that is broad enough to stand a chance to be adopted by all the countries in the EU and would be flexible to adjust to existing local context across European countries. There is already hard law and soft law regarding illegal online hate speech present, but there is opportunity to lobby for soft law regarding harmful hate speech as well.
5. One needs to be aware of the politicization of hate speech; citizens and groups of people who are already inclined to distrust authorities will only feel reinforced in believing in conspiracy theories. Legal action therefore can come across as censorship and generate more hate as a counter effect. More awareness raising about this is needed, as well as more research into how to prevent this from happening while at the same time protecting minority groups.
6. Governments in general, and politicians especially, need to be aware of the exemplary role they play ethically. The use of disinformation, conspiracy theories

and hate in order to attract voters has been normalized. Politicians know where to stay within the line of what is legal while at the same time contributing to the division and distrust between citizens and governments. It has been an accepted strategy. There is room for civil society to call governments out on their responsibility.

Policy suggestions

We will end this part by giving a few policy suggestions that can contribute to the solution of the problems that surround disinformation, specifically regarding the intertwined relationship between online hate speech and disinformation.

- Educational programmes and projects from an early age about digital media literacy are pivotal in raising citizens who will be able to know the difference between trustworthy news sources, be able to recognize patterns of hate and recognize when disinformation campaigns are happening. It is pivotal to create a society where citizens will be able to protect democracy.
- Second, support is needed for fact checking NGOs because their work is pivotal. Society needs organizations that keep highlighting what is incorrect, what is fabricated or a symptom of an underlying strategy.
- Third, extreme pressure on platforms is needed to not boost disinformation via their algorithms. The business models of social media platforms are based on the content that attracts extreme emotions and will be pushed to the top of everyone's feeds to keep users longer on the platforms. Disinformation cannot be used for that purpose. If it does not get removed by platforms, the very least they can do is make sure it does not get amplified by the algorithms.
- Finally, awareness is needed of the politicization of hate speech; citizens and groups of people who are already inclined to distrust authorities will only feel reinforced in believing in conspiracy theories. Legal action therefore can come across as censorship and generate more hate as a counter effect. More awareness raising about this is needed, as well as more research into how to prevent the spread of disinformation and hate speech and protecting minority groups.

INACH's policy recommendations

Now that we have reviewed the different issues, we will describe the policy recommendations we have. But first, let us review the policy recommendations we had in 2021.

These were our policy recommendations from 2021.

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.

Recommendation 2 regarding AI and moderation, we would like to highlight the importance of it and underline that this is still very much up to date. But since 2021 many technological developments have been taking place on AI, regulations from governments and international organizations are very much needed. These regulations should focus on putting in place ethical standards that AI would need to live up to. This to ensure that AI functions without any biases before going public. The European AI Act is an opportunity for this. We underline the need for ethical standards that ensure the training data that is used does not contain any biases and that the AI is tested on possible biases and efforts have been made to remove any biases before using it in any form.

Recommendation 3 regarding staying up to date on new developments, INACH has started with organizing member webinars to share knowledge. It also organizes Roundtables together with the INACH members and social media platforms to exchange knowledge and discuss issues.

Recommendation 7 regarding the adherence of the social media companies to the CoC is still very valid. However, we are looking forward to an updated CoC that should improve the process of addressing online hate speech.

Recommendation 8 and 9 regarding the EU considering taking tougher measures if the CoC does not have the intended results and the new German law, does not apply really anymore since the introduction of the DSA.

Recommendation 11 regarding training moderators of social media companies is still very much important today. This training should not only focus on illegal online hate speech but also include legal harmful hate speech, disinformation, and conspiracy theories.

A few policy recommendations should be added to the list:

1. Regarding the DSA, funding is needed for those NGOs that decide to be a Trusted Flagger since it implies a lot of extra work for them. Most organizations are small and understaffed and need funds to be able to comply with the new regulations.
2. More transparency is needed as soon as possible on the Trusted Flagger system. It is pivotal for NGOs to receive more clarity on the application process in order to have time to be able to work with it.
3. There needs to be oversight as to how organizations are chosen on the national level, in order to ensure that they are chosen in a fair manner and that diversity is

guaranteed. And to ensure that there is no abuse of power taking place by governments.

4. Regarding IBSV, awareness is needed. Education on what IBSV is and everyone's role in it. For example, not everyone who receives and shares pictures in WhatsApp or Telegram groups realizes that they are part of the problem when they do so. It should be addressed that sharing intimate pictures has grave consequences. Also, as a society it is important to be educated on the dangers of victim blaming. In general, we need to accept that sexting is part of modern flirtation culture, but the sharing of it and the hate that follows, is not.
5. Moreover, laws are necessary to ask for age and identity verification of the person depicted when creating, uploading, and distributing pornography (National Center on Sexual Exploitation).
6. Social media platforms play an essential role here and they could do a lot more. For instance, they should be pressured to have algorithms in place that can identify content that has already been removed once and not let it be published again. Platforms could also have harsher punishments for users who keep regularly sharing illegal content.
7. With IBSV the porn platforms are a huge problem. So far, despite being some of the most visited websites in the world, the DSA has not included porn platforms as VLOPs, so the rules do not apply to them. Of course, they should be.
8. Platforms like Telegram are not based in the EU or do not have a contact person to reach out to. We recommend them to have a contact person so that it will become easier to discuss matters with them.
9. Regarding AI, regulations from governments and international organizations are very much needed. These regulations should focus on putting in place ethical standards that AI would need to live up to. This to ensure that AI functions without any biases before going public. The European AI Act is an opportunity for this. We underline the need for ethical standards that ensure the training data that is used does not contain any biases and that the AI is tested on possible biases and effort has been made to remove any biases before using it in any form.
10. Much more research and development are needed of AI tools that can help us to identify hate patterns in order to be able to counter it. But also, in order to be able to identify the silent majority groups who need to be convinced of stepping up against online hate speech. Of course, NGOs can carry out the work, but funds provided by governments and international organizations are needed in order to help out with that.
11. Educational programmes and projects from an early age about digital media literacy are pivotal in raising citizens who will be able to know the difference between trustworthy news sources, be able to recognize patterns of hate and recognize when disinformation campaigns are happening. It is pivotal to create a society where citizens will be able to protect democracy.
12. Support is needed for fact checking NGOs because their work is pivotal. Society

needs organizations that keep highlighting what is incorrect, what is fabricated or a symptom of an underlying strategy.

13. Extreme pressure on platforms is needed by International Organisations and governments to not boost disinformation via their algorithms. The business models of social media platforms are based on the content that attracts extreme emotions and will be pushed to the top of everyone's feeds to keep users longer on the platforms. Disinformation cannot be used for that purpose. If it does not get removed by platforms, the very least they can do is make sure it does not get amplified by the algorithm.
14. Awareness about the politicization of hate speech is needed; citizens and groups of people who are already inclined to distrust authorities will only feel reinforced in believing in conspiracy theories. Legal action therefore can come across as censorship and generate more hate as a counter effect. More awareness raising about this is needed, as well as more research into how to prevent the spread of disinformation and hate speech and protecting minority groups.

So, summarizing these are our policy recommendations of 2023:

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. Regarding AI, regulations from governments and international organizations are very much needed. These regulations should focus on putting in place ethical standards that AI would need to live up to. This to ensure that AI functions without any biases before going public. The European AI Act is an opportunity for this. We underline the need for ethical standards that ensure the training data that is used does not contain any biases and that the AI is tested on possible biases and effort has been made to remove any biases before using it in any form.**
3. **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.**
4. **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.**
5. **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.

6. **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. We are looking forward to the new Code of Conduct that is being negotiated with the platforms.**
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.**
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.**
9. **Regarding the DSA, funding is needed for those NGOs that decide to be a Trusted Flagger since it implies a lot of extra work for them. Most organizations are small and understaffed and need funds to be able to comply with the new regulations.**
10. **More transparency is needed as soon as possible on the Trusted Flagger system. It is pivotal for NGOs to receive more clarity on the application process in order to have time to be able to work with it.**
11. **There needs to be oversight as to how organizations are chosen on the national level, in order to ensure that they are chosen in a fair manner and that diversity is guaranteed. And to ensure that there is no abuse of power taking place by governments.**
12. **Regarding IBSV, awareness is needed. Education on what IBSV is and everyone's role in it. For example, not everyone who receives and shares pictures in WhatsApp or Telegram groups realizes that they are part of the problem when they do so. It should be addressed that sharing intimate pictures has grave consequences. Also, as a society it is important to be educated on the dangers of victim blaming. In general, we need to accept that sexting is part of modern flirtation culture, but the sharing of it and the hate that follows, is not.**
13. **Moreover, laws are necessary to ask for age and identity verification of the person depicted when creating, uploading, and distributing pornography (National Center on Sexual Exploitation).**
14. **Social media platforms play an essential role here and they could do a lot more. For instance, they should be pressured to have algorithms in place that can identify content that has already been removed once and not let it be published again. Platforms could also have harsher punishments for users who keep regularly sharing illegal content.**
15. **With IBSV the porn platforms are a huge problem. So far, despite being some of the most visited websites in the world, the DSA has not included porn platforms as VLOPs, so the rules do not apply to them. Of course, they**

should be.

16. **Less mainstream platforms are often not based in the EU or do not have a contact person to reach out to. We recommend them to have a contact person so that it will become easier to discuss matters with them.**
17. **Educational programmes and projects from an early age about digital media literacy are pivotal in raising citizens who will be able to know the difference between trustworthy news sources, be able to recognize patterns of hate and recognize when disinformation campaigns are happening. It is pivotal to create a society where citizens will be able to protect democracy.**
18. **Support is needed for fact checking NGOs because their work is pivotal. Society needs organizations that keep highlighting what is incorrect, what is fabricated or a symptom of an underlying strategy.**
19. **Pressure by International Organisations and governments is needed on platforms to not boost disinformation via their algorithms. The business models of social media platforms are based on the content that attracts extreme emotions and will be pushed to the top of everyone's feeds to keep users longer on the platforms. Disinformation cannot be used for that purpose. If it does not get removed by platforms, the very least they can do is make sure it does not get amplified by the algorithm.**
20. **One needs to be aware of the politicization of hate speech; citizens and groups of people who are already inclined to distrust authorities will only feel reinforced in believing in conspiracy theories. Legal action therefore can come across as censorship and generate more hate as a counter effect. More awareness raising about this is needed, as well as more research into how to prevent the spread of disinformation and hate speech and protecting minority groups.**

Reference List

Alba, Davey, 'OpenAI chatbot spits out biased musings, despite guardrails', 8 December 2022, Bloomberg, ChatGPT, Open AI's Chatbot, Is Spitting Out Biased, Sexist Results - Bloomberg, accessed on 19-07-2023

Amnesty International, Global: Social media companies must step up crisis response on Israel-Palestine as online hate and censorship proliferate - Amnesty International, October 27 2023

Budek, Konrad, 'How Artificial Intelligence can fight hate speech in social media', Deepsense.ai, How artificial intelligence can fight hate speech in social media - deepsense.ai, accessed on 19-07-2023

Das, Mithun, Pandey, Saurabh Kumar & Mukherjee, Animesh, 'Evaluating ChatGPT's performance for multilingual and emoji-based hate speech detection', Indian Institute of Technology, 2305.13276.pdf (arxiv.org), accessed on 19-07-2023

Dwoskin, Elizabeth, Tiku, Nitasha, Timberg, Craig, ' Facebook's race-blind practices around hate speech came at the expense of black users, new documents show', 21 November 2021, Washington Post, Facebook knew its algorithms were biased against people of color - The Washington Post, accessed on 19-07-2023

Eliot, Lance, 'How hard should we push generative AI ChatGPT into spewing hate speech, asks AI ethics and AI law', 5 February 2023, Forbes, How Hard Should We Push Generative AI ChatGPT Into Spewing Hate Speech, Asks AI Ethics And AI Law (forbes.com), accessed on 19-07-2023

European Commission on the DSA, The Digital Services Act: ensuring a safe and accountable online environment (europa.eu), accessed on 19-07-2023

European Commission on the DSA, DSA: Very Large Online Platforms and Search Engines (europa.eu), accessed on 19-07-2023

European Commission on Gender Based Violence, Ending gender-based violence (europa.eu), accessed on 19-07-2023

European Commission, '2022 strengthened code of practice on disinformation', 16 June 2022, 2022 Strengthened Code of Practice on Disinformation | Shaping Europe's digital future (europa.eu), accessed on 23-07-2023

Fowler, Olivia, 'How misinformation, disinformation and hate speech threaten human progress', Impakter, 20 June 2023, How Misinformation, Disinformation and Hate Speech Threaten Human Progress (impakter.com), accessed on 19-07-2023

Future of Life Institute, 'What is the EU AI Act?', The Artificial Intelligence Act |, accessed on 19-07-2023

Getahun, Hanna, 'ChatGPT could be used for good, but like many other AI models, its rife with racist and discriminatory bias', Insider, ChatGPT Is Like Many Other AI Models: Rife With Bias (insider.com), accessed on 19-07-2023

Hameleers, Michael, van der Meer, Toni & Vliegenthart, Rens, 'Civilized truths, hateful lies? Incivility and hate speech in false information - evidence from fact checked statements in the US', 10 February 2021, Information, Communication & Society, Full article: Civilized truths, hateful lies? Incivility and hate speech in false information – evidence from fact-checked statements in the US (tandfonline.com), accessed on 19-07-2023

Havlicek, Sasha, 'We need transparency, not censorship, to address hate speech and other harms on social media', 17 APril 2023, Institute for Strategic Dialogue, We need transparency, not censorship, to address hate speech and other harms on social media - ISD (isdglobal.org), accessed on 19-07-2023

Huber, Antoinette, 'A shadow of my old self: The impact of Image Based Sexual Abuse in a digital society', Sage Journals, 29 April 2022, 'A shadow of me old self': The impact of image-based sexual abuse in a digital society - Antoinette Huber, 2023 (sagepub.com), accessed on 19-07-2023

Hundt, Andrew, Kacianka, Severin, Agnew, William, Gombolay, Matthew & Zeng, Vicky, 'Robots enact malignant stereotypes', 21-24 June 2022, FAccT, https://dl.acm.org/doi/pdf/10.1145/3531146.3533138, accessed on 22-07-2023

INACH, 'INACH Monitoring Report 2022', compiled by Adinde Schoorl & Maia Feijoo and edited by Tamas Berecz, ME and shadow ME 2022 – INACH, accessed on 4 August 2023
Jugendschutz.net, 'Right-wing Extremism on the internet', November 2021, 2020/2021 Report – Right-wing Extremism on the Internet – INACH, accessed on 23-072023

Keller, Daphne, 'What does the DSA say?', 25 April 2022, The Center for Internet and Society, What Does the DSA Say? | Center for Internet and Society (stanford.edu), accessed on 31-07-2023

Mohammed, Linah, Jarenwattananon & Summers, Juana, 'An open letter signed by tech leaders, researchers propose delaying AI development', heard on 'All Things Considered', 29 March 2023, NPR, An open letter signed by tech leaders, researchers proposes delaying AI development : NPR, accessed on 19-07-2023

National Center on Sexual Exploitation, '5 types of Image-Based Sexual Abuse you should know about' 11 April 2023, 5 Types of Image-Based Sexual Abuse You Should Know About (endsexualexploitation.org), accessed on 19-07-2023

National Center on Sexual Exploitation, Image-Based Sexual Abuse - NCOSE (endsexualexploitation.org), accessed on 19-07-2023

Nextias 'Does ChatGPT have an ethical problem?', Does ChatGPT have an Ethical Problem? - NEXT IAS, accessed on 19-07-2023

Piantadosi, Steven, on Twitter steven t. piantadosi on Twitter: "Yes, ChatGPT is amazing and impressive. No, @OpenAI has not come close to addressing the problem of bias. Filters appear to be bypassed with simple tricks, and superficially masked. And what is lurking inside is egregious. @Abebab @sama tw racism, sexism. https://t.co/V4fw1fY9dY" / Twitter, accessed on 19-07-2023

Robinson, Nick, 'How the tide of migration is changing European politics', BBC, 15 July 2023, How the tide of migration is changing European politics - BBC News, accessed on 22-07-2023

Samurai Labs, The World's First Cyber Guardian by Samurai (samurailabs.ai), accessed on 22-07-2023

Serhan, Yasmeen, 'When the Far Right picks fights with a teen', 14 August 2021, The Atlantic, How the Right Lost Its Mind Over Greta Thunberg - The Atlantic, accessed on 22-07-2023

Solon, Olivia, 'Facebook ignored racial bias research, employees say', NBC News, 23 July 2020, Facebook ignored racial bias research, employees say (nbcnews.com), accessed on 21-07-2023

Waters, John, 'Carnegie Mellon uses AI to counter hate speech with 'hope speech'', 15 January 2020, Pure AI, Carnegie Mellon Uses AI To Counter Hate Speech with 'Hope Speech' -- Pure AI, accessed on 22-07-2023

Wiggers, Kyle, 'How bias creeps into the AI designed to detect toxicity', 9 December 2021, Venture Beat, How bias creeps into the AI designed to detect toxicity | VentureBeat, accessed on 19-07-2023

Wolf, Zachary B., 'AI can be racist, sexist and creepy. What should we do about it?', 18 March 2023, CNN Politics, AI can be racist, sexist and creepy. What should we do about it? | CNN Politics, accessed on 19-07-2023