BREAKING DOWN THE SOCIAL MEDIA DIVIDES

A GUIDE FOR INDIVIDUALS AND COMMUNITIES TO ADDRESS HATE ONLINE

A project on
Communication Rights
by the **European Region of the World Association for Christian Communication (WACC Europe)**

WACC Europe
communication for all

This project was made possible with financial support from the Otto per Mille fund (OPM) of the Waldensian Church in Italy.

Research consultant: Francesca Pierigh.

Project Steering Committee: Stephen Brown (WACC Europe President), Ralf Peter Reimann (WACC Europe Vice-President, Evangelical Church in Rhineland), Torsten Moritz, (General Secretary, Churches' Commission for Migrants in Europe), Timo Versemann (Protestant Academy of Berlin), Agnieszka Godfrejów-Tarnogórska (Evangelical-Augsburg Church in Poland), Sara Speicher (WACC Global regional liaison for Europe).
Netzteufel summary translations from German and workshop outline by Jane Stranz.

WACC Europe promotes communication as a basic human right, essential to people's dignity and community. Rooted in Christian faith, WACC works with all those denied the right to communicate because of status, identity, or gender.

contact@wacceurope.org
www.wacceurope.org

Photo Credits:
Page 3: Albin Hillert/WCC
Page 5: images supplied by contributors; Photo of Anna and Karol Wilczynska by Albin Hillert.
Page 6, 16, 27, 30: Adobe Stock,
Page 15: Freepik, Page 32: Shutterstock.

Design by: *Tick Tock Design, www.ticktock-design.co.uk*

CONTENTS

Introduction **3**
The focus of this project 4
Case study contributors 5

Understanding hate speech **6**
What is hate speech? 7
How much do we hate? 7
Why do we hate online? 9
Hate speech and hate crimes 11
Freedom of expression 13
News you can trust 14
Fake news and disinformation campaigns 15

Responding to hateful content online **16**
Legislation and voluntary codes of conduct 17
Education and media literacy 17
Counter-speech and other counter-actions 18
Counter-narrative campaigns 28
Evaluating the impact of counter-strategies 29

Conclusion **30**

Resources **32**

Hope not hate workshop **36**

Additional references **37**

INTRODUCTION

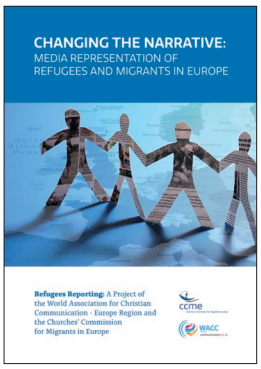
> *"When people are attacked, physically, verbally or on social media because of their race, religion, or ethnicity, all of society is diminished. It is crucial for us all to join hands, stand up, and defend the principles of equality and human dignity.*
>
> *António Guterres, United Nations Secretary General, 2019 International Day for the Elimination of Racial Discrimination*

All people have the right to live in dignity, free from discrimination. This applies everywhere, including in our online interactions. Unfortunately, intolerance and hate speech online are both widespread and dangerous in today's world. Hate speech goes far beyond disagreement and threatens democratic societies because it attacks and silences people.

Encountering hate and division online can be distressing and hurtful. Sometimes we try to engage in conversation; sometimes we avoid an online argument. As social media has become a fixed feature of our lives, we as individuals and as communities need to find ways to break down divides, to build conversation, and to promote diversity and respect online.

The focus of this project

In 2017, our report, *Changing the Narrative: Media Representation of Refugees and Migrants in Europe* assessed how migrants and refugees were represented in the news media in seven European countries. It found that their representation was often characterized by simplification and even invisibility.



This project made a major contribution by analysing mainstream newspapers and media Twitter feeds. However, the absence of a deeper analysis of social media was sorely felt. The current project, *Breaking Down the Social Media Divides*, attempts to bridge that gap. As an extension of Refugees Reporting, this project focuses exclusively on social media. It addresses the proliferation of hate speech and negative narratives on online platforms, and it suggests ways to counter these narratives.

Economic downturns and the increases in arrivals of refugees to Europe, as well as the widespread coverage of the so-called crisis of 2015, put refugees and migrants in the spotlight and made them particularly vulnerable to scapegoating, hateful messaging, and outright hate speech.

According to a December 2018 survey by the European Commission, xenophobia (including anti-migrant hatred) was the most common type of hate speech reported to social media companies. Hatred against people with different sexual orientations, and against Muslims followed closely. These results, which are very similar to the findings of the preceding year, confirm the widespread existence of racist hatred against ethnic minorities, migrants, and refugees.

While the starting point for this project was hate speech against refugees and migrants, in practice, it became clear that this narrow focus was not always achievable or helpful. Therefore, at times, the report adopts a broader lens and discusses hateful content on social media in general.

Fundamentally, this project is about what can be done to address hate speech and create a space for conversation, one in which all people are able to express their voices in a respectful and dignified manner.

This report is particularly aimed at individuals, small organisations, and community groups, such as churches, which may not have extensive expertise in managing social media, but want to start engaging when they see hateful content on the internet.

From case studies, research, and experience, we have put together tips and strategies, as well as useful resources, for those who want to speak up in support of targeted groups, especially, but not limited to, refugees and migrants.

How this report is organised

This report is divided into three sections. The first provides an overview on the topic of hate speech and its potential dangers. The second deals with strategies to counter hate speech: counter-speech, education, and legislation. The third and final part is a collection of useful resources to further your understanding and to help you find the strategies which work best for you.

We have included tips for managing social media throughout the report. Additionally, we have used real life case studies to provide concrete and authentic examples of how people have responded to hateful online content. The case studies also include positive examples of creating a space for dialogue on social media.

Very much like society, the internet is what we make of it. It does not need to be a place of fear, dominated by trolls and haters, where the only escape is to disconnect. If what we want is an open space where information is shared openly and freely, and where everyone feels safe to express their opinions in a respectful manner, it is up to all of us to work towards this goal.

Case study contributors

Meet the people who have shared examples of how they have countered hateful content and engaged in creating a better internet. We use their first names throughout the report.

Annegret Kapp is a communication officer for the World Council of Churches (WCC) based in Geneva, Switzerland and an active participant in the European Christian Internet Conference. Moderating the social media channels of the WCC, she regularly deals with hateful content.

> I think that what we are trying to do in responding to hateful content is for the benefit of the bystanders, the silent readers. To show them that there is a proper way of dealing with hate, and to set the right example.

Dóra Laborczi is a journalist who worked for progressive Christian newspapers in Hungary for many years. In her country, the hate towards those perceived as more progressive members of society is part of a broader societal crackdown. Dóra's experience of being a target of hateful commentaries is, unfortunately, quite wide.

> Almost every time I published something 'hot-topic,' there was a backlash. On the first day I would usually get a nice, respectful response, but on the second and third day, the trolls would come in at full force, especially when writing about migration-related issues. Hate against migrants is very ingrained in the society.

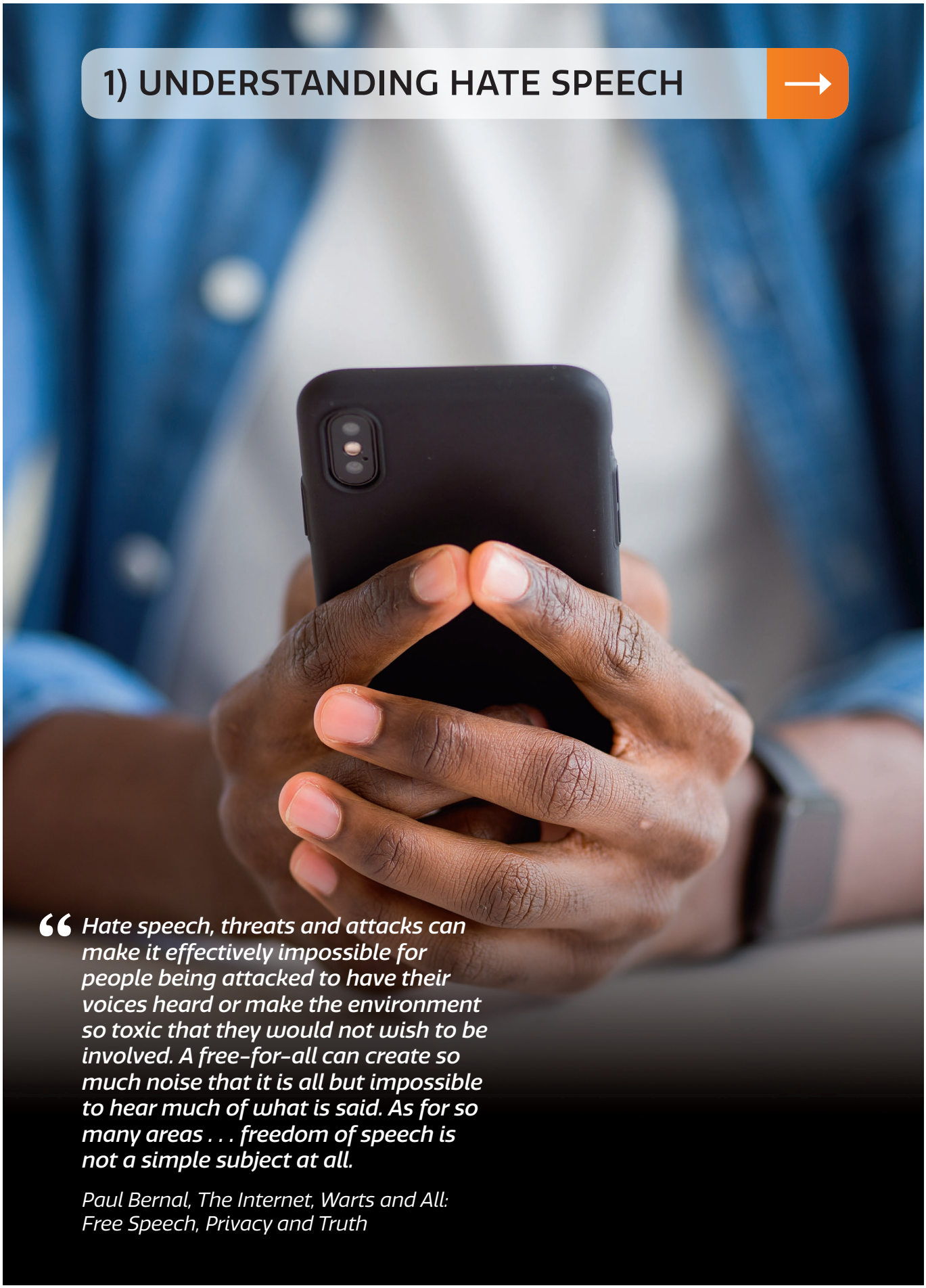
Anna Wilczyńska and **Karol Wilczyński** are journalists from Poland. Karol previously wrote for a Catholic website, Deon (https://www.deon.pl), and he collaborates with Anna on her project, Islamista (https://islamistablog.pl), an independent blog about Islam, migration, and Muslims in Poland. The experience of writing for a Catholic website and a private blog are very different, but in both cases, Karol and Anna have been subjected to hateful comments.

> Everything that is published about the Middle East or Muslims gets a lot of negative attention. Other than that, the topic that is hugely controversial in our country at the moment is the question of LGBTQ rights: homophobia is rampant and gay people are perceived as 'threatening' the very concept of Polish identity and Catholicism.

Timo Versemann is project coordinator for the Protestant Academy of Berlin. In this capacity, he contributed to the development of the *NetzTeufel* project, which included a series of multiplicators' workshops on #HopeSpeech, tackling the topic of hate speech from a Christian perspective.

> When we do the workshops, we are not trying to educate people to do something good, we are trying to empower them to be able to have broader conversations, to engage with others. It is not just about hate speech, it is about how we see society around us.

1) UNDERSTANDING HATE SPEECH



> *Hate speech, threats and attacks can make it effectively impossible for people being attacked to have their voices heard or make the environment so toxic that they would not wish to be involved. A free-for-all can create so much noise that it is all but impossible to hear much of what is said. As for so many areas . . . freedom of speech is not a simple subject at all.*
>
> *Paul Bernal, The Internet, Warts and All: Free Speech, Privacy and Truth*

UNDERSTANDING HATE SPEECH

What is hate speech?

Everyone's perception of what constitutes online hate, what is permitted, and what is too extreme, is different. It does not help that there is no clear-cut definition of hate speech in international law. If we consider not just hate speech alone, but also hateful content in general, the issue becomes even more complicated.

What hate speech and hateful content have in common, though, are the targets: a group, or members of a group, who share a particular characteristic, such as race, gender, or opinion. In the case of migrants and refugees, this may be their national origin, religion, or the fact that they are not citizens of a specific country.

Hate seems to require deeming some individuals as *other*—as essentially different or inferior in some way. It is built upon negative stereotypes derived from this judgment.

> *"Hate speech is "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."*
>
> *Committee of Ministers of the Council of Europe, 1997*

How much do we hate?

To understand how widespread online hate speech is, we have to begin by looking at how frequently it is reported.

However, many people feel that instances of online hate speech are vastly underreported.

And, indeed, there are many reasons why victims or witnesses of hate speech on social media may not report it. They may think reporting is pointless. They may have had previous negative experiences with reporting. They may lack confidence in the justice system. The police themselves may not be well equipped to deal with online hate speech.

The lack of reliable data makes it very difficult to quantify how much hate speech exists online, which in turn makes it hard to address it effectively. However, studies that have attempted to quantify online hate speech, have generally found it to be on the rise.

Annegret: The problem with social media is that people get more extreme just by talking with like-minded people, so countering that is very tedious, slow work.

WHAT IS YOUR EXPERIENCE WITH REPORTING HATEFUL COMMENTS?

Dóra: I have never reported any comments to the social media platforms. Several times, however, a comment disappeared from the platform, so it is possible that it was reported by another reader and was subsequently deleted.

I have also never reported any comment to the police. In one case, I received a very real and personal rape threat from someone who claimed to know me. That was of course very distressing. I replied to this person saying that what they had published could be reported to the police and punished by law. The person deleted their comment and never posted anything after that, so I can only speculate, and hope that they learned a lesson.

Anna and **Karol:** We have mixed experience with reporting hateful content to the social media companies. At times, we reported content which we clearly deemed as violating the platform community standards, but it was not always taken down.

It also happened that the reaction from the platform was very slow, which made us think it could have been quicker to deal with the comment ourselves. However, reporting can be a teachable moment. If the person who wrote a hateful comment receives a notification from Facebook saying that it violates their community standards, we can hope that they learn something.

Annegret: I have reported hateful comments to the platforms, but the results have been mixed at best. I do not remember seeing any effect on Facebook, while on Twitter it has been a bit better. It is quite unpredictable, but sometimes they say that a specific content was violating their rules.

When I am in Germany, I have not yet dared to report anyone because you get a message that looks a bit frightening when you want to report something – it is different from how it works in Switzerland. So I was too afraid that an important working tool will stop working properly to engage in reporting in Germany (For background on the strict law in Germany on hate speech online and the obligations of social media companies, see the case study on the German NetzDG law on page 13).

TIP

WHEN SHOULD I REPORT?

Social media platforms have "community standards" (though they may have slightly different names) and if you feel an account or a comment violates these standards you can report it to the platform.

To report a post or person on Facebook, click on the three dots to the right of their name or on the post itself, and choose "report post".

If you are not sure that a comment is violating the standards, or feel unsure about reporting, you can consider these other options:

- Respond to the post
- Hide the post (this will delete it from your own timeline but will still be available in your friends' newsfeeds)
- Unfollow the person
- Block the person

Why do we hate online?

Understanding why online hate is so common and widespread helps us to counter it. Many studies have explored the differences between interactions in the real world (offline) and on the internet (online). While this is a complex topic, some of the facts below may help to explain why hate, bigotry, and vulgarity are so pervasive online.

OFFLINE	ONLINE
If we say something hateful or vulgar about someone else, we often do it in the face of the person we are targeting. Most likely there are others around. This can increase our sense of responsibility, and consequently, restraint.	On the internet, we can be anonymous. There seem to be no consequences for any message that we send out in the online world. This lowers our constraints, and makes it easier to say things we would not say to someone's face.
When we talk with someone face-to-face, we see their immediate reactions, spoken and unspoken. Nonverbal communication—body language, facial expressions, tone of voice—is a very large part of communication activity.	The whole nonverbal element of communication is entirely lost to us online. We cannot see the other person, which has huge potential for loss of understanding. Because we cannot see the reactions of the other person, it is easier to use language that we would not use in a face-to-face interaction.
In non-digital mass communication, considerable effort and access to specific technology —such as printing presses, cameras, radio, and television broadcasts—is required to write, produce video, and disseminate information to large numbers of people. In the past, this limited hateful messaging by individuals, although it was still used as a tool for propaganda by those with power.	Posting or sharing a hateful comment or content on social media is a quick, impulsive, and generally effortless decision, which people may find extremely satisfying. On the other hand, engaging in counter-speech requires a conscious decision and involves considerable effort. This may explain why there is so much hateful content online and relatively little counter-speech.
The role played by media and politicians is key. The way media covers specific events, and which events it covers, has an impact on the way the audience perceives them. Likewise, the way in which politicians talk—or do not talk—about specific groups or events has a strong impact on how the general public perceives them. Studies show that when certain behaviours are sanctioned by authorities, people will act on their prejudices in the worst ways.	The more we see and hear hateful content, the more we become desensitized to it. Rather than shocking us, hate becomes normalised, a feature of everyday life.

FROM HATE SPEECH TO HOPE SPEECH

The Protestant Academy of Berlin developed workshops on responding to hate speech from a Christian perspective. Timo, the project coordinator, explains how they work.

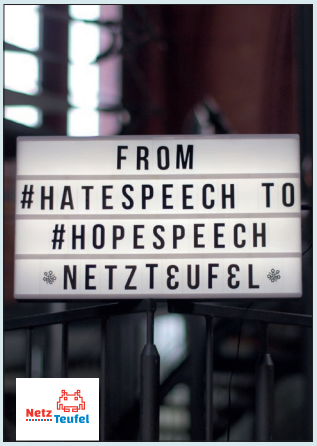
Timo: One of the first actions we took in relation to hate speech was to carry out an analysis of hate speech and toxic narratives on Facebook within a Christian framework. We received excellent feedback, but those who read the report found it insufficiently concrete for people to actually put in practice. So we decided to create workshops as a way for people to reflect on the issue of hate speech in their own personal circumstances. The workshops take place offline, in order to facilitate interaction among participants.

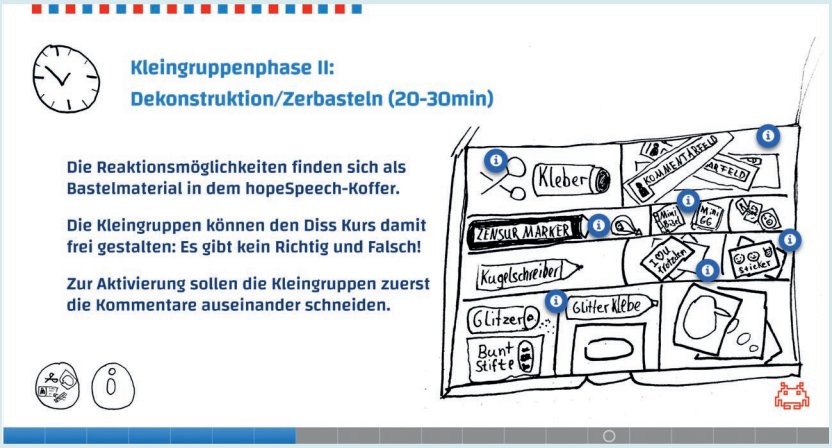
For the workshops, we use a fake social media platform (which we called *Diss Kurs*) and we display real comments and posts from Facebook. We facilitate the ensuing discussion, as participants reflect on the comments and share their own personal experiences with similar content, online or offline. They also share their strategies for dealing with this content. Together we reflect on the contributions brought by everyone and add some suggestions.

We then invite a broader reflection on social media and the way we communicate. We want to move from an attitude of "There is nothing that can be done about it" to finding new ways, new possibilities for replying to hateful content. We also discuss the audience people are trying to communicate with: the authors of hateful content themselves, or the broader audience, the silent bystanders?

The workshops are targeted at multipliers, people who can replicate the experience in their own communities. The programmes are easily scalable, and can accommodate from 8 to more than 30 people. We've had workshops with pastors who can replicate them in their parishes, and also with teenagers, who can do them again in their own communities. This way, so far, we have been able to reach about 1,000 people. An example of a half-day workshop structure can be found on page 36.

For more information on the project, read here:
http://www.wacceurope.org/projects/social-media-divide/hope-not-hate/





The "Netzteufel" (NetDevil) initiative focuses on helping people in civil society and church networks to engage with the web and develop skills, know-how and stories to become positive "web devils". The workshops are based on the understanding that what takes place in the virtual world is real and needs to be taken seriously by both civil society and the churches.

Hate speech and hate crimes

Hate speech is an act of intolerance, which, if not addressed, can provoke hate crimes—acts of conflict and violence.

The latest migration monitoring report released by the Fundamental Rights Agency of the European Union provides an overview of hate crimes against refugees and migrants in the member states. These crimes are often spurred on and accompanied by hate speech.

Hate speech on the internet can and does have effects in real life. A 2018 study analysed the correlation between Facebook usage and violent crimes against refugees in Germany. Examining anti-refugee sentiments expressed on the Facebook page of the extreme right *Alternative für Deutschland* (AfD) party, the study found that crimes against refugees increased disproportionately in areas in Germany with a high use of Facebook and during times of strong online anti-refugee sentiments. Thus, according to the authors, social media was a medium of propagation of anti-refugee sentiments, which could and did lead to actual violent crimes against refugees.

Outside of Europe, in the case of violence against the Rohingya population in Myanmar, similar correlations between increased spreading of hateful messages on Facebook and real-life violence have been documented. The same was true for the 2019 attack against Muslims in Christchurch, New Zealand, and the attack against Jews in Pittsburgh, United States, in 2018.

These and other similar cases make visible the connection between hate speech on social media and hate-motivated crimes. Notably, while not all instances of hate speech lead to hate crimes, all hate crimes involve previous instances of hate speech.

In the worst scenarios, when hate speech and hate crimes are left unchecked, situations can degenerate still further and lead to crimes against humanity and genocide. That was the case in Nazi Germany, in the former Yugoslavia, and in Rwanda: in these places, the target groups were first vilified, denigrated, and then dehumanised in the press, by politicians, and by the broader society.

UNITED NATIONS AND HATE SPEECH

Even though there are difficulties in defining and quantifying hate speech, the consensus about its danger for society is wide. In a Strategy and Plan of Action on Hate Speech released in 2019, United Nations Secretary General António Guterres called it "a menace to democratic values, social stability, and peace." He further stated:

> *"Addressing hate speech does not mean limiting or prohibiting freedom of speech. It means keeping hate speech from escalating into something more dangerous, particularly incitement to discrimination, hostility, and violence.*

The United Nations has repeatedly called for attention to this issue. In September 2019, 26 experts signed an open letter expressing grave concern at the rise of hate in the world against migrants and other minorities, highlighting the connection between hate speech and hate crimes and urging public officials, politicians, and the media to promote tolerant and inclusive societies.

TIP

When writing something on social media, consider whether you would be happy to say it to the person directly or see it printed publicly with your name attached to it.

The Pyramid of Hate is a visual depiction of how hate speech can degenerate into hate crimes.

While all hate speech is potentially dangerous, not all hate speech is equally dangerous. It is important to recognise its diverse forms and degrees, as well as to consider its broader societal context. For example, a hateful comment posted by a user in a small closed group will not have the same impact as a comment shared by a politician with thousands of followers.

Understanding the different degrees of hate speech is critical for identifying the strategy that will work best in a given situation. In some cases, education and awareness raising are key. In others, reporting to social media platforms or even the police may be in order.

The Pyramid of Hate

Pyramid of Hate, Anti-Defamation league, ADL.org. Published with permission.

Freedom of expression

> *Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive, and impart information and ideas through any media and regardless of frontiers.*
>
> *United Nations Universal Declaration of Human Rights, Article 19*

Freedom of expression is a widely recognised right, enshrined in the United Nations Universal Declaration of Human Rights and in further treaties, among which are the International Covenant on Civil and Political Rights and the European Convention on Human Rights. Freedom of expression is a pillar of democracy, vital to prevent censorship and promote accountability, and indispensable for an effective and free media.

However, freedom of expression is often invoked by those spreading hateful content as the right which makes it permissible to engage in hateful discourse. It may seem that freedom of expression implies that hate speech must be tolerated.

However, one person's freedom of expression should not stifle another's. Freedom of speech is a right, but speech that incites hatred or violence against a person or a community is subject to legal sanction.

That said, this becomes more complicated as states attempt to legislate the matter. The fact that most social media companies are headquartered in the United States presents an additional dilemma; the US has a much more lenient approach to free speech than Europe, and it is very difficult to enforce standards prohibiting hate speech across different jurisdictions.

GERMANY AND THE NetzDG LAW

Germany was one of the first countries to take a strong legal stance against hate speech. In 2018, it adopted the Network Enforcement Act, known as NetzDG. Broadly speaking, the law obliges social media companies to remove illegal content in less than 24 hours, or face potential fines of up to 50 million euros. The deadline can be extended for cases requiring further investigation.

The law is not, however, without criticism. Organisations such as Human Rights Watch and Reporters Without Borders have raised concerns. Human Rights Watch argues that the law places an undue burden on social media companies, which are private companies and not fit to be the judges of whether content is illegal. According to the organisation, the law effectively encourages companies to suppress speech, even if it is not necessarily illegal, to avoid large fines.

Human Rights Watch also criticises the lack of appeal provisions in the law, which means that users whose published comment or content has been blocked cannot ask the social media companies to review the decision. The lack of accountability and oversight is a worrying aspect of the NetzDG law.

Perhaps still more concerning is the example that the law sets for other countries with less stable democracies. Russia has already passed a law explicitly referencing the German law. Passage of such a law in a non-democratic society clearly has different consequences. Other countries, including some which notoriously infringe on freedom of expression, such as Venezuela and the Republic of the Philippines, have also hailed the German law as a positive example.

News you can trust

> *Political polarisation has encouraged the growth of partisan agendas online, which together with clickbait and various forms of misinformation is helping to further undermine trust in media—raising new questions about how to deliver balanced and fair reporting in the digital age.*
>
> *Nic Newman, Digital News Report 2019, Reuters Institute*

Traditional media (newspapers, TV, and radio) were for a long time the only sources for news. With the advent of the internet, the landscape completely changed: information from many sources, in multiple formats, became available 24 hours a day to those with an internet connection.

Social media brought further changes. Not only did even more sources of information become readily available, social media offered the opportunity to create information easily. Barriers that previously existed to access and to produce information have almost disappeared. Almost anyone can set up social media accounts, websites, or blogs, and immediately start communicating with an audience.

These changes in how we find and communicate information have contributed greatly to the democratisation of speech and freedom of expression. Producing and disseminating content to the public is no longer only the domain of those with power, access, or expertise. This means, however, that there is also little or no control over the veracity of the information shared on the internet. Hate speech and other forms of dangerous or hateful content can propagate easily across platforms, users, and voices. Platforms such as Facebook, Instagram, and YouTube also use algorithms to reward posts that receive a lot of engagement by placing them at the top of feeds. This works to the advantage of highly entertaining or extreme content.

Conversely, the broadening of the communications landscape has also triggered a diminishing trust in the media. According to a 2019 global study by the Reuters Institute, only around 42% of the people surveyed said they trust the news, including those sources they themselves use. Trust in social media is even lower, at around 23%.

A Pew Research Center study analysed how people get their news in a number of European countries. While the most established news outlets were mentioned, many people also named Google and Facebook as their sources for news. In the countries surveyed, a consistent number of people reported regularly getting their news from social media. Where that was the case, Facebook was mentioned as the most used social media source of information, with Twitter a distant follower. Younger users (18-29 years) are even more likely to use Facebook as a source of information, according to the findings.

When looking for news on social media, a high number of those surveyed also mentioned that they do not pay attention to the source of the news items shared on the platform. This is a worrying sign in light of the disinformation campaigns and false news spreading all over the internet and particularly on social media.

TIP

Support the news you trust.
Financially, if you can.

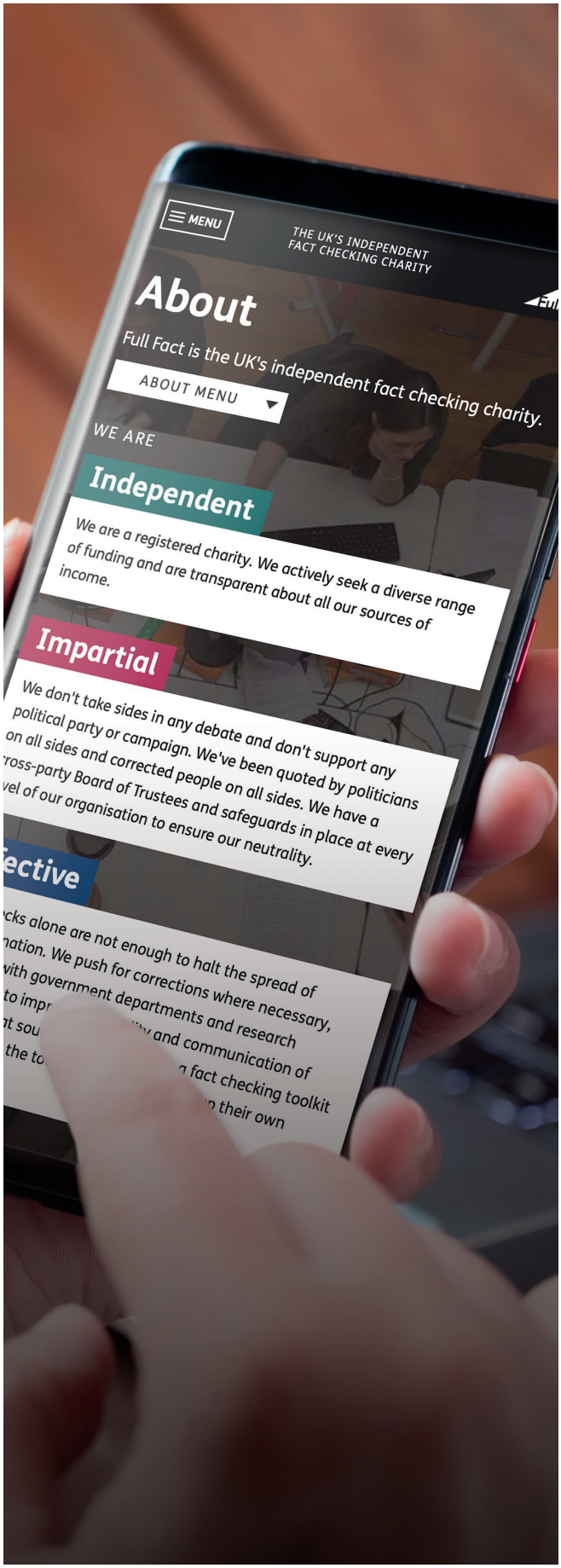
Fake news and disinformation campaigns

Charges of fake news and exposure of strategic disinformation campaigns make it all the more imperative that we critically evaluate the news that we read and see, and share only what we trust, or have verified through a fact-checking service such as Full Fact.

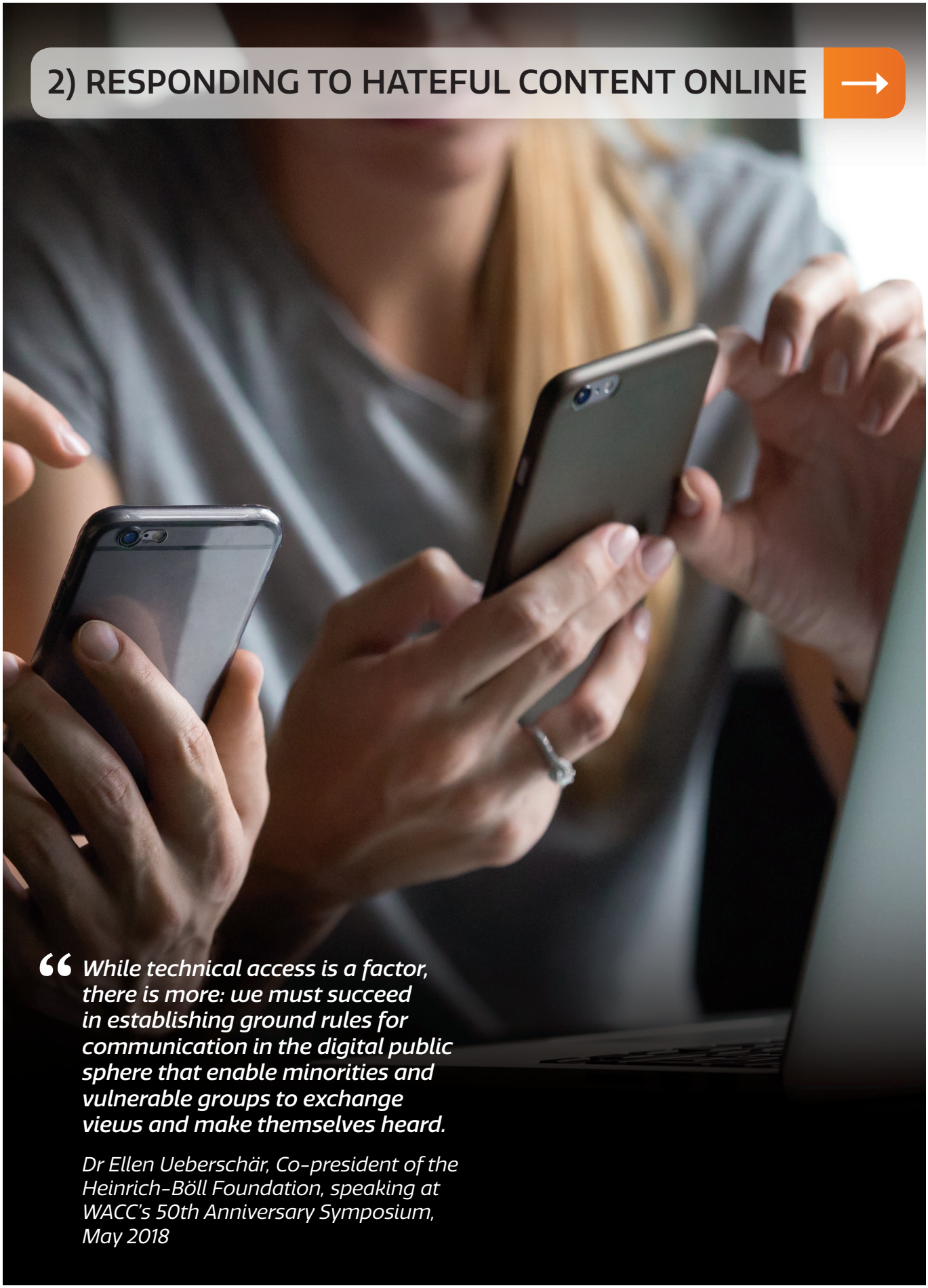
Incorrect information should be corrected. However, it may be even more important, especially in the case of deliberate disinformation, to consider what its purpose is. In many cases, those posting do not intend people to believe the information. Their intention is, rather, to sow mistrust in the news in general or in the authorities, and to direct public behaviour in a particular way. Hence, our responses should promote behaviour and processes that respect facts, people's rights, and democratic societies.

TIP **THINK BEFORE YOU SHARE**

Is this item from a trusted news source?
Will it promote positive action?
Make sure YOU can be a trusted source.



2) RESPONDING TO HATEFUL CONTENT ONLINE



> ***While technical access is a factor, there is more: we must succeed in establishing ground rules for communication in the digital public sphere that enable minorities and vulnerable groups to exchange views and make themselves heard.***
>
> *Dr Ellen Ueberschär, Co-president of the Heinrich-Böll Foundation, speaking at WACC's 50th Anniversary Symposium, May 2018*

RESPONDING TO HATEFUL CONTENT ONLINE

How can we deal with hate online? There are, broadly speaking, three main areas of action:

a) Advocating for appropriate legislation and voluntary codes of conduct

b) Supporting education and media literacy efforts

c) Engaging in counter-speech and other counter-actions

While all three approaches are necessary, the key is to adapt the strategy depending on the different manifestations of hate speech and the changing nature of digital technology. Familiarize yourself with these approaches and adapt them for use in the situation at hand.

Legislation and voluntary codes of conduct

Binding legislation and voluntary codes of conduct are based on the belief that social media companies should be held responsible for content posted and shared on their platforms. Hate speech legislation is, however, a very challenging matter, as demonstrated by the German NetzDG law case study (see page 13).

Deciding what role social media companies should play in moderating content and hate speech is complicated. When they are expected to set their own standards and police their own users, they may lean towards removing flagged content out of pressure to comply with national laws, to avoid fines, or simply to seem worried about the spreading of hateful content on their platforms.

This in turn may lead to inadequate protection of freedom of expression as the 'private censorship' of these platforms may be more restrictive than that imposed by international human rights law. ARTICLE 19, a civil society organisation specialising in freedom of expression, analysed the community standards of Facebook and Twitter, and found that both fall below international standards on freedom of expression, especially in regards to hate speech. Lack of transparency and accountability in the removal process and lack of appeal provisions for users whose content is deleted are also concerns.

In May 2016, the European Commission worked with Facebook, Microsoft, Twitter, and YouTube on a voluntary Code of Conduct on Countering Illegal Hate Speech Online, which was later additionally signed by Google, Instagram and Snapchat. The companies agreed to evaluate notifications by users within 24 hours and to remove content deemed illegal. Progress on the code is monitored through periodic exercises and evaluations. The most recent evaluation, published in February 2019 by the European Commission Directorate-General for Justice and Consumers, states that "IT companies are now assessing 89% of flagged content within 24 hours and 72% of the content deemed to be illegal hate speech is removed, compared to 40% and 28% respectively when the Code was first launched in 2016. However, companies need to improve their feedback to users."

Tackling hate speech with legislation and guidelines, whether binding or voluntary, is one way to attempt to deter the problem. However, the very nature of hate speech on the internet, its volume, its reach, and its transnational nature, are considerable complications. Legislation on hate speech is one avenue that can be pursued. At the same time, other measures aimed at addressing the issue are equally important.

Education and media literacy

Education is key to counter hate speech and hateful online content. As a preventive tool, education and awareness raising are fundamental in increasing our understanding of how hateful content spreads on the internet, and how we can double-check the information we find. The more we are able to do so, the less hate speech will have a free pass.

Media literacy is particularly important in addressing and countering hateful online content. Essentially, it is about developing critical thinking and "critical clicking". It is a conscious use of social media, which allows individuals to identify and question hateful content, to understand the prejudices underneath it, and to develop arguments to confront it.

As the nonprofit European Association for Viewers Interests (EAVI) puts it, media literacy is not about how to technically use media and social media platforms, but about how to "critically evaluate and analyse numerous sources of information simultaneously. This skill requires traditional literacy, reasoning, social injunction, and the ability to decipher symbolic and cultural codes and conventions."

One of the key projects in this regard is carried out by the Council of Europe No Hate Speech Movement, a youth campaign to combat hate speech and promote human rights online. Also, in 2019, the European Commission initiated the EU Media Literacy Week to highlight the importance of media and information literacy as a key factor enabling digital citizens to take informed decisions, online and offline.

Counter-speech and other counter-actions

Counter-speech is a term that includes all activities aimed at responding to hateful content online. If you see a violent or vulgar comment and engage with it, you are doing counter-speech.

Counter-speech can also expand to become a counter-narrative or campaign. These are especially useful if you are working in a group or for an organisation. Both counter-narratives and campaigns are larger-scale activities, and require more planning, time, and resources. The suggestions for individual action also apply to those handling a group's social media accounts. If you are interested in developing a counter-campaign, you will find more information in the resources section of this report.

WHAT DO YOU DO WHEN YOU RECEIVE A HATEFUL COMMENT ON SOCIAL MEDIA?

Dóra: When I receive a hateful comment on social media, I ask myself two questions: Do I have time to deal with this now? And, is there an actual question or comment to reply to?

If it is a pointless attack that is not leading to a conversation, I ignore it. This is also the case when the comment is made from a fake account.

On the other hand, if I can find some sort of contribution underneath the vulgarity, I usually engage. It may be useless for the person who wrote the comment, but it may have an impact on those who read it.

Anna: When I see a vulgar or hateful content or comment, I read it first, with attention. If the comment has a potential for discussion, I reply, even if I do not agree with the views expressed.

On the other hand, if there is no potential at all for discussion, I delete the comment and/or block the user. I do not want to encourage hate, and some people are just looking for a reaction of any kind; this is when I delete and disregard the comment entirely. But it is always very important to be aware of our own biases: I will not delete a comment simply because it expresses views that I am opposed to.

If I see that someone is using a fake account, I block them. I run my website under my real name and my personal social media accounts, and I do not think it is fair to me or to the other participants to have fake accounts in the discussion.

WHAT ADVICE DO YOU HAVE FOR PEOPLE WHO ARE NEW TO COUNTERING HATE SPEECH, BUT WOULD LIKE TO START?

Annegret:

- Stay calm and friendly when correcting inaccurate information.
- Make your message short, and include a link to the correct information.
- If you are not an expert in dealing with hateful content, it is a good idea to take a break and then reread your message one more time before you publish it. This will help to ensure that it is informative rather than heated.
- Try not to get angry; your matter-of-fact, factual comment will be much more effective.

#IAMHERE. CREATING A CIVILISED SPACE FOR DISCUSSION ON SOCIAL MEDIA.

Swedish journalist Mina Dennert wanted to improve the tone of comments on Facebook and to establish a respectful dialogue on the platform. In 2016, she created the Facebook group **#jagärhär, #IamHere**, which currently has almost 75,000 members.

When a group member encounters hateful content in the comments section on Facebook, they respond and call others into the conversation with the hashtag **#IamHere**. The aim is to insert facts and reasonable viewpoints in the conversation, so that other social media users will see a balance of opinions.

#IamHere is not about changing the political views of society. Indeed, it welcomes members from the most diverse political views. It is, rather, about changing the way we debate on the internet. Its intention is to promote respect and civility to stop hate and disinformation, improve the debate, and ensure that more people are able to express their opinions without fear, so long as these views are not violent or threatening.

The **#IamHere** network has spread all over the world and has thousands of volunteers. You can find more information about your country group here: **https://www.jagarhar.se/kolumnen/the-iamhere-network/**

The **#IamHere** network has also created a list of tips for engaging in counter-speech on Facebook to create a civilised dialogue with those we do not agree with. Tips include the following:

- Assume the other person means well, and listen to what they have to say.
- Try to find a common ground, and to understand if words mean the same for both of you—maybe they do not!
- Do not attack, and keep your feelings under control.

Read the full list at: **https://www.jagarhar.se/kolumnen/best-practices-counter-speakers/**

IF YOU ARE TAKING ACTION AS AN INDIVIDUAL:

Your decision to engage in counter-speech activities, and how much time you dedicate to this work, is **entirely up to you**. Countering hateful content can be difficult, so it is important to be mindful of the tips below, to ensure that you remain as safe and healthy as possible:

- **Do not work on counter-speech activities alone.** Make sure that you are in a supportive environment, whether online, offline, or both.
- **Do it for only a limited amount of time every day/week.** Limit the amount of time you engage in countering hate speech and take breaks.
- **Your mental health is the most important parameter.** If you feel that you have had enough, or that you cannot take it anymore, stop. Get up and do something pleasant. Only come back to counter-speech activities when you feel safe and grounded.

Although counter-speech is often advocated as the best way to deal with hate speech, it is important to be aware of the power imbalance between those who post hateful content, and those who engage in counter-speech. Engaging in counter-speech requires much more mental and emotional effort than posting or sharing something hateful, which is why so many people remain silent in the face of hate. Furthermore, harassment of a particular group may prompt other members of that group to remain silent, for fear of facing the same harassment.

This is also why engaging in counter-speech activities—as scary and as unrewarding as that may be—is so important. Especially if you are not from one of the targeted groups, you can remind the authors of hateful messages that those who are being targeted are also human beings. In this way, you can help decrease the feeling of isolation that targeted people frequently suffer from and you can promote a more respectful online debate.

Many times, the hate aimed at one specific group or topic may actually be about something else. Migrants and refugees are often targeted because of their visibility; however, migration is sometimes a smokescreen used to get attention, while people really want to talk about something else. In these cases, it is useful to identify the real issues behind the nastiness and vulgarity and address these issues rather than the emotions expressed through the smokescreen of migration.

> **Remember:** Even if your objective in engaging in counter-speech is not achieved (for example, if the author of the hateful content does not remove it or does not apologise), there is a much broader audience of silent readers who may place great value in your counter-speech. Do not be discouraged!

If you feel ready to engage with hateful online content, the tips on the following pages may be useful for you. As every expression of online hate is different, every possible counter-action is also different. Use the suggestions to find a way to respond that works for you.

> **Remember:** It is very important throughout this process to be mindful of your own biases. Not everyone who disagrees with you is a hater. Be mindful of this, and be open to different viewpoints. You may learn something new!

IF YOU ARE THE TARGET OF HATE:

- **Evaluate** whether it is worth engaging. Is the comment nothing but hateful? Then it could be better to ignore, delete it, or even block the author. Is there an actual message or question in the comment? Then you may want to respond.
- **Remind** the author of the consequences of their words. This may be the harm their speech causes you or others. The fact that this content will be visible for an indefinite time may also affect the writer by negatively impacting their relationships or future employment opportunities.
- **Report** the hateful comment to the social media platform. This may not have the immediately desired effect, but the more hateful content is reported, the more we can measure and understand it. Plus, this experience may educate the author and have positive longer-term effects.
- **If** you know the author of a hateful comment or post, reach out to them privately and let them know that you are uncomfortable with what they wrote and why. Doing this before debating it publicly may give the author a moment to reflect rather than feeling attacked and retreating into their initial hateful position. This will give them a chance to re-think their post and maybe even edit or delete it.
- **Do** not use hateful or vulgar tones in your replies. Replying to hate with hate only generates more hate, and that may be exactly what the author wanted in the first place.
- **Use** normal language, the same as you use when speaking with friends. When we bring normality back into the discourse, we can establish a human connection and may initiate a dialogue.
- **Speak** to the underlying objective of the comment, not to the overt negative narrative.
- **Humour** may defuse the situation.
- **Use** visuals in your replies. An image or short video can sometimes go much further than a written reply.
- **Ask** for help. It is fine if you do not want to read hate anymore. You can ask someone else to go through comments for you, deleting ones that are pointless or just hateful, so you do not have to read them.
- **Even** if you are not able to change someone else's mind, remember that there is a vast audience of passive social media users. These are people who do not engage in the conversation but read the comments. Your response to a hateful comment may not be useful to the one who wrote it, but it could have an impact on others who read your reply.

Anna and **Karol:** We have a good working relationship with the police in Krakow and once, when we received a comment saying that 'migrants bring crime to Poland so they should not be accepted', we replied saying 'What are your concerns about safety and security in Krakow? The police can answer!' and tagged the police into the comment. Of course, this only works if the police actually reply.

IF YOU ARE WITNESSING SOMEONE ELSE BEING TARGETED BY HATE:

- **Be** supportive, both of the person or group targeted, and of counter-speakers. Engaging in counter-speech is difficult, so an alternative to replying directly is to support those who do speak out. "Like" their comment, share their post, or write some words of support. This will not only make the counter-speaker feel supported, it will increase the reach of the post/comment.
- **Report** the hateful comment to the social media platform. This may not have the immediate desired effect, but the more that hateful content is reported, the more we can measure and understand it. Plus, this experience may educate the author and have positive longer-term effects.
- **Remind** the author of the consequences of their words. This may be the harm their speech causes you or others. The fact that this content will be visible for an indefinite time may also affect the writer by negatively impacting their relationships or future employment opportunities.
- **If** you know the author of a hateful comment or post, reach out to them privately and let them know that you are uncomfortable with what they wrote and why. Doing this before debating it publicly may give the author a moment to reflect rather than feeling attacked and retreating into their initial hateful position. This will give them a chance to re-think their post and maybe even edit or delete it.
- **Change** the tone of a hateful conversation to a more empathetic one. For example, find some common ground with the writer—that may have nothing to do with the topic of the hate speech.
- **If** there are threats of violent actions, and they seem credible, inform the police.
- **Do** not use hateful or vulgar tones in your replies. Replying to hate with hate only generates more hate, and that may be exactly what the author wanted in the first place.
- **Speak** to the underlying objective of the comment, not to the overt negative narrative.
- **Humour** may defuse the situation.
- **Use** visuals in your replies. An image or short video can sometimes go much further than a written reply.
- **Join** organised counter-speech activities, such as the Facebook group #IamHere.
- **Even** if you are not able to change someone else's mind, remember that there is a vast audience of passive social media users. These are people who do not engage in the conversation but read the comments. Your response to a hateful comment may not be useful to the one who wrote it, but it could have an impact on others who read your reply.

WHAT IS IMPORTANT FOR YOU WHEN DEALING WITH HATEFUL CONTENT?

Annegret: What is important to me is to remember that people are more likely to spread hate on social media than face to face, and that it would be too much to expect to be able to change their minds. I think that what we are trying to do in responding to hateful content is for the benefit of the bystanders, the silent readers—to show them that there is a proper way of dealing with hate, and to set the right example.

Sometimes just stepping back is helpful. There is really no point in having a long discussion; if the aim is to talk to a third

IF YOU ARE WORKING FOR AN INITIATIVE OR AN ORGANISATION:

- **Have** clear Terms and Conditions or Engagement Guidelines on your social media or website. The guidelines should clearly state what type of language is permitted and what type of content will not be tolerated. This way, you will be able to link back to the guidelines whenever you need to reprimand a user for their conduct or delete their comments altogether.
- **Develop** a strategy with ready-made actions and answers to be used on different platforms.
- **If** you delete a comment or block a user, be sure to post a standard explanation and refer to your engagement guidelines.
- **Take** a screenshot of the post and username before deleting, for your records.
- **In** case of someone posting misinformation, post a link to correct information and state that the previous information is incorrect.
- **Always** be polite and friendly.
- **Be** sure you have a clear understanding of organisational policy and practice in determining what content is allowed, and that you have sufficient human resources to be actively and responsibly engaged.
- **Work** on creating a counter-narrative: What is the positive message that your organisation/initiative wants to promote? Once you have developed your counter-narrative, you can always come back to it when you reply. This way, you are not just replying to a hateful comment, but actively promoting your own message. Promoting your own narrative may also be more effective than simply countering the hateful content with, for example, fact-checking.
- **If** you know a post will spark hate-filled responses, consider settings that will allow you to moderate the comments before they are visible.
- **Where** possible, bring in trusted partners who can support your statement, for instance, by tagging or linking to external sources of information.
- **If** you have the means, consider engaging with a well-respected personality in your context. Often, famous people can become speakers on a topic, and influence a vast public who would not otherwise pay much attention to your organisation or initiative.
- **Even** if you are not able to change someone else's mind, remember that there is a vast audience of passive social media users. These are people who do not engage in the conversation but read the comments. Your response to a hateful comment may not be useful to the one who wrote it, but it could have an impact on others who read your reply.

WHAT IS IMPORTANT FOR YOU WHEN DEALING WITH HATEFUL CONTENT? (cont.)

reader, then it is also important to know that they will not read beyond a couple of messages anyway.

One important aspect to take into consideration is how much time you have. There is always a balancing exercise between how likely it is that other people would see the hateful comment, and would be influenced in a negative way, and how much time it would take you to deal with it. There is a need to maximise the impact while at the same time minimising the resources needed.

IF **YOU** ARE **TIRED** OF SEEING SO MUCH HATE ONLINE AND WOULD LIKE A NICER, **SAFER ONLINE ENVIRONMENT** FOR **ALL:**

- **"Like"** and share positive news and resources that promote inclusive communities, tolerance, respect, and dignity.
- **Ensure** that your own online posts are models of good practice. Be considerate, respect privacy, and share news and information only from trusted and/or verified sources.
- **Be** supportive of people who become targets of hateful content. "Like" their original comment; share their post; write some words of support. This will not only make the person targeted feel supported, it will increase the reach of their post/comment.
- **Develop** some standard replies to potential hateful content so that you can respond more quickly and without as much drain on your energy and emotions.
- **Try** to find common ground, both online and offline, between people or groups of people who are divided. The common ground may have nothing at all do with the topic you are discussing. For example, you may be discussing the NGO rescue boats in the Mediterranean with someone whose views are completely opposed to yours. But what if you both have a keen interest in gardening? Finding the common ground may help establish a basic level of trust, which can be used to move the dialogue forward.

HOW DO YOU HANDLE HATE ON SOCIAL MEDIA?

Annegret: The WCC has developed a strategy with some ready-made answers for certain cases. This has evolved over the years. First, we started dealing with hate on a case-by-case basis, then when the same issue would come up, we would deal with it in the same way, and at some point, this became a document so that other people could use it. The pre-made answers are also a way to be transparent about the way we operate.

So it depends on whether there is a question being asked, or an inaccurate statement about what WCC is doing, or a clear case of a hateful message. In the case of a clear hateful message without any other content, I block the person, delete the comment and write a new comment saying that we deleted the content and explaining why.

This is the message that I would post in this case:

> *"Note: some comments had to be removed from this thread due to the WCC's policy to remove comments containing hate speech or inciting violence. The WCC welcomes comments; however, it reserves the right to delete comments that are vulgar, defamatory, clearly spam (including self-promotion), or in general, not contributing to the ongoing discussion."*

I want people who saw the negative comment to know that they do not have to pay attention to it because it is just hate. The idea is also to educate people on how to behave correctly, and on what constitutes problematic behaviour on social media. I also always take a screenshot of the comment and username when I block someone and keep it in the records.

HOW DO YOU HANDLE HATE ON SOCIAL MEDIA? (cont.)

On Twitter you cannot really delete a comment, so I try to reply with the ready-made answer and/or block the person. Sometimes, however, it feels that by replying to a comment, I am actually giving it more visibility, and a platform. That happens especially on Twitter, where if you reply, you make a comment more visible. So in many cases I do not reply from the WCC main account, but instead use my personal account which has fewer followers.

As a consequence, this has also made me the target of hate. At that point, it's a personal discipline not to look at the hate in the wrong moments, to preserve some rest time. You can mute a person, so you do not see their comments for a certain amount of time.

In the case of someone posting inaccurate information, I provide a link to the correct reading material and state that the previous information is incorrect. (We have a pre-formulated answer for this case also.) In this case, often people do follow up, mostly because they do not believe that they are wrong. The discussion can then progress, though if there is hateful content in the follow-up, I would block and delete.

Sometimes it is not so easy to decide when to delete or to respond to a comment, so we have to discuss it case by case.

If someone makes multiple comments and they tend to be hateful, then it is better to block the person, because it saves time. I have to say that banning is quite an effective way of reducing the hate, of not being a platform for hate. It has happened that people come back under different accounts, but only rarely.

In the beginning I also wrote messages to people directly, but then it became impossible because of the workload. Now I only do that if there is a starting point that I can use and I think there is a possibility to educate the person. I hope that by engaging this way there is a little seed planted in the brain, but I have not checked the response, so I could not say if this works or not.

On Facebook, because we already know that some topics are going to provoke incendiary reactions, the comments are hidden by default according to a long list of keywords. I then read the comments one by one and decide if they can be published. It is very time-consuming work, but we opted for this rather than receiving a lot of hate and having to delete it. We regularly post this comment to make this practice transparent:

> *"Please note: The WCC welcomes discussion; however, it reserves the right to delete comments that are vulgar, defamatory, clearly spam (including self-promotion), or in general not contributing to the on-going discussion. As some topics tend to attract hateful comments, and the WCC doesn't wish to be used as a platform for spreading hate, the settings of this page are such that some comments will be hidden based on keywords until they can be reviewed by a moderator. This thread has attracted a high number of comments, some of them not very clear or unlikely to advance the discussion in a positive way. In the interest of maintaining a focus on positive contributions, we regret not to be able to react to such comments. Thank you for your understanding."*

HOW DO YOU HANDLE HATE ON SOCIAL MEDIA?

Anna and **Karol:** We moderate every single comment on Islamista, which means we are able to see them before they are published. Sometimes we hide an aggressive or vulgar comment and write a personal message to the author letting them know that if they want their post published, they need to edit it to comply with our standards of use.

Sometimes this strategy works, and the author edits the comment and then we publish it; sometimes they delete the comment themselves. Other times, they do not change anything, so we delete the comment.

On the blog we can also track the authors of the comments, so if the comment is particularly hateful, we may even write to the author something like: "Hate speech on the internet is trackable, I can go to the police with your IP address."

Moderating the comments and explaining why certain comments are not tolerated has also been an educational experience for the audience. The group of readers is in a way "educating" itself, is understanding what hate speech is, what type of language to use, and is increasingly able to have a civilised discussion.

Dóra: When I was writing for a collaborative blog, we adopted a supportive strategy. Instead of reading the comments under our own articles, we read the comments under our colleagues' articles. That way, we could filter out all the comments that were simply hate or personal attacks, and leave the ones which were genuine contributions or questions for the author to reply to. We adopted this strategy after realising it was too disturbing for one person to handle it alone. Feeling supported and not isolated is really important when dealing with hateful comments and content.

It is also important to have a public engagement policy to explain what type of content will be tolerated and most importantly, which content will not—such as personal attacks, vulgar language, and hate speech. It is important to have this policy because then you can always refer to it. If you decide to delete or hide a comment, you can link the policy below, so the author of the comment can understand why the comment was removed.

If I know personally the person who wrote a hateful comment, I reply directly to them—just as I would in an offline interaction with an acquaintance or friend.

I also always try to bring the discourse back to reality. Migration issues get a lot of attention, but there are actually very few migrants in Hungary. So I try to ask commenters about other issues where they live: in a country where millions live in poverty, surely there are bigger problems than migrants?

I do not like to ban people from the platforms. I believe that we all live together as a society, off- and online, and I treat the people who post hate online as people who are looking for answers or help. I try to see the people behind the comment, and banning people does not bring anything positive. Our society is already so divided; we should try to build bridges instead.

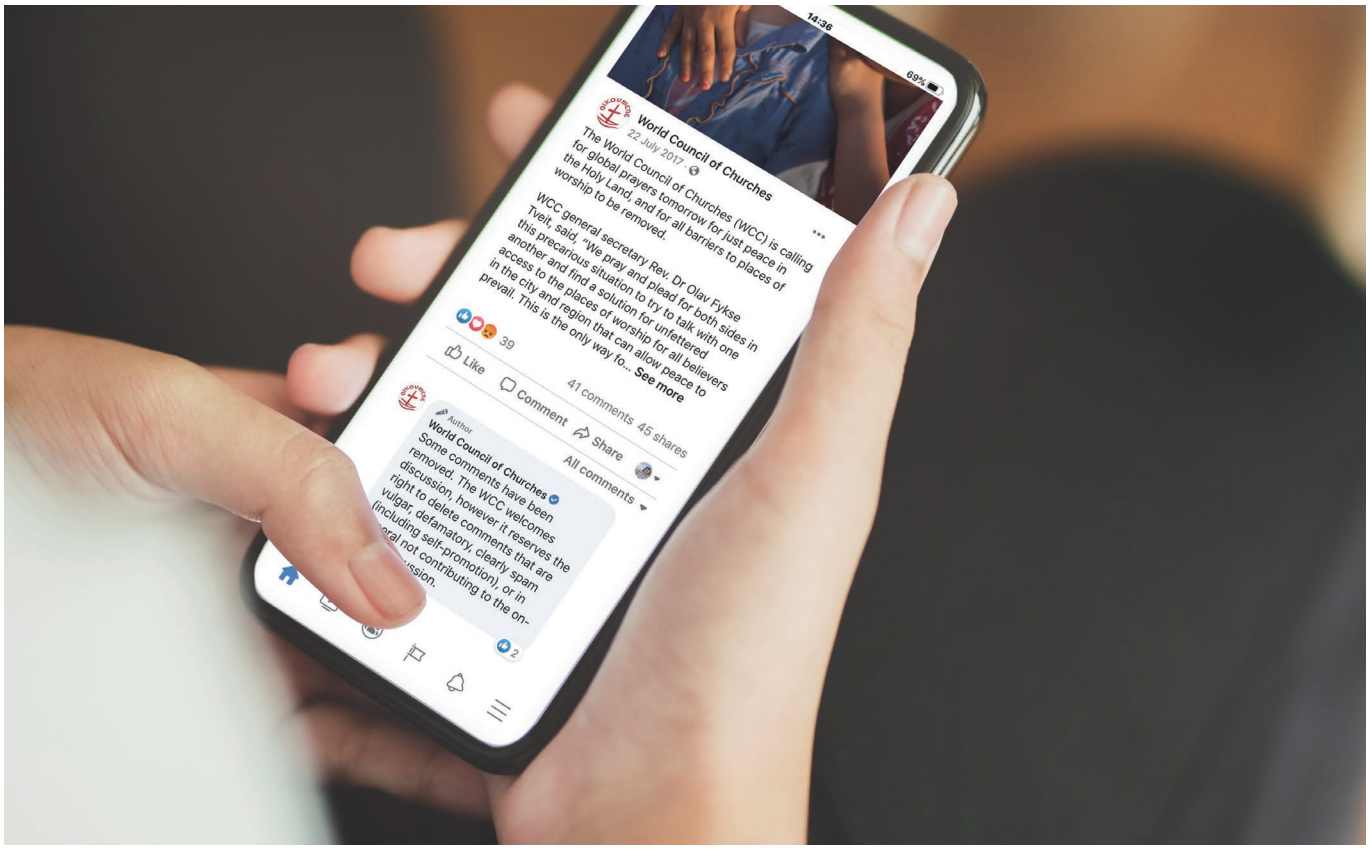
CAN YOU SHARE A CASE IN WHICH YOU WERE ABLE TO TURN THE HATEFUL COMMENTS AROUND?

Dóra: I once shared an article from my personal Facebook account about the fact that my hometown in eastern Hungary was founded by Slovak migrants, and about the value of migrants and migration for our society. The original settlers of the town are very proud of their Slovak origins. I did not write the story, I simply shared it, but that did not seem to matter.

I received many comments, especially from acquaintances from the same town, disbelieving the article and questioning it. People were writing, "We can see on our public service media what migrants are, what they are doing in Western Europe, so how can we possibly be the same? How dare you claim this?"

I replied by sending private messages to the commenters, because I knew them personally. To one of them, I sent the definition of the word *migrant* from Wikipedia. After the commenter read it, they got back to me saying "You are right. I am a migrant too."

Annegret: Sometimes it actually happens that a person says "Hey, I really didn't know that, thanks." In any case, even if that does not happen, when I reply, it is mostly for the benefit of others who read. And I think it is always important to stay polite and friendly and always assume the best of the other person.



Counter–narrative campaigns

For those who are developing strategic counter-narrative campaigns, the Radicalisation Action Network (RAN) Centre for Excellence has developed very useful guidelines for monitoring and evaluation, including tips like the following:

- Make an evaluation plan in advance of the campaign.
- Use realistic indicators.
- Monitor the campaign and adjust as necessary.
- At the end of the campaign, evaluate your success in reaching your goal.

The RAN Centre also created a checklist for planning a counter-narrative campaign, according to its GAMMMA+ Model. The key elements are: Goal, Audience, Message, Messenger, Medium, and Action, plus Monitoring and Evaluation. Below are some of the essential points. (For full list see https://bit.ly/RAN_GAMMMA)

- Effective communication campaigns have goals that are clear, realistic, and measurable.
- The promoted messages are relevant and the target audience considers the messengers credible.
- The campaign works with the target audience's preferred medium or online platforms, and is also present when the audience communicates offline.
- Narrative campaigns in the form of monologues are unlikely to meet the needs of an audience that wants to talk, or is upset or outraged about a real or perceived injustice.
- Campaigns should offer a call to action for those wishing to become involved in the issue at hand, which will facilitate monitoring and evaluation.
- Campaigns aiming to change minds and behaviours offer opportunity for sustained dialogue (both online and offline) with those in their audience who wish to talk.
- Campaigns which ensure they have monitoring and evaluation components in place from the start can then adjust ongoing activities if needed, and once completed, can learn whether they had the desired impact.
- Campaigns that produce a constant stream of content for their target audience to interact with increase their chances of having an impact. Authenticity and quantity are more relevant than technical quality.
- Alternative narratives promote positive alternative perspectives, courses of action, and role models, and foster critical thinking. Counter-narratives, which aim at debunking extremist propaganda, should only be directed at a well-researched and understood audience which is already engaged with extremist content.
- Prepare for success and remember to take into consideration all security risks for your organisation and partners.

More information on monitoring & evaluating counter- and alternative narrative campaigns is available at https://bit.ly/RAN_MEcampaigns

THE DIFFICULTY OF EVALUATION

Timo: Evaluating the impact of the workshops is very challenging. How can you say that a workshop has been successful? Is it about having people replying more frequently to hateful comments? Is it about the quality of the replies? How can this quality be judged?

At the end of the workshops, we always ask participants if they will do anything differently in the future. The response we usually receive is mixed. Some still find it very difficult to impossible to reply to hate online. Others say that they will engage more, not so much for the haters, but for all the passive communicators present online.

I do not see a major problem in being unable to evaluate the workshops' impact. What we are trying to do is not so much to educate people to do something right, as to empower people to have conversations online—to engage with others. It is not just about hate speech, it is about how we see society around us. Why do we ignore certain issues? This is something I would like to be able to focus on more in the future: our positions of privilege and how we relate to others in society.

Evaluating the impact of counter-strategies

Evaluating the impact of strategies to counter hateful online content is a challenging enterprise. As social media constantly changes and evolves, it is difficult not only to keep track of it, but also to evaluate the effectiveness of counter-strategies.

How can you evaluate the effect of your actions on social media? Many times, it may feel like you cannot. However, it is also possible that you will see an immediate result: someone deleting their hateful comment after you have responded to it for example. If you keep engaging, especially if you are working in the context of an organisation, you may feel that the climate of comments is improving in the long term, and that the readers are understanding why certain words and expression are hateful and should not be used. And that may be a success in itself!

THE IMPACT WE DESIRE

Annegret: We engage in counter-speech also because we want to educate people on how to behave correctly online and help them understand what constitutes problematic behaviour on social media.

CONCLUSION

CONCLUSION

The presence of hate on the internet, and its increasing volume and reach, are facts of our everyday life. It may feel like we have little control over this; however, how we choose to deal with it is entirely up to us.

In a world that is increasingly divided, where people retreat into their filter bubbles and refuse to have conversations with those who do not share their views, there is a strong and urgent need to engage. We need to break down the divides we see on social media and in life, and talk with each other. The risks involved in ignoring division and hatred are extremely high. Consequences manifesting across the world include populist leaders taking charge and spreading hateful messages against demonised communities.

Migrants and refugees are one such community, omnipresent in European politics and news, and consequently, easily targeted on social media. The experiences of migrants and refugees as targets of online hate is also transferable to many other marginalised communities.

The fact is that we are all human beings—those who spread hate and those who are the targets of that hate. Reflecting on our commonalities, looking for what unites us, allows us to start having conversations with those with views diametrically opposed to ours.

Our living together depends on our ability to respect each other, and to be able to disagree with each other without using hateful, vulgar, or threatening language. There is always a person on the other side of a hateful comment. We most likely would not purposefully hurt that person in real life. Why should we do that on the internet?

This report has shown that there is not one simple way to hate on the internet. Each case is specific, and can be addressed in many different ways. However you choose to respond, your engagement in the matter is important. The number of haters out there is small in absolute terms, but they are very vocal. Using our voices to support the causes we believe in, and the targeted groups we work with, helps to demonstrate that haters are a minority.

This is how we move away from being silent bystanders. This is how we confront online hate. This is how we bring respect and civility into the dialogue and break down the social media divides.

3) RESOURCES

RESOURCES

The following is a non-exhaustive list of resources that may be useful to those interested in understanding more about hate speech and willing to engage in counter-speech activities. Some of the resources have a specific target audience. Unless specified in the short description, all resources are available in English.

Explaining hate speech

- ARTICLE 19, **Hate Speech Explained: A Toolkit.** The toolkit provides a guide to identifying hate speech and countering it effectively, while protecting the rights to freedom of expression and equality. Available at: https://www.article19.org/resources/hate-speech-explained-a-toolkit/
- Quaker Council for European Affairs, **Anti-Migrant Hate Speech**. The report provides an overview of what hate speech is, how significant it is in Europe, and of initiatives countering hate speech. Available at: http://www.qcea.org/wp-content/uploads/2018/06/Hate-Speech-Report_final.pdf
- Facing Facts Online, **free online courses** available on hate speech, advocacy against hate speech, and monitoring and countering hate speech. Available at: https://www.facingfactsonline.eu/

Guides for counter–speech

- Anti-Defamation League, **Hate Symbols Database**. A database of hateful images, symbols and content commonly used online, to help you understand their symbolism and meaning. Strongly US-focused. Available at: https://www.adl.org/hatesymbolsdatabase
- **Get the Trolls Out** is a project of the Media Diversity Institute and other partners to combat discrimination and intolerance based on religious grounds in Europe. They have a number of helpful resources for different levels of social media expertise, in multiple languages, such as **Fantastic trolls and how to fight them**, and **Stopping hate on Twitter**. All resources available at: https://www.getthetrollsout.org/
- Institute for Strategic Dialogue, **The Counter-Narrative Handbook**. The Handbook provides civil society, youth and NGO-led online initiatives with the tools to develop effective counter-narratives and strategies to push back against hateful and extremist narratives. The Handbook provides insightful advice and suggestions on how to create, launch, and evaluate a counter-narrative campaign. It also includes a list of tools that can be used to create and manage social media content, a list of counter-narrative campaign case studies, and a bibliography with further resources. Available at: https://www.isdglobal.org/wp-content/uploads/2016/06/Counter-narrative-Handbook_1.pdf
- **The Counter-Narrative Toolkit** (funded by Facebook). Online tool which helps you to create a counter-narrative campaign and guides you through all its steps. Requires registration to access the resources and planning tools. Includes a list of counter-narrative campaigns. Available at: http://www.counternarratives.org/
- The Dangerous Speech Project (https://dangerousspeech.org/) has published a series of **Considerations for Successful Counter-speech**, available at: https://dangerousspeech.org/considerations-for-successful-counterspeech/. The short document focuses on strategies that have been successful as a direct response to a hateful post or comment on Twitter, with real life examples, and lists actions which are not helpful at all when countering hate speech. It also has produced **Counterspeech DOs and DON'Ts**. These are downloadable graphic

tips of how to engage in counter-speech, and what behaviours are better avoided. Available at: https://dangerousspeech.org/counterspeech-tips/.

- Radicalisation Action Network (RAN) Centre for Excellence is a European network of policymakers and practitioners exchanging good practice and developing together responses to preventing and reversing radicalisation of individuals and communities. Their working group on **Communication and Narratives** has collected research and resources on extremist narratives and counter strategies, including resources on how to develop, implement and evaluate counter-narrative campaigns. Available at: https://ec.europa.eu/home-affairs/what-we-do/networks/radicalisation_awareness_network
- **Social media safety guides:** user-friendly information on how to use different platforms' reporting and privacy tools, for Facebook, Twitter, Tumblr, Reddit and Youtube. Available at: https://iheartmob.org/resources/safety_guides

Particularly for working with young people

- Media Diversity Institute and others, **Silencing Hate: How to Report Migration and Counter Hate Speech against Migrants and Refugees**. The short report is an overview of the issues, for students. Includes tips for video-making, how to engage in mobile journalism, and how to develop relationships with the media. Available at: https://www.media-diversity.org/resources/silencehate-countering-hate-speech-against-migrants/
- The project SELMA (Social and Emotional Learning for Mutual Awareness) seeks to empower young people (ages 11-16) to tackle the problem of online hate and build mutual awareness, tolerance, and respect. Their **Hacking Hate Toolkit** is a compendium of resources on hate speech and strategies to tackle it. Available at: https://hackinghate.eu/toolkit/
- Council of Europe No Hate Speech Movement, **Bookmarks**, is a manual for combating hate speech through human rights education, specifically created to support the No Hate Speech Movement. The manual presents activities designed for young people aged 13 to 18, adaptable for other age groups. Available in several languages at: https://www.coe.int/en/web/no-hate-campaign/bookmarks-connexions
- Council of Europe, **Compendium of resources on hate speech** that accompanied the No Hate Speech Youth Campaign. Available at: https://www.coe.int/en/web/no-hate-campaign/publications-education

Particularly for journalists

- Ethical Journalism Network, **5-Point Test for Hate Speech**. The resource highlights some questions to be asked in the gathering, preparation, and dissemination of news, to help journalists and editors place what is said and who is saying it in an ethical context. Available at: https://ethicaljournalismnetwork.org/resources/publications/hate-speech
- Media Against Hate, **How to Counter Hate Speech and Manage an Online Community**. For journalists, bloggers, media activists, social media managers, and professionals involved in countering online hate speech. Available at: http://europeanjournalists.org/mediaagainsthate/wp-content/uploads/2018/08/EFJ_module4_def.pdf

- Media Against Hate, **Media against Hate Speech: Training Module**. The module aims to help media regulators and law enforcement authorities to carry out their mission while respecting international freedom of expression standards. Available at: http://europeanjournalists.org/mediaagainsthate/wp-content/uploads/2018/08/EFJ_module3_def.pdf
- Center for Countering Digital Hate, **Don't Feed the Trolls**. Short practical guide for public figures and journalists on how to deal with hateful trolls on social media. The suggestions aim to limit the impact of trolls in the public discourse and to protect the targeted individuals and the broader society. Includes a further bibliography. Available at: https://www.counterhate.co.uk/dont-feed-the-trolls

For further reading

See the articles below as well as the additional references for more resources on each section of this report.

- **"A beginner's guide to fact-checking"**, Orna Young. Available at: https://coinform.eu/a-beginners-guide-to-fact-checking/
- **Conversations with People Who Hate Me**, podcast by Dylan Marron. Dylan interviews individuals who posted hateful comments about him on social media and engages them in conversations to understand their motives. Available at: http://www.dylanmarron.com/podcast
- **"Hate Speech on Social Media: Global Comparisons"**, Zachary Laub, 2019. Available at: https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons
- **"Our experiments taught us why people troll"**, several authors, 2017. Available at: https://theconversation.com/our-experiments-taught-us-why-people-troll-72798
- **"Susan Benesch on Dangerous Speech Project"**, Ethan Zuckerman, 2014. Available at: https://dangerousspeech.org/
- **"The challenge of drawing a line between objectionable material and freedom of expression online"**, Philippa Smith, 2019. Available at: http://theconversation.com/the-challenge-of-drawing-a-line-between-objectionable-material-and-freedom-of-expression-online-108764
- **"The Future of Free Speech, Trolls, Anonymity and Fake News Online"**, Pew Research Center, 2017. Available at: https://www.pewinternet.org/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/
- **"The Future of Truth and Misinformation Online"**, Pew Research Center, 2017. Available at: https://www.pewinternet.org/2017/10/19/the-future-of-truth-and-misinformation-online/
- **"The myth of the free speech crisis"**, Nesrine Malik, 2019. Available at: https://www.theguardian.com/world/2019/sep/03/the-myth-of-the-free-speech-crisis

HOPE NOT HATE WORKSHOPS

The Hope not Hate project (see page 10) developed a number of hands-on communications workshops which were offered in a variety of different settings. The workshops built capacity in social media communications techniques and also offered training for those working in church-related, social, health and youth work, to develop strategic approaches to hope speech in specific situations, to try to change the narrative. Below is an example of a half-day workshop that aims to develop a nine-point plan to deal with online hate speech. In evaluating the workshops, participants highlighted that it was empowering to work on the case study situations, rather than immediately on their own situations.

Introduction

Offer an introduction to online hate speech, what it is, and how it can go viral.

Group work

The workshop is divided into groups, each of which deals with a fictional case study related to hate speech. The online workshop outline offers four examples as starting points (see box). However, those facilitating the workshop are encouraged to get participants to think up their own case studies, either in the plenary session or in each small group. This makes the process more creative and relevant to the participants.

Each small group is asked to put together a simple nine-point strategic plan around three key areas:

- Crisis communication – three points to deal with the immediate issues in a strategic and clear way.
- Follow up – three points that check the initial strategy is working and can evolve as needed.
- Strategic preventative measures – three points that could be developed within the institution to avoid this happening in future.

Discussion and presentation

Following the group work, each group presents their strategic plans in a plenary session, with the opportunity for discussion and critical feedback.

A final information session points to further resources on issues of communication and hate speech, as well as to education material which can be adapted to develop workshops elsewhere. A further module could be developed encouraging participants to work on a strategic plan for their own contexts.

Material is available online, enabling people to train others in their own contexts. An overview can be found at: http://www.wacceurope.org/projects/social-media-divide/hope-not-hate/

CASE STUDY OUTLINES: WHAT DO YOU DO WHEN... ?

1. A handout on finding ways to deal with extreme right-wing views in the work place, produced for staff training in a nursery, is given to the local newspaper which reprints parts of it out of context. Over the next 48 hours the modest Facebook page the organisation normally uses to post job offers is inundated with targeted racist comments.

2. Your church-run care home for elderly people rents much-needed extra space in a nearby building, only to later discover it is owned by a politician with known neo-Nazi views. The politician uses the rental agreement to publicize their credentials, meanwhile your organization is accused on social media platforms and in the local press of cooperating with Nazis.

3. Your new youth centre for work with young people from a migration background begins to attract a growing number of critical comments on its social media platforms. Why are you only doing things for migrants? Why aren't you doing something for homeless people or older people?

4. You run a small family guidance centre in a provincial town. At an open day you present some of the work at the centre, including its work with women and girls in situations of domestic violence. An older man asks some rather strange questions. In the weeks following the open day a growing number of very hostile comments about victims of domestic abuse are left on the organisation's social media platforms, all the comments mention you personally. It would seem to be the person who attended the open day.

ADDITIONAL REFERENCES

1. Understanding hate speech

What is hate speech? / How much do we hate?

European Commission Against Racism and Intolerance. "Hate Speech and Violence." Council of Europe, n.d. Available at: https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence

Why do we hate online?

Bojarska, Katarzyna. "The Dynamics of Hate Speech and Counter-Speech in the Social Media: Summary of Scientific Research." Centre for Internet and Human Rights, Europa-Universität. October 2018. Available at: https://cihr.eu/wp-content/uploads/2018/10/The-dynamics-of-hate-speech-and-counter-speech-in-the-social-media_English-1.pdf

Fiske, Susan T. "Look Twice." *Greater Good Magazine*. UC Berkeley Greater Good Science Center. 1 June 2008. Available at: https://greatergood.berkeley.edu/article/item/look_twice

Friedman, Richard A. "The Neuroscience of Hate Speech." *New York Times*. 31 October 2018. Available at: https://www.nytimes.com/2018/10/31/opinion/caravan-hate-speech-bowers-sayoc.html

United Nations General Assembly. *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. 7 September 2012. Available at: http://daccess-ods.un.org/access.nsf/Get?Open&DS=A/67/357&Lang=E

Hate speech and hate crimes

Davis, Alan. "How Social Media Spurred Myanmar's Latest Violence." Institute for War & Peace Reporting. 12 September 2017. Available at: https://iwpr.net/global-voices/how-social-media-spurred-myanmars-latest

Guterres, António. Foreword to *United Nations Strategy and Plan of Action on Hate Speech*. May 2019. Available at: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

"Hate Speech Exacerbating Societal, Racial Tensions with 'Deadly Consequences Around the World', Say UN Experts." UN News. 23 September 2019. Available at: https://news.un.org/en/story/2019/09/1047102

Ingram, Mathew. "Facebook Now Linked to Violence in the Philippines, Libya, Germany, Myanmar, and India." *Columbia Journalism Review*. 5 September 2018. Available at: https://www.cjr.org/the_media_today/facebook-linked-to-violence.php

McGonagle, Tarlach. "The Council of Europe Against Online Hate Speech: Conundrums and Challenges," n.d. Available at: https://rm.coe.int/16800c170f

Migration: *Key Fundamental Rights Concerns*: 1.1.2019–31.3.2019. European Union Agency for Fundamental Rights. 2019. Available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-migration-bulletin-2_en.pdf

Müller, Karsten, and Carlo Schwarz. "Fanning the Flames of Hate: Social Media and Hate Crime." 3 November 2019. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972

"Secretary-General Launches United Nations Strategy and Plan of Action against Hate Speech, Designating Special Adviser on Genocide Prevention as Focal Point." United Nations Press Release. 18 June 2019. Available at: https://www.un.org/press/en/2019/pi2264.doc.htm

Freedom of expression

Council of Europe. "Freedom of Expression and Information," n.d. Available at: https://www.coe.int/en/web/freedom-expression/freedom-of-expression-and-information-explanatory-memo

European Court of Human Rights. *European Convention on Human Rights*, 1 June 2010. Available at: https://www.echr.coe.int/Documents/Convention_ENG.pdf

Ó'Siochrú, Seán. *Assessing Communication Rights: A Handbook*. Communication Rights in the Information Society (CRIS) Campaign. September 2005. Available at: https://waccglobal.org/wp-content/uploads/2020/07/Assessing-Communication-Rights.pdf

United Nations. *Universal Declaration of Human Rights*. 10 December 1948. Available at: https://www.un.org/en/universal-declaration-human-rights/

United Nations, Office of the High Commissioner. *International Covenant on Civil and Political Rights*, 16 December 1966. Available at: https://www.ohchr.org/EN/ProfessionalInterest/Pages/CCPR.aspx

Germany and the NetzDG Law

Human Rights Watch. "Germany: Flawed Social Media Law." 14 February 2018. Available at: https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

Reporters without Borders. "Russian Bill Is Copy-and-Paste of Germany's Hate Speech Law," 19 July 2017. Available at: https://rsf.org/en/news/russian-bill-copy-and-paste-germanys-hate-speech-law

Reporters without Borders. "The Network Enforcement Act Apparently Leads to Excessive Blocking of Content," 3 August 2018. Available at: https://rsf.org/en/news/network-enforcement-act-apparently-leads-excessive-blocking-content

News you can trust

Mitchell, Amy, Katie Simmons, Katerina Eva Matsa, Laura Silver, Elisa Shearer, Courtney Johnson, Mason Walker, and Kyle Taylor. "Many Western Europeans Get News via Social Media, but in Some Countries, Substantial Minorities Do Not Pay Attention to the Source." Pew Research Center. 14 May 2018. Available at: https:/www.journalism.org/2018/05/14/many-western-europeans-get-news-via-social-media-but-in-some-countries-substantial-minorities-do-not-pay-attention-to-the-source/

Newman, Nic, Richard Fletcher, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. *Digital News Report 2019*. Reuters Institute. 2019. Available at: http://www.digitalnewsreport.org/survey/2019/

2. Responding to hateful content online

a) Legislation and voluntary codes of conduct

ARTICLE 19. "The Compatibility of Facebook's Community Standards with International Standards on Freedom of Expression. 30 July 2018. Available at: https://www.article19.org/resources/facebook-community-standards-analysis-against-international-standards-on-freedom-of-expression/

ARTICLE 19. "The Compatibility of Twitter's Rules, Policies, and Guidelines with International Standards on Freedom of Expression. 5 September 2018. Available at: https://www.article19.org/resources/twitter-rules-analysis-against-international-standards-on-freedom-of-expression/

European Commission. "Code of Conduct on Countering Illegal Hate Speech Online: Questions and Answers on the Fourth Evaluation." 4 February 2019. Available at: https://europa.eu/rapid/press-release_MEMO-19-806_en.htm

Facebook. *Community Standards Enforcement Report*. November 2019. Available at: https://transparency.facebook.com/community-standards-enforcement#hate-speech

Jourová, Vera. "Code of Conduct on Countering Illegal Hate Speech Online: Fourth Evaluation Confirms Self-Regulation Works." European Commission. February 2019. Available at: https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf

Varner, Madeleine, Ariana Tobin, Julia Angwin, and Jeff Larson. "What Does Facebook Consider Hate Speech?" *ProPublica*. 28 December 2017. Available at: https://projects.propublica.org/graphics/facebook-hate

b) Education and media literacy

Council of Europe. "What is the No Hate Speech Movement?" n.d. Available at: https://www.coe.int/en/web/no-hate-campaign

European Association for Viewers Interests (EAVI). "Media Literacy," n.d. Available at: https://eavi.eu/media-literacy/

European Commission. "European Media Literacy Week." Available at: https://ec.europa.eu/digital-single-market/en/news/european-media-literacy-week

Bateman, Jessica. "'#IAmHere': The people trying to make Facebook a nicer place." BBC News. 10 June 2019. Available at: https://www.bbc.com/news/blogs-trending-48462190

WACC Europe
communication for all

BREAKING DOWN THE SOCIAL MEDIA DIVIDES

A guide for individuals and communities to address hate online

www.wacceurope.org

9 782492 052019

Publisher: WACC Europe
10 rue de Versoix
01210 Ferney Voltaire, France
Printed in the United Kingdom
Published: August 2020
Dêpot legal: 3e. trimestre 2020

Otto per mille ITALY
WALDENSIAN CHURCH
UNION OF METHODIST AND WALDENSIAN CHURCHES

Project supported with funds from Otto Per Mille of the Waldensian Church of Italy