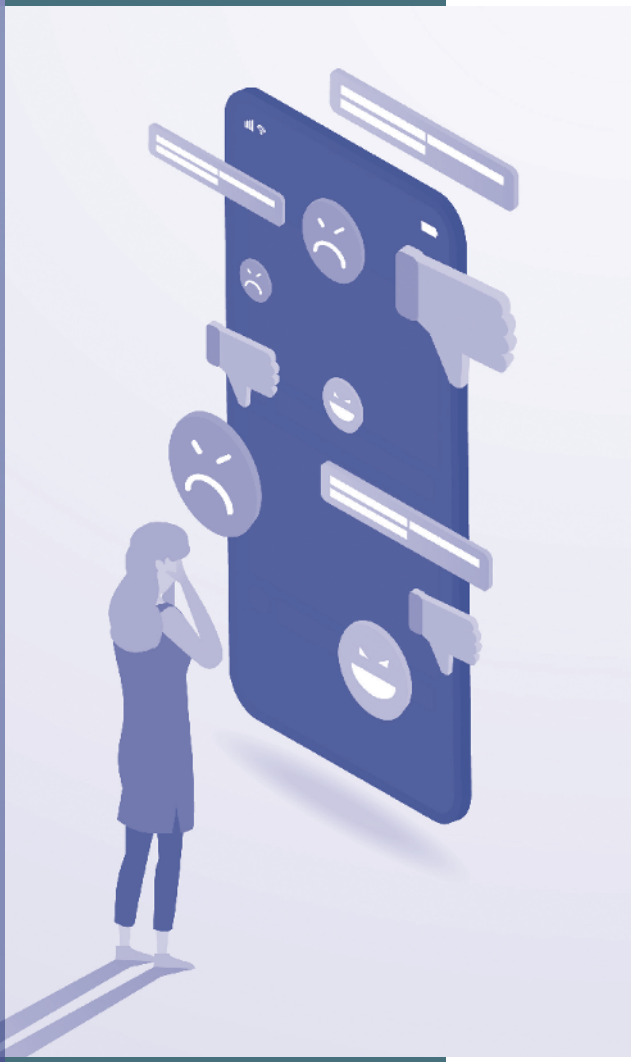
FRA EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS

ONLINE CONTENT MODERATION

CURRENT CHALLENGES IN DETECTING HATE SPEECH

REPORT



FRA

Online Content Moderation – Current challenges in detecting hate speech

Vienna, 2023

Luxembourg: Publications Office of the European Union, 2023

PRINT	ISBN 978-92-9489-259-1	doi:10.2811/332335	TK-04-23-883-EN-C
PDF	ISBN 978-92-9489-258-4	doi:10.2811/923316	TK-04-23-883-EN-N

Photo credits:

Cover: © Lerbank-bbk22 / Adobe Stock
Page 7: © CarlosBarquero / Adobe Stock
Page 10: © Lumeez / peopleimages.com / Adobe Stock
Page 13: © Amgun / Adobe Stock
Page 17: © Stokkete / Adobe Stock
Page 22: © Monster Ztudio / Adobe Stock
Page 25: © Aniqpixel / Adobe Stock
Page 29: © Deagreez / Adobe Stock
Page 45: © LIGHTFIELD STUDIOS / Adobe Stock
Page 47: © Icons gate / Adobe Stock
Page 51: © DimaBerlin / Adobe Stock
Page 75: © Drobot Dean / Adobe Stock
Page 80: © ParinPIX / Adobe Stock
Page 87: © Paul prescott / Adobe Stock

Foreword

Connecting with the world online can be a wonderful way to engage with others and bring us closer as a society. But we know the internet has a darker side, as a space for hate and division. People use online platforms to insult and offend, to harm and threaten.

Hate speech remains a worrying and shameful reality in our societies, and all instances of it – whether online or offline – must be denounced. Increased digitalisation brings new challenges. Online platforms can exacerbate the situation as hate speech spreads rapidly, simply with a click of a button. It is already a serious problem. Innovative solutions are required to tackle this scourge and protect people online.

FRA's research looks in-depth at social media posts directed against women, people of African descent, Jews and Roma. These groups are often the targets of online hate. The research looks at major online platforms and covers four EU languages.

Disturbingly, a large amount of misogyny and racism slips through content moderation systems designed to prevent it. Human content checkers can miss online hate. Also, algorithms are prone to errors. They may multiply errors over time, and may even end up promoting online hate. This is deeply worrying.

Tackling online hate is about protecting the rights of victims of hate speech. With new content generated every second, the sheer scale of moderation is overwhelming and online platforms are struggling to keep up. It is imperative that human rights voices are centrally involved in the design and implementation of moderation measures.

Striking the right balance between freedom of expression and monitoring online content is a challenge. It forces us to reconcile the tension between combatting hate speech and preserving our right to freedom of expression. It is extremely complex. For this reason, we need appropriate and accurate content moderation.

Investing in a variety of measures to tackle hate speech is critical. Only then can we genuinely safeguard people's rights online and create spaces for people to connect, learn, share thoughts and join discussions online.

Let us create a digital environment that protects human rights. We should aspire to build one that enables us to enjoy our rights online rather than curtail them.

Michael O'Flaherty
Director

Abbreviations

AI	artificial intelligence
API	application programming interface
DSA	Digital Services Act
ECHR	European Convention on Human Rights
ECRI	European Commission against Racism and Intolerance
ECtHR	European Court of Human Rights
FRA	European Union Agency for Fundamental Rights
ICCPR	International Covenant on Civil and Political Rights
VLOP	very large online platform
VLOSE	very large online search engine

Contents

Figures and tables

Key findings and FRA opinions

The emergence of online platforms and social media has transformed modern communication. Online platforms provide many opportunities to express opinions and participate in public and political discussions. However, just as offline discussions are replicated or amplified online, so are expressions of hate. This is of increasing concern.

The EU has updated its laws and implemented policies to tackle illegal content online, such as through the Digital Services Act (DSA), to more effectively regulate online content, including hate speech. However, these changes are relatively recent. In addition, there are still uncertainties concerning how to better protect human rights online, with regard to combating online hate while protecting freedom of expression, and how to efficiently implement existing and newly developed laws.

This report aims to better understand whether standard tools to address online hate speech, hereafter referred to as 'online hate', are effective by looking at manifestations of online hate after social media platforms have applied their content moderation controls. This report presents findings covering four social media platforms – Telegram, X (formally Twitter), Reddit and YouTube. The platforms were selected based on their accessibility for research purposes, their popularity (i.e. audience reach) and the assumed magnitude of hate speech on them.

The report aims to achieve the following:

- ⋆ It provides a targeted overview and analysis of online manifestations of misogyny and of hate against people of African descent, Jewish people and Roma. It is based on data collected on four online platforms, in four languages, and encompasses four EU Member States: Bulgaria, Germany, Italy and Sweden.
- ⋆ Based on data analysis of social media posts to identify potential online hate that has already gone through the online platforms' content moderation systems, the report offers a critical assessment of the limitations of online content moderation tools in detecting online hate against specific groups. At the same time, the report highlights these very challenges as they relate to researching and measuring online hate.

The study focuses specifically on online hate in social media posts targeted at women, people of African descent, Jews and Roma to explore the limits of online content moderation and the extent of hate speech against these groups. The study collected social media posts over 6 months, using specific keywords that could indicate potential online hate against these target groups (see Annex 1 for example keywords).

The research covers Bulgarian, German, Italian and Swedish. This essentially encompasses the four EU Member States where these languages are mainly spoken: Bulgaria, Germany, Italy and Sweden. These Member States were chosen based on considerations linked to their national policy situations, potential similarities and differences that may have enriched the comparative analysis, and the feasibility of conducting the research. Their selection also considered the inclusion of languages that are less frequently studied in relation to online hate and its moderation.

In total, 344 132 online posts and comments were collected from social media platforms by detecting and filtering posts and comments that contained specific keywords. Inclusion of these keywords may indicate the presence of hate speech. The data collection included original posts and comments reacting to other posts, including comments on YouTube videos.

From these, close to 400 posts were randomly selected for each of the four Member States. Selection used stratified sampling by platform and target group, ensuring that each platform and target group was represented in the posts (1 573 posts in total). Human coders trained to identify potential online hate assessed the posts and categorised them as involving offensive language; denigration; negative stereotyping; or incitement to violence, discrimination and/or hatred (see Chapter 2 and Annex 1).

Half of the posts were assessed by two coders to establish the level of certainty and consistency of the coding. Legal experts also assessed a small subset of the posts to determine whether they would meet the legal definition of incitement.

More than half of the 1 573 relevant and manually analysed posts (53 %) were considered hateful in the sense of falling into at least one of the categories identified. These include elements of incitement to violence, discrimination or hatred; denigration; offensive language; negative stereotyping; or any other hateful content, such as supporting hateful ideologies. Of all posts coded as hateful, almost 85 % were coded as containing offensive language. The coders considered 55 % of hateful posts to express hatred of people based on protected characteristics.

The data collection aimed to mainly search for misogyny (discrimination on the ground of sex) and hatred of people of African descent, Jewish people and Roma (discrimination on the grounds of race, ethnic origin and/or religion). These categories were selected based on the level of online hate against certain groups, as indicated in existing literature and experts' assessments carried out within the framework of this study. The data collection also considered the comparability between the countries where certain groups are targeted. However, the data collection captured instances of hatred of groups on other grounds as well, including hatred of other ethnic groups and regarding political opinion and disability.

Only a limited selection of posts is included in this report, as the language is very offensive. But it is important to include some examples to illustrate the kind of language people face and the types of post identified for the purposes of this report. Offensive language has not been reproduced and has been redacted.

The European Union Agency for Fundamental Rights (FRA) research was only able to analyse posts that had already been through platforms' own content moderation processes. Therefore, it was to be expected that a significant amount of hateful content had already been removed in line with platforms' terms and conditions, including the prohibition of hate speech and harassment. These moderation processes appear to work to some extent, but many posts still get through and could, after further scrutiny, be considered online hate.

The platforms' online content moderation approaches and algorithms are not open to researcher scrutiny. But the fact that the research was able to flag posts that could potentially be categorised as online hate does indicate that these content moderation systems are not capturing all forms of hate. Given that posts were assessed after having supposedly gone through platforms' content moderation systems, it is clear that these tools fail to correctly identify some posts that could constitute online hate, especially misogyny. However, the methodology does not allow conclusions to be drawn about how well current content moderation systems work and how systems vary across platforms.

When it comes to removing content, moderation can make two types of error: online hate not being detected on the one hand and legal content being deleted on the other hand. It is important to address both errors.

The current research does not offer a legal analysis of freedom of expression with regard to the current practices and limitations of platforms' online content moderation processes. Rather, the report focuses on online content that can be categorised as potential hate speech and still exists after each platform's own content moderation process. The report provides some insights into how hate is expressed and how challenging it is to assess and tackle it on current online platforms.

Regulators and online platforms should step up their efforts to offer a safer and non-discriminatory online environment for women and other groups

Misogyny is the most prevalent form of online hate across all the platforms the research covers. For example, the number of posts targeted at women is almost three times that of those targeted at people of African descent across the four countries covered. Still, many hateful posts express hatred of people of African descent, Jewish people and Roma. Posts targeted at Jewish people and Roma were covered in only two of the four countries.

Posts targeted at women most often include denigrating language, which in this context means comparing people to objects or animals. There are also higher levels of incitement to violence against women compared to the other groups, with online violence against women most often based on sexualised violence. Posts targeted at people of African descent, Jews and Roma most often contain negative stereotyping.

FRA OPINION 1

Online platforms should have specific regard to protected characteristics of users in the context of their terms and conditions, content moderation practices and monitoring policies, including addressing sexist online hate. Performance indicators should be in place to record the volume of misogyny online and the effectiveness of content moderation, looking at developments over time.

For very large online platforms (VLOPs), such as X and YouTube, misogyny should be one of the systemic risks considered in the context of the risk assessment and risk mitigation measures required by Articles 34 and 35 of the DSA.

The Council of Europe Convention on Preventing and Combating Violence against Women creates a coherent legal framework for the prevention, support and protection of women from violence – both offline and online. The EU is now party to this convention. EU Member States that have not yet signed and ratified the convention are urged to do so.

Overall, the posts captured in this data collection rarely include more than one target group. Thus, they provide limited evidence on the intersectionality of online hate. Only 5 % of all posts coded as hateful of any of the four target groups include more than one target group.

Almost half of all hateful posts (47 %) were classified as harassment – that is, harassment perceived to be targeted at one or more specific individual(s). Women are particularly often targets: two thirds (67 %) of all hateful posts targeted at women were found to be harassment.

Currently, no specific legal instrument addresses violence against women at the EU level, despite obviously high levels of (online) misogyny. The European Commission, however, adopted a **proposal for a directive on combating violence against women and domestic violence** in March 2022. The draft instrument aims to ensure a minimum level of protection against such violence across the EU. This includes protection from forms of online violence such as intentional non-consensual sharing of intimate or

manipulated material, cyberstalking, cyber harassment and cyber incitement to violence or hate (Articles 7–10).

In addition, the EU ratified the Council of Europe Convention on Preventing and Combating Violence against Women on 28 June 2023. Requirements include that parties encourage the ICT sector and the media to implement policies, set guidelines and introduce self-regulatory standards that will help prevent violence against women. Currently, 37 countries have ratified the convention, with the EU being the 38th party to ratify. Despite the EU's ratification, six EU Member States have not yet ratified the convention: Bulgaria, Czechia, Hungary, Latvia, Lithuania and Slovakia.

FRA called for the ratification of the convention in 2014 when highlighting the magnitude of **violence against women in the EU**. Given the lack of legal clarity around hate speech in the EU, both mentioned developments are important steps towards safeguarding women from online violence.

In addition, the DSA requires VLOPs to assess systemic risks stemming from their design or the functioning of their services (Articles 34 and 35). This includes risks in relation to fundamental rights and gender-based violence, as manifested online. They must also put reasonable, proportionate and effective measures in place to mitigate those risks. Two of the four platforms analysed – X and YouTube – are considered VLOPs under the **DSA's first classification**.

The EU has **proposed adding hate crimes**, including hate speech and gender-based violence, to the areas of crime listed in Article 83(1) of the Treaty on the Functioning of the European Union. The DSA and the proposed directive on combating violence against women and domestic violence are important steps towards a safer online environment for women. However, as they are still in their infancy, with one being a proposal, their **effects are yet to be seen**.

The degree of misogyny found during the research for this report offers a clear indication that both regulators and online platforms should strengthen their efforts to ensure a safer and non-discriminatory environment for women online, including more targeted (legal) action.

A variety of approaches to tackling online hate are needed, including different ways to detect hate speech

FRA OPINION 2

The European Commission and national governments should support, practically and financially, the creation of a wide and heterogeneous network of organisations acting as trusted flaggers to ensure that different forms of online hate are widely and reliably detected. Organisations representing groups with limited resources should not be put at a disadvantage in combating online hate. Users need to be made aware of easy ways to notify companies of hate speech, in line with Articles 16 and 22 of the DSA.

Given that views on what constitutes online hate may differ, a variety of measures to detect and report hate speech are needed. These include training on legal thresholds for identifying online hate for police, content moderators and trusted flaggers. The training could also ensure that platforms do not over-remove content.

The threshold for and description of what constitutes illegal online hate need more clarity. The changing landscape of expressions of hate and the magnitude of online hate mean that those involved in detecting and those responsible for addressing hate speech should not be left in any doubt regarding the rules. The EU and national legislators should consider clearer guidance and rules on what kind of online hate is illegal.

There is no commonly agreed definition of 'hate speech' in EU law, and interpretations vary at the international, EU and national levels. This report shows that objectively categorising hate speech was very challenging for both trained coders and legal experts in the context of this research. The difficulty with categorisation is related to several factors, such as the importance of context in interpreting speech and its intention.

Furthermore, hatefulness is often interpreted differently, so the same speech may be assessed in different ways. The coder's background, for example their sex, age and ethnic origin, can also affect whether they code certain types of speech as hateful or not. For example, **research indicates** that men often deem online content targeted at women to be less offensive than women do.

The level of agreement in the categorisation of posts for this study is relatively high, but the trained coders faced uncertainty when categorising online hate in around one out of five posts. In addition, there was some disagreement between legal experts and coders about the extent to which certain posts might fall under the definition of incitement to violence. Legal experts had a higher threshold for labelling content as incitement. In this regard, the need to uphold freedom of expression while at the same time removing content that can be categorised as online hate is a delicate exercise.

Given the challenge of objectively assessing online hate, a variety of approaches to detecting and responding to online hate have emerged to try to ensure that fundamental rights have a high level of protection. These include engaging with several stakeholders in efforts to detect and report potential hate speech in addition to the efforts of platforms and public authorities, and also involving platforms' users. If provided with adequate resources, civil society organisations can support efforts and contribute to safer online spaces by notifying platforms or authorities of this content.

Article 14 of the DSA lays down obligations for online platforms regarding the content moderation practices set out in their terms and conditions. These obligations include that platforms must act with due regard to the fundamental rights of users. Therefore, Article 14 provides an important basis for providing online platforms with more guidance on how to address fundamental rights in their content moderation practices.

Article 16 of the DSA lays out notice and action mechanisms for online platforms with regard to content. Article 22 provides further provisions for 'trusted flaggers', entities to which the digital service coordinators of the Member States have awarded trusted status. They are meant to reliably notify platforms of abusive and illegal expressions online. However, given the difficulty of assessing whether a post contains online hate, errors in (not) deleting online posts will remain.

The use of artificial intelligence to detect and moderate online hate must comply with fundamental rights

Of the 344 132 online posts collected based on keywords for this report, 1 573 were randomly selected for detailed manual coding. These covered the four languages and all of the countries, platforms and groups included in the research.

The coding took place after the platforms had undertaken their own standard form of online content moderation. Yet online content that can be assessed as potential online hate remains undetected and unactioned. The volume of potential misogyny that has slipped through the platforms' online content moderation indicates that discrimination against women – in the form of online hate – is inherent in the algorithms, as they fail to pick up this form of content.

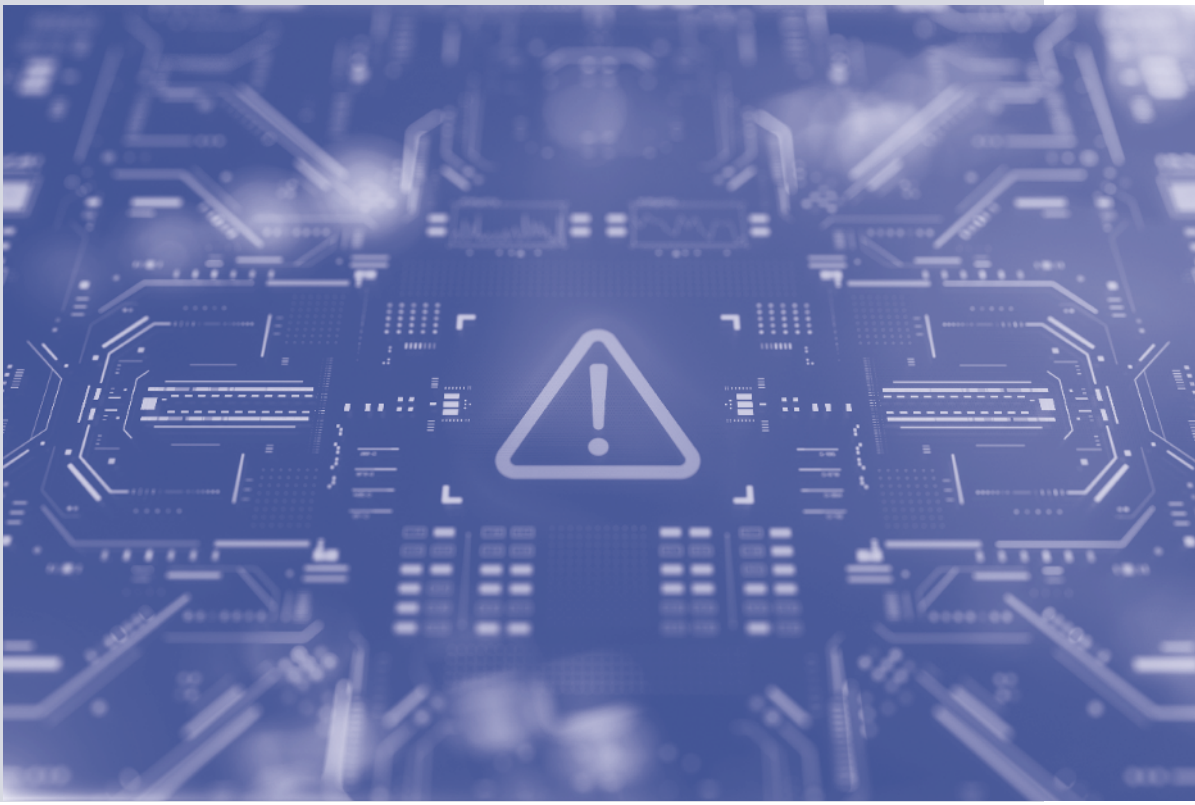
The manual human-based assessment of posts is costly and time consuming and requires extensive training of coders. This was reflected in the current research. At the same time, manual coding is also subject to human disagreement and error, including as a result of bias, as noted in FRA opinion 2. Therefore, it is not a panacea for the potential failings of artificial intelligence (AI).

FRA OPINION 3

It must be ensured that AI-supported online content moderation decisions are not discriminatory. Providers and users (i.e. platforms) must assess the fundamental rights compliance of any AI system in line with the DSA and current and developing standards regulating the use of AI.

The EU should ensure that applicable EU law, such as the DSA, appropriately addresses potential discrimination through the use of AI content moderation and requires that these systems are not used in a discriminatory way.

Data from online platforms have shown that automated detection of online hate has increased considerably in the past few years. Online platforms should ensure that any automated tool deployed for content moderation is as reliable as possible to minimise the rate of errors, in line with the DSA (recital 26). The use of AI may increase the efficiency of content moderation and can scale up tasks that would be impossible to undertake through human review alone. However, online content moderation decisions based on automated means that utilise AI can miss hateful content and be discriminatory.



The 2020 **FRA report on AI and fundamental rights** underlines these issues with AI-based content moderation. Namely, users and providers of AI technologies may not always be aware of the biases and potential discrimination that can result from AI use. These biases are also inherent in AI models for detecting potentially hateful online content.

There are clear limitations and biases inherent in offensive speech detection algorithms, **FRA's 2022 report on bias in algorithms** shows. These need to be taken into consideration and rectified with regard to online content moderation.

Regulators and online platforms must ensure that independent researchers have easier access to information required to analyse online hate

This report provides a snapshot of online hate based on data collected from the selected platforms in four EU Member States. The results provide only a limited and targeted overview of online hate. This is partly due to the online hate being detected using a set of keywords that were the focus of the research.

At present, there is no methodology that provides a comprehensive picture of online hate. Therefore, different means of detecting online hate are **prone to selection bias**. Moreover, access to data for analysing online hate is limited to certain platforms due to online platforms' very restrictive policies and practices regarding access to data.

Pursuant to Article 34 of the DSA, VLOPs and very large online search engines have to undergo risk assessments with respect to systemic risks, including fundamental rights. In line with this, the analysis of fundamental rights infringements on platforms can be based on various methods. These include, but are not limited to, using data and indicators that online platforms provide as well as directly collecting data from online platforms through application programming interfaces and web scraping.

The research for this report involved directly accessing data available on platforms (i.e. data that had already been through platforms' own content moderation processes). FRA faced issues with accessing data on certain platforms, notably Facebook and Instagram, because the company responsible for providing access for research purposes did not accept new users during the research period. This meant that some major platforms could not be included in the research.

In addition, other methodologies for gathering hard evidence on how platforms' conduct influences the enjoyment of fundamental rights are needed. This involves carrying out surveys among social media users about their experiences, potentially combining this information with social media data, and conducting experimentation and testing. These methods, especially when used in conjunction, can help identify online hate more consistently.

FRA OPINION 4

The European Commission should ensure that risk assessments under the DSA – including with regard to online hate – are complemented by extensive independent research using a variety of methods to ensure the accuracy and diversity of assessments. This is necessary, as any single method for analysing online hate remains limited. Only a variety of approaches and tests will provide a fuller picture of the challenges linked to identifying and combating online hate. Independent research can offer critical views and the appropriate methodologies required to further the understanding of the complex and fast-changing landscape of online hate and content moderation.

The European Commission should ensure that independent research institutes and academic researchers can access the data of online platforms without burdensome administrative procedures or other potential obstacles and in line with data protection safeguards. This will allow researchers to better investigate:

- what kinds of online hate are not identified and appropriately taken down;
- where content that is not illegal online hate is taken down;
- what online hate's impact is with regard to fundamental rights.

1

CHALLENGE OF ONLINE HATE AND HOW TO RESEARCH IT

1.1. NEED TO ADDRESS ONLINE HATE

The past few decades have seen unprecedented change in the way the world is connected and how people communicate. There was a shift from posting letters in the 1990s to communicating online at the beginning of the 2000s. In the 2010s, it became possible for most people to communicate with a multitude of others on a daily basis. They can send messages via a variety of services and post statements, opinions and comments on online platforms that can be seen and shared by a potentially unlimited number of people.

This new era of communication has many positive societal effects. It allows people to stay in close contact with friends and family, which is especially important during crisis situations such as a pandemic or a war. These new ways of communicating also allow people to contribute to political discussions and express their views more easily, and can even support political movements against repression in totalitarian states. Moreover, the digital traces of online communication offer the opportunity to study and understand how people communicate and behave in these online spheres.

Yet there are also downsides. The services may be used to illegally influence the political attitudes of people, as the Cambridge Analytica scandal highlights ([1]). Platforms may be abused to incite hatred of political opponents and ethnic minorities ([2]).

In addition, hatred is expressed and promoted much more easily online and can contribute to increasing offline violence ([3]). Expressing online hate may be illegal under national laws, but such expressions are difficult to identify and take down. It may also be legal in some countries. Regardless of legality, online hate has negative impacts and chilling effects on those who wish to join online conversations.

Hatred is expressed online in a variety of ways, including widespread cyber harassment, where people are directly addressed in a hateful manner. The European Union Agency for Fundamental Rights (FRA) Fundamental Rights Survey showed that more than 14 % of people in the EU-27 have experienced cyber harassment in the past 5 years. This includes incidents where people received emails or text messages or when someone posted comments about them that were offensive or threatening. The percentage increases to 27 % when looking at people aged between 16 and 29: 28 % for men and 25 % for women ([4]).

While comments that are offensive, threatening and insulting are a major problem online, special attention must be paid to hate targeted at people based on protected characteristics, such as that targeted at people of African descent, Roma and Jews. For example, on average, 12 % of Roma aged 16–24 experienced cyber harassment because of being Roma in the five Member States surveyed in 2018 and 2019 for the FRA Roma and Travellers Survey. The survey covered Belgium, France, Ireland, the Netherlands and Sweden. In the Netherlands, one in three Roma and Sinti aged 16–24 reported having experienced cyber harassment in the past 5 years because of being Roma or a Traveller ([5]).

However, online hate is a problem not only when directed at individuals, but also when expressed more generally. For example, most Jews (80 %) have seen antisemitic comments online at least occasionally, according to FRA's 2018 survey among Jews in the EU. Most Jews consider the amount of antisemitism on the internet to be increasing ([6]).

Changes in the way we communicate through online services have created considerable challenges for service providers and policymakers. The sheer magnitude of online content makes it very difficult to monitor. Major service providers use algorithms to scale up their efforts to deal with harmful online content. Algorithms are used to rank content and flag it for potential deletion or other actions.

Policymakers have reacted to these new challenges. The EU has updated its laws and implemented policies that tackle illegal content online, for example through the European Commission's Code of Conduct on Countering Illegal Hate Speech Online (hereafter 'the code of conduct') ([7]). The EU also adopted the Digital Services Act (DSA) in November 2022. The DSA addresses the conduct of online platforms ([8]).

However, these developments are relatively recent and there are still major uncertainties with regard to how to better protect human rights online and how to efficiently implement existing and newly developed laws. Is too much content deleted, thus diminishing people's right to freedom of expression and their right to information? Or is too much illegal and harmful content still present online, thus promoting hatred, discrimination and violence, particularly against vulnerable groups? Answering these questions in detail is beyond the scope of this report. This report instead covers the following.

— **A targeted overview and analysis of manifestations of online hate.** This is based on data collected from four online platforms in four languages encompassing four EU Member States – Bulgaria, Germany, Italy and Sweden. The data relate to hate directed at women, people of African descent, Jews and Roma.

— **A critical assessment of the limitations of online content moderation tools in detecting online hatred of specific groups.** This is based on data analysis of social media posts to identify potential online hate. These posts had already gone through online platforms' own content moderation systems. At the same time, the report highlights challenges related to researching and measuring online hate.

The twin focuses of this report are built on a critique of the fundamental rights compliance of online platforms. Specifically, the critique focuses on their ability to detect and remove online hate, in line with EU law and Member States' international obligations (see next section).

The scope of the research delimited the results. However, the results serve to increase our understanding of how online hate is expressed, how it interferes with freedom of expression and how difficult it is to measure and analyse online hate. They focus on four languages that are relatively under-researched in this field. The results aim to assist in finding solutions that can achieve fundamental rights-compliant online content moderation in the EU, particularly in terms of countering sexist, racist and antisemitic speech online.

The report sets out ways forward on how online hate needs to be addressed by EU and national policymakers, other stakeholders – notably online platforms – and civil society. The findings contribute to an increased understanding of existing challenges and provide possible solutions from a fundamental rights perspective.

Finally, a central aspect of the added value of the report is that it highlights challenges in assessing online hate. There are several ways to research social media platforms. All approaches have their drawbacks and advantages, but none is easily and quickly carried out. None currently offers a comprehensive picture of online hate.

This report is also based on an extensive literature review and expert interviews conducted during the background research phase. It discusses the challenges of its methodology; this discussion also highlights the limitations of content moderation. Biased data collection and speech detection, subjectivity in assessing content and the scale of online content present real challenges in ensuring people's right to freedom of expression in modern communication.

In addition, the landscape of online platforms and how online hate is expressed keep changing. This is why the ways in which online hate can be addressed remain a moving target. This report contributes to understanding this very modern challenge.

The DSA – a modern framework for the protection of fundamental rights online

The DSA was adopted to better protect people's fundamental rights online, to establish a transparency and accountability framework for online platforms and to foster innovation, growth and competitiveness within the single market.

It includes several provisions and approaches that apply to assessing the fundamental rights compliance of online platforms. Examples of provisions to safeguard fundamental rights in the DSA include:

- **terms and conditions with due regard for fundamental rights**, including restrictions imposed on service use in light of the freedom of expression, freedom and pluralism of the media and other rights enshrined in the Charter of Fundamental Rights of the European Union (the Charter) (Article 14 of the DSA);
- **fundamental rights risk assessments by VLOPs and VLOSEs** in relation to privacy, data protection, non-discrimination, freedom of expression, the rights of the child, human dignity and the creation of mitigation measures (Article 34);
- **recommender system transparency**, especially in view of the main parameters that lead to content recommendations (Article 27);
- **data access and scrutiny** to allow for external scrutiny of the societal and fundamental rights risks of online platforms (Article 40).

Source: **Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Act)** (OJ L 277, 27.10.2022, p. 1).

1.2. ONLINE HATE – LEGAL FRAMEWORK AND DEFINITIONS

1.2.1. Overview of the legal framework

States are bound to protect human rights – offline as well as online. One of the main human rights treaties, the International Covenant on Civil and Political Rights (ICCPR) ([9]), enshrines the protection of free speech (Article 19) and requires the prohibition of any advocacy of national, racial or religious hate that constitutes incitement to discrimination, hostility or violence (Article 20(2)). The European Union and its Member States are bound to protect fundamental rights, as enshrined in the Charter. The Charter covers the rights included in the European Convention on Human Rights (ECHR), which was acceded to by all EU Member States and hence also applies to areas outside EU law. Any human rights-based assessment of online content needs to consider the protection of freedom of expression against the infringement of any other right, such as the right to privacy (Articles 7 and 8 of the Charter) and the right to non-discrimination (Article 21 of the Charter).

The UN published a strategy and plan of action on hate speech in 2019, which is aimed at addressing a wider range of hate speech than just incitement 'in line with international human rights norms and standards, in particular the freedom of opinion and expression' ([10]). The plan clarified that only a very narrow category of hate speech should be criminalised. Based on Article 20(2) of the ICCPR, advocacy is narrowly defined as 'explicit, intentional, public and active support and promotion of hate towards the target group' ([11]). Moreover, UNESCO has published guidelines for regulating digital platforms, which aim to safeguard freedom of expression and access to information online in the context of regulatory processes and were subject to a broad consultation process ([12]).

The Council of Europe is also very active in the area of hate speech. It has issued several relevant recommendations on this issue, such as its 1997 Recommendation No (97) 20 ([13]) and its updated 2022 Recommendation CM/Rec (2022) 16 of the Committee of Ministers to member states on combating hate speech ([14]). The latter also contains a broader definition of hate speech that goes beyond incitement and aims at combating hate speech in a comprehensive way, including in the online environment.

With respect to online content and hate speech, the EU and its Member States have to hold companies to account in case of conduct that infringes fundamental rights. Importantly, on the one hand, governments can request that companies remove illegal content. However, they cannot require companies to remove content that is considered legal, despite being potentially harmful. On the other hand, companies are free to remove content that they do not deem to comply with their terms and conditions.

Yet a large amount of hateful content may be considered to be on the fringes of legality, or within its bounds. Such content, for example negative stereotyping, still harms targeted groups. It can also create an atmosphere of hate that may prevent people from joining online conversations, and thus has a chilling effect on freedom of expression. It may also have an effect on other users who feel encouraged to express hate, including illegal hate speech. Hateful speech, even in its legal and protected form, may contribute to increasing more severe forms of speech and, indeed, actions against protected groups ([15]).

However, while the existence of online hate may create a chilling effect on those affected, by – for example – discouraging them from joining or taking part in online discussions, removing too much content can also be a problem. FRA's *Fundamental Rights Report* of 2023 highlighted that some civil society organisations expressed concerns that the regulation of illegal content may be detrimental to fundamental rights, mostly the freedom of expression, as companies may err on the side of caution and over-remove content to avoid negative sanctions ([16]).

Any action against online hate needs to respect people's right to freedom of expression and information, guaranteed in Article 11 of the Charter, which corresponds to Article 10 of the ECHR, without compromising any other right of the Charter. Limitations that may be imposed on people's right to freedom of expression and information may not exceed those provided for in Article 10(2) of the ECHR. As outlined in a factsheet produced by the European Court of Human Rights (ECtHR) on hate speech-related case-law, it is crucial to note that:

'[f]reedom of expression constitutes one of the essential foundations of [a democratic] society, one of the basic conditions for its progress and for the development of every man. Subject to paragraph 2 of Article 10 [of the ECHR], it is applicable not only to "information" or "ideas" that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population. Such are the demands of that pluralism, tolerance and broadmindedness without which there is no "democratic society". This means, amongst other things, that every "formality", "condition", "restriction" or "penalty" imposed in this sphere must be proportionate to the legitimate aim pursued' ([17]).

In addition, '[t]olerance and respect for the equal dignity of all human beings constitute the foundations of a democratic, pluralistic society. That being so, as a matter of principle it may be considered necessary in certain democratic societies to sanction or even prevent all forms of expression which spread, incite, promote or justify hate based on intolerance ..., provided

that any "formalities", "conditions", "restrictions" or "penalties" imposed are proportionate to the legitimate aim pursued' ([18]).

International and regional human rights instruments and courts have recognised that freedom of expression can be limited and have detailed different tests to assess whether limitations are indeed acceptable in specific circumstances.

When dealing with online hate cases, the ECtHR uses two approaches. The court considers exclusion from the protection of the ECHR justified when expressions amount to hate speech and negate the fundamental values of the ECHR (this approach is provided for by Article 17 on the prohibition of abuse of rights). In addition, the court considers it appropriate to set restrictions on protection (through Article 10(2)) if they are deemed necessary in the interest of national security, public safety, the prevention of disorder and crime, the protection of health or morals, and protection of rights and freedoms of others ([19]). When assessing whether the removal of a post would infringe the freedom of expression, the following questions need consideration: (1) Was the removal prescribed by law? (2) Was it in pursuit of one or more legitimate aims? (3) Taking all the relevant circumstances into account, was the removal necessary and proportionate in a democratic society? ([20]) The underlying principle of the balance to strike in legislating in a manner compliant with the freedom of expression is set out in ECtHR case-law ([21]). For example, in the case of the platform Delfi, in *Delfi v Estonia* ([22]), the Estonian government had imposed a fine on a news portal for not removing defamatory comments quickly enough. Delfi claimed that this fine infringed Article 10 of the ECHR. The ECtHR rejected this claim and ruled that the domestic legislation and case-law 'made it clear that a media publisher was liable for any defamatory statements made in its publication' ([23]). This legislation pursued the legitimate aim of protection of the reputation and rights of others ([24]). Finally, the court considered the interference with Article 10 to

be proportionate, as the 'company had been able to exercise a substantial degree of control over readers' comments' and by allowing comments by non-registered users 'there had been no realistic opportunity of bringing a civil claim against the actual authors of the comments' ([25]). A more recent Grand Chamber judgment in May 2023, *Sanchez v. France* ([26]), further increased the liability of third parties, finding that requiring a politician to delete hate speech comments to his Facebook post does not violate his freedom of expression. This judgment highlights some of the current challenges of regulating hate speech online.

At the EU level, in line with Article 52(1) of the Charter, any interference with the freedom of expression needs to be provided for by law. In this context, however, there is currently no EU law instrument that would address online hate in a comprehensive manner and provide a harmonised definition. Certain EU laws establish specific forms of speech that are illegal and not protected by freedom of expression. Notably, the Council framework decision on combating certain forms and expressions of racism and xenophobia by means of criminal law of 2008 ('the framework decision') defines a common EU-wide criminal law approach to countering severe manifestations of racism or xenophobia. It requires Member States to legislate not only against the public expression of hateful speech related to the protected grounds but also against the dissemination of that speech. The protected grounds are race, colour, religion, descent and national and ethnic origin. The framework decision also applies to cases where the conduct is committed through an information system (Article 9). However, even after more than a decade, several Member States have yet to fully and correctly transpose the framework decision on racism and xenophobia into national law ([27]).

In December 2021, the European Commission proposed to extend the list of EU crimes in the Treaty on the Functioning of the European Union to also include hate speech and hate crime. This would allow, in the future, to adopt EU legislation establishing minimum rules on the definition of and sanctions against hate speech ([28]).

Besides hate speech, there are other grounds for rendering online content illegal under EU law, including terrorist content, child sexual abuse material and intellectual property violations ([29]).

In the field of combating terrorism, for example, the terrorism directive defines relevant offences, in particular the public provocation of terrorism ([30]), while the regulation on addressing the dissemination of terrorist content online creates an obligation to take down corresponding online content.

Services provided by online platforms are essential services available to the public and open to any person ready to subscribe to the terms and conditions needed to open an account. As a consequence, access to such services fall within the scope of both the racial equality directive ([31]) and the gender goods and services directive ([32]), prohibiting any direct or indirect discrimination on grounds of ethnicity or gender. 'Access' is to be understood broadly, for example covering the deletion of posts and suspension or termination of accounts, as this directly affects access to the service. However, the gender goods and services directive does not apply to media and advertising. Member States can still choose to address these areas in national law, going beyond the directive's minimum requirements. Furthermore, certain limitations in terms of the grounds of discrimination are covered by these directives. At present, neither the racial equality directive nor the gender goods and services directive directly addresses discrimination based on sexual orientation, gender identity or religion. A prohibition of discrimination based on sexual orientation and religion currently exists in the employment context.

As already mentioned in Section 1.1, the DSA is a relatively new EU regulation that updated EU law to better protect people's fundamental rights online, to establish a transparency and accountability framework for online platforms and to foster innovation, growth and competitiveness within the single market. Article 3(g) of the DSA defines illegal content as 'any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law'. While not defining what illegal content is in itself, the DSA harmonises due diligence requirements for online service providers including as regards the way they should tackle the dissemination of illegal content.

The text box in Section 1.1 mentions a few selected provisions in the DSA that are relevant for the protection of fundamental rights. Some relevant provisions are discussed further in **Chapter 3** in view of the issues covered in this report.

Monitoring how well companies remove illegal online hate related to racism and xenophobia

The European Commission has developed soft law instruments to combat online hate, namely the voluntary Code of Conduct on Countering Illegal Hate Speech Online (code of conduct) (*). The code of conduct was developed by the European Commission in 2016 and signed by a number of major online platforms, such as Facebook, X and YouTube, among others. Reddit and Telegram did not join the code of conduct. Perceived illegal material is flagged to the platforms by a number of organisations across Europe, and the companies have pledged to remove the material in a timely fashion. Their performance is monitored regularly, and reports are produced outlining how much illegal material has been removed. The code of conduct states that the majority of notifications shall be assessed and potentially removed within 24 hours (**).

The seventh monitoring report of November 2022 has shown that on average the platforms are assessing 64.4% of the flagged material within 24 hours, and 63.3 % of notified content is deleted from their platforms. This presents a decrease compared to the previous years, with 81 % of flagged content reviewed in 2021 and 90.4 % in 2020.

Reported grounds of hate speech in the seventh round of evaluation monitoring exercise include the following:

— xenophobia (including hate speech against migrants, 16.3 %)
— anti-Gypsyism (16.8 %)
— afrophobia (4.9 %)
— ethnic origin (6.2 %)
— race (5.2 %)
— gender-based hate speech (4.1 %)
— national origin (4.6 %)
— religion (2.6 %)

The focus of the code of conduct is illegal hate speech based on the 2008 Council framework decision on combating certain forms and expressions of racism and xenophobia by means of criminal law, which defines a common EU-wide criminal law approach to countering severe manifestations of racism or xenophobia, which is why gender-based hate speech may not be among the most prominent grounds identified.

(*) European Commission (2016), ***Code of Conduct on Countering Illegal Hate Speech Online***

(**) European Commission (2022), ***'Countering illegal hate speech online: 7th evaluation of the Code of Conduct'****, November 2022.*

National legislation, besides implementing those categories regulated by EU law, can play an important role in defining other forms of illegal online hate currently not covered by EU law. The national legislation of the four Member States covered in this report does not specifically address or define hate speech, beyond the obligations specified by the Council framework decision. Although none of the four Member States' national legislation provides a specific definition of incitement to violence and/or hate, they all criminalise incitement to violence or hate based on race, colour, religion, descent or national or ethnic origin via the transposition of Article 1(1)(a) of the framework decision. While Bulgarian and Italian criminal codes criminalise incitement to discrimination, violence and hate based only on these grounds, German and Swedish criminal codes contain more exhaustive lists of protected grounds. The German legislation prohibits incitement to violence and hatred of a national, racial, religious or ethnic group, against parts of the population or against an individual because of his or her membership of a specified group or part of the population, and attacks on the human dignity of others. The Swedish legislation prohibits threatening and expressing contempt for a population group by allusion to race, colour, national or ethnic origin, religious belief, sexual orientation or transgender identity. Much action is taken in several EU Member States to tackle online hate. A detailed analysis and overview goes beyond the scope of this report.

1.2.2. Defining a multifaceted phenomenon

There is no commonly agreed legal definition of the term 'hate speech'. There are many legislative and policy frameworks at the international, regional and national levels that cover (forms of) online hate. These offer definitions and protection based on certain grounds.

On a very general level, hate speech can be understood as the advocacy of hate based on one of the protected grounds. It encompasses any public expressions that spread, incite, promote or justify hate, discrimination or

hostility towards a specific group. It is dangerous, as it contributes to a growing climate of intolerance against certain groups ([33]).

The UN strategy and plan of action on hate speech ([34]) defines hate speech as 'any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor' ([35]).

The Council of Europe recommendation on hate speech of 1997 ([36]) states that 'the term "hate speech" shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin' ([37]).

The updated recommendation on combating hate speech of 2022 distinguishes between making racist, xenophobic, sexist and LGBTI-phobic threats and public insults, publicly inciting hatred or crimes against humanity and distributing materials containing expressions of hate speech ([38]). The first additional protocol to the Council of Europe's Convention on Cybercrime extends the scope to also cover offences of racist or xenophobic propaganda. This improves the ability of parties to the convention to make use of international cooperation in this area as well ([39]).

At the EU level, the framework decision defines a common EU-wide criminal law approach to countering severe manifestations of racism or xenophobia. Member States must punish intentional conduct that amounts to 'publicly inciting … violence or hatred directed against a group' (Article 1(1)(a)) and the 'public dissemination' of material that amounts to public incitement to hatred of a group (Article 1(1)(b)) ([40]). The protected grounds are race, colour, religion, descent and national and ethnic origin.

The term 'hate speech' can refer to expressions and acts that are illegal. It can be prohibited and punishable under criminal law or rendered illegal and sanctioned under civil or administrative law. However, hate speech may also be used as an umbrella term with multiple meanings. In this sense, it may also include expressions of hate speech that are not illegal under national or international law.

Therefore, this report works with a generic definition of online hate. It covers hate speech defined as (written) expressions on social media platforms that are found to incite violence, hatred or discrimination against a person or a group of persons, or that denigrate them, include negative stereotypes or offensive language, or are otherwise hateful, by reason of their real or attributed personal characteristics or status. The report collected online expressions (posts or comments on posts) through a dedicated keyword search on four online platforms. The results can only be understood in light of the research's data collection methodology, which **Section 1.3** summarises and **Annex 1** discusses in detail. Much of the analysis and discussion in this report deals with legal but potentially harmful speech.

This study developed a typology for categorising the posts found through the keyword search. Legal definitions and previous research informed the typology. It was validated during a dedicated expert meeting with those working in this field. The text analysis identified different elements of hateful characteristics, including:

— incitement to violence, discrimination or hatred, defined more narrowly to include only speech that directly calls for action;

— denigration, including attacks on the capacity, character or reputation of one or more persons in connection with their (perceived) membership of a particular group;

— offensive language, defined as hurtful, derogatory or obscene language, such as insults referring to protected characteristics;

— negative stereotyping, defined as negative traits and characteristics assigned to a social group and to its individual members in relation to protected characteristics;

— other hateful content– a residual category that may include support for hateful ideologies (such as nationalism) or Holocaust denial.

These categories are not mutually exclusive. Expressions of online hate may fall into several categories. Hate speech may lead to criminal liability (especially in cases including incitement) or civil and administrative liability (for instance, in cases involving offensive and denigrating language) or be otherwise hateful (such instances may often fall within freedom of speech limits). Yet this distinction is not clear-cut, and assessments of legality need to be made in context. Expressions including elements of incitement may still be covered in some instances by freedom of expression, whereas some cases involving denigrating language may lead to criminal liability. This report did not assess the legality of posts. However, for a selected number of posts, legal experts assessed the posts against existing definitions in the Council framework decision.

1.3. SCOPE AND METHODOLOGY

The current section offers a summary of the research methodology. **Annex 1** provides a more detailed description of the methodology.

Online hate is notoriously difficult to measure. This is related to the sheer number of posts shared online, the different ways in which it is expressed and the challenge of defining online hate. Identifying online hate strongly depends on the context of its expression.

This report is based on a standard and widely used methodology for collecting data on online hate. It is based on keywords considered to indicate hate speech, harassment and incitement to hatred and violence against women and people of African descent in four languages (Bulgarian, German, Italian and Swedish) as well as against Jews (in German and Swedish) and Roma (in Bulgarian and Italian) ([41]). The research covered the platforms Reddit, Telegram, X (formally Twitter) and YouTube.

The study used words from the project The Weaponized Word ([42]). The keywords were a set of discriminatory and derogatory words, which were:

— based on existing lists of words that researchers working in the natural language processing field use to detect online hate;

— refined further at an expert group meeting in March 2022 to match the needs of this study.

The use of keywords is frequently applied in research on online hate. However, it comes with limitations, as discussed further below.

Data collection from Telegram was based on country experts' selection of certain groups ('channels'). The experts chose these after carrying out background research on Telegram groups that are likely to include misogyny and hatred of ethnic minorities. **Annex 1** includes more details on the keywords and data collection.

This research collected 344 132 posts in total from the selected online platforms. Of those, 1 573 were randomly selected across the four countries covered and manually coded with regard to online hate. This smaller number of posts was selected to allow for further analysis of the posts. In addition, a selection of 40 of these 1 573 posts were used to exemplify some of the online hate identified (10 posts for each country). These 40 posts were analysed in more detail to highlight what these expressions of hate may look like and to to showcase the challenges of coding and assessing posts, also in view of legal definitions.

The posts collected for this report cover January–June 2022. The research collected posts in Bulgarian, German, Italian and Swedish to limit the scope to four Member States. However, the search for keywords in certain languages is only a rough indicator of whether or not the posts were posted in those four countries. **Annex 1** provides further explanation of the strengths and limitations of this approach.

The selected languages cover a heterogeneous group of countries in relation to their national policy situations regarding regulating online content and hate speech. Selection also considered the methodological feasibility of carrying out research in those countries.

The selection tried to include languages that are less frequently used in studies concerning online hate. The study of online hate is still predominantly undertaken in English.

The keywords developed capture misogyny and hatred of people of African descent in all four countries. In addition, the data collection used keywords indicating hatred of Jews in German and Swedish, and Roma in Bulgarian and Italian. The team selected target groups based on considerations linked to the assumed or known level of hatred that the groups experience in the countries– that is, their vulnerability. Selection also considered policy relevance, the comparability of the groups between countries and the budget available for the study.

The identification of online hate through keywords is not 100 % accurate. Therefore, the data collection also captured some posts with hateful elements targeted at other groups. Some of these posts contained hate based on other protected characteristics. Some contained hate that did not target people based on protected characteristics.

Posts were collected from Reddit, Telegram, X and YouTube comments. Platform selection considered the feasibility of accessing the data for research purposes, the number of platform users in the countries covered and the potential prevalence of online hate. The initial plan was to cover Facebook. However, the platform was not accessible to this study.

In addition, it should be noted that the data collected consist of content that stayed online after platforms' standard content moderation efforts. Access to platforms' data prior to content moderation is not possible for research purposes outside the confines of the platforms themselves. This refers to platforms' immediate content moderation efforts and actions. Content moderation also happens when posts are published on platforms and users or organisations then report them.

Close to 400 posts were randomly selected for each of the four countries (languages) from the full dataset of over 344 000 posts. These 1 573 selected posts were randomly sampled, while ensuring that each platform and target group was well represented in the dataset. Trained expert coders in each

of the countries then coded the posts. A guide to coding the posts was developed and validated in an expert workshop.

Coders were national experts on online hate. They underwent training on how to code the posts, completing several iterations of the training session to harmonise their approaches. The coding aimed to categorise each post with regard to the type and selected elements of hate expressed.

Several posts did not contain any hateful elements, although keywords were flagged, and had to be coded accordingly. These included posts that were considered irrelevant to the study of online hate. Others contained incidents of counterspeech in which users of online services expressed anger about someone else's hate.

The following steps were taken for control, consistency and to understand the challenges in categorising and monitoring online hate.

— For half of the posts, two coders coded each post in order to understand how consistently they applied the codes. **Section 2.5** provides the results of this consistency analysis.

— Legal experts reviewed a selection of 10 posts per country to determine whether a post flagged as potential hate speech would be likely to meet the legal threshold for being categorised as online hate.

Data were collected retrospectively from the online platforms covered. This indicates that platforms' content moderation systems fail to remove a number of potentially hateful posts.

Coding means that the posts were assessed and categorised according to predefined characteristics, as described in the next chapter. These included characteristics of online hate and other features of the posts, such as whether

they are targeted at specific individuals, whether they seem to be intended as humour and whether they may constitute counterspeech.

The 1 573 coded posts were distributed more or less equally between the four Member States covered, meaning close to 400 coded posts per country. Slightly more than half of the coded posts (53 %) came from X, 20 % from Reddit, 17 % from Telegram and 9 % from YouTube. Due to rounding, these values total 99 %. Most posts were collected from X and the others were less well represented in the collection. A limited number of YouTube comments were found for some languages.

FRA commissioned a research consortium, led by RAND Europe, to perform data collection and research ([43]). FRA prepared this report based on the data and analysis RAND Europe provided.

The methodology was selected based on considerations of feasibility and added value. Using keywords to capture online hate is a well-established methodology. One benefit is that various other research projects have tested the methodology. It offers clarity and transparency compared with, for example, the non-transparent algorithms that platforms use to capture online hate and it easily allows the collection of posts in different languages.

The methodology was assessed against other ways of collecting data, such as targeting networks or groups that are known for expressing a disproportionate amount of hatred of the groups covered or asking online platform users to report hateful incidents for the purpose of the study. In addition, previous FRA research on bias in algorithms ([44]) has already tested the application of algorithms with regard to offensive speech detection. That research analysed the extent to which bias occurs with regard to offensive speech detection, focusing on terms related to ethnicity and gender. Predictive algorithms were built for that purpose and tested for bias.

These methodologies can address types of online hate that may not be covered in the keyword searches used for this research. However, they also suffer from limitations, such as known and unknown biases in algorithms. These biases can be replicated or amplified during the process of running a predictive model, as FRA research shows.

Using keywords is, however, also known for its limitations. The method misses certain instances of online hate because people expressing hatred online often use different words to escape content moderation filters. These filters also depend on keywords to some extent. Hatred is often expressed in subtle ways using expressions that are particularly difficult to capture. The keywords also capture content that is irrelevant and are thus limited in their efficiency. Finally, keyword searches are necessarily limited to communication expressed in text and exclude other forms, such as images and audio.

When reading the results of this report, it is important to keep in mind that the data collection captured a fraction of online hate. The present data collection:

- is limited to selected international platforms;
- is based on a large but limited set of keywords that may indicate the presence of hateful speech;
- includes several posts that target people but are not based on protected characteristics;
- covers posts after platforms have applied their content moderation and hence does not capture already removed posts;

— does not cover images, audio/voice and video material;

— cannot offer detailed explanations of the posts' contexts.

These limitations also mean that the posts found are not fully representative of the distribution of forms of online hate. Different keywords, platforms, human coders and data collection periods might have led to different results. These elements of limited representativeness are common in data collection on topics that are difficult to measure and in projects with limited resources.

However, efforts were made to mitigate the limitations inherent in the methodology. The list of keywords was based on existing lists and further refined with experts considering the platforms, national context and target groups covered. In addition, coder bias was mitigated through discussions at an expert workshop, repeated training of coders, consistency checks and discussions among coders.

In addition, the analysis in this report carefully interprets the differences found in the statistical analysis to take account of the limitations discussed above. This report makes all those challenges visible and therefore provides insights into the challenges of measuring, understanding and moderating online hate. It offers valuable insights into the presence of online hate targeted at women, people of African descent, Jews and Roma. The report presents some results in a descriptive manner so that they can serve as a basis for discussion of the interpretation and explanation of online hate. They serve to highlight substantial statistical differences and offer an opportunity to discuss any lack of clarity linked to the interpretation of results and categorisation of posts.

As mentioned, the limitations of the analysis highlight the challenges and difficulties involved in addressing online hate. Tackling online hate requires a multitude of approaches to dealing with an old problem, hatred of vulnerable people, in the relatively new and quickly changing online environment.

Lack of access to platforms for research purposes – worrying developments

Exploring hate speech on online platforms is challenging due to limited access to relevant data, with some platforms remaining impossible to research. This inaccessibility is an impediment to grasping the full extent of online hate. Online platforms have the discretion to grant access to researchers. This has created a discrepancy in which platforms are accessible and featured in studies, leading to a limited understanding of online hate on other platforms.

For instance, X has practised an open access policy for researchers with its X application programming interface (API). An API is an application that communicates directly with the platform in question and consequently facilitates direct access to data from these platforms. Due to the user-friendly X API, past research has predominantly focused on X data. This may have contributed to an imbalance in the platform's representation in research. However, Twitter recently announced the end of its free and full API access (*).

To gain access to other platforms, such as Facebook and Instagram, researchers can use Meta's own platform CrowdTangle (**). However, CrowdTangle provides access to only a limited number of organisations. CrowdTangle did not accept any new applications during this report's research period.

Web aggregator tools, such as Brandwatch (***), offer alternative ways into APIs for data collection. For some platforms, mainly X, Reddit and YouTube, Brandwatch provides the tools necessary for the large-scale collection of posts, including comments.

However, for Facebook, this web aggregator is tailored to marketing needs, rather than scientific research. Including data from Facebook and Instagram is consequently impossible due to the lack of tools for collecting the necessary data using a keyword-based search approach. For example, it cannot collect data from comments on posts and publications in public groups.

Web aggregation does not cover social media messaging platforms, such as Telegram. Identifying relevant channels with a query-based search is challenging and risks missing out relevant channels and, consequently, data. Data collection on these platforms requires specially programmed tools, such as those used for Telegram in this report.

Moreover, user agreements can limit access to certain types of data (e.g. geolocation). They can also limit the amount of information that can be accessed. For example, the message parsing limit in the Telegram API restricts automated data collection from multiple channels.

*(*) X (2023), '**Getting started – About the Twitter API**'.*

*(**) CrowdTangle (2023), '**About us**'.*

*(***) See the **Brandwatch website**.*

Endnotes

(1) Kanakia, H., Shenoy, G. and Shah, J. (2019), '**Cambridge Analytica – a case study**', *Indian Journal of Science and Technology*, Vol. 12, No 29, pp. 1–5.
(2) Amnesty International (2022), ***Myanmar: The social atrocity – Meta and the right to remedy for the Rohingya***, London; Alaniz, H., Dodson, K. D. and Dmello, J. R. (2021), '**Race, rallies, and rhetoric: how Trump's political discourse contributed to the capitol riot**', *Journal of Criminal Justice and Law*, Vol. 5, No 1.
(3) See, for example, Williams, M. et al. (2020), 'Hate in the machine: anti-black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime', *British Journal of Criminology*, Vol. 60, pp. 93–117.
(4) See FRA (2020), '**Fundamental Rights Survey: experiences of cyberharassment in the past 5 years – a_har5y_cyb**'.
(5) FRA (2021), '**Roma and Travellers Survey: experiences of cyberharassment because of being Roma/Sinti in the past 5 years**' (pw_vh_eth_5y_cyber).
(6) FRA (2018), ***Experiences and Perceptions of Antisemitism – Second survey on discrimination and hate crime against Jews in the EU***, Publications Office of the European Union, Luxembourg.
(7) European Commission (2016), **Code of Conduct on Countering Illegal Hate Speech Online**.
(8) Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Act) (OJ L 277, 27.10.2022, p. 1).
(9) UN, **International Covenant on Civil and Political Rights**, 1966.
(10) UN, ***United Nations strategy and plan of action on hate speech***, May 2019.
(11) UN, Human Rights Council, ***Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression***, 7 September 2012.
(12) UNESCO (2023), ***Guidelines for Regulating Digital Platforms***.
(13) Council of Europe (1997), ***Recommendation No. R (97) 20 of the Committee of Ministers to member states on 'hate speech'***.
(14) Council of Europe, Committee of Ministers (2022), ***Recommendation CM/Rec(2022)16 of the Committee of Ministers to member states on combating hate speech***.
(15) This pyramid of hate or scale of prejudice was outlined decades ago by Gordon Allport. Allport, G. (1954), *The Nature of Prejudice*.
(16) FRA (2023), ***Fundamental Rights Report – 2023***, Publications Office of the European Union, Luxembourg.
(17) European Court of Human Rights (ECtHR), Handyside v. the United Kingdom, judgment of 7 December 1976, § 49.
(18) European Court of Human Rights (ECtHR), Erbakan v. Turkey, judgment of 6 July 2006, § 56.
(19) European Court of Human Rights (ECtHR) (2023), '**Factsheet – Hate speech**'.
(20) Council of Europe, Steering Committee for Media and Information Society (CDMSI) (2021), ***Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation – Guidance note***, Strasbourg, p. 29.
(21) European Court of Human Rights (ECtHR) (2023), '**Factsheet – Hate speech**'.
(22) European Court of Human Rights (ECtHR), ***Delfi AS v. Estonia, No. 64 569/09***, 16 June 2015.
(23) *Ibid*., para 62.
(24) *Ibid*., para 63.
(25) *Ibid*., para 65.
(26) European Court of Human Rights (ECtHR) ***Sanchez v. France***, No. 45 581/15, 2023.
(27) For example, in January 2023 the **Commission sent additional letters of formal notice to Estonia, Finland and Poland, and reasoned opinions to Greece and Poland for failing to comply with the framework decision**. **Other infringements** in relation to the framework decision were still active at the beginning of 2023 as well.
(28) European Commission (2021), **A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime** (COM(2021) 777 final).
(29) De Streel, A. et al. (2020), *Online Platforms' Moderation of Illegal Content* – Law, practices and options for reform European Parliament, Luxembourg.
(30) Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA (OJ L88, 31.3.2017, p. 1).
(31) Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (racial equality directive) (OJ L180, 19.7.2000, p. 22).
(32) Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services (gender goods and services directive) (OJ L373, 21.12.2004, p. 37).
(33) FRA and Council of Europe (2018), *Handbook on European Non-discrimination Law*, 2018 edition, p. 86.
(34) UN (2019), **United Nations strategy and plan of action on hate speech**.
(35) UN (2019), **United Nations strategy and plan of action on hate speech**, p. 2.
(36) Council of Europe, Committee of Ministers (1997), ***Recommendation No. R (97) 20 of the Committee of Ministers to members States on 'hate speech'***, Strasbourg.
(37) See 'Scope' in Council of Europe, Committee of Ministers (1997), ***Recommendation No. R (97) 20 of the Committee of Ministers to member states on 'hate speech'***, Strasbourg.
(38) Council of Europe, Committee of Ministers (2022), ***Recommendation CM/Rec(2022)16 of the Committee of Ministers to member states on combating hate speech***, Strasbourg.
(39) Council of Europe (2003), *First additional protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems*, European Treaty Series No 189, Strasbourg.
(40) Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law (OJ L 328, 6.12.2008, p. 55).
(41) For a collection of keywords and related studies, see, for example, hatespeechdata (n.d.), '**Hate speech dataset catalogue**'.
(42) See the website for **The Weaponized Word**.
(43) See the website for **RAND Europe**.
(44) FRA (2022), *Bias in Algorithms – Artificial intelligence and discrimination*, Publications Office of the European Union, Luxembourg.

2

HOW ONLINE HATE MANIFESTS ITSELF – A SNAPSHOT

The research collected examples of online hate against women, people of African descent, Jews and Roma from selected online platforms in four EU Member States, encompassing four languages. Posts and comments were collected through a special keyword search that aimed to cover hate speech on X, in YouTube comments and on Reddit. In addition, data collection covered selected open chat groups on Telegram.

The data collection provides examples of manifestations of online hate. It is important to note that the collection is by no means exhaustive or representative of online hate on the selected platforms in the countries covered. It is technically impossible to collect all incidents due to the variety of ways in which hate speech is expressed online.

This report analyses a snapshot of hate expressed on selected online platforms and this must be understood in light of the report's data collection methodology.

Section 1.3 provides an overview of the methodology and **Annex 1** provides a detailed description.

Subjective elements prevail in assessments of online hate

People assessing online content and speech may arrive at different conclusions. Potential disagreements in relation to assessing posts according to certain characteristics depend on people's perspectives and the depth of information available when assessing a post.

For example, people with different backgrounds may arrive at different judgements, as certain speech will affect them differently. Men may show a higher threshold for judging sexist content as offensive than women do, especially in borderline cases (*).

In addition, information on the positions of the speaker and the recipient, the social and political context and the intent of the speaker must be taken into account for speech to be assessed exhaustively. However, this information may not be fully available and requires time and resources to assess. These issues are best addressed through qualitative research. The coding for this report could not consider information about the speaker and recipient for resource and privacy reasons.

The study developed clear guidance for coding and dedicated training and ensured continuous exchanges among coders to ensure the high quality of assessments (i.e. coding) of posts. For half of the posts, two people annotated each post to understand the reliability of their assessments – that is, inter-coder reliability. Coders also reported their certainty when assessing posts.

In addition, 40 posts underwent further assessment. The subsequent sections present and discuss these posts.

The report makes the possible consequences of the uncertainty and potential disagreements explicit. These are important aspects to keep in mind when discussing and understanding the limits of the moderation of online hate.

(*) *See, for example, Wojatzki, M., Horsmann T., Gold D. and Zesch T. (2018), 'Do women perceive hate differently: examining the relationship between hate speech, gender, and agreement judgments',* Proceedings of the 14th Conference on Natural Language Processing (Konvens 2018), *Vienna, Austria, 19–21 September 2018; Sachdeva, P. S., Barreto, R., von Vacano, C. and Kennedy, C. J. (2022), 'Assessing annotator identity sensitivity via item response theory: a case study in a hate speech corpus',* FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, *Seoul, South Korea, June 2022, pp. 1585–1603.*

2.1. CHARACTERISTICS OF ONLINE HATE

Out of the 1 573 posts that the trained coders analysed, 1 050 posts (67 %) are relevant to the analysis of online hate. The remaining third are not considered relevant, despite being captured through the keyword search. These posts include people discussing other topics using the keywords. This also highlights the challenge of using keywords alone to identify online hate. **Annex 1** discusses this in more detail.

Some posts include offensive language and words that are not used in a hateful way, according to the coders. For example, often people describe situations where others expressed inappropriate insults and the users cite or repeat the language in their posts. One example is someone telling the story of a person insulting his girlfriend on the street. **Figure 1** includes only posts the coders perceived to be potentially 'threatening, abusive, insulting, or likely to offend, humiliate or intimidate'.

Types of online hate

This study used background research and an expert workshop to develop a typology based on the hateful character of post content for the purpose of categorising the data. The study distinguishes five main categories of online hate.

- **Incitement to violence, discrimination or hatred.** This is defined more narrowly to include only speech that directly calls for action.
- **Denigration.** This includes attacks on the capacity, character or reputation of one or more people in connection with their (perceived) membership of a particular group. This covers objectifying language, dehumanising language and other attacks on the capacity, character or reputation of a person in connection with their membership of a group with protected characteristics.
- **Offensive language.** This is defined as hurtful, derogatory or obscene language, such as insults referring to protected characteristics. The targeted person judges whether specific words or language is offensive, and this is highly context dependent.
- **Negative stereotyping.** This is defined as negative traits and characteristics assigned to a social group and to its individual members in relation to protected characteristics.
- **Other hateful content.** This is a residual category that may include support for hateful ideologies or Holocaust denial.

These categories are not mutually exclusive. Expressions of online hate may fall into several categories. It is also important to note that the categorisation of posts based on those characteristics remains subjective, particularly in a research setting where many posts are coded by people who are not directly affected. For example, while a coder may not consider a post offensive, the person targeted may still perceive the language used as offensive. Therefore, the results from coding reflect the settings of research projects and online content moderators, rather than assessments in courts.

Overall, more than half of the coded posts (53 %, meaning 836 posts) were considered hateful by coders in the sense of including elements of incitement to violence, discrimination or hatred; denigration; offensive language; negative stereotyping; or any other hateful content, such as supporting hateful ideologies (see text box '**Types of online hate**' for definitions). Posts could be coded into several of the categories/types. **Figure 1** presents the distribution of posts coded as hateful across types of online hate.

Of all posts coded as hateful, almost 85 % (708 of 836 hateful posts) were coded as using offensive language. Hateful posts most often include offensive language. By contrast, the next largest categories are denigration, at 39 % of all posts coded as hateful (323 of 836 posts), and negative stereotyping, at 29 % of all hateful posts (241 of 836 posts). Given that most posts were coded as using offensive language, these proportions clearly substantially overlap. Of the 836 hateful posts, 59 % (491) exhibit more than one type of online hate.

Many posts containing offensive language targeted at women refer to sexualised violence using strong language. This chapter shows some anonymised and translated examples of posts. These illustrate some of the report's findings, including the ways online hate is expressed and the difficulty and lack of clarity in coding some posts based on the text alone. The number of posts shown is minimal to avoid unnecessarily repeating and reproducing hateful posts including offensive language.

Denigration insults people by comparing them to objects or animals, for example. Denigration may also be coded as another kind of attack in connection with a person's membership of a particular group.

Example post – denigration where a person is compared to a monkey

@ @ I don't understand if when he speaks he moves his lips or if he is moving his ass, the monkey

Italian, X

Example post – denigration of people of African descent

Ten little N******

German, Telegram

Importantly, not all of the hateful posts identified target people based on protected characteristics. The coders considered only 55 % of hateful posts to express hatred of people based on protected characteristics. Despite searching mainly for hatred targeted at people based on gender (women) and race and ethnic origin or religion (people of African descent, Roma and Jews), the data collection also found instances of hatred of other groups. **Section 2.3** provides a more detailed analysis of the target groups.

Example post – targeting people based on protected characteristics

@@ I'm not a racist but g****** into soap 😤 The f****** were also imported from Paris, there were no f****** during the Communism and I have irrefutable proof for this.

Bulgarian, X

Example post – targeting people but not based on protected characteristics

Yes maybe it is like that. I would certainly have used violence if anybody had hurt my woman. But in this case it is pretty bizarre, especially that no one says anything. So if you do not like a joke then just hit that b*****d.

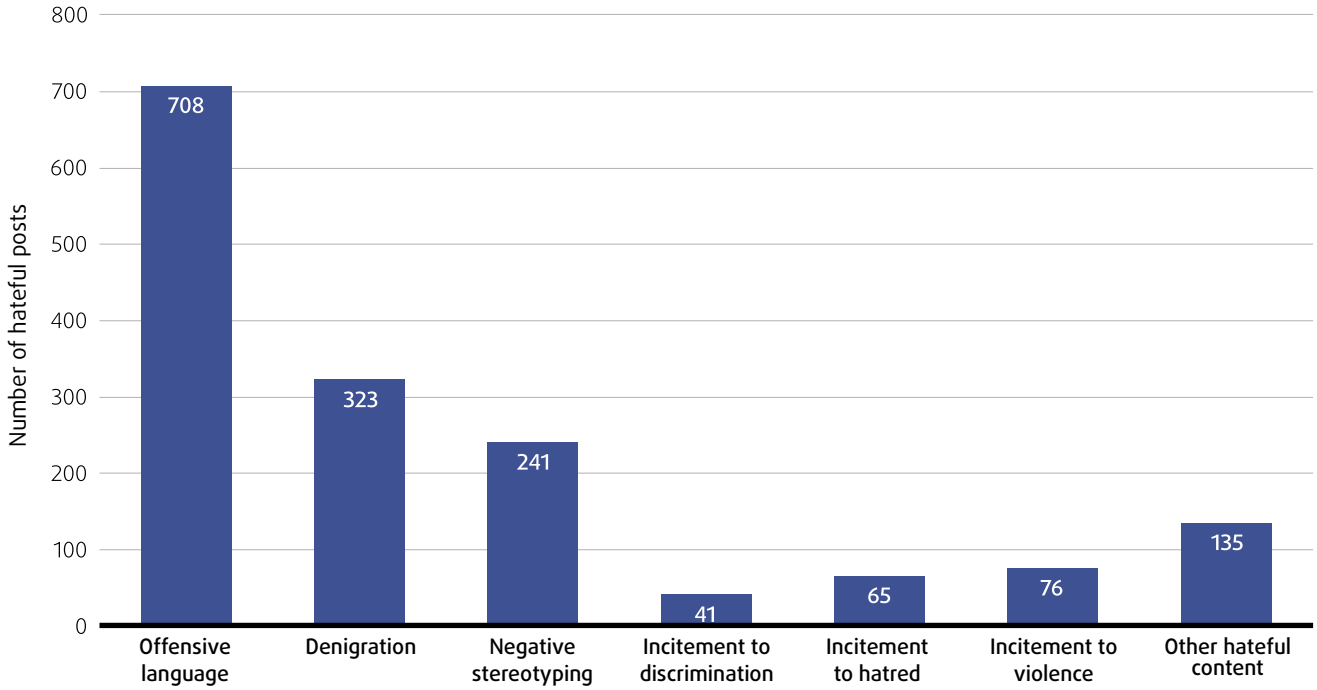
Swedish, X

Example post – targeting people but not based on protected characteristics

[Politician] has been normalised and this bristle-covered animal has found a new master to lick his hand

Italian, YouTube

FIGURE 1: TYPES OF HATEFUL POSTS



Source: FRA (2023), online hate dataset.

N = 836 posts. Several of the posts fall into more than one of the categories.

Overall, 16 % of posts coded as hateful in the sample exhibit incitement to violence, discrimination or hatred.

Coders coded 137 out of the 836 hateful posts in the sample (16 %) as incitement to violence, discrimination or hatred. Of these three types of incitement, incitement to violence is most prevalent, with just under 80 instances (**Figure 1**). This represents 9 % of all posts coded as hateful.

It is important to keep in mind that the coding instructions included a narrower definition of incitement than that in the framework decision. The instructions define incitement as including a call for action (see box '**Types of online hate**'). Incitement was limited to posts making calls for action to increase consistency in coding. Incitement without an explicit call for action is more difficult to code, especially with limited contextual information on the posts.

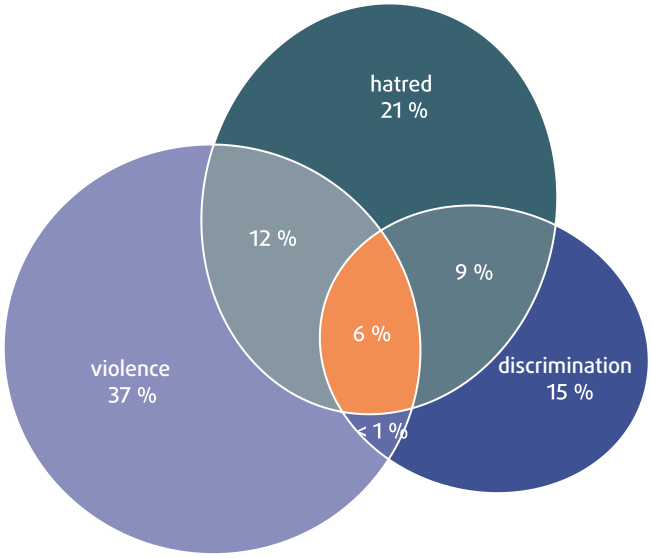
Incitement to hatred is almost as prevalent as incitement to violence, with coders categorising 8 % of all posts coded as hateful (65 of 836 posts) as incitement to hatred. By contrast, coders considered only 5 % of all posts coded as hateful (41 of 836) to be incitement to discrimination.

Several posts fall into two or more of the categories. 21 % of posts coded as incitement contain two forms of incitement (29 of 137 posts). Almost 6 % of posts contain all three forms of incitement (8 of 137 posts).

Figures 2 and **3** show how the different types and elements of hateful posts overlap.

Figure 2 shows how different posts coded as including incitement overlap. It shows that incitement to violence and incitement to discrimination rarely occur together. Coders judged less than 1 % of posts to include incitement to violence and discrimination, 12 % to include incitement to violence and hatred, 9 % to include incitement to discrimination and hatred and 6 % of posts to include all three forms of incitement.

FIGURE 2: INTERSECTION OF FORMS OF INCITEMENT



Based on 137 posts considered to include at least one form of incitement.

Source: FRA (2023), online hate dataset.

Coding incitement to violence appears to be most difficult. This is largely because incitement is often implicitly included in the text and the target groups are not necessarily spelled out directly. The following post was coded as incitement to violence, discrimination and hatred.

Example post – incitement to violence, discrimination and hatred

GERMANY – GERMANIC SPARTA! ⚜️😤⚜️🌨️⚜️ NO TO IMMIGRATION! The (Coronaspook) of the puppet regimes of Europe has, besides the totalitarian empowerment by the cold kitchen (IfSG), the function of a side show of war to divert attention from the mass alienation of Europe! The settlement of foreigners is a biological-ethnic atomic bomb and wanted by global politics for the slavery of the free peoples! The older video from 2016 should make us aware of this once again. Then as now – For the preservation of the great European identities!NO to the settlement of foreigners!Stand ready and steel your body and mind! Y O U ARE GERMANY @

German, Telegram

Online hate of women is very often of a sexual nature. The following post exemplifies how men use sexual violence when addressing women online. Such posts usually use very strong and explicit language. The post below was written by someone who apparently disagreed with a woman's statement about a football team and used the post to aggressively attack the woman.

Example post – incitement to violence

The little refugee b**** does not even know 5 players of [football team]. Better to f*** in the ***, that's what she is better at.

German, Reddit

This report will not show many example posts, as the language is very offensive. It is, however, important to show what kind of language women face online, given that most online hate identified in FRA's research targets women. This hate is often of an extremely sexually violent nature.

Note that the posts coded for this research had already been through platforms' standard content moderation processes. This indicates that sexist online hate too often falls through the net.

The next post does not target women. It shows an example coded as incitement to violence. However, the coder expressed uncertainty about whether or not the post fell into this category. Thus, the post is illustrative of the challenges of coding.

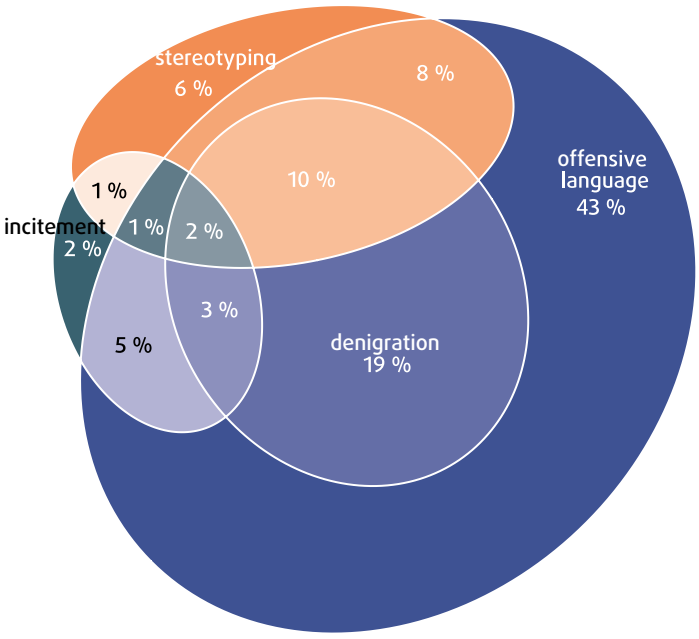
Example post – incitement to violence (uncertain assessment by coder)

If it was like in the USA, I would shoot with a legally owned gun at the g****** who had turned the volume up of chalga music under my window 🤦 but I need an idea how to stop the chalga music without being in the USA.

Bulgarian, X

Figure 3 shows how the different types of hatred intersect. Unsurprisingly, the largest category of offensive language overlaps with and includes other forms most often. It includes all forms of denigration and most posts coded as negative stereotyping and incitement. Its dominance also comes from the data collection methodology, as several of the keywords were examples of offensive language.

FIGURE 3: INTERSECTION OF TYPES OF HATRED



Based on 836 posts considered to include at least one type of hatred.

Source: FRA (2023), online hate dataset.

While all denigration is included in offensive language, negative stereotyping does not always use offensive language. The post below is an example

expressing antisemitic views in relation to Ukraine followed by another example of an antisemitic post.

Example post – antisemitism

yes that is exactly what it is ... also USA is packed full of jews who emigrated from the former Russian empire (which includes Ukraine) and the soviet union (also including Ukraine). An awful lot of them come from the territory known today as Ukraine, their dream is to avenge wrongs committed by Russians against their forefathers whilst simultaneously profit on the riches in that part of the world. [Female name] is a case in point. Her forefathers come from the regions west of Odessa. She is a 100 % jewish. Ukrainian nationalism is inexplicitly her passion.

Swedish, Telegram

Example post – antisemitism

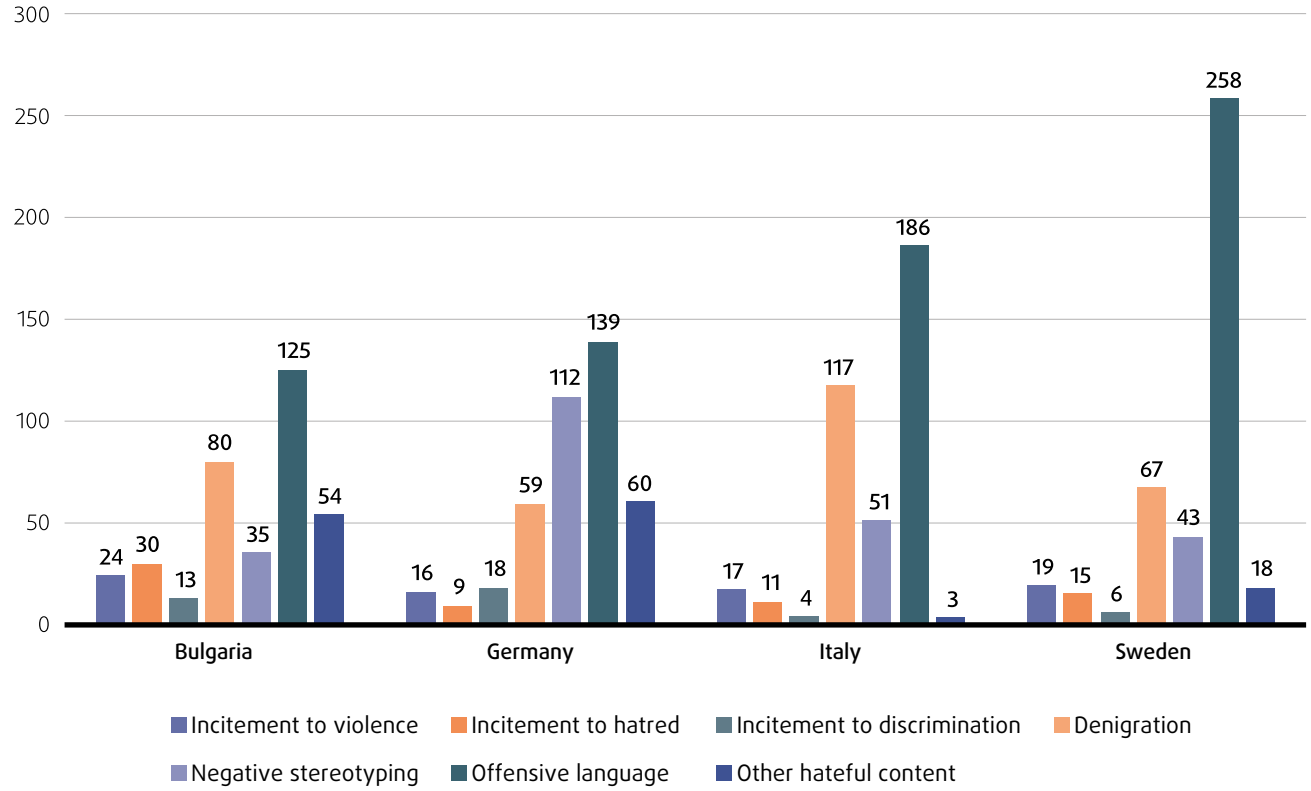
@ @ @ You are as much Jew as Aryan your ancestors were saved by my ancestors at the cost of their lives, but nothing matters to you. Leave Leopard 3 to himself, a real Aryan

German, X

The same number of posts were coded for each country. Therefore, the numbers do not reflect the actual relative numbers of posts in the countries. It may be that countries with a higher incidence of hatred in the coded sample also have a higher incidence of online hate. However, such inferences should not be made due to the unrepresentative nature of the purposive sample.

Nevertheless, differences in the percentages of types of hateful posts still provide insights into potential differences in the way hatred is expressed between countries. **Figure 4** shows the number of posts by type of hate in the four countries.

FIGURE 4: NUMBER OF CODED POSTS, BY TYPE OF HATE AND BY COUNTRY



Source: FRA (2023), online hate dataset.

N = 836 posts. Several of the posts fall into more than one of the categories.

Offensive language is the predominant type of online hate in all countries. However, countries differ in the extent to which offensive language dominates. Sweden has the largest number of hateful posts that exhibit offensive language, with 258 posts classified as offensive language. By comparison, Italy has 186 posts identified as containing offensive language, Germany has 139 and Bulgaria has 125.

Denigration is relatively more prevalent in Bulgaria and Italy.

Bulgaria and Italy follow similar profiles of hate, with similar levels of denigration in the sample posts. For Italy, 60 % of posts are classified as denigration. For Bulgaria, 47 % of posts are classified as denigration. In both countries, denigration is substantially more common than in the German and Swedish data.

Negative stereotyping is more common in Germany than in the other countries. Negative stereotyping makes up almost 56 % of all hateful posts in Germany and occurs almost as often as offensive language.

The levels of incitement to violence, discrimination or hatred are lower than those of other types of online hate in all four Member States. Between 20 and 40 posts are categorised as any form of incitement in all countries. This suggests that absolute levels are broadly similar between the countries. Bulgaria has more posts featuring incitement to hatred (30 posts) than other countries (between 9 and 15 posts).

Other undefined types of online hate are more prevalent in Bulgaria and Germany.

Bulgaria and Germany have non-negligible numbers of posts exhibiting other hateful content. This category is the third most prevalent category for both Germany and Bulgaria. This pattern could be linked to the slightly higher prevalence of posts targeted at Jews and Roma. These posts tend to exhibit relatively more instances of other types of hate than posts targeted at women do.

2.1.1. Online hate across platforms

This report focuses on online platforms that operate internationally. It does not include platforms that only have a national or regional user base, given the scope of this report.

Providing an overview of the prevalence of online hate across platforms is notoriously difficult due to the challenges of systematically capturing online hate and the different types of users and content posted across platforms. Online hate may be more prevalent on a small number of international platforms, primarily Facebook and X, research from 2018 suggests (1). YouTube, Instagram, Reddit, Telegram and Tumblr are other sources of online hate, according to the research (2). Hateful content is widely and publicly available in both YouTube videos and comment sections, another study demonstrates (3).

However, online hate is common on national and niche platforms as well, a group of studies finds. National platforms associated with online hate are VK.com (4) for Russian-speaking audiences and Jeuxvideo in France (5).

In addition, niche platforms such as 4Chan, Parler and Gab have frequent expressions of online hate, according to previous research (6). Critically, these niche platforms are gaining users' interest, evidence suggests, as the most popular platforms are increasing their moderation mechanisms to prohibit hateful conduct and the spread of online hate. For instance, extremist groups are increasingly seeking out social media platforms and online forums with low levels of content moderation to evade hate speech regulation, the Platforms, Experts, Tools: Specialised Cyber-Activists Network reports (7).

With regard to evasion of content moderation, it is worth mentioning the use of Discord, 8kun, Telegram, WhatsApp and DLive alongside the continued existence of other, more traditional online fora and of the dark web, which is made up of hidden sites (8).

In Finland, only a small fraction of the many hate speech incidents come from international platforms, one study points out. Most are from national discussion platforms (9).

Other studies have demonstrated the presence of online hate on platforms whose primary uses are not social networking. For instance, there is evidence of online hate on the online music and podcast streaming platform Spotify, one study notes (10). Another study found evidence of young people being exposed to hateful content on Pinterest, an image-sharing platform (11).

More evidence being available on the magnitude of online hate on international platforms such as X and Facebook might be the result of publication bias, as most research focuses on those platforms. X appears to be the most examined platform, followed by Facebook, YouTube and Reddit (12).

This might be because global-level platforms have a wider user base than niche, national or regional platforms. Academic studies may focus on platforms that are more popular and have a wider audience. Conversely, there may

be less evidence available on the magnitude of online hate on national and smaller global platforms. Another reason may be that X used to be more accessible to researchers (see text box 'Lack of access to platforms for research purposes – worrying developments' in **Section 1.3**).

Moreover, platforms that are predominantly based on text are more often researched because research based on text is much easier than research based on images, video or audio/voice content. Multimodal detection of hate is still an emerging field [13]). In addition, international platforms more often include content in English. This is easier to study for most researchers, as tools for language detection and analysis are more readily available in English [14]).

To reiterate, the data collected for this report come from four internationally acting platforms: Reddit, Telegram, X and YouTube. The results provide a snapshot of hate speech in the online space.

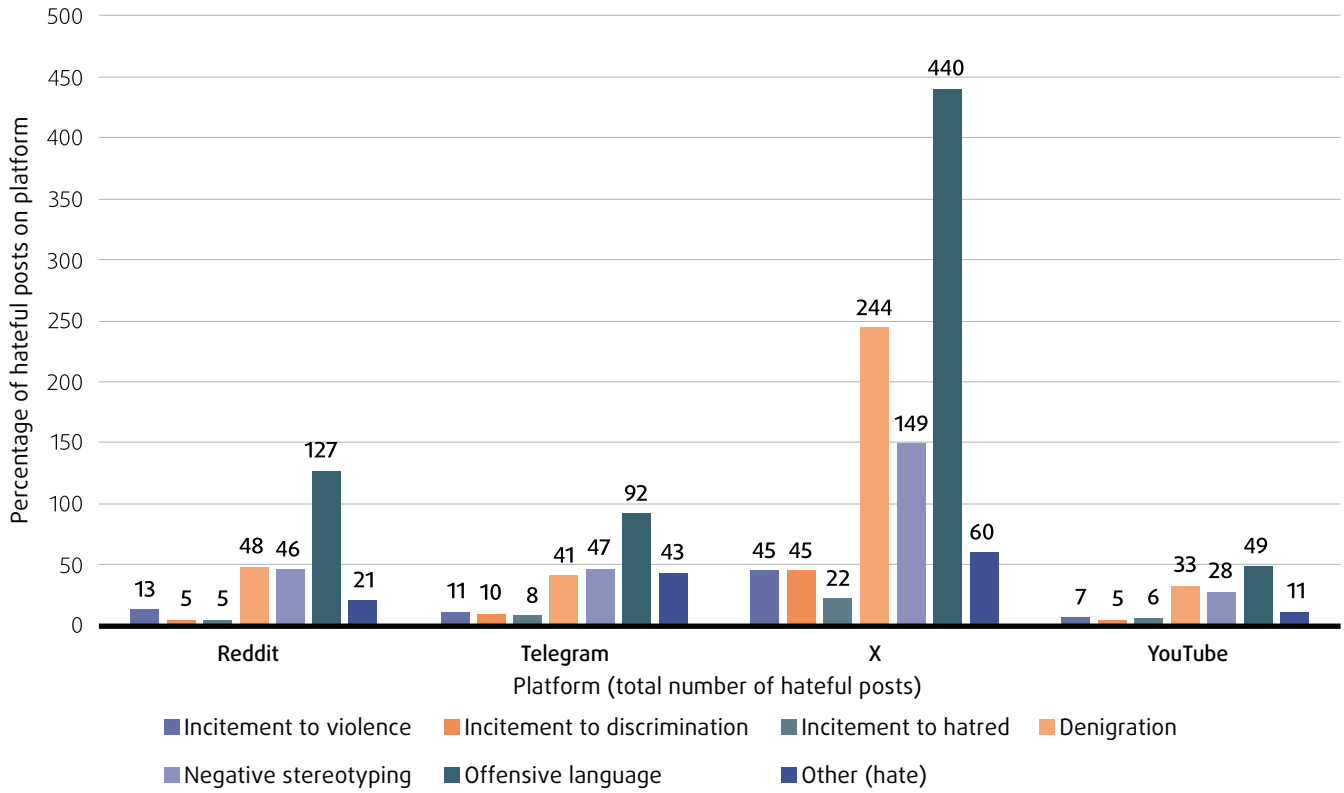
Across all platforms, offensive language, denigration and negative stereotyping are the most prevalent types of online hate.

Figure 5 shows the distribution of posts between the types of hate by platform. On all platforms, the most prevalent type of online hate is offensive language. Denigration and negative stereotyping are close second and third categories on most platforms.

The exception is Telegram, where other hateful content is found in every third hateful post. This might be a result of the different methodology used for collecting the data from Telegram. Data were collected from specific discussion groups and not through a direct keyword search.

The keyword search explains the high level of offensive language used in many posts on X, Reddit and YouTube. However, the percentage of offensive language is also very high in hateful Telegram posts.

FIGURE 5: TYPES OF ONLINE HATE, BY PLATFORM



Source: FRA (2023), online hate dataset.

N = 836 posts.

Levels of incitement are broadly similar across all platforms.

Incitement to violence, discrimination or hatred make up a minority of all hateful speech across all platforms. As a proportion of all hateful posts, the levels of incitement to violence, discrimination or hatred are broadly similar across the four platforms.

Incitement to violence is the most prevalent of the three types of incitement analysed on most platforms. On X, incitement to violence and hatred are equally common. For both X and Reddit, the number of posts categorised as incitement to violence is at least double that of incitement to discrimination. The differences are smaller for both Telegram and YouTube.

On all platforms except Reddit, the numbers of posts exhibiting incitement to violence or incitement to hatred are broadly similar. It is possible, however, that these categories overlap, although only 7 % of posts coded as incitement (9 of 137 posts) were coded as both incitement to violence and incitement to hatred. It should again be noted that the numbers do not include posts coders assessed as not 'threatening, abusive, insulting, or likely to offend, humiliate or intimidate'.

Example post – incitement to violence, discrimination and hatred

A g**** and beating are two compatible things.

Bulgarian, X

Telegram is unusual in that it has a higher share of hateful posts containing other hateful content.

Telegram has twice as many posts coded as other hateful content compared with the other three platforms. This may, in part, be due to the different way of collecting data through selecting channels. However, if may also be the result of the unmoderated nature of Telegram potentially allowing a greater variety of types of online hate. Telegram is also fundamentally organised around communities of interest (the channels).

2.1.2. Spread and reach of hateful messages

Social media and other online platforms allow people to spread messages to a potentially unlimited number of people. While some messages and comments may not be seen by many, a post may be disseminated widely depending on certain factors.

On X, the number of followers a person has is one factor influencing the reach of their posts. On Telegram, the size of a group increases the audience. On Reddit, the prominence and use of discussion threads are determining factors. On YouTube, if hate is expressed in comment sections, its reach depends on how widely a video is seen.

Importantly, other people can promote any post or comment through likes, upvotes and sharing the content further on the same or even other platforms. This user behaviour may indicate support for a post in the case of likes, upvotes or approving comments. On engagement-based platforms, platforms' algorithms can process this as an indication of a preference for similar content ([15]).

Consequently, if other people see a post that contains a hateful message, it may also shape the online environment of those who engage with it. Those who generate hate messages may do so to gain reward, express their antagonism and connect with others who feel similarly about the hated targets, some researchers suggest. This may reinforce their animosity ([16]).

The current report included platforms with millions of users and thus any post could potentially reach a wide audience. In reality, it was estimated that most posts were seen by only a few people. The 499 posts that included hateful content were seen on average by slightly more than 2 400 people, according to information X provided.

However, a few more prominent posts influence this average and half of the posts were seen by no more than 228 people. Yet some of the posts were seen by more than 100 000 people, with the most prominent reaching over 266 000 people. Almost a quarter of the hateful posts found on X (24 %) were likely to have been seen by more than 1 000 people. Eleven posts were seen by more than 10 000 people.

Platforms call the number of people who have seen the content the post's 'impressions'. Interestingly, average impressions are not directly linked to the population size of the countries. The average number of impressions of hateful posts is highest in Italy (3 861), followed by Sweden (3 161), Germany (1 043) and Bulgaria (815).

Hate is, unfortunately, attracting more attention. Of the types of hate identified, the 45 posts found on X and coded as incitement to violence show much higher impressions. On average, posts coded as incitement to violence receive over 5 900 impressions. Emotionally charged posts are generally shared more often and rapidly than content that is considered neutral, according to other similar findings ([17]).

However, the number of retweets is rather modest. Retweets are when people share content from another user through their networks. The almost 500 hateful posts on X received an average of 1 retweet and only slightly more than 1 % of the posts received more than 10 retweets. The maximum number of retweets is 141.

2.2. FORMS OF HARASSMENT – ONLINE HATE DIRECTED AT INDIVIDUALS

Almost half of all hateful posts were classified as harassment.

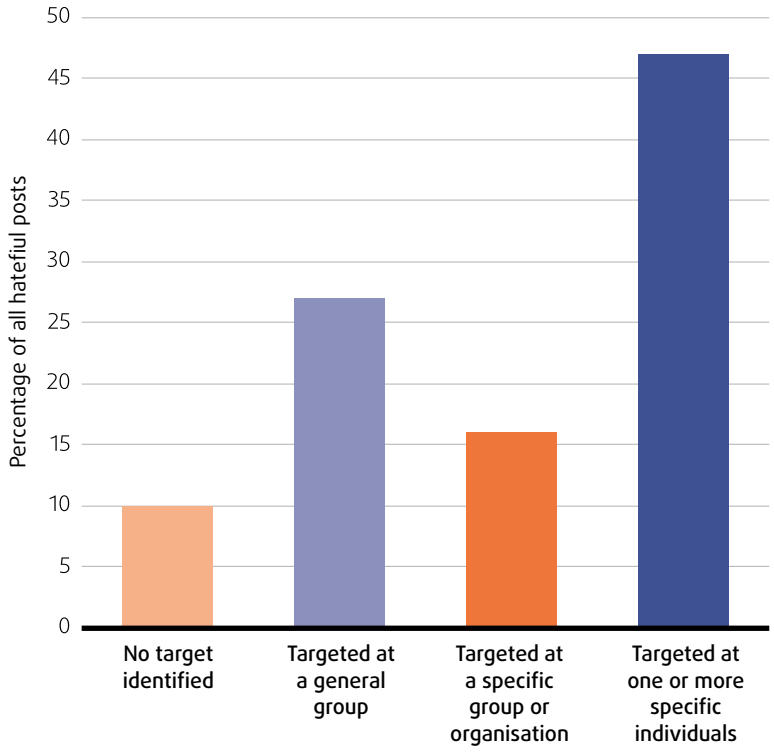
Almost half of all hateful posts were classified as harassment targeted at an individual (47 % or 397 of 836 posts). This means the posts were coded as hateful and targeted a specific individual. **Figure 6** shows the percentages of all hateful posts that were also coded as harassment targeted at an individual.

Example post – harassment

@ @ @ @ You probably know whose tweet came first. I will give you a hint: it was not mine. So forget the fairy tale about the poor, innocent witch. No one here will believe you.

German, Telegram

FIGURE 6: TARGETS OF HATEFUL POSTS

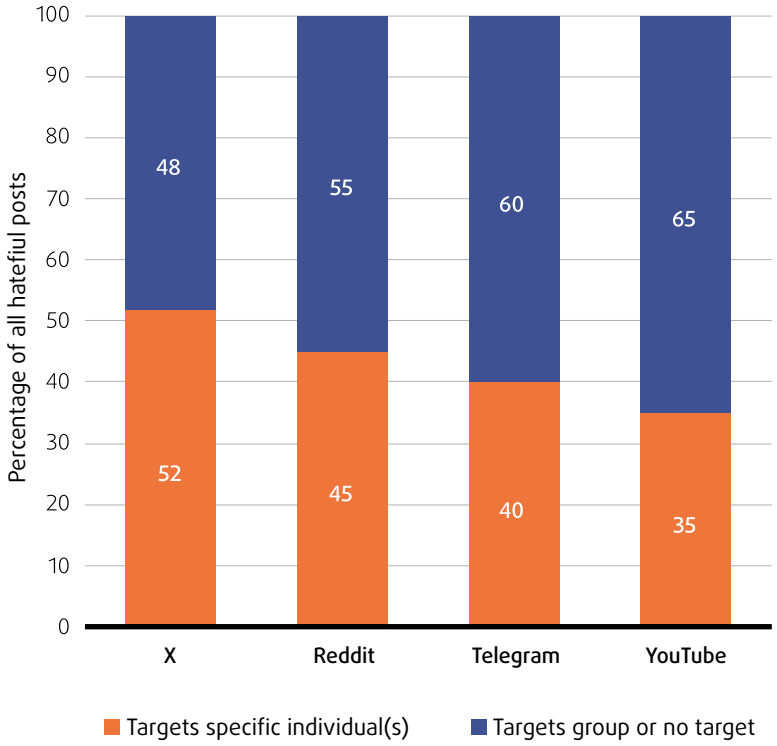


N = 836 posts.

Source: FRA (2023), online hate dataset.

As **Figure 7** shows, X hosts the largest percentage of online harassment, with 52 % of hateful posts targeted at one or more specific individuals. The posts collected from YouTube show the least amount of online harassment. However, one in three hateful posts still target one or more specific persons. The prevalence of harassment on Reddit and Telegram is around 45 % and 40 %, respectively.

FIGURE 7: PERCENTAGE OF HATEFUL POSTS TARGETED AT ONE OR MORE INDIVIDUALS, BY PLATFORM



N = 836 posts.

Source: FRA (2023), online hate dataset.

In Bulgaria, Italy and Sweden, between 50 % and 60 % of all hateful posts can be classified as harassment. In Germany, only 30 % of posts can be classified as harassment. These findings may also relate to the fact that posts targeted at women are relatively more prevalent in Bulgaria, Italy and Sweden and harassment is more prevalent in posts targeted at women.

2.3. COUNTERSPEECH

There is limited counterspeech across all platforms.

Searching for online hate through keywords will also capture counterspeech. Counterspeech can be defined as 'any direct response to hateful or harmful speech which seeks to undermine it' ([18]). Any content moderation efforts need to distinguish between counterspeech and online hate. The difference may be difficult to detect, as counterspeech may include the same words and sentences used in online hate and may be polemical in nature.

The number of posts classified as counterspeech is low across all four platforms. Overall, only 61 posts were considered counterspeech. This is less than 4 % of all coded posts.

The level of counterspeech is lowest for Telegram. The platform registers just two posts classified as counterspeech. Telegram's nature may explain this, as it organises online interactions around groups with shared interests. This means it could create echo chambers. These spaces do not generate much meaningful counterspeech.

For X, coders classified 36 posts as counterspeech, the highest of all four platforms. Unlike Telegram, the non-group-oriented nature of X may arguably lend itself to more combative social interaction. This potentially facilitates a greater degree of online speech that directly counters other content. Coders classified 16 posts on Reddit and 7 posts on YouTube as counterspeech.

Counterspeech is most prevalent in German posts.

Counterspeech across all four Member States is relatively low, but it is comparatively more prevalent in Germany. The level of counterspeech in Germany is double that in the other three countries. However, this starts from a relatively low base, with less than 5 % of all hateful posts classified as counterspeech in Bulgaria, Italy and Sweden. Just over 10 % were classified as counterspeech in Germany.

The higher levels of counterspeech in Germany could be due to the relative predominance of posts targeted at people of African descent or Jewish people. Counterspeech is more prevalent in response to posts targeted at Jewish people.

2.4. TARGET GROUPS AND INTERSECTIONALITY

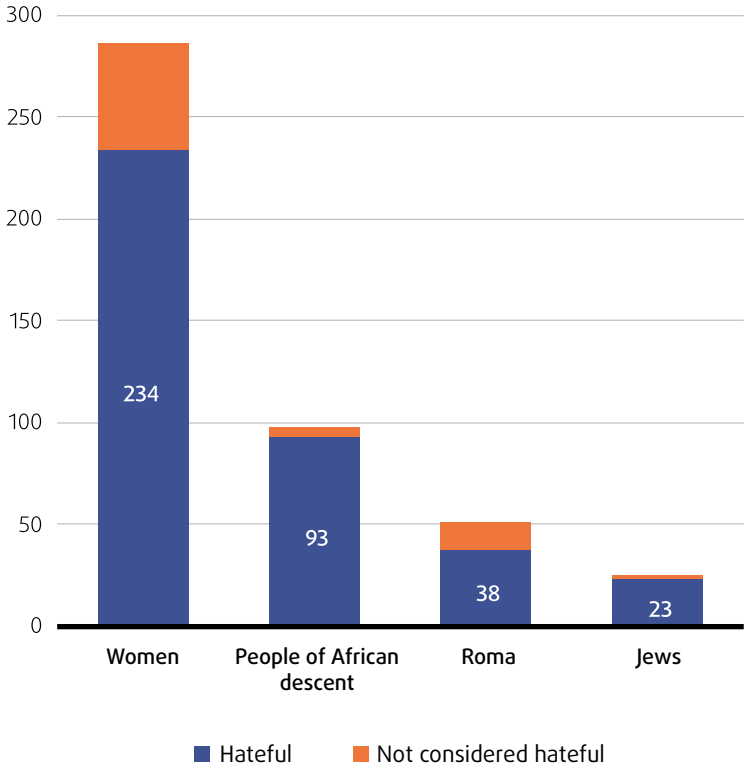
Hateful posts, in absolute terms, largely targeted women.

Only 35 % of all relevant posts target one of the four target groups (367 of 1 050 posts). An additional 20 % of posts target those with another protected characteristic (210 of 1 050 posts).

Overwhelmingly, most online hate in the data collected for the report – by volume of posts – targets women, as depicted in **Figure 8**.

There were almost three times as many relevant posts targeted at women as those targeted at people of African descent. Overall, 286 posts target women and 98 posts target people of African descent. Of these relevant posts, women face more than double the number of hateful posts of any other target group. There are 234 hateful posts targeted at women and 93 targeted at people of African descent.

FIGURE 8: NUMBER OF POSTS, BY TARGET GROUP



N = 367 posts. Several of the posts fall into more than one of the categories.

Source: FRA (2023), online hate dataset.

Posts targeted at people of African descent are most likely to be hateful.

The absolute number of hateful posts targeted at women is considerably higher than those of posts targeted at others. However, the percentage of hateful posts relative to the overall number of posts is highest among posts targeted at people of African descent. **Figure 8** shows the percentage of all posts coded as hateful, by target group. Almost 95 % of all posts targeted at people of African descent are categorised as hateful (93 of 98 posts).

Posts targeted at Jewish people are almost as likely to be hateful, with 92 % of relevant posts (23 of 25 posts) being hateful. By contrast, 82 % of posts targeted at women (234 of 286 posts) and almost 75 % of posts targeted at Roma (38 of 51 posts) are hateful.

Only 5 % of all posts coded as hateful towards any of the four target groups exhibited intersectionality.

Most hateful posts target only one of the four target groups, as shown in Figure 9. Of all posts coded as hateful and targeted at one of the four target groups, 92 % of them (358 of 388 posts) pertain to only a single target group. That value rises as high as 97 % for all hateful posts targeted at women (226 of 234 posts). Intersectionality, as a percentage of all hateful posts, is higher for the other three focus target groups.

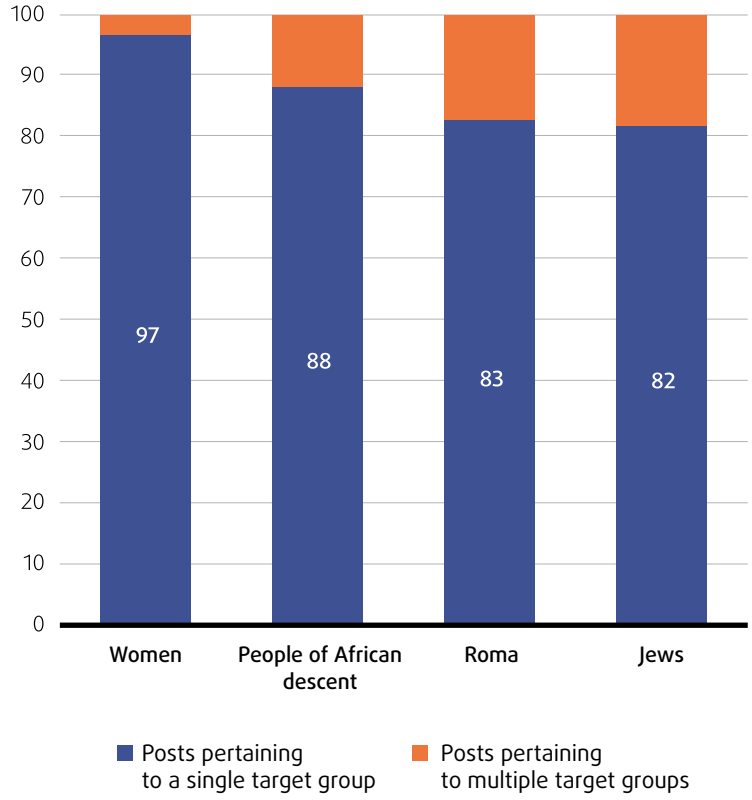
At a more granular level, just over 9 % of all posts targeted at people of African descent (9 of 98 posts) also target either Jewish people or Roma. Approximately 6 % of posts targeted at those of African descent also target women (6 of 98 posts).

Example post – intersectionality – women and religion

You are psychotically retarded, don't f****ng believe you dare to show your face, lol, you do not have a nice future ahead, there will be significant consequences for traitors to the country and Islamic-hugging wh***s. Just wait.

Swedish, X

FIGURE 9: PERCENTAGE OF ALL HATEFUL POSTS THAT PERTAIN TO ONLY ONE TARGET GROUP



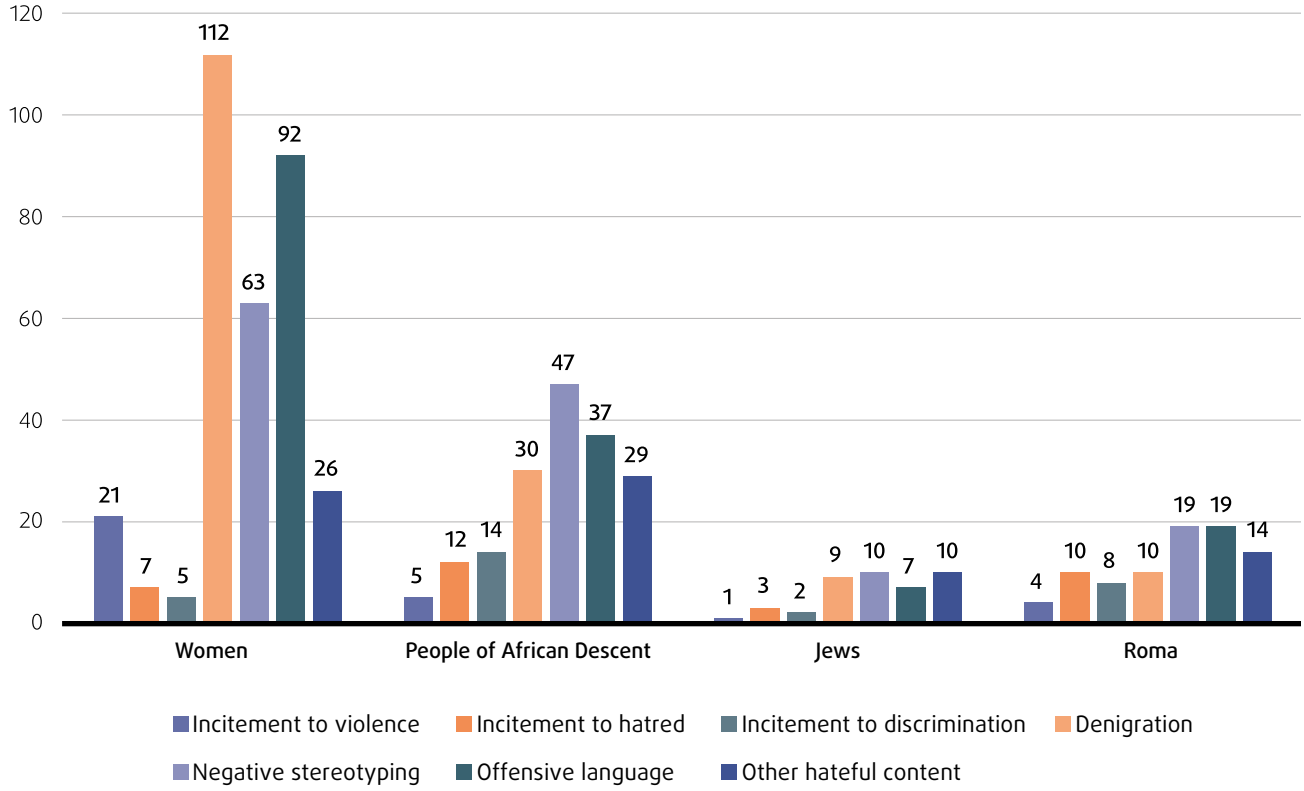
N = 367 posts.

Source: FRA (2023), online hate dataset.

Offensive language and denigration are most prevalent in posts targeted at women.

The profile of hate differs markedly by target group, as **Figure 10** shows. Levels of denigration are highest among hateful posts targeted at women.

FIGURE 10: NUMBER OF POSTS CODED AS HATEFUL, BY TYPE OF HATE AND BY TARGET GROUP



Source: FRA (2023), online hate dataset.

N = 367 posts. Several of the posts fall into more than one of the categories.

Negative stereotyping is most prevalent in posts targeted at people of African descent, Roma and Jews.

Negative stereotyping makes up approximately half of all hateful posts targeted at people of African descent and Roma. It is the most prevalent type of hate for people of African descent. For Roma, the level of offensive language equals that of negative stereotyping.

People of African descent's experiences of harassment on social media in the EU

Hatred of people of African descent is widespread on social media – even after content moderation. However, personal experiences with offensive comments against individuals vary considerably between Member States and age groups. Mainly younger people face such harassment.

Results from the largest EU-wide survey FRA has conducted regarding immigrants and descendants of immigrants support these findings. The survey asked immigrants and descendants of immigrants with different groups of origin about their experiences of cyber harassment.

The FRA survey covers selected groups of immigrants and descendants of immigrants in selected EU Member States. This includes people from sub-Saharan African and from north African countries, depending on the size of the groups in each country. The group of north Africans in the Netherlands was sampled through social media channels. This may explain the higher levels of engagement on social media in this group than in other groups (see **Figure 11**).

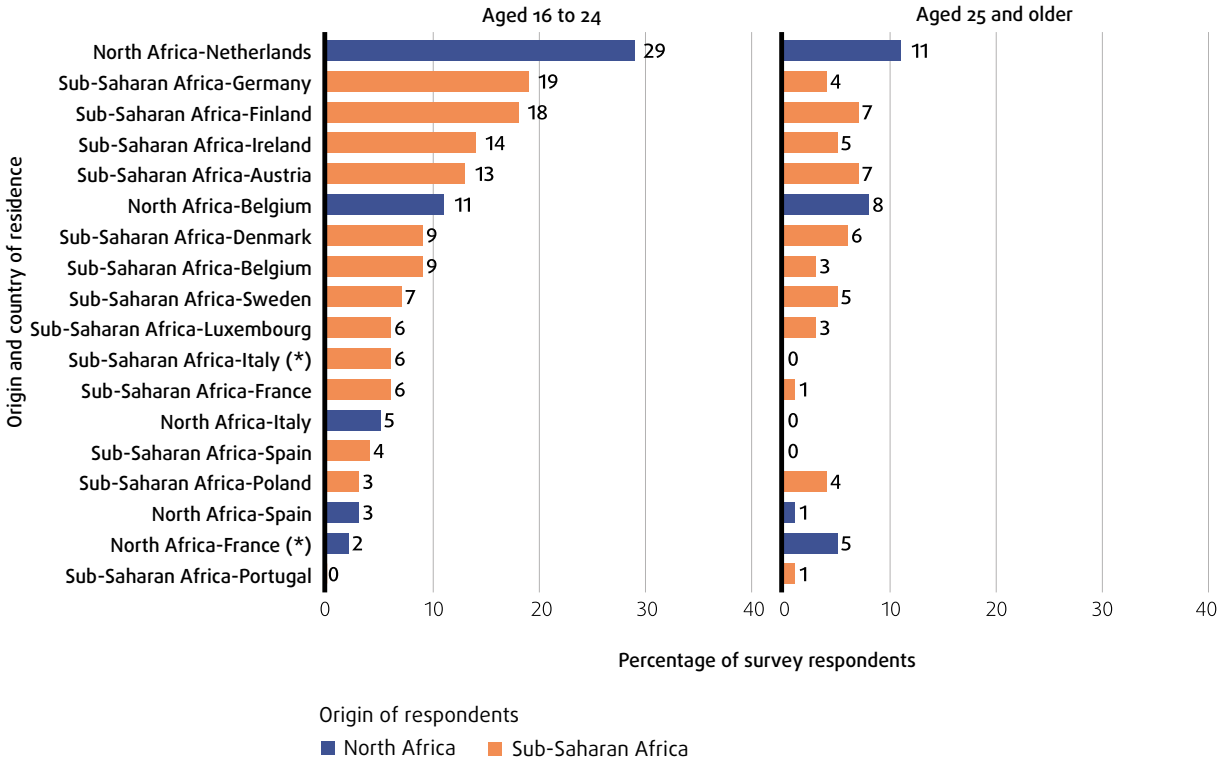
Among people of African descent aged 16–24, it is particularly people from north Africa living in the Netherlands who frequently experience people posting offensive comments about them. Almost one in three people interviewed (29 %) had seen offensive comments about themselves. Among this age group, percentages are also high for people from sub-Saharan Africa in Germany (19 %), Finland (18 %), Ireland (14 %) and Austria (13 %).

Most respondents who had experienced offensive comments thought it was because of their ethnic origin. Only about 25 % of them did not think so.

The results in **Figure 11** cover only respondents who have access to the internet. However, how (much) respondents in each country actually engage with social media and current ongoing political discussions also influences the results.

Question was 'Specific experiences of harassment in [country] in past 5 years: Posted offensive comments about you on the internet, for example on Facebook, Instagram, X, WhatsApp or TikTok?' Data cover respondents who answered 'Yes' to the question and who have access to the internet. N = 8 082. The survey covers selected groups of immigrants and descendants of immigrants in a selection of Member States.

FIGURE 11: PEOPLE OF AFRICAN DESCENT WHO HAVE EXPERIENCED HARASSMENT ON THE INTERNET, BY COUNTRY



Source: FRA (2022), EU Survey on Immigrants and Descendants of Immigrants.
Note: Results marked with an * are based on 20 to 49 unweighted observations and are less reliable.

For Jewish people, the category 'other hateful content' is as dominant as negative stereotyping. Both negative stereotyping and other types of hate make up a similar proportion of all hateful posts. 43 % of all hateful posts fall into each category (in both cases, 10 of 23 hateful posts). Holocaust denial may partly drive the predominance of other types of hate.

Women face higher levels of incitement to violence but lower levels of incitement to hatred and discrimination.

Women face the lowest relative levels of incitement to hatred and discrimination of the groups studied. Posts coded as misogynistic exhibit the lowest levels of incitement to hatred and discrimination in relative terms. Moreover, they have lower absolute levels of incitement to hatred and discrimination, as **Figure 10** makes clear.

Absolute levels of incitement to hatred and discrimination in posts targeted at people of African descent are double that found in posts targeted at women. This is despite the aggregate number of hateful posts targeted at women being three times that of hateful posts targeted at people of African descent. In relative terms, incitement to hatred and discrimination is 4 to 5 times higher for people of African descent and up to 10 times higher for Roma than it is for women.

Violence against women is often of a sexualised nature.

Women face higher levels of incitement to violence, however. In relative terms, almost 9 % (21 of 234) of all hateful posts targeted at women are classified as incitement to violence. Only Roma face similar levels of incitement to violence, with 10 % of all hateful posts classified as incitement to violence. Posts targeted at people of African descent or Jewish people exhibit approximately half the level of incitement to violence, in relative terms, found in posts targeted at women or Roma.

This suggests that women face misogynistic posts that are more violent in nature than the posts targeted at other target groups. The actual number of posts of this nature is relatively small (21). However, it is noteworthy from both criminal law and fundamental rights perspectives given that it represents an extreme manifestation of hate – incitement to violence.

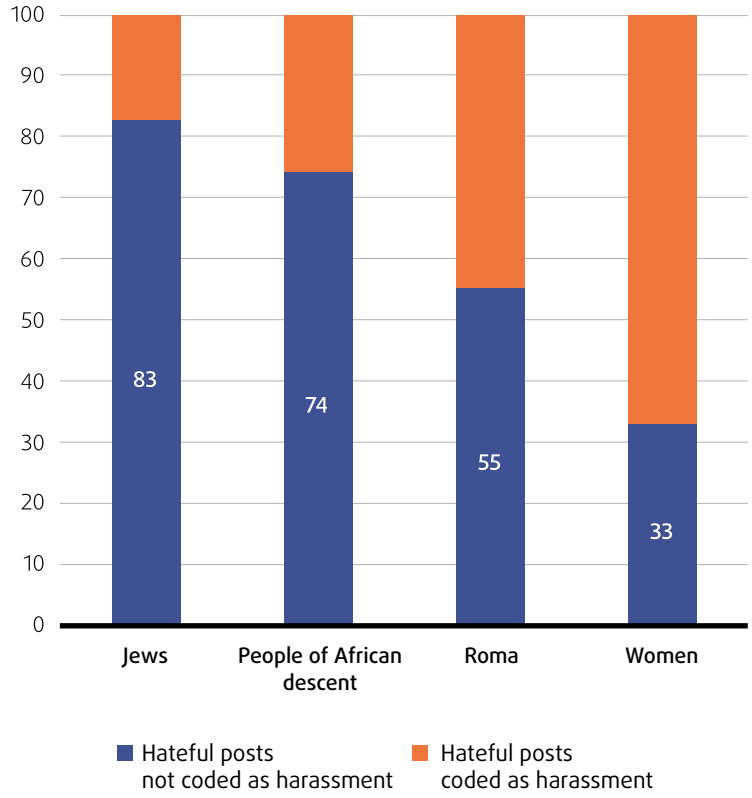
Importantly, violence against women incited in online posts involves references to sexualised violence. Most posts including elements of incitement to violence are of a sexualised nature. They often attack women directly through calls for them to be sexually abused.

Women face more harassment than any other target groups, at two thirds of all hateful posts.

Women face substantially more harassment targeted at individuals than any other target group covered. This means that the posts target one or more specific individuals more often than they target women in general. A recent UN Women and World Health Organization report outlining the severity of technology-facilitated violence against women supports this finding. ([19])

Figure 12 shows the percentage of hateful posts – classified as harassment – that target individuals, by target group. The level of harassment that women face exceeds the levels that other focus target groups face. Two thirds (67 %) of all the hateful posts targeted at women were found to be harassment. Roma experience the next highest level of harassment, in percentage terms, at 45 %. Jewish people experience the lowest levels of harassment, at 17 %.

FIGURE 12: PERCENTAGE OF HATEFUL POSTS CODED AS HARASSMENT, BY TARGET GROUP



N = 367 posts.

Source: FRA (2023), online hate dataset.

Counterspeech is more dominant in posts targeted at Jewish people.

Counterspeech is largely absent in hateful posts. However, where it is present, it is more likely to occur as a counternarrative to antisemitic online hate.

The rate of counterspeech in posts targeted at Jewish people is almost double that of the other three target groups. This could be linked to the prevalence of other types of hate directed towards this group, given that Holocaust denial tends to fall under the category of other hateful content. Given the small sample sizes, however, any such analysis should be considered exploratory, as differences are too small to be significant.

Women face a different distribution of types of online hate

Overall, the types of online hate prevalent in posts targeted at women are markedly different from the types of online hate in posts targeted at the three ethnic, racial and religious target groups. Posts targeted at women are more likely to be characterised as denigration and more likely to include incitement to violence. Posts targeted at those of African descent, Jewish people or Roma are more likely to be characterised as negative stereotyping. These posts are more likely to be classed as incitement to both hatred and discrimination.

Misogynistic posts may be more violent in nature, given the higher levels of incitement to violence and harassment targeted at individuals. Hateful speech against the ethnicity or race-based target groups may be more discriminatory. However, given the small sample sizes, such interpretations should only be considered exploratory. This particularly applies to hatred of those of African descent, Jewish people and Roma.

Women are the group facing the most hateful speech in all countries except Germany.

The incidence of hatred is highest for women in all countries except Germany. Women are the dominant target group for hate, **Figure 13** shows. For Sweden and Italy, the incidence of hatred targeted at women is more than double that of any other target group.

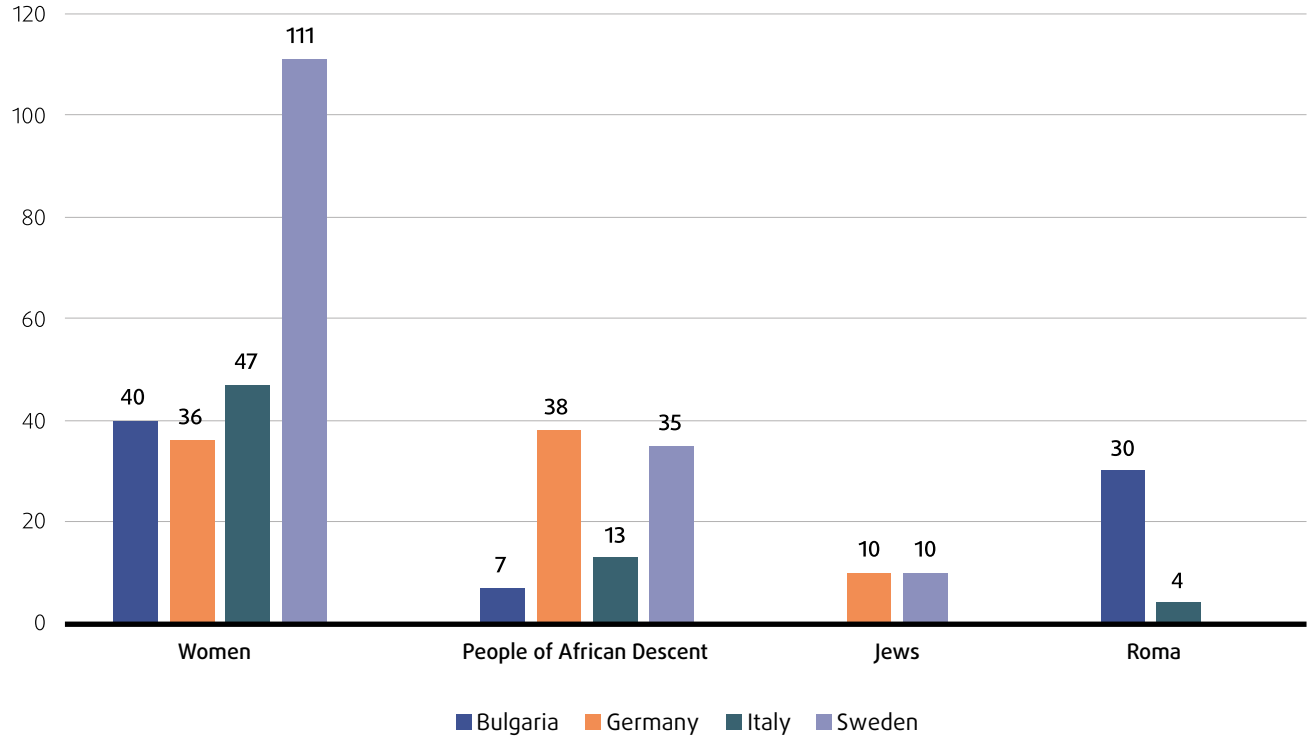
Germany is unusual: the incidence of hatred of people of African descent is marginally higher than the incidence of hatred of women. Analysis of the German posts shows a relatively more hateful space for people of African descent online than for women. In all other Member States, the coded posts for this study show that the online space is more hateful for women.

The context of this research cannot explain the very large number of hateful online posts targeted at women in Sweden, relative to the three other Member States covered. However, in this context, it is worth noting the results of FRA's EU-wide survey on violence against women. This survey collected data through face-to-face interviews with some 42 000 women in 2012 and the results were published in 2014.

The survey results for Sweden showed among the highest rates of self-reported experiences of violence against women in comparison with the other countries surveyed. The EU average for physical and/or sexual violence by any partner or non-partner was 33 %, whereas the figure for Sweden was 46 %. Only Denmark and Finland had higher rates.

Sweden also had one of the highest rates of sexual harassment in this survey. Notably, Sweden, alongside Denmark, has the highest rate of women having experienced cyber harassment among all EU Member States. Nearly one in five women surveyed (18 %) had experienced some form of cyber harassment in Sweden since the age of 15 ([20]). This may be linked to generally higher rates of harassment in Sweden but needs further investigation and analysis.

FIGURE 13: NUMBER OF POSTS CODED AS HATEFUL, BY TARGET GROUP AND COUNTRY



Source: FRA (2023), online hate dataset.

N = 367 posts. Several of the posts fall into more than one of the categories. Some posts cover more than one target group. Note that hateful posts against Jews were not specifically searched for in Bulgaria and Italy and hateful posts against Roma were not specifically searched for in Germany and Sweden.

Intersectionality in coded posts is low across all countries.

There is a low proportion of intersectional posts in the coded sample. Across all countries, rates of intersectionality are below 4 %, with over 96 % of all hateful posts coded as relevant to only one target group. Given the uniformly low levels of intersectionality across all countries, no clear patterns in intersectionality could be established.

2.5. UNCERTAINTY AND CONSISTENCY IN CODING ONLINE HATE

The above analysis gives a general sense of what kind of hate can be found online when searching for misogyny and hatred of people of African descent, Jewish people and Roma on the selected platforms. However, it is obvious that coding text into categories depends on a variety of factors, including information not available to the coders, and entails a certain level of subjectivity. Coders for this report received dedicated training to increase the reliability and quality of coding. They also engaged in group discussions to align their coding.

2.5.1. Agreement of trained coders' and legal experts' assessments

To understand how reliably the text was coded, two people coded each of about 200 of the 400 posts per country. They were a main coder and a second coder for control and verification purposes. In Italy, the two coders discussed each of the posts to arrive at agreement. In the other three countries, the coders discussed ways of coding but assessed posts independently.

The results show a very high level of agreement for several of the codes. In Bulgaria, for 93 % of the posts the coders agreed that the post was hateful. In Sweden, the rate was 92 % and in Germany it was 89 %. The level of

agreement about whether or not a post contained incitement (to violence, discrimination or hatred) was even higher, at 95 % to 97 %.

The lowest level of agreement occurred in relation to denigration, at 81 % in Bulgaria. Importantly, both of the coders in Italy were men. All other coders were women. This may also have had an influence on the results, as the introduction to **Chapter 2** discusses (see text box 'Subjective elements prevail in assessments of online hate').

In addition, coders were asked to indicate how certain they felt about their codes. The scale was 'Certain', 'Fairly certain', 'Fairly uncertain', 'Uncertain' and 'Impossible to code'. Generally, the coders were fairly certain about their coding. On average, the coders in the four countries indicated that they were certain or fairly certain about their coding in 78 % of cases.

The level of certainty was lowest in Bulgaria, where the main coder indicated being (fairly) certain about only 66 % of the codes. In Germany, the coder was (fairly) certain about 91 % of the codes. In Italy this was 81 % and in Sweden it was 76 %. In Sweden, the coder indicated that 8 % of posts were 'impossible to code' due to a lack of information on the posts' contexts.

The second coders were generally less certain about their assessments. Coders agreed on their levels of certainty for 77 % of the double-coded posts.

2.5.2. Differences in trained coders' and legal experts' assessments

Additional analysis of 40 selected posts – 10 in each country – shows the challenge of assessing online hate. The main coder made a more detailed assessment, which national legal experts cross-checked, based on the definitions found in the framework decision. The coders used a lower threshold for classifying content as inciting violence than the legal experts did, according to the analysis. This is based on both the legal analysis of the illustrative posts the national legal experts provided and further discussion.

For example, the coders coded a Swedish post on X as hateful and offensive and even considered it to appear threatening; this was the example post shown in **Section 2.4**. However, the national legal expert was reluctant to conclude on whether this post constitutes actual incitement to violence under the framework decision without knowing the post's full context.

Another example is a Bulgarian post containing an example of negative stereotyping of Roma, as shown in **Section 2.1.1**. It equates being Roma with beating other people. The study coders considered this to be hateful negative stereotyping and incitement to violence. However, the national legal expert did not think the post met the threshold for incitement to violence.

Interpreting incitement to violence and hatred appears particularly challenging. The coders coded the Swedish example shown in **Section 2.4**, for example, as misogynistic ('wh***s') and it appears to also target people on the basis of religion ('Islamic-hugging'). However, the national legal expert did not consider this post to directly incite others to commit hateful acts. People can interpret (inciting) hatred differently, demonstrating the subjectivity that such assessments can entail, as the Swedish post illustrates.

As another example, coders saw the following Italian post as inciting violence and hatred. Yet it does not fall under the definitions provided in the framework decision, according to the national legal expert. This seems a somewhat contradictory finding at first instance. However, it does not fit the definition

because the hatred is directed at the person due to the crime she committed, not because of her gender, the national legal expert notes.

Example post – incitement to violence and hatred

This b**** must suffer every day she lives. She must starve and thirst her like she did to the 16-month-old baby! Damn [apparent target name] [URL about an event reported in the national crime news]

Italian, X

This challenge of consistently assessing online hate is not only a result of a lack of context when assessing text. It is also because there is a level of subjectivity and interpretation in the definition of online hate. It also shows that, in assessing the legality of posts, legal expertise is needed. Employing legal experts to assess large numbers of posts is, however, very resource-intensive and does not reflect the reality of coding posts in real online content moderation settings (and research studies). This study did not assess posts with respect to their legality. However, this brief examination of the extent to which posts may fall under legal definitions of incitement to violence, hate or discrimination indicates the challenges of doing so in practice. **Chapter 3** discusses how such challenges can be addressed in content moderation efforts.

Endnotes

(1) O'Regan C. (2018), 'Hate speech online: an (intractable) contemporary challenge?', *Current Legal Problems*, Vol. 71, No 1, pp. 403–429.
(2) O'Regan C. (2018), 'Hate speech online: an (intractable) contemporary challenge?', *Current Legal Problems*, Vol. 71, No 1, pp. 403–429.
(3) Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K. and Zollo, F. (2021), 'Dynamics of online hate and misinformation', *Scientific Reports*, Vol. 11, 22083.
(4) Zueva, N., Kabirova, M. and Kalaidin, P. (2020), 'Reducing unintended identity bias in Russian hate speech detection', *Proceedings of the fourth workshop on online abuse and harms*, Association for Computational Linguistics, pp. 65–69.
(5) Hamelmann, M (2018), ***Antigypsyism on the Internet***, Platforms, Experts, Tools: Specialised Cyber-Activists Network.
(6) Hamelmann, M (2018), ***Antigypsyism on the Internet***, Platforms, Experts, Tools: Specialised Cyber-Activists Network.
(7) Platforms, Experts, Tools: Specialised Cyber-Activists Network (2019), ***Beyond the 'Big Three' – Alternative platforms for online hate speech***.
(8) Kfir, I. (2021), ***Online Radicalisation and Algorithms***, The Henry Jackson Society, London.
(9) Kettunen, L. and Paukkeri, M.-S. (2021), *Utilisation of Artificial Intelligence in Monitoring Hate Speech* (***Tekoälyn hyödyntäminen vihapuheen seurannassa***), Finnish Ministry of Justice (Oikeusministeriö justitieministeriet), Helsinki.
(10) Tan, J. A. (2019), '**Censoring hate in the music industry: shifting perspectives in pursuit of cultural equity**', *Backstage Pass*, Vol. 2, No 1, Art. 21.
(11) Reichelmann, A., Hawdon, J., Costello, M., Ryan, J., Blaya, C., Llorent, V., Oksanen, A., Räsänen, P. and Zych, I. (2021), 'Hate knows no boundaries: online hate in six nations', *Deviant Behavior*, Vol. 42, No 9, pp. 1100–1111.
(12) Matamoros-Fernández, A. and Farkas, J. (2021), 'Racism, hate speech, and social media: a systematic review and critique', *Television & New Media*, Vol. 22, No 2, pp. 205–224.
(13) See, for example, Chen, Y. and Pan, F. (2022), '**Multimodal detection of hateful memes by applying a vision-language pre-training model**', *PLoS One*, Vol. 17, No 9, e0274300.
(14) This point was also made in FRA (2022), *Bias in Algorithms – Artificial intelligence and discrimination*, Publications Office of the European Union, Luxembourg.
(15) Center for Humane Technology (n.d.), '**Make Facebook #OneClickSafer**'.
(16) Walther, J. B. (2022), '**Social media and online hate**', *Current Opinion in Psychology*, Vol. 45, 101298; Bail, C. (2021), *Breaking the Social Media Prism: How to make our platforms less polarizing*, Princeton University Press, Princeton, New Jersey.
(17) Stieglitz, S. and Dang-Xuan, L. (2013), '**Emotions and information diffusion in social media – sentiment of microblogs and sharing behavior**', *Journal of Management Information Systems*, Vol. 29, No 4, pp. 217–248.
(18) See Dangerous Speech Project (n.d.), '**Counterspeech**'.
(19) UN Women and World Health Organization, joint Programme on Violence against Women Data (2023), ***Technology-facilitated Violence against Women: Taking stock of evidence and data collection***.
(20) FRA (2014), ***Violence against Women: An EU-wide survey – Main results***, Publications Office of the European Union, Luxembourg, p. 105.

3

ADDRESSING ONLINE HATE

There is a lot a hate found on online platforms and it often targets women and ethnic minorities, empirical results in **Chapter 2** show. This chapter highlights points of action needed to address online hate from a fundamental rights perspective in view of the report's findings.

It discusses measures needed to detect and take action against online hate and outlines measures that can enable a better understanding of the conduct of platforms in order to tackle existing interference with fundamental rights.

In practice, it is very hard to identify and decide on the legality of online content. This is not only the result of the magnitude of online content, the difficulty in detecting it and some content's potential lack of context. It may also come from a lack of legal clarity of what actually constitutes illegal content.

The EU and Member States have defined what constitutes illegal online hate at the EU and national levels. The EU and Member States also created laws that regulate how illegal content should be handled to restrict its dissemination.

Most importantly, the landmark DSA entered into force on 16 November 2022. One of the objectives of the DSA is to create a safer digital space in which all users' fundamental rights are protected. To achieve this, an important shift was made towards creating more responsibilities, but also obligations, for intermediary service providers of online content.

Service providers will only be held liable for the content on their service if they have actual knowledge of the illegal activity (Articles 5(1)(e) and 6(1) of the DSA). Service providers are permitted, but not required, to carry out their own voluntary investigations or to take other measures aiming to detect, identify and remove or disable access to illegal content (Article 7 of the DSA). Finally, service providers are obligated to assess and potentially remove or disable content flagged to them as illegal through notice and action mechanisms (Articles 16 and 17 of the DSA).

Other relevant provisions include legally binding obligations related to the fundamental rights-compliant application and enforcement of terms and conditions (Article 14), transparency in reporting (Articles 15, 24 and 42), notice and action mechanisms for illegal content (Article 16), complaint handling (Articles 20 and 21), the transparency of recommender systems (Article 27), the protection of minors online (Article 28).

There are also provisions on risk assessment and mitigation efforts for VLOPs and VLOSEs regarding fundamental rights (Articles 34 and 35). Of the platforms analysed for this research, X and YouTube are considered VLOPs in the first assessment round of the DSA ([1]). This means that at least 45 million users are active on the platforms each month. VLOPs have to comply with

additional requirements, such as the provision of annual assessments of the platform's conduct in relation to systemic risks, including but not limited to those regarding fundamental rights.

The DSA includes a variety of measures that may contribute to the protection of fundamental rights online. It specifies ways of tackling illegal content, without defining illegal content in itself, as this is defined in other specialised EU law and national law, as discussed above. In addition, the DSA has promising provisions protecting fundamental rights in view of legal but harmful online hate. The most notable are its provisions for risk assessments and platforms' regard for fundamental rights in their terms and conditions.

In all the selected Member States except Bulgaria, there is also a move towards a legal framework within which platforms can be held responsible for the content they host.

The German Network Enforcement Act establishes obligations for providers of commercial social networks. Not complying with these and failing to remove content can result in fines of up to EUR 5 million ([2]).

In Sweden, Section 7 of the law on responsibility for electronic bulletin boards ([3]) sets out when platforms are responsible for content. The provider of an interactive website may, under certain special conditions, be held criminally liable for third-party content if the provider does not remove the content and the content falls under the Swedish Criminal Code (Chapter 16, Section 8).

In Italy, an administrative bill is currently being discussed that would create penalties for websites that do not remove hate content from their web pages ([4]).

Social media companies play an active role in combating online hate on their platforms. To ensure a safe environment for users, social media companies often establish guidelines of acceptable and prohibited behaviour that are specific to the platforms and their audiences. These are also known as terms and conditions.

Governments can ask platforms to delete illegal content. They cannot require companies to keep (i.e. not delete) content the companies deem inappropriate in view of their business models. However, the DSA has now added provisions that require companies, most notably VLOPs, to step up their efforts in relation to fundamental rights protection. This includes efforts in terms and conditions and risk assessments.

The important role of civil society organisations in addressing online hate

Numerous civil society organisations at the EU and national levels work to promote civil engagement with regard to online hate. The organisations represent different perspectives.

At the EU level, some notable civil society organisations that engage with online hate are European Digital Rights and Access Now. European Digital Rights is a network of non-governmental organisations, experts, advocates and academics working to defend and advance digital rights across the continent (*).

Access Now is a non-profit organisation with a mission to defend and extend the digital civil rights of people around the world. It has created a comprehensive roadmap called *26 Recommendations on Content Governance – A guide for lawmakers, regulators, and company policy makers* to guide decision-makers towards human rights-centred content governance policies. For instance, state regulatory models should focus on expressly illegal content and avoid defining evolving online societal phenomena. Furthermore, any state regulation addressing these online societal phenomena must always be grounded on solid evidence (**).

An example of a relevant civil society organisation at the national level is the German non-profit organisation HateAid. It campaigns for human rights in the digital space and combats digital violence and its consequences at the social and political levels. HateAid offers support, advice and funding for legal costs to those digital violence affects (***).

Another example is the Italian National Network to Fight Hate Speech and Hate Crime (****). It aims to merge the experiences of stakeholders working on hate crime.

In Sweden, the Institute of Law and the Internet works on internet-related rights issues, with a focus on freedom of expression and integrity (*****).

In Bulgaria, the MultiKulti Collective (******) is part of the International Network Against Cyber Hate. The collective focuses on the integration of migrants and refugees, community development and human rights.

In addition, civil society organisations contribute to safer online spaces through their efforts to detect illegal content and notify platforms of this. For example, civil society organisations carry out work as 'trusted flaggers'. These are entities with trusted status whose reports to companies about illegal content are treated with higher trust and priority.

Civil society organisations also contribute through carrying out research and investigations to highlight platforms' potential misconduct.

All these important tasks require proper funding, as civil society organisations frequently raise issues linked to underfunding.

*(*) See the **European Digital Rights** website.*

*(**) Access Now (2020), **26 Recommendations on Content Governance – A guide for lawmakers, regulators, and company policy makers**.*

*(***) See the **HateAid** website.*

*(****) See the Italian National Network to Fight Hate Speech and Hate Crime website, **Rete Nazionale per il Contrasto ai Discorsi e ai Fenomeni d'Odio**.*

*(*****) See the website of the Institute of Law and the Internet (**Institutet för Juridik och Internet**).*

*(******) See the **MultiKulti Collective** website.*

3.1. ADDRESSING FUNDAMENTAL RIGHTS-COMPLIANT CONTENT MODERATION OF ONLINE HATE

A significant amount of online hate remains online even after platforms' content moderation efforts, empirical analysis shows. Most of it is probably not illegal content and the level of offensiveness perceived may vary considerably between people. This highlights the difficulties of analysing and categorising online hate and the importance of having context to be able to make such assessments. Assessing each of the posts in the full context of the expression, such as information on the positions of the speaker and recipient, was beyond the scope of this quantitative research.

Furthermore, large portions of online hate are directed at specific individuals and misogyny is the most prevalent form. Only a small portion of this online hate constitutes incitement to violence, discrimination or hatred. Other forms of online hate may still be harmful and infringe on fundamental rights.

Building on these findings, this section highlights measures that are promising in tackling online hate and suggests some elements that can be strengthened or incorporated. Ultimately, a multifaceted approach to online hate is needed. It should feature a combination of measures that consider the important elements of prevention, action to address online hate, monitoring and ways of ensuring access to an effective remedy. In this regard, the DSA's effective implementation will be an important driving force of the fostering of fundamental rights-compliant online content moderation.

3.1.1. Detecting online hate

Detecting (illegal) online hate is notoriously difficult, analysis shows. There is no single methodology that reliably detects all online hate. Several actors must make efforts within their mandates to detect online hate.

Illegal content and legal content that is not in line with platforms' terms and conditions may be identified through the following routes.

- Public authorities and, most notably, law enforcement authorities can investigate platforms for potential illegal online hate. There are clear limits to the power law enforcement authorities (should) have to monitor online communication. Hence, they can only contribute to detecting illegal content online to some extent.
- Individual users can report content to platforms or other bodies, such as the police. Because of the difficulty in assessing whether a post contains hate, it is particularly important to allow individuals and users to submit alerts of online hate, in line with Articles 16 and 22 of the DSA. Articles 16, 17 and 22 only apply to illegal content. However, reporting is also relevant to detecting other harmful content and to taking appropriate and proportionate measures in response.
- (Civil society) organisations may act as bodies that collect input from individuals and have a more structured approach to detecting and reporting illegal and otherwise problematic content. The DSA calls such entities trusted flaggers. Digital service coordinators in each of the Member States appoint trusted flaggers, in line with the DSA. A wide and heterogeneous network of organisations acting as trusted flaggers is needed to ensure that online hate is widely detected and not disproportionately detected for specific groups.
- Platforms can take proactive measures to detect illegal content and content that goes against their terms and conditions. Companies' efforts vary depending on their terms and conditions and their general approach and policies towards online hate. Article 14 of the DSA lays down obligations for providers of intermediary services, such as online platforms, regarding certain content moderation practices stated in their terms and conditions. This includes algorithmic decision-making and human review. For the platforms covered in this report, the following points are noted.
    - Telegram demonstrates limited efforts to detect online hate (see Section 3.1.3). This probably contributes to the higher levels of online hate detected on Telegram in this report.
    - Other platforms, such as YouTube and X, use algorithms to detect potential online hate to complement their efforts to proactively detect users' breaches of their terms and conditions. The use of algorithms for content moderation of online hate is discussed in the text box below.

The use of algorithms to detect online hate and illegal content on platforms

Artificial intelligence (AI) usually refers to the use of data to create algorithms that automate or at least support tasks that previously required humans. Large online platforms are at the forefront of developing AI. They are the providers of some of the most widely used AI tools, such as Google's Tensor Flow (*) software or Facebook's Torch library (**).

The use of AI to detect online hate has increased considerably in the past decade. Facebook's 'proactive rate' shows the percentage of posts that Facebook 'actioned' – meaning removed or flagged. It is considered to be mainly based on automated detection systems. Some types of content that go against Facebook's community guidelines are fully automatically detected, such as spam. Other content, such as harassment, are not detected automatically that often and detection relies on user reporting.

Facebook's most considerable development in increased proactivity in automated detection was in relation to online hate. The rate of automated detection increased from 24 % in the fourth quarter of 2017 to 82 % in the fourth quarter of 2022 (***). **Figure 14** shows this development.

YouTube also relies heavily on the use of AI to detect online content that violates their guidelines, such as online hate. In the 3 months from October to December 2022, over 1.9 billion comments were removed. 99.3 % of these were found through automated flagging. This is, however, mainly linked to posts that are considered spam. Still, in the same period, over 52 million comments were removed for harassment and cyberbullying, and over 26 million for being hateful or abusive (****).

Algorithms and AI may be used to support content moderation decisions at scale. However, the technology is still far from being used reliably for automated decision-making. The main challenge of this technology is that AI is strongly biased and often embeds societal stereotypes and prejudice.

Offensive speech detection algorithms, even if highly sophisticated, rely too strongly on certain words that appear in text, as FRA's report on algorithm bias shows. For example, the simple existence of the word 'Muslim' can make algorithms predict text as offensive and can flag it for moderation. This AI needs to be scrutinised in detail for biases before it can be used to support online hate detection (*****).

*(*) See the **TensorFlow** website.*

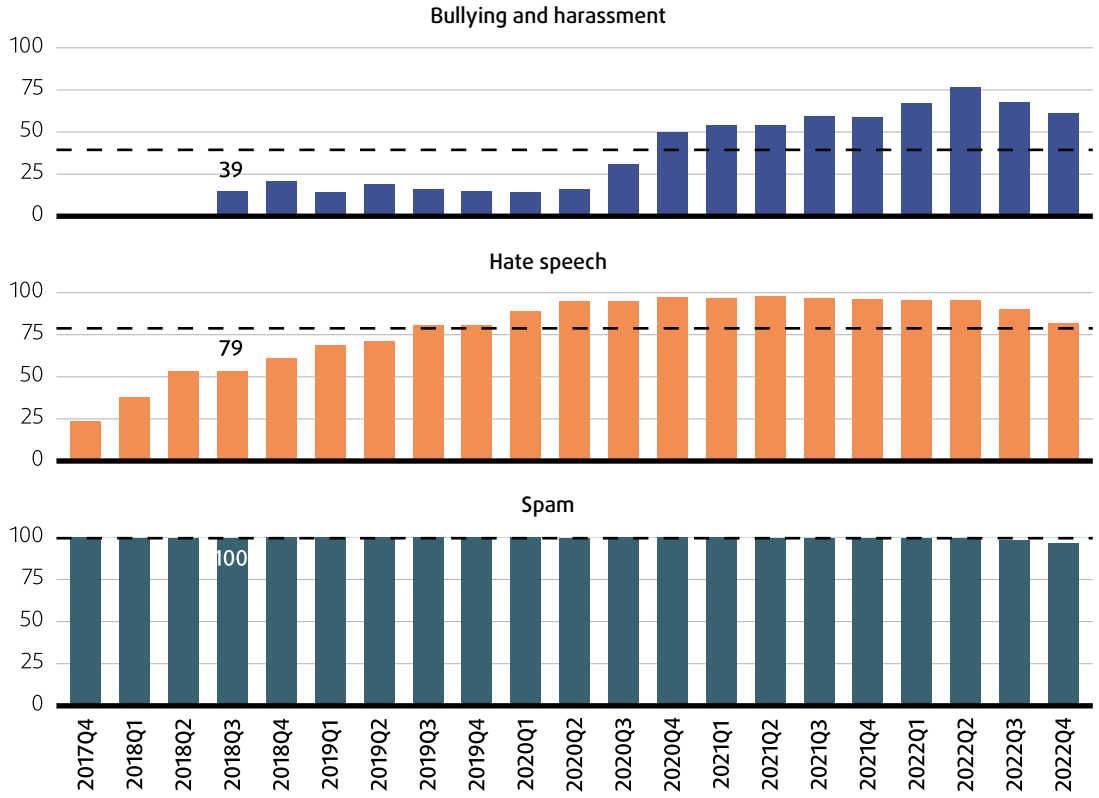
*(**) Meta (n.d.), '**Meta Research – Torch**'.*

(***) Meta (2023), **Community Standards Enforcement Report – May 2023 quarterly report**.

*(****) Google (2023), '**Transparency Report –YouTube Community Guidelines Enforcement**'.*

(*****) FRA (2022), **Bias in Algorithms – Artificial intelligence and discrimination**, *Publications Office of the European Union, Luxembourg.*

FIGURE 14: FACEBOOK'S RATE OF CONTENT ACTIVELY IDENTIFIED WITHOUT ANYONE ELSE REPORTING IT ACROSS ALL CONTENT ACTIONED, 2017–2022 (%)



Source: FRA (2023), based on data from Meta (2023), ***Community Standards Enforcement Report – May 2023 quarterly report***.

▲

The dashed line indicates the average. The average is also indicated as a number in the figure. The percentage gives the percentage of content actively identified by Facebook, which involves the use of algorithms for detection. It means, for example, that virtually all Spam is automatically detected.

3.1.2. Taking action on online hate

Different measures can be taken to address online hate, depending on whether it passes the threshold of illegality or not.

Governments can alert platforms of illegal content for removal, as mentioned previously.

Reddit received 214 requests from governments across the globe to remove content during the first half of 2022, according to its mid-year transparency report for 2022. Most came from non-EU countries, but 23 requests came from EU Member States, including France (14), Belgium (4), Germany (3), and Denmark (2). These requests from EU Member States led to Reddit removing 53 pieces of content or communities (i.e. discussion groups) from the platform. In addition, in the same period, Reddit registered 103 private party legal removal requests, of which 28 came from EU Member States, of

which 11 came from Germany ([5]). These removal requests are not limited to online hate, but also include other illegal content, such as copyright violations.

Removal of illegal content is essential. However, the assessment of what constitutes illegal content is very difficult, especially in the absence of clear definitions. The decisions often lie in the hands of privately employed content moderators at platforms. They must quickly decide whether content is illegal, does not respect the company's terms and conditions or whether it is fine to keep.

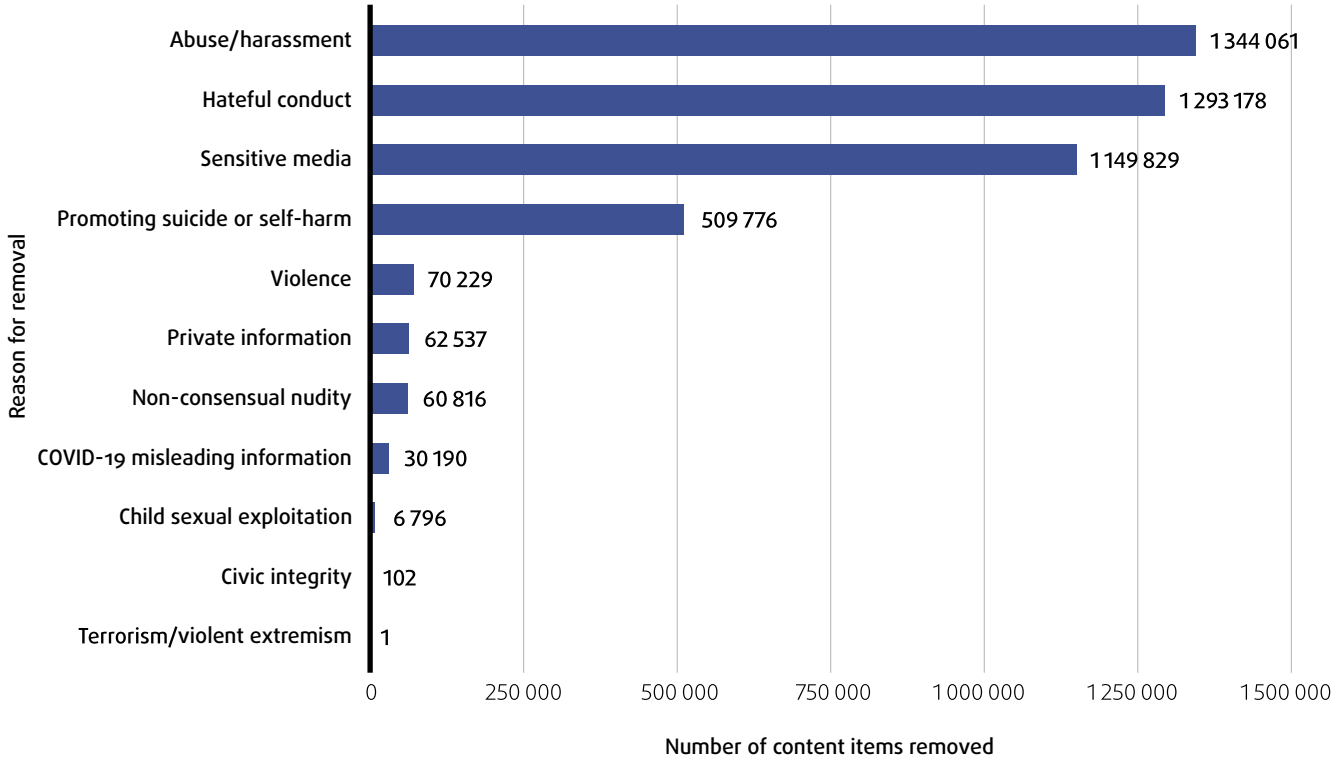
In addition to the requested removals mentioned above, in 2022 Reddit removed over 120 million posts and comments due to content policy violation, most of which consisted of spam. In 2022, Reddit removed 79 316 types of hateful posts and comments, out of which 39.4% were flagged by Reddit automation systems and 60.6% were actioned after user reports. Reddit also takes action on separate categories such as harassment and violent content. These two categories made up an additional 387 313 posts and comments removed in view of content policy violations ([6]).

The analysis of this report was based on content that was found to be likely to constitute online hate after platforms' had applied initial content moderation. This means it covered speech that may have slipped through the moderation system. This may also be a reason why the search found relatively few posts classified as incitement. A huge number of content items were removed from online platforms for including online hate.

As part of X's user agreement in 2022, the platform requires users to comply with its community standards and content policies ([7]). X reviews posts for hateful behaviour through a system of automated detection and users' reports ([8]). X's enforcement of its policy on hateful conduct is dependent on several factors, including the severity of the violation and an individual's record of rule violations. While X removes posts containing online hate, it immediately and permanently suspends users sharing direct violent threats against identifiable people or groups ([9]).

From July to December 2021, globally 11.6 million accounts were reported to X, mainly for hateful conduct. Some action was taken on 4.3 million accounts, of which 1.3 million were suspended. In those 6 months, X removed over 4.5 million pieces of content. The largest percentages were removed for hateful conduct and abuse or harassment, as **Figure 15** shows ([10]).

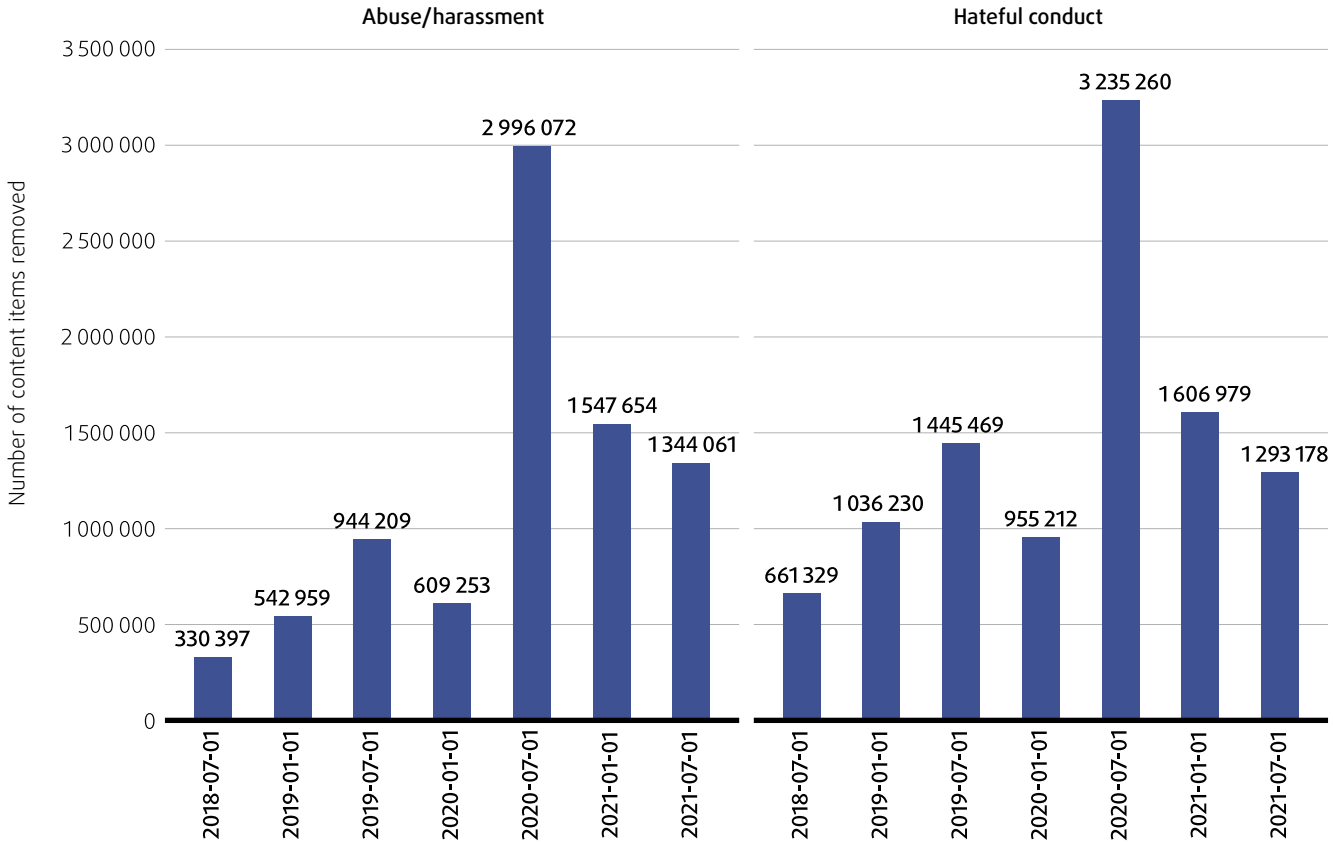
FIGURE 15: CONTENT REMOVED ON X, BY REASON FOR REMOVAL, JULY–DECEMBER 2021



Source: FRA (2023), based on data from Twitter, Transparency Center (2022), '**Rules enforcement**'.

Removals for hateful conduct and abuse or harassment peaked in the second half of 2020, as **Figure 16** shows. The considerable increase in removals of hate in the second half of 2020 is likely to be a result of an increase in online hateful posts and harassment during the pandemic ([11]). Increased use of the internet due to the restrictions on freedom of movement that were imposed during lockdowns may have driven the increase in online hate/harassment.

FIGURE 16. CONTENT REMOVED FROM X FOR ABUSE OR HARASSMENT AND FOR HATEFUL CONDUCT, 2018–2021



Source: FRA (2023), based on data from Twitter, Transparency Center (2022), '**Rules enforcement**'.

Changing content moderation policies has an impact on fundamental rights

Elon Musk took over as Twitter's (now X) new chief executive officer in October 2022. His approach to managing the company and regarding content moderation, specifically less moderation, has had a huge impact on the company (*). This report analysed Twitter data before his takeover.

According to the Institute for Strategic Dialogue, antisemitic content has more than doubled since Musk took over Twitter. He introduced changes to its content moderation policies, such as reinstating previously banned accounts and laying off staff working on content moderation. Between 1 June 2022 and 9 February 2023, 146 516 accounts posted 325 739 tweets identified as antisemitic, with the majority posted after Musk's acquisition of the company, according to the institute (**). In addition, there was an increase in newly created accounts that shared antisemitic or otherwise hateful content.

Musk's takeover enabled an unrestrained environment that is attractive to users with extreme right-wing, misogynistic or other harmful agendas, these findings suggest (***). This hateful environment may be a lucrative business model for the platform provider, according to the Center for Countering Digital Hate. Just 10 of the reinstated accounts of users known for harmful content could generate up to USD 19 million in advertising revenue for Twitter per year, the Center for Countering Digital Hate estimates (****). This highlights the need to further investigate the extent to which platforms profit from hateful content online.

*(*) Jones, B. (2022), **'Inside Elon Musk's takeover of Twitter'**, The New York Times, 11 November 2022.*

(**) Miller, C., Weir, D., Ring, S., Marsh, O, Inskip, C. and Prieto Chavana, N. (2023), **Antisemitism on Twitter before and after Elon Musk's acquisition**, *Institute for Strategic Dialogue and CASM Technology.*

*(***) Institute for Strategic Dialogue (2023), **'BBC Panorama research: Misogyny and abuse on Twitter before and after Elon Musk's takeover'**, blog post, 6 March 2023.*

(****) Center for Countering Digital Hate (2023), **Toxic Twitter – How Twitter generates millions in ad revenue by bringing back banned accounts**.

Removing content remains essential for illegal content. However, there are also other ways of dealing with (potentially illegal) online hate. For example, research has shown that warning users to 'think twice' before posting potential online hate is an effective additional measure to reduce online hate ([12]).

Nevertheless, platforms must continue removing online hate from their platforms to address illegal content, ensure users align with their terms and conditions and create a safer and friendlier online environment. Article 14(4) of the DSA sets out the rules regarding the application and enforcement of terms and conditions. This includes that platforms must act with due regard to the fundamental rights of the service's users.

To enforce its terms and conditions, X removes posts containing online hate and immediately and permanently suspends users sharing direct violent threats against identifiable people or groups ([13]).

YouTube also enforces its own rules. If the platform finds infringing content, it removes the content and sends the user a warning ([14]). After a second warning, YouTube issues a strike against the user's channel. If a channel gets three strikes within 90 days, the platform terminates the channel ([15]).

Reddit reserves its right to review, screen, edit and monitor content and to remove it for any reason, including violations of the user agreement. However, it does not indicate the mechanisms it uses to enforce its content policy ([16]).

These platforms will now have to act with due regard for fundamental rights when enforcing their terms and conditions, in line with Article 14(4) of the DSA. What this means in practice is yet to be clearly established ([17]).

One crucial aspect related to fundamental rights-compliant conduct with regard to online hate is paying attention to protected characteristics. Section 3.2.3 focuses on this issue.

3.1.3. Paying attention to protected characteristics and vulnerable groups

Hate expressed online does not necessarily consistently address people based on protected characteristics, such as gender or ethnic origin, as the analysis above shows. However, from a fundamental rights perspective, the expression of hate against people because of their protected characteristics is a key consideration in determining if content is online hate under legal thresholds. Expressing hatred of a person because she is a woman or because they are of African descent is a key fundamental rights concern.

The Charter's list of protected characteristics contains a variety of attributes. It prohibits discrimination on the basis of sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation. The grounds of online hate that the platforms included in this report cover vary (see **Table A1** in **Annex 1**).

- X prohibits online hate. This is defined in X's policy on hateful conduct as users' not being allowed to promote 'violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease' ([18]).
- Telegram's terms of service prohibit the promotion of violence on publicly viewable Telegram channels and bots ([19]). The platform's terms of service do not stipulate any guidance regarding online hate. Nor do they reveal how Telegram takes actions against violations of its terms and conditions.
- YouTube's terms and conditions bind users to comply with its policy on online hate ([20]). The policy identifies a wide range of prohibited content ([21]). Online hate is not allowed on YouTube. The platform states it will remove content promoting violence against or hatred of individuals or groups based on age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation or veteran status or against victims of a major violent event and their kin.
- Reddit's content policy stresses that the platform is a place of discussion free of threats of violence and harassment ([22]). Reddit prohibits content that incites violence or promotes hate based on identity or vulnerability ([23]). These groups include groups based on their actual and perceived race, colour, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy or disability.

None of the platforms refers to political opinion as a basis for online hate.

It is important to highlight that children need particular attention to allow for their safe participation in the online world. This report does not address questions related to the protection of children's well-being and online hate, but acknowledges that this topic requires special attention, analysis and

efforts. The United Nations Committee on the Rights of the Child issued its General Comment No 25 (2021) on children's rights in relation to the digital environment, which explains how countries should implement the Convention on the Rights of the Child, including guidance on relevant legislative, policy and other measures ([24]). The EU is actively developing an EU code of conduct on age-appropriate design. This will build on the DSA to create and monitor a code of conduct to ensure children's privacy, safety and security when using digital products ([25]).

Paying attention to vulnerable groups and protected attributes is necessary because these groups often face historical inequalities and protected attributes are the basis for prejudice and hate among the general population. Hatred of people based on their protected characteristics can harm the people concerned even in legal forms, such as in jokes or stereotypical thinking.

Interestingly, vulnerable groups appear to report fewer experiences of online harassment than the general population does. As indicated in **Chapter 1**, men indicate a slightly higher percentage of cyber harassment than women, as the FRA Fundamental Rights Survey shows. Ethnic minorities often indicate very low levels of harassment experienced online, FRA's surveys with vulnerable groups show (see **Chapter 1**).

However, this result may indicate an already lower level of engagement in online discussions. There may be a variety of reasons influencing this, including the danger and threat of being exposed to harassment. Hence, there may be a chilling effect of not engaging in online commentary or having an online presence as a way to avoid online hate.

3.1.4. Other measures needed to protect fundamental rights online in view of online hate

Detecting and taking action on online hate by various actors, with a view to protecting vulnerable groups and those with protected characteristics, is thus essential to protect people's fundamental rights. However, more action is needed. This section highlights additional measures and aspects that are relevant to fundamental rights-compliant online content moderation, without being exhaustive. The following subsections highlight the importance of efficient complaint mechanisms, the role of digital literacy and counterspeech to complement current content moderation efforts.

Complaint mechanisms

Decisions platforms take will necessarily include errors. There are errors with regard to not taking down content, as this report shows. All data included in this report were found online after platforms applied their moderation mechanisms. This means that there is often online hate that should have been removed.

However, there are also cases of legal content being erroneously removed. Since it was not possible for this report to analyse what posts platforms had already taken down, we do not have any sense of how much content is erroneously deleted. Nevertheless, an appropriate complaint mechanism must be in place so that users know where and how to easily complain and seek redress.

Effective dispute settlement options must be in place for user complaints, including non-judicial and judicial avenues, in line with Articles 20 and 21 and recital 55 of the DSA. Every online user must have the ability to go to a court over a decision (not) being taken. The ultimate authorities deciding the illegality of online content are the courts, not platforms.

Increasing digital literacy through formal and informal education

Online hate often targets individuals in connection with their membership of a particular group. Users play an essential role in identifying and reporting online hate and can potentially react with counterspeech. Therefore, it is important that social media users are well informed about online hate, platform properties and their rights. Thus, systematic digital education programmes and non-formal training can effectively complement regulations such as the DSA in combating online hate online ([26]).

Young people can especially benefit from digital education, as they spend more time online than any other age group and can be particularly affected by hateful content ([27]). As young people increasingly network online, digital education may lead to more equitable conversations online and, potentially, stronger civic participation.

Findings from Greece highlight this. A group of young Roma was equipped with non-formal critical digital literacy, non-violent communication, active participation and human rights training. This was shown to be effective in strengthening their resilience and critical awareness to safely navigate online content. It also deconstructed adversarial narratives about their group ([28]).

The Council of Europe's publication *Bookmarks – A manual for combating hate speech online through human rights education* is an excellent resource for education in this respect ([29]).

Counterspeech

Data collection searching for online hate may also pick up counterspeech. That is, it may pick up people calling out those expressing hatred or people reacting to online hate. This speech is important to counteract the negative and hateful environment created for vulnerable groups.

This report briefly discusses counterspeech and the role of counterspeech is an additional promising asset, specifically its spread and reach ([30]). Counterspeech also often includes elements of anger and hate, this report shows. This may be understandable in some situations. However, research has pointed out that empathy-based counterspeech is more effective than counterspeech that contains insults of its own ([31]).

3.2. UNDERSTANDING PLATFORMS' CONDUCT IN TERMS OF SAFEGUARDING FUNDAMENTAL RIGHTS

At present, there is a lack of knowledge on how well platforms perform in terms of protecting fundamental rights. It is the responsibility of the authorities to oversee and control platforms' implementation of laws to safeguard fundamental rights. This calls for efforts to increase knowledge on hate that exists online and also to learn about what kind of content platforms delete.

Platforms' handling of online hate should be effectively monitored through a variety of indicators. This includes the transparency reports required under Articles 15, 24 and 42 of the DSA and the risk assessment procedure in Article 34 of the DSA.

The risk assessment procedure requires VLOPs and VLOSEs to carry out assessments at least once every year to identify, analyse and assess systemic risks such as the dissemination of illegal content through their services (Article 34(1)(a)) and any actual or foreseeable negative effects for the exercise of fundamental rights (Article 34(1)(b)). Under the terms of the DSA, the assessment of the effect on fundamental rights should consider the right to human dignity, right to privacy, right to the protection of personal data, right to freedom of expression, right to non-discrimination and rights of the child.

The assessment of systemic risks is not limited to illegal content. It also concerns content that is legal but contributes to a systemic risk to fundamental rights. Misogyny should be one of the systemic risks taken on board in the context of the risk assessment procedure given its prevalence, as this report highlights.

If risks are identified, VLOPs and VLOSEs must 'put in place reasonable, proportionate and effective mitigation measures' tailored to the systemic risk identified, 'with particular consideration to the impacts of such measures on fundamental rights' (Article 35). Articles 34 and 35 of the DSA offer an important basis for achieving a clearer picture of the risks online hate poses to fundamental rights and taking measures to address these. Of the four selected platforms, X and YouTube are subject to this obligation. Reddit and Telegram were not VLOPs at the time of writing this report.

Moreover, the role of algorithms needs to be better analysed through learning about the extent to which algorithms may promote online hate.

Holding platforms accountable, especially under the DSA, requires the use of various measures and data. In that context, Article 40 of the DSA provides a basis for researchers to better access data needed to conduct 'research that contributes to the detection, identification and understanding of systemic risks

in the Union' and 'the assessment of the adequacy, efficiency and impacts of the risk mitigation measures pursuant to Article 35'.

Measures to improve access to platforms are required, as platforms are restricting researchers' access. This makes them less transparent and less accountable to independent review. However, assessments cannot only rely on data obtained from platforms, but require additional data and analysis for a comprehensive critique. This research, monitoring and analysis includes:

- using data and indicators online platforms provide, for example on removals and complaints;
- directly accessing content of online platforms through using APIs and web scraping to detect online hate and carry out other analysis of online content;
- obtaining data from users of online platforms, including through access to their social media data and interviewing them about their experiences;
- using experiments and tests to obtain hard evidence on how platforms' conduct influences the enjoyment of fundamental rights.

There is an urgent need to clarify the legal framework and options for better access to data. However, questions that require answers in relation to fundamental rights compliance should drive the need for data. It is crucial to focus on the conduct of platforms in order to implement the DSA. What design features, measures and circumstances of the platforms contribute to fundamental rights breaches?

For example, the following non-exhaustive list of questions need answers to safeguard fundamental rights.

- In terms of platforms trying to increase screen time through using recommendation algorithms, how strongly is hate promoted over other less hateful content?
- Do content moderators and/or algorithms used consistently or systematically miss online hate against certain groups?
- Do platforms' terms and conditions cover all relevant protected attributes as described in Article 21 of the Charter?
- Do users sufficiently read and understand terms and conditions? What percentage of users do not read and understand terms and conditions? What are the differences between groups and how does this change over time?
- What kind of content is deleted and how often do people complain about it (not) being deleted?
- How often do people feel they are not able to express themselves freely on certain platforms and what are the differences in (lack of) engagement between groups?
- How frequently do users see content that offends them? What are the differences between groups and how does this change over time?

Answering these questions, among others, will be crucial for the successful protection of fundamental rights online. They require combined evidence based on data from platforms, user surveys and researcher experiments. The next step is to address measures to mitigate the challenges identified. This is beyond the scope of this report.

Endnotes

(1) European Commission (2023), '**Digital Services Act: Commission designates first set of very large online platforms and search engines**', press release, 25 April 2023.
(2) Germany, **Act to improve law enforcement in social networks (Network Enforcement Act – NetzDG)** (*Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG)*), 1 September 2017, Section 4(2).
(3) Sweden, Law (1998:112) on responsibility for electronic bulletin boards (***Lag (1998:112) om ansvar för elektroniska anslagstavlor***), 12 March 1998.
(4) Italy, **Administrative proposal – measures to prevent and combat the spread of hate manifestations through the internet** (*Proposta di Legge – Misure per la prevenzione e il contrasto della diffusione di manifestazioni d'odio mediante la rete internet*), Camera dei Deputati, No 2936, 10 March 2021.
(5) Reddit (2022), '**Mid-year transparency report 2022**'.
(6) Reddit (2022), '**2022 transparency report**'.
(7) Twitter (2022), '**Twitter user agreement**'.
(8) X (2023), '**Hateful conduct**'.
(9) X (2023), '**Hateful conduct**'.
(10) FRA calculations based on data from Twitter, Transparency Center (2022), '**Rules enforcement'**. Downloaded in February 2023.
(11) See, for example, Toliyat, A. Levitan, S. I., Peng, Z. and Etemadpour, R. (2022), '**Asian hate speech detection on Twitter during COVID-19**', *Frontiers in Artificial Intelligence*, Vol. 5, 932381.
(12) See, for example, Yildirim, M. M., Nagler, J., Bonneau, R. and Tucker, J. A. (2021), '**Short of suspension: how suspension warnings can reduce hate speech on Twitter**', *Perspectives on Politics*, Vol. 21, No 2, pp. 651–653.
(13) X (2023), '**Hateful conduct**'
(14) YouTube (n.d.), '**YouTube Help – hate speech policy**'.
(15) YouTube (n.d.), '**YouTube Help – hate speech policy**'.
(16) Reddit (n.d.), '**Reddit user agreement**', version of 21 September 2021.
(17) Quintais, J. P., Appelman, N. and Fahy, R. (forthcoming), '**Using terms and conditions to apply fundamental rights to content moderation**', *German Law Journal*.
(18) Twitter (2022), '**Hateful conduct**'.
(19) Telegram (n.d.), '**Terms of service**'.
(20) YouTube (2022), '**Terms of service**'.
(21) YouTube (2022), '**YouTube Help – hate speech policy**'.
(22) Reddit (2022), '**Reddit content policy**'.
(23) Reddit (2022), '**Reddit content policy**'.
(24) United Nations (2021), **General comment No. 25 (2021) on children's rights in relation to the digital environment**.
(25) European Commission (n.d.), '**Special group on the EU Code of conduct on age-appropriate design**'.
(26) Buckingham, D. (2020), '**Epilogue: rethinking digital literacy: media education in the age of digital capitalism**', *Digital Education Review*, Vol. 37, pp. 230–239; European Roma Rights Centre (2023), '**Media literacy is a vaccine against disinformation about Roma**', 23 May 2023.
(27) Statista (2023), '**Average daily time spent using the internet by online users worldwide as of 4th quarter 2022, by age and gender**'; UNICEF (2017), ***How does the time children spend using digital technology impact their mental well-being, social relationships and physical activity? An evidence-focused literature review***, Innocenti Discussion Paper 2015-02.
(28) Agapoglou, T., Mouratoglou, N., Tsioumis, K. and Bikos, K. (2021), '**Combating online hate speech through critical digital literacy: reflections from an emancipatory action research with Roma youths**', *International Journal of Learning and Development*, Vol. 11, No 2, pp. 104–120.
(29) Council of Europe (2020), *Bookmarks – A manual for combating hate speech online through human rights education*, revised edition, Strasbourg.
(30) Ozalp, S., Williams, M. L., Burnap, P., Liu, H. and Mostafa, M. (2020), '**Antisemitism on Twitter: collective efficacy and the role of community organisations in challenging online hate speech**', *Social Media + Society*, Vol. 6, No 2.
(31) Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Dirksen, L., Hall, A., Jochum, M, Murias Munoz, M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F, Gilardi, F. and Donnay, K. (2021), 'Empathy-based counterspeech can reduce racist hate speech in a social media field experiment', *Proceedings of the National Academy of Science of the United States of America*, Vol. 118, No 50, e2116310118.

WAYS FORWARD

Online hate speech targeted at specific groups in a variety of ways is prevalent on many online platforms. It is easy to detect some incidents of online hate. But analysing systemic infringements of fundamental rights through online hate remains very difficult.

This report shows that it is easy to find a significant amount of misogyny and hate against people of African descent, Jews and Roma when searching online platforms Reddit, X and YouTube with selected keywords, and Telegram in selected open groups. The data collection covered the countries Bulgaria, Germany, Italy and Sweden by limiting the search to the main language spoken in those countries. Out of 1 573 posts that trained coders analysed, 53 % we assessed as hateful. Hateful posts found most often include offensive language; they frequently involve denigration and negative stereotyping. Less often, hateful posts are assessed as incitement to violence, hate or discrimination.

Hate against women is found to be most prevalent across all platforms covered. Most often, posts targeted against women include denigrating language, meaning comparing people to objects or animals. Women also face higher levels of incitement to violence than other groups, with online violence against women most often based on sexualised violence. Posts against people of African descent, Jews and Roma most often contain negative stereotyping.

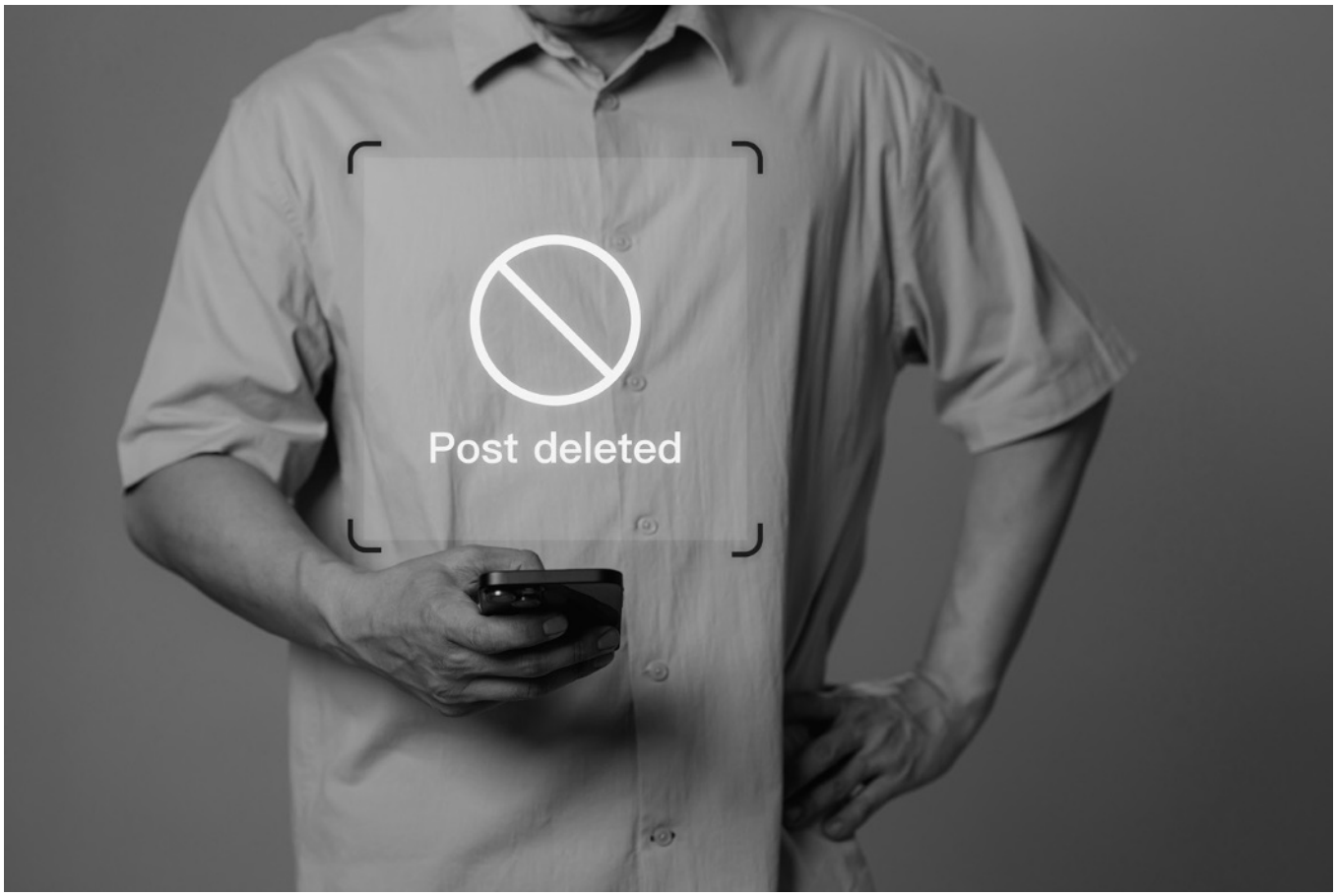
Overall, the posts captured in this data collection rarely include more than one target group and, as such, provide limited evidence on intersectionality with respect to online hate. Only 5% of all posts coded as hateful against any of the four target groups include more than one target group.

Almost half of all hateful posts (47 %) are classified as harassment; that is, harassment perceived to be targeted at one or more specific individual(s). Women are particularly often targets: two thirds (67 %) of all hateful posts targeted at women were found to be harassment.

The analysis does not include an assessment of the legality of the posts. However, for a selected number of posts (10 per country) coders assessed the extent to which the post might be categorised as incitement, as defined in the Council framework decision. In addition, legal experts assessed these posts. The results show that legal experts usually have a higher threshold and consider few posts to potentially fall under the definition. This shows how difficult and uncertain assessments with respect to the legality of posts are, especially when it comes to practical content moderation and research efforts that require a quick assessment of posts.

The results are necessarily limited because use of keywords can detect only certain types of online hate and will miss other types, particularly online hate that uses more subtle or masked language.

However, no methodology provides flawless and comprehensive results. Current research on technology-facilitated violence against women is not only scarce but marked by methodological flaws, as a recent UN Women and World Health Organization publication emphasises ([1]). The main issues are linked to a lack of standardisation and a lack of methodologies for the measurement of different phenomena.



In addition, this report only captures posts that platforms had not already taken down. Exploring content moderation practices within platforms remains even more difficult, as relevant information is very difficult to access. The challenges of assessing the legality of posts in practice and determining who should be conducting such assessments at what scale and how quickly remain.

Furthermore, the phenomenon of online platforms is relatively recent and the ways in which online hate is expressed are developing over time. The detection of, analysis of and fight against online hate will remain a moving target in the years to come. Consequently, a variety of methodologies are needed to understand interferences with fundamental rights – including freedom of expression – due to online platforms' content moderation practices.

However, such an analysis is difficult due to limited access to data. Platforms remain very protective when it comes to providing access to their data for a variety of reasons, which may include the protection of business models, privacy concerns and potential fears of revealing flaws in platforms' content moderation practices. While their behaviour is understandably cautious, it is not acceptable for the public and regulators to remain in the dark regarding these practices that may put fundamental rights at risk.

One important way forward is to crack open the parts of platforms' conduct relevant to understanding how fundamental rights are put at risk online.

Article 40 of the DSA offers an opportunity to push for algorithmic and online platform transparency, thus enabling researchers to gather the evidence needed to grasp the scale of online hate independently, regardless of the varying access policies of online platforms.

Currently, developments are, however, going in the other direction. The need to provide data access becomes ever more important in light of the changes to X's data access model and challenges in accessing data from Facebook and Instagram.

Platforms may make efforts to protect access to their valued data. However, regulators must protect the fundamental rights to which they have committed. It may very well be the case that some elements of business models, such as promoting online hate through non-transparent algorithms, go against the protection of fundamental rights. Consequently, they must be scrutinised and, if necessary, stopped.

An understanding of potential fundamental rights violations on online platforms will only come from independent scrutiny by relevant authorities and only be possible with the support from independent research on online platforms. The promotion of a variety of opportunities to research platforms will provide regulators with the required evidence in the years to come.

Endnotes

(1) UN Women and World Health Organization, Joint Programme on Violence against Women Data (2023), ***Technology-facilitated Violence against Women: Taking stock of evidence and data collection***.

ANNEX 1: TECHNICAL DETAILS OF THE METHODOLOGY

DATA COLLECTION

Data were collected from X, Reddit, YouTube and Telegram. These platforms were selected based on their relevance across the four countries and feasibility considerations regarding access to data. The data collected only include publicly available posts from pages, channels or groups.

Data from X, Reddit and YouTube were collected using Brandwatch ([1]), a social media analytic tool. Brandwatch archives social media data, allowing users to download public data on a set of online platforms. Data from Telegram were collected using web scraping on Python.

Data collection took place over the 6 months from 25 January to 25 July 2022. The data collection period was selected to ensure a sufficient sample size across platforms in all four Member States and to ensure equal posts for all target groups in the Member States.

Data collection for X, Reddit and YouTube filtered data to find posts that matched any of the keywords specified. Brandwatch provides an interface where posts can be filtered by keywords and downloaded. The query for each country contained two lists, following the logic of The Weaponized Word ([2]) – an online lexicon including lists of words that indicate hate and malicious speech. The study team acknowledges the valuable input and consultation of Timothy Quinn from The Weaponized Word database and the Dark Data Project.

The query distinguished between discriminatory terms and derogatory terms. Discriminatory terms are terms referring to (perceived) characteristics of a target group. Derogatory terms are terms that are broadly insulting regardless of the recipient's identity and might be found in combination with discriminatory terms.

In some cases, a word is both discriminatory and derogatory. If so, the word was included in both lists.

For illustrative purposes, examples of discriminatory keywords in German include the word *Juden****, which is a slur meaning 'Jewish pig'. Others are words that refer to prostitutes or female genitals in offensive ways or discriminatory terms targeted at people of African descent, such as 'N****'. Some words are a combination of other words, such as *Migratten*, a mixture of the word migrant and rats. These words were searched for in combination with derogatory words, such as 'shoot dead' or 'son of a b****'.

Country experts from the four Member State supported the compilation of lists of key words for the collection of online hate. The lexicon of The Weaponized Word was the basis of the lists. The country experts and country leads then refined the list for their country based on other literature, expert inputs and their understanding of the national context.

The approach to compiling these lists was the same in all four Member States. However, due to the national context, the lists varied in the number of words and the nature of the terminology.

A group of experts held two workshops to review the lists the national researchers had created. The experts included country experts working on the study and additional experts on online hate regarding the target groups covered. The lists were updated based on the experts' feedback.

Following the workshops, the keyword search queries were tested to ensure that they generated content relevant to the study aims. The lists of keywords were then adapted following the review of results, including the frequencies of relevant words in the text. Adapting the search queries in this way increased the number of relevant posts in the dataset.

For Telegram, the team used API to acquire data from a list of expert-identified Telegram groups and chats, while staying within the restrictions of the user agreement. The study team used Python to search Telegram in an analogous way to the Brandwatch queries. The study team used the queries on Python to select observations corresponding to the keywords selected for the research.

However, Boolean queries on Python could not fully match the Brandwatch queries. For example, 'OR' and 'NOT' operators could not be used and the 'AND' operator could not be used in conjunction with brackets. Therefore, Telegram data collection included observations with only discriminatory or derogatory terms. This posed challenges for sampling data from Telegram for annotation.

The country experts identified Telegram channels to search. They conducted a preliminary review of evidence from grey literature on Telegram channels that reports and the news had mentioned being linked to extremist groups and hate speech.

In addition, country experts searched for Telegram channels using a combination of discriminatory and derogatory terms and mentions of Telegram on Google. They looked at the results generated on Telegram Web. For example, country experts searched for the discriminatory word 'gypsy' combined with derogatory words.

The above data collection yielded over 344 000 posts. From those, approximately 400 posts were sampled per country to generate a dataset for hand coding.

The study team used stratified random sampling without replacement to split the posts collected into 12 strata for each Member State of focus. Each stratum was the sample of observations for the intersection of an online platform and a target group. There were 12 strata for each country because data were collected for four online platforms and three target groups, as follows:

— Germany: women; people of African descent; Jews.

— Sweden: women; people of African descent; Jews.

— Italy: women; people of African descent; Roma.

— Bulgaria: women; people of African descent; Roma.

The sampling procedure was designed to optimise the equal representation of target groups within the platforms of interest and the disproportionate representation of online platforms with small sample sizes compared with X.

However, the sample sizes for target groups and online platforms differed between countries. Therefore, the team made practical adjustments to achieve the desired sample size of approximately 400 observations per country. If a platform with a smaller sample size could not reach the target, the next smallest online platform was oversampled within a country. The same procedure was adopted for target groups with small sample sizes.

LABELLING ONLINE HATE

This section describes how the team analysed online hate for this report. The results of the report can only be read and understood against the background of its data analysis methodology. This section first describes how the team categorised the collected data.

Two coders for each country/language were trained on how to carry out the coding in a workshop on two days, with a week between sessions to allow practice. The coding workshop was delivered by the project manager in English to ensure consistency between countries and languages. Coders practiced coding during the workshop and between the workshop days, with consistency checking. The second session discussed areas of inconsistency and the coders' feedback helped create further iterations of the coding grid and how to code guide.

Coding and training took place in three steps. First, two coders – a main coder and a second coder – coded the first 50 posts. Then consistency checks took place and the coders had an opportunity to discuss posts identified as inconsistently coded. The consistency checks focused on the coding of the two typologies: target group typology and hateful characteristics typology. Once a sufficient level of consistency was achieved, the two coders coded a further 150 posts, which went through the same consistency checking.

Assessment of the inter-coder reliability of the coding decisions of the two coders for each country combined comparing percentage agreement and using Cohen's kappa as a statistical measure of the reliability. The main coder coded the remaining 200 posts on their own. The final dataset for each country includes about 200 double-coded posts and 200 single-coded posts.

The data cleaning stage discovered that the sample included some duplicate posts and these were removed. That is why the final number of posts is a little less than 400 posts per country.

CATEGORISATION OF ONLINE HATE

This study used background research and an expert workshop to develop a typology based on the hateful character of post content for the purpose of categorising the empirical analysis. This study distinguishes five main categories of online hate:

— incitement to violence, discrimination or hatred

— denigration

— offensive language

— negative stereotyping

— other hateful content.

These categories are not mutually exclusive. Expressions of online hate may fall into several categories.

Incitement to violence, discrimination or hatred. Different EU Member States have interpreted incitement in different ways. However, incitement is acknowledged as including dissemination or distribution of material to incite violence or hatred ([3]). This study uses a definition of incitement based on the European Commission against Racism and Intolerance (ECRI) recommendation (see Section 1.2).

It refers to content that clearly encourages or urges the audience to:

— commit violence;

— act in a discriminatory manner, which means treating someone differently because of a (perceived) protected characteristic;

— act in a hateful manner, including speaking or writing.

There are other ways of distinguishing between kinds of incitement. Article 1 of the framework decision distinguishes between kinds of incitement (inciting to violence or hatred) and the public dissemination of material that amounts to incitement. Similarly, the International Convention on the Elimination of All Forms of Racial Discrimination distinguishes between dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination and incitements to acts of violence ([4]). The ICCPR distinguishes between 'incitement to discrimination, hostility or violence' ([5]).

The recent recommendation from the Council of Europe also expanded the scope of online hate beyond incitement ([6]). In the recommendation, hate speech is understood as 'all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as "race", colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation.'

Denigration. This means an attack on the capacity, character or reputation of one or more people in connection with their (perceived) membership of a particular group or, as the Council of Europe recommendation states, 'by reason of their real or attributed personal [protected] characteristics' ([7]). This study took other types of denigration into consideration, including:

- objectifying language, meaning treating a person as a tool or toy, as if they have no feelings, opinions or rights of their own ([8]);
- dehumanising language, meaning describing a person or group as somehow less than human, for example by comparing them to animals, parasites or disease ([9]);
- other kinds of attack on a person's capacity, character or reputation in connection with their membership of a particular group with protected characteristics ([10]).

Negative stereotyping. Negative stereotyping is when certain negative traits and characteristics are 'negatively valenced and attributed to a social group and to its individual members' in relation to protected characteristics ([11]).

Offensive language. Offensive language means 'hurtful, derogatory or obscene' language ([12]), such as insults referring to protected characteristics. The targeted person judges whether specific words or language is offensive and this is highly context dependent.

Other hateful content. This residual category may include support for hateful ideologies or Holocaust denial. This is also covered in Article 1 of the framework decision. It covers hateful content such as 'publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes … directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin when the conduct is carried out in a manner likely to incite to violence or hatred against such a group or a member of such a group'.

Incitement to violence or acting in a discriminatory manner are considered hateful as a matter of course. However, denigration, negative stereotyping or offensive language may not always be considered hateful. This will depend on the target's perception and the author's intention.

For instance, offensive language may not be considered hateful when intended and perceived as playful, for example among acquaintances. However, this is a very subjective assessment.

Given this, offensive language may not always be equated with hatred ([13]). It is important to note that this distinction between hateful and non-hateful entails subjective assessments. In the case of this study, the coders made those assessments.

When making a judgement call, different factors may be taken into account. For instance, online hate definitions require consideration of various factors, such as context or the form of the content ([14]). The Rabat plan of action gives six aspects to be taken into account when considering incitement to hatred: (1) the social and political context, (2) the status of the speaker, (3) the intent to incite the audience to action against a target group, (4) the content and form of the speech, (5) the extent of its dissemination and (6) the likelihood of harm, including imminence.

This elaborate consideration of factors was not possible within the scope of this study and when categorising a significant number of posts. The focus of the analysis is whether the content in the post expresses hatred. Not considering the wider context in which a post is placed is a limitation of this study, but at the same time reflects how posts on online platforms are assessed. Assessment using automation and limited human coding does not typically cover context.

TARGET GROUPS AND VICTIMS OF ONLINE HATE

International and EU instruments, online platforms and researchers all discuss online hate through reference to the target groups of the material. This study focuses on four target groups covering different protected characteristics. The groups are women (gender or sex), people of African descent (ethnic origin), Jews (religion and/or ethnic origin) and Roma (ethnic origin).

Institutions typically organise these target groups into categories when discussing them, depending on their relevance to the institutions' mandates. For instance, the UN Committee on the Elimination of Racial Discrimination focuses on racist hate speech ([15]) and the Committee on the Elimination of Discrimination against Women focuses on misogyny ([16]). The ICCPR primarily focuses on racist hate speech ([17]) and ECRI focuses on racist, xenophobic and antisemitic content ([18]).

Online platform user agreements often list specific target groups in the terms of service when explaining the type of content that is prohibited (**Table A1**). Similarly, signatories to the EU code of conduct provide the European Commission with reports of online hate based on reference to target group characteristics ([19]).

TABLE A1: PLATFORMS LISTING PROTECTED CHARACTERISTICS IN THEIR TERMS OF SERVICE

Platform	Protected characteristics listed in terms of service
Reddit (*)	'people that incite violence or that promote hate based on identity or vulnerability will be banned. Marginalized or vulnerable groups include, but are not limited to, groups based **on their actual and perceived race, colour, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability.**'
Telegram (**)	'You … agree not to … promote violence on publicly viewable Telegram channels.' No protected characteristics mentioned.
Twitter (***)	'You may not promote violence against or directly attack or threaten other people on the basis of **race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.**'
YouTube (****)	'We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: **age, caste, disability, ethnicity, gender identity, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their family members, and veteran status.**'

*(*) Reddit (n.d.), '**User agreement**', as of 21 September 2021.*

*(**) Telegram (n.d.), '**Terms of service**'.*

*(***) Twitter (2022), '**Hateful conduct**'. The terms have changed since 2022, as some parts were cut in the* **2023 version**.

*(****) Google (n.d.), '**Transparency Report – Featured policies: Hate speech**'.*

NB: Emphasis added.

In a similar vein, researchers tend to focus on specific categories when exploring online hate. For instance, they may focus on content that is antisemitic ([20]), xenophobic ([21]), sexist (or misogynistic) ([22]) or Islamophobic ([23]).

In addition to the target group, it is possible to categorise the attributes of the targeting itself. This includes the level of targeting – whether it is directed at individuals, specified groups or unspecified groups. Meta's policy, for instance, defines hate speech as 'a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease' ([24]).

Finally, it should be acknowledged that not all hateful content has a specific target. In some cases, content may be in support of an ideology that is generally deemed to be hateful.

ADDITIONAL CATEGORISATIONS AND CONSIDERATIONS

The current study has focused on the above typology of online hate. In addition, we note the following possible approaches to categorising online hate that may be of interest in this context.

— **Whether content can be classified as harassment.** While the framework decision does not refer to harassment, the racial equality directive and the employment equality directive do ([25]). They define harassment as 'unwanted conduct' with a bias motivation taking place 'with the purpose or effect of violating the dignity of a person and of creating an intimidating, hostile, degrading, humiliating or offensive environment'. When Member States prohibit this conduct in their criminal legislation, harassment can be considered a hate crime ([26]). Harassment, as defined in these directives, can specifically target individuals.

— **Original content or posts about original content.** The literature distinguishes hateful content from content that represents counterspeech ([27]), journalism or commentary / awareness raising. For example, this covers when the author of a post reports someone else's online hate with the purpose of informing the public (e.g. journalism), commenting on it (e.g. commentary) or raising awareness.

— **Whether criminal, subject to civil and administrative liability or permissible as freedom of expression.** Only a very narrow category of hate speech should be criminalised, according to the UN action plan ([28]), but the instruments described in this report set out a wider scope for what is not permissible (see **Section 1.2**).

— **Connected to an event.** Research reveals some patterns of online hate, as it tends to spike after certain 'trigger' events, such as terrorist attacks, migration flows, demonstrations, riots and elections ([29]). Similarly, a study on online hate in the context of the refugee crisis observed an increase in the expression of online hate against

migrants in several EU Member States, such as Austria, the Netherlands and Spain. Members of the International Network against Cyber Hate participated in the study, along with project partners from Austria, Belgium, France, Germany, the Netherlands and Spain ([30]).

— **Whether humorous/sarcastic or not.** Online hate can contain humorous or sarcastic elements. The study team considered the Cambridge Dictionary definitions of humour and sarcasm in the analysis phase of the current study ([31]). It is acknowledged that humour and sarcasm may or may not be linked to or be perceived as hatred.

— **Whether public or private.** Online hate can be public or private, for example through private correspondence in emails. The framework decision specifically focuses on public instances of hatred. There is clearly a category of private hatred, but this is outside the scope of the current study.

— **Format.** Online hate may be expressed in a text-based format or in other formats, such as images, video, emojis or sound. The framework decision refers to 'public dissemination or distribution of tracts, pictures or other material' ([32]). However, the scope of the present research is limited to online hate in word form.

DATASET DESCRIPTION

From 1 January 2022 to 30 June 2022, 344 132 posts were collected from the four platforms. **Table A2** breaks down the posts by language and platform.

Most posts were in German (39.49 %, or 135 882 posts) or Italian (58.26 %, or 200 500 posts). Swedish posts made up only 1.96 % (6 745 posts) of posts and Bulgarian was used in fewer than 1 % of posts (1 005 posts). This is linked to the size of the language/country, as the numbers of posts in German and Italian were much larger.

The numbers of posts collected do not directly reflect the prevalence of online hate in the countries. Other factors also influence the numbers, such as how widely the platform is used in each country.

TABLE A2: NUMBER OF POSTS COLLECTED, BY LANGUAGE AND PLATFORM

Platform	Bulgaria	Germany	Italy	Sweden	Total
Reddit	74	10 964	2 050	439	**13 527**
Telegram	694	4 687	97 640	486	**103 507**
X	232	118 639	99 432	5 796	**224 099**
YouTube	5	1 592	1 378	24	**2 999**
Total	**1 005**	**135 882**	**200 500**	**6 745**	**344 132**

By and large, the four countries are represented in order of how widely spoken the language is as a mother tongue in Europe, but reversed for German and Italian. German is more widely spoken as a mother tongue ([33]) but more posts were collected in Italian. This is largely due to the differences in the number of Telegram posts. The different sampling technique for Telegram may mean the numbers of Telegram posts are not truly indicative of the presence of discriminatory and derogatory words in online posts.

Posts were collected based on the language they were written in rather than their geographical origin. Not all posts have a geographical marker, so collecting data based on country of origin alone would have substantially reduced the sample size. The remaining analysis uses the language of the posts as a proxy for origin.

Language may be a suitable proxy for place of origin for most languages that are less widely spoken, such as Bulgarian or Swedish. Others, such as German, are both more widely spoken as a second language and an official language of multiple countries. Therefore, posts assigned to Germany may have easily stemmed from Austria, Switzerland, Luxembourg or any state where German is widely spoken.

X is the dominant source of posts collected, making up 65 % of all posts. In Germany and Sweden, X is the primary source of posts, making up 87 % (118 639 of 135 882) and 86 % (5 796 of 6 745) of all posts, respectively. For Bulgaria

and Italy, however, Telegram played a more prominent role. Telegram is the primary source of posts (69 %, or 694 of 1 005 posts) in Bulgarian and X and Telegram are equally represented in Italian (50 % or 99 432 and 49 % or 97 640 of the 200 500 posts, respectively).

The approach to collecting Telegram data differed substantially from the approach used for X. Therefore, these headline figures for Telegram are not necessarily suggestive of a difference in platform use in Bulgaria and Italy. Restricting attention to the three platforms for which data collection used Brandwatch (Reddit, X and YouTube), all countries follow the same pattern: most posts come from X, with smaller numbers from Reddit and even smaller numbers from YouTube.

TABLE A3: NUMBER OF POSTS COLLECTED, BY TARGET GROUP AND PLATFORM

Platform	Women	People of African descent	Jewish people	Roma
Reddit	4 024	4 132	1 077	116
Telegram	7 228	2 559	125	1 413
X	112 371	88 820	15 534	4 745
YouTube	769	1 586	176	91
Total	**124 392**	**97 097**	**16 912**	**6 365**

Posts containing language targeted at women represent over a third of all posts collected (**Table A3**). Posts were classified based on whether they contained language associated with the four focus target groups. This does not mean that they contain online hate that targets any of those groups, but that they contain keywords that potentially indicate online hate.

Of the 344 132 posts collected, 36 % (124 392) contain discriminatory words targeted at women. Over 90 % of those posts (112 371 of the 124 392 posts) come from X. Telegram dominates the remaining 10 %.

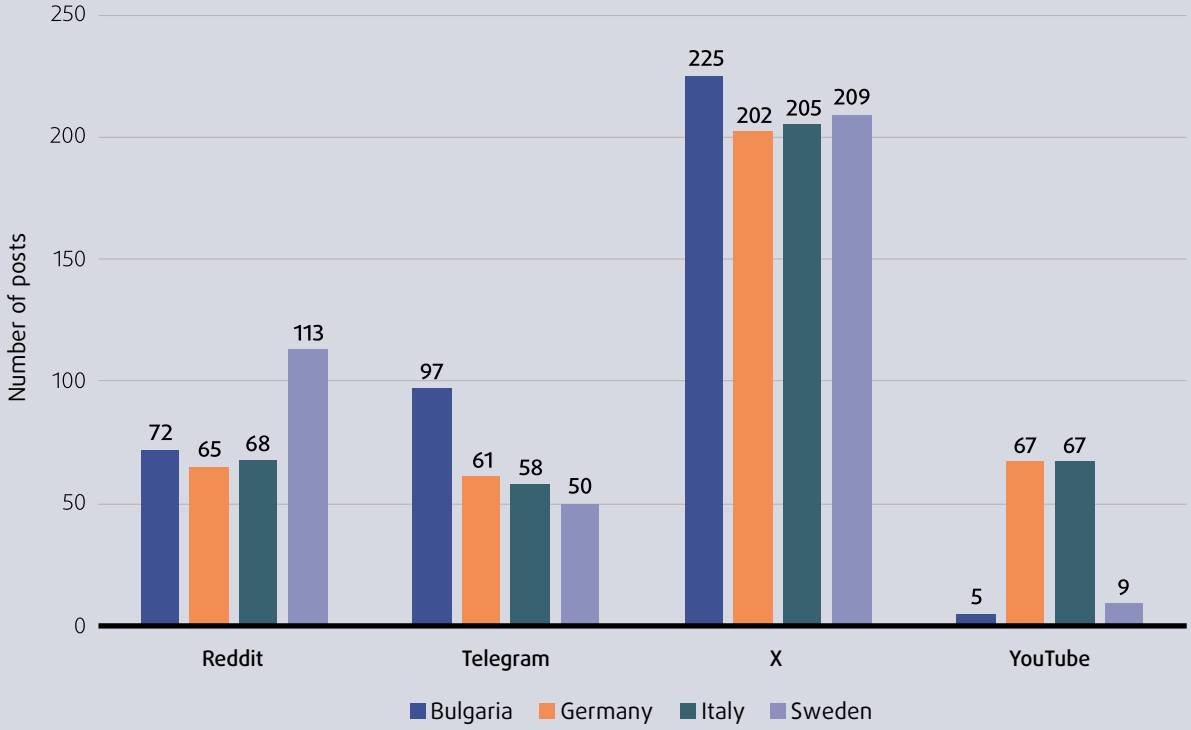
A broadly similar proportion of posts contain language targeted at people of African descent, with over a quarter of all posts targeted at them. Approximately 28 % of all posts analysed (97 097 of the 344 132 posts) contain discriminatory words targeted at people of African descent. The bulk of these posts stem from X.

X accounts for over 91 % of all posts targeted at people of African descent (88 820 of the 97 097 posts). This is broadly in line with patterns observed for posts containing language targeted at women. However, Reddit, rather than Telegram, dominates the remaining 9 % of posts. Reddit accounts for approximately half (4 132) of the remaining 8 277 posts.

For each country, coders handcoded close to 400 posts (**Figure A1**). As X dominated the number of posts in the collected data, the selection of 400 posts for coding was restricted to approximately 50 % to ensure a good representation of other platforms in the coded data.

Purposive sampling of hand-coded data ensures that a sufficient number of posts is examined for each platform. Therefore, the analysis could build an understanding of any differences in the manifestation of hatred between platforms. By nature, however, purposive sampling is not representative of the content on the target platforms. This limits the extent to which wider conclusions can be drawn.

FIGURE A1: OVERVIEW OF CODED POSTS, BY LANGUAGE AND PLATFORM



Source: FRA (2023), online hate dataset.

The detection of online hate through keywords is also prone to error. Thus, the data collection also captured posts that were not relevant to the research. **Figure A2** depicts the share of posts containing key words indicating possible online hate. These posts were referred to as 'relevant' posts.

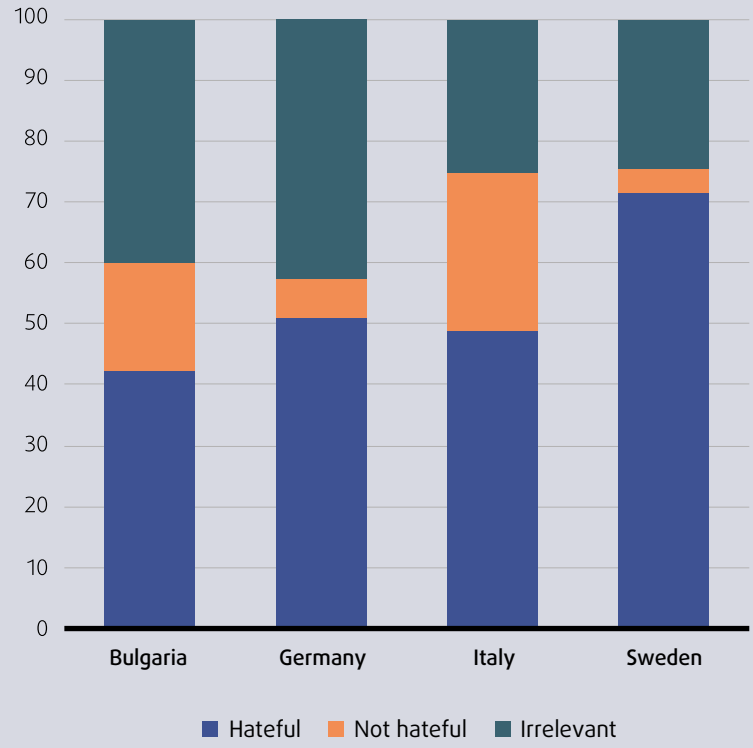
The team established the share of actually hateful posts among those relevant posts. Both Sweden and Italy have similar numbers of relevant posts, at around three quarters of all posts. This means that the keywords were better at capturing relevant posts in those countries.

Among all relevant posts, Sweden has a higher incidence of hateful posts, with over 70 % of all coded posts marked as hateful (272 of 381 posts). By contrast, less than 50 % of coded posts in Italy are categorised as hateful (194 of 398 posts).

Germany and Bulgaria exhibit smaller numbers of relevant posts, both with just under 60 % of all coded posts. Despite this, Germany has a larger proportion of online hate than Italy, with approximately 51 % of all German posts in the sample categorised as hateful (201 of 395 posts). Bulgaria has both a smaller number of relevant posts and a lower level of hateful speech, with only 35 % of all coded posts categorised as hateful (169 of 399 posts).

These results are only very soft indications of potential differences between countries. The selection of different keywords for each country influences the results. So do potential variations in coding methodology. However, the research team made an effort to harmonise both keywords and coding methodology.

FIGURE A2: PERCENTAGE OF CODED POSTS CATEGORISED AS RELEVANT AND HATEFUL, BY COUNTRY



N = 1 573 posts.

Source: FRA (2023), online hate dataset.

Endnotes

(1) See the Brandwatch website.
(2) For more details, see the website for **The Weaponized Word**.
(3) See European Commission (2021), ***Report on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law***, COM(2014) 027 final.
(4) UN (1965), **International Convention on the Elimination of All Forms of Racial Discrimination**, General Assembly Resolution 2106 (XX), 21 December 1965.
(5) UN (1966), **International Covenant on Civil and Political Rights**, General Assembly Resolution 2200A (XXI), 16 December 1966.
(6) Council of Europe (2022), ***Recommendation CM/Rec(2022)16 of the Committee of Ministers to member states on combating hate speech***, Strasbourg.
(7) Council of Europe (2022), ***Recommendation CM/Rec(2022)16 of the Committee of Ministers to member states on combating hate speech***, Strasbourg.
(8) Cambridge Dictionary (n.d.), '**Objectify**'.
(9) Enge, E. (2014), '**Dehumanization as the central prerequisite for slavery**', seminar paper, *GRIN Verlag*, Munich.
(10) Council of Europe, ECRI (2022), ***ECRI General Policy Recommendation No. 15 on combating hate speech***, CRI(2016)15, Strasbourg, explanatory memorandum.
(11) Voci, A. (2014), '**Negative stereotypes**', in Michalos, A. C. (ed.), *Encyclopedia of Quality of Life and Well-being Research*, Springer, Dordrecht.
(12) Black's Law Dictionary (2022), '**Offensive language**'.
(13) Jeswani, J., Bhardwaj, V., Jain, B. and Singh Kohli, B. (2022), 'Negative sentiment analysis: hate speech detection and cyber bullying', *International Journal for Research in Applied Science and Engineering Technology*, Vol. 10/5.
(14) UN, Office of the High Commissioner for Human Rights (2021), '**The Rabat plan of action**'.
(15) UN, Committee on the Elimination of Racial Discrimination (2013), ***General Recommendation No. 35 – Combating racist hate speech***, CERD/C/GC/35.
(16) UN, Committee on the Elimination of Discrimination against Women (2022), ***CEDAW/C/NOR/CO/9: Concluding observations on the ninth periodic report of Norway***, CEDAW/C/NOR/CO/9, para. 21.
(17) UN (1966), **International Covenant on Civil and Political Rights**, General Assembly Resolution 2200A (XXI).
(18) Council of Europe, ECRI (2000), ***ECRI General Policy Recommendation No. 6 on combating the dissemination of racist, xenophobic and antisemitic material via the internet***, CRI(2001)1, Strasbourg.
(19) See 'Grounds for reporting hatred' in European Commission (2021), '**Countering illegal hate speech online: 6th evaluation of the Code of Conduct**', factsheet, 7 October 2021.
(20) Ozalp, S., Williams, M. L., Burnap, P., Liu, H. and Mostafa, M. (2020), '**Antisemitism on Twitter: collective efficacy and the role of community organisations in challenging online hate speech**', *Social Media + Society*, Vol. 6, No 2.
(21) Arcila-Calderón, C., Blanco-Herrero, D., Frías-Vázquez, M. and Seoane, F. (2021), 'Refugees welcome? Online hate speech and sentiments in Twitter in Spain during the reception of the boat *Aquarius*', *Sustainability*, Vol. 13, No 5.
(22) Frenda, S., Ghanem, B., Montes-y-Gómez, M. and Rosso, P. (2019), 'Online hate speech against women: automatic identification of misogyny and sexism on Twitter', *Journal of Intelligent and Fuzzy Systems*, Vol. 36, No 5, pp. 4743–4752.
(23) Allen, C. (2021), *Islamophobia in the media since September 11th*, Forum against Islamophobia and Racism.
(24) Meta, Transparency Center (n.d.), '**Facebook community standards**'.
(25) **Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin** (OJ L 180, 19.7.2000, p. 22), Art. 2(3); **Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation** (OJ L 303, 2.12.2000, p. 16), Art. 2(3).
(26) FRA (2018), ***Hate crime recording and data collection practice across the EU***, Publications Office of the European Union, Luxembourg, p. 14.
(27) Binny, M., Punyajoy, S., Hardik, T., Subham, R., Prajwal, S., Suman, M., Goyal, P. and Animesh, M. (2019), '**Thou shalt not hate: countering online hate speech**', International AAAI Conference on Web and Social Media 2019, Munich, 11–14 June 2019.
(28) UN (2019), **United Nations strategy and plan of action on hate speech**.
(29) Castano-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T. and Lopez, H. M. H. (2021), 'Internet, social media and online hate speech: systematic review', *Aggression and Violent Behavior*, Vol. 58, 101608.
(30) International Network against Cyber Hate (2016), '**Relevance of cyber hate in Europe and current topics that shape online hate speech**'.
(31) Cambridge Dictionary (n.d.), '**Humor**'; Cambridge Dictionary (n.d.), '**Sarcasm**'.
(32) Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law (OJ L 328, 6.12.2008. p. 55).
(33) European Commission, Directorate-General for Communication and Directorate-General for Education, Youth, Sport and Culture (2012), ***Europeans and Their Languages***, Special Eurobarometer 386.

Getting in touch with the EU

In person
All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (**european-union.europa.eu/contact-eu/meet-us_en**).

On the phone or in writing
Europe Direct is a service that answers your questions about the European Union. You can contact this service:
— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696,
— via the following form: **european-union.europa.eu/contact-eu/write-us_en**

Finding information about the EU

Online
Information about the European Union in all the official languages of the EU is available on the Europa website (**europa.eu/european-union/index_en**).

EU publications
You can view or order EU publications at **op.europa.eu/en/publications**. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (**european-union.europa.eu/contact-eu/meet-us_en**).

EU law and related documents
For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EURLex (**http://eur-lex.europa.eu**).

Open data from the EU
The portal **data.europa.eu** provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

FRA EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS

PROMOTING AND PROTECTING YOUR FUNDAMENTAL RIGHTS ACROSS THE EU

Online hate speech is a growing problem in our digitalised societies. Women, Black people, Jews and Roma are often targets of online hate speech. Online hate proliferates where human content moderators miss offensive content. Also, algorithms are prone to errors. They may multiply errors over time and may even end up promoting online hate.

This report presents the challenges in identifying and detecting online hate. It explores the difficulties in researching online hate and the complex task that policymakers and technology platforms face in trying to tackle it.

Hate of any kind should not be tolerated, regardless of whether it is online or offline. The report discusses the implications for fundamental rights to support creating a rights-compliant digital environment.

FRA – EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS
Schwarzenbergplatz 11 – 1040 Vienna – Austria
TEL. +43 158030-0 – FAX +43 158030-699

fra.europa.eu

facebook.com/fundamentalrights
twitter.com/EURightsAgency
linkedin.com/company/eu-fundamental-rights-agency

Publications Office of the European Union