Draft Study on effectiveness, risks, and potentials of using counter and alternative narratives in combating hate speech

Co-funded by the European Union

EUROPEAN UNION

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Co-funded and implemented by the Council of Europe

Draft Study on effectiveness, risks, and potentials of using counter and alternative narratives in combating hate speech

Authors: Justice for Prosperity Foundation

This publication was produced with the financial support of the European Union and the Council of Europe. Its contents are the sole responsibility of the author(s). Views expressed herein can in no way be taken to reflect the official opinion of the European Union or the Council of Europe.

This publication has not been copyedited by the SPDP Editorial Unit to correct typographical and grammatical errors. Cover and layout: Document and Publications Production Department (SPDP), Council of Europe

Photo: Shutterstock

Executive Summary

During the past decade, minorities have been facing a rise in hate speech globally. Hate speech causes its victims distress and can lead to severe mental health issues including depression, anxiety, post traumatic disorder, and even suicidal behaviour. Hate speech additionally creates conditions for discrimination, human rights violations, and marginalisation to take place. In order to maintain democratic values and stability, it is crucial to effectively combat this unprecedented rise in hate speech.

This study provides CSOs and other stakeholders with new research data, studies, and information, offering insights into the effectiveness, risks, and potentials of using CANs to combat hate speech. It also provides information on how to meaningfully measure campaign impact and effectiveness. This study is divided into 3 sections. Sections 1 and 2 were written after rigorous desk research and section 3 was written after conducting an analysis of Counter and Alternative Narrative (CAN) communication campaigns and combined quantitative and qualitative elements from those campaigns.

Section 1 addresses 'Hate Speech, who it targets and legal and non-legal measures to address it'. It defines what hate speech is, and rigorously explains its categories, means of spread (both online and offline), aim, risk factors, main target groups, perpetrators, and drivers. Section 1 also dives into the contextual challenges, consequences and impact of hate speech, as well as the different legal and non-legal measures to address it. Furthermore, it displays the legal framework through which hate speech (and hate crimes) are addressed, making distinctions between the UN, EU, and Council of Europe. This section then explains different existing methodologies that can combat hate speech, including CANs. Finally, it provides contextualisation on the state of research of CANs up to now and explains how this study attempts to fill certain research gaps.

Section 2 addresses 'The Impact of CANs to Combat Hate Speech'. It uncovers the intricacies of 'CAN' as a methodology to combat hate speech and looks at the specific risks and potentials that this methodology entails to effectively and impactfully combat hate speech. The findings are based on existing literature gathered during months of desk research exploring communication science (and persuasion theories), public health communication papers, institutional studies, political studies, and security papers - which mainly analyse the potential of CANs to reduce extremism. This section starts by defining narratives, counter-narratives and alternative-narratives. Then, it describes some of the potentials of using CANs to combat hate speech and provides some multi-disciplinary insights on how to change or influence the behaviour of members of society. Following this it presents some of the main risks and aspects to consider when conducting a CAN to combat hate speech. Furthermore, it presents an applied case study, where all the knowledge and theories previously presented are applied in practice. Finally, it issues a practical guide on how to create an effective CAN to combat hate speech.

Section 3 addresses 'Observations from six CSOs who used CANs as a methodology to combat hate speech across five EU countries'. Justice for Prosperity (author of this study) developed two monitoring tools which serve to measure the impact of Civil Society Organisations (CSOs) using CANs as a methodology to combat hate speech. These are a quantitative monitoring tool (with over 130 data inputs) and a qualitative 'End of Impact Study Evaluation Report'. Through these, the six participant CSOs were able to monitor and evaluate the results of their campaigns in a highly effective manner. They were able to extract lessons learnt and identify avenues for improvement for future campaigns. Section 3 of this report starts by presenting the exact methodology and steps involved to observe the CSO campaigns. It then introduces the CSOs and their

campaigns, including their working methods, specific campaign characteristics, and target audiences. Furthermore, it presents the findings from the data, both on a CSO case-by-case basis and in an aggregated manner. It includes insights into some of the main challenges experienced by each CSO when developing their campaign(s), recommendations and conclusions, and the different avenues that each will explore for future campaigns. Additionally, this section presents some interesting insights into platform-specific (audience) engagement. The section then answers critical questions on the effectiveness and impact that CANs can have on combating hate speech, particularly regarding educating instead of countering, campaign formats, campaign language, campaign timing, platform-specific considerations, target audience, risk mitigation, and meaningful evaluations of CAN campaigns (especially through behavioural changes). It also presents some interesting findings about how to best engage with the media and how to react to criticism. Finally, the section ends with some recommendations for online and offline campaigns and gives suggestions for where future research should be directed.

Content

Introduction

Hate creates harm to the individuals targeted and provides space for discrimination, silencing, disempowerment, marginalisation, and human rights violations. Additionally, it leads to increased insecurity, with political divisions and polarisation creating space for domestic terrorism (Piazza 2020: 449). This negatively affects the freedom of speech and pluralistic inclusion in democracy (Jääskeläinen 2020: 347).

Amongst those who suffer the most from hate speech are victims of discrimination, as well as those who witness other members of their own community being victimised. Hate speech poses a threat to the population's overall well-being as its victims often suffer anger, shame and fear, and can result in depression, anxiety, post-traumatic stress, or suicidal behaviour (Cramer et al. 2020). Furthermore, when hate speech occurs due to, or in the context of, structural discrimination, it can also increase socio-economic disparities.

It is therefore important to study the impacts of hate speech, and most importantly how best to prevent and combat it. Addressing hate speech is a form of supporting freedom of expression, democracy and public health and wellbeing (Cramer et al. 2020). Due to its significant impact on society, there is a need for interdisciplinary, multilevel research to better understand hate speech reduction efforts. This is what this Study seeks to accomplish. To do so it explores the 'effectiveness, risks and potentials of using CANs to combat hate speech'. This Study will deliver insights into meaningful ways CSOs and practitioners can determine the impact and effectiveness of their CANs.

This Study has been written in a context which lacks available, comprehensive, significant and comparable data on hate speech incidents and impactful methodologies to combat it. Performing this study has therefore been a complicated task due to this lack of data and any previous research, and the subjective nature of such a study. This Study therefore addresses the above-mentioned research gaps and provides CSOs and other stakeholders with new research data, studies, and information, offering insights into the effectiveness, risks, and potentials of using CANs to combat hate speech. It includes a multi-sectoral/faceted gender and age disaggregated perspective, by explaining how different groups are particularly targeted by hate speech.

Section 1 of this study explores what the world of hate speech entails and how different institutions attempt to address it. The first chapter outlines different existing definitions and categorisations of hate speech. For the purpose of this research, this Study will adopt the following Council of Europe definition of hate speech: "understood as all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as "race[1] colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity, and sexual orientation" (Recommendation CM/Rec (2022)16 on Combating Hate Speech, 2022). The second chapter categorises the different forms of hate speech (i.e. online and offline hate speech), differentiates between hate crime and hate speech, and provides key terms strongly related to hate speech, namely misinformation; disinformation; mal-information and propaganda. The third chapter explores the various propagation channels of hate speech and how it is mainly

[1] Since all human beings belong to the same species, the Committee of Ministers rejects, as does the European Commission against Racism and Intolerance (ECRI), theories based on the existence of different "races". However, in this document, the term "race" is used in order to ensure that those persons who are generally and erroneously perceived as "belonging to another race" are not excluded from the protection provided for by the legislation and the implementation of policies to prevent and combat hate speech. (CM/Rec(2022)16 on combating hate speech)

disseminated online through memes, songs, tweets, pictures, videos, speech by politicians and public figures or mainstream media. Subsequently, the fourth chapter provides a contextualisation of what has happened in the past five to 10 years and the contextual challenges provided by contemporary communication channels. The fifth chapter provides a multi-sectoral approach and demonstrates how different target groups have suffered from a rise in hate speech and hate crime. This chapter also briefly dives into gender-based hate speech, and how female journalists and women in positions of power are particularly exposed to hate speech.

The sixth chapter identifies the aims and main perpetrators of hate speech. The seventh chapter delves into the consequences and impact of hate speech. It explains how hate speech, while directly harmful in itself, can also create conditions that exacerbate harm in the long-term, leading to deeper consequences such as human rights violations, discrimination, acts of violence and/or socio-economic disparities. Chapter eight overviews past and existing legislation at the international and European level with regards to hate speech. This chapter additionally critically explores the legislation, and points out some areas of concern, for example, hate speech being defined very broadly leaves room for states interpretation and manipulation. Furthermore, it also provides an overview of the legislative measures and policies against hate speech targeting specific groups. The ninth chapter critically expands non-legal measures to prevent and combat hate motivated behaviour and hate speech. This includes community-based policing, education and training strategies, and other more recent developments using CANs or exposing foreign influence actors who aim to destabilise democracies. Finally, the last chapter places the study in the context of previous CAN research and further explains how this Study fills the previous literature gap.

Section 2 analyses, through desk research, the effectiveness of CANs to combat (online) hate speech. The first chapter explores the persuasive nature of hate narrative's, looking at their organisation, how the information is transferred, and how this renders its content and meaning more structured and imaginable. It explains how hate narratives are dangerous due to the emotions they cause, reducing reactance and rationality from individuals, which encourages them to form discriminatory and anti-democratic ideas. The chapter defines counter and alternative narratives and explores their complementarity. It shows that to spark change, direct and short-term response are required as well as the development of alternative stories in order to have a long-term impact.

The second chapter explains how and why CANs can be effective in combating hate speech. It begins by looking at behavioural theories and why narratives can be effective in changing people's minds. The theories covered include Transportation Theory, Identification Theory, Parasocial Interaction and Processing Fluency. The chapter goes on to explore persuasive narration techniques such as Perceived Realism and Utopian Narratives. Then, the chapter discusses which emotions are recommended to be used within narratives by analysing the use of humour, regret, empathy and fear. The end of the chapter oversees other variables or methods that need to be taken into consideration when creating CAN campaigns, namely entertainment education, the medium, the length, the characters involved, the level of interactivity, education, and technology and social media companies.

The third chapter is an overview of the risks that come with using CANs to combat hate speech. It summarises Reactance and Cognitive Dissonance Theories which explain why it is complicated to change an individual's mind. CANs can be risky if certain aspects are not taken into account. For example, if the messenger of the campaign lacks credibility, or the target audience is not segmented correctly and the information being shared does not connect with that particular audience, if there is a lack of effective evaluation, if the

resources are limited, or if the campaign loses control of the message in the (social) media. The fourth chapter is a case study on the use of empathy in a CAN campaign. The fifth chapter is a practical guide on how to create effective CAN campaigns to counter hate speech based on the risks and potentials. The sixth chapter is a summary of the main findings.

Section 3 also provides insights for helping practitioners and civil society understand how CANs can effectively combat hate speech online. The section details the observations from six CSOs who used CANs as a methodology to develop campaigns to combat hate speech. These six CSOs are based in Greece, Ireland, Italy, the Netherlands, and Germany. Section 3 details the methodology used for the Study and report and explains all the steps involved. It additionally mentions all the tools that were created (available as additional material) as part of this Study and the training given to CSOs (formal and informal). At the end of the methodology section, there is also a brief subsection on strengths and limitations of the methodology chosen for this Impact Study.

The next chapter introduces the CSOs, their campaigns, and details of the campaign results. It addresses the CSOs working methods to develop their campaigns, and specifies campaign characteristics including their duration, geographic coverage, objective, topics, and target audiences. The second of this chapter presents a shortened version of the (raw) data for each CSO, based on their inputs to the monitoring tools. There is a case-by-case analysis of the data and the impact that the campaigns had. It also presents the main challenges each CSO had when delivering their campaign, and there are CSO-specific recommendations and conclusions for future campaigns. There is also an aggregated analysis of all the CSO's quantitative and qualitative data which provides interesting conclusions on platform specific audience engagement. The last part of the chapter presents some aggregated results on the CSOs general satisfaction levels towards their campaigns, their perceptions of their campaign's effectiveness, and some of the most common recurring challenges they faced when designing their campaigns.

The next chapter delves into further observations from the study, and answers critical questions on the effectiveness and impact that CANs can have on combating hate speech, particularly regarding, educating instead of countering, campaign format, campaign language, campaign timing, platform-specific considerations, target audience, risk mitigation, and meaningful evaluation (especially of behavioural changes). This information and analysis are built on the knowledge acquired by Justice for Prosperity through individual meetings with the CSOs, as well as by CSOs filling in the 'End of Impact Study Evaluation Report'. This chapter additionally presents some further observations on behavioural changes, strategic partnerships, how to react when receiving criticisms of the organisation or campaign, and how best to deal with the media. The final chapter presents some key take-aways and recommendations based on the results from the reports of each campaign and the analysis conducted. It presents recommendations for online and offline campaigns, as well as online-specific campaign recommendations. Furthermore, some future research recommendations based on the study's findings are presented. Finally, the Section ends with some concluding remarks.

Glossary of Terms

Hate Speech:
"understood as all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as "race"[2] colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation" (Council of Europe, Recommendation CM/Rec(2022)16 on Combating Hate Speech, 2022).

Gender-based hate speech:
"any supposition, belief, assertion, gesture or act that is aimed at expressing contempt towards a person, based on her or his sex or gender, or to consider that person as inferior or essentially reduced to her or his sexual dimension" (Combating hate speech, Council of Europe, 2016, p.2).

CAN - Counter and Alternative Narratives:
"narratives that are designed to combat hate speech by discrediting, deconstructing and condemning the narratives on which hate speech is based by reinforcing the values that hate speech threatens, such as human rights and democracy. Counter- and alternative narratives to hate speech also promote openness, respect for difference, freedom, and equality. While counter speech is a short and direct reaction to hateful messages, alternative speech usually does not challenge or directly refer to hate speech but instead changes the frame of the discussion (see Council of Europe manual We CAN! 2017). The use of counter- and alternative speech forms are particularly important for addressing hate speech that does not reach the severity level for being addressed via criminal, civil or administrative procedures (see §§ 3 and 4 of the Recommendation)." (Council of Europe, Explanatory Memorandum of Recommendation CM/Rec(2022)16 on Combating Hate Speech, 2022)

Hate Crime:
In CM/Rec(2024)4 on combating hate crime, "hate crime" is understood as a criminal offence committed with a hate element[3] based on one or more actual or perceived personal characteristics or status, where:
a. "hate" includes bias, prejudice or contempt;
b. "personal characteristics or status" includes, but is not limited to, "race", 2 colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender, sexual orientation, gender identity and expression, and sex characteristics.

Hate manifests itself with different degrees of severity. The occurrence of hate crime can be a direct consequence of the escalation of hate speech. The criminalised forms of hate speech, which are listed in § 11 of Recommendation CM/Rec(2022) 16 on Combating Hate Speech are also covered by the hate crime definition. Measures to prevent and combat hate speech will often contribute to preventing and combating

[2] Since all human beings belong to the same species, the Committee of Ministers rejects, as does the European Commission against Racism and Intolerance (ECRI), theories based on the existence of different "races". However, in this document, the term "race" is used in order to ensure that those persons who are generally and erroneously perceived as "belonging to another race" are not excluded from the protection provided for by the legislation and the implementation of policies to prevent and combat hate speech. (CM/Rec(2022)16 on combating hate speech)

[3] The "hate element" is a general term used to ensure compatibility with a variety of different legal traditions. Moreover, the definition encompasses, but is broader than, "bias motivation," a term which has so far been used by a significant number of organisations and member States as an operational framework for hate crime.

hate crime and vice versa (Council of Europe, Recommendation CM/Rec(2024)4 on combating hate crime §2, and its explanatory memorandum).

Hate Group:
A group "whose goals and activities are primarily or substantially based on a shared antipathy towards people of one or more other different races, religions, ethnicities/ nationalities/national origins, genders, and/or sexual identities. The mere presence of bigoted members in a group or organization is typically not enough to qualify it as a hate group; the group itself must have some hate-based orientation/purpose." (Council of Foundations n.d.)

CSO - Civil Society Organisation:
"Civil society refers to all forms of social action carried out by individuals or groups who are neither connected to nor managed by state authorities. A civil society organisation is an organisational structure whose members serve the general interest through a democratic process, and which plays the role of mediator between public authorities and citizens." (European Commission n.d.)

Council of Europe:
"The Council of Europe is the European continent's leading human rights organisation. It includes 46 Member States, 27 of which are members of the European Union." (Who we are - The Council of Europe in brief, n.d.)

EC - European Commission:
"The European Commission is the executive body of the European Union. Its main roles include proposing new laws and policies, monitoring their implementation, and managing the EU budget. The Commission also ensures that EU policies and laws are correctly applied across Member States, negotiates international agreements on behalf of the EU, and allocates funding. Additionally, it represents the interests of the EU on the global stage, ensuring a coordinated approach among EU countries." (European Commission n.d.)

EP - European Parliament:
"The European Parliament is an important forum for political debate and decision-making at the EU level. The Members of the European Parliament (MEPs) are directly elected by voters in all Member States to represent people's interests with regard to EU law-making and to make sure other EU institutions are working democratically. The Parliament acts as a co-legislator, sharing with the Council the power to adopt and amend legislative proposals and to decide on the EU budget. It also supervises the work of the Commission and other EU bodies and cooperates with national parliaments of EU countries to get their input." (About Parliament n.d.)

EU - European Union:
"The EU is a unique economic and political partnership between 27 European countries that together cover much of the continent. It was created in the aftermath of the Second World War. The first steps were to foster economic cooperation: the idea being that countries who trade with one another become economically interdependent and so more likely to avoid conflict. In 1951, six countries founded the European Coal and Steel Community, and later, in 1957, the European Economic Community and the European Atomic Energy Community. A further 22 countries have since joined the EU, including a historic expansion in 2004 marking the re-unification of Europe after decades of division. As of 1st of February 2020 the United Kingdom is no longer part of the European Union." (European Commission n.d.)

EU MS - EU Member States:
"The 27 Member States which are part of the European Union (EU), and are bound by common EU/Schengen acquis regarding management and control of the EU external borders (Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain and Sweden). There are also 4 Associated Countries or SAC (Switzerland, Iceland, Norway and Liechtenstein)." (European Commission n.d.)

JFP - Justice for Prosperity Foundation:
The NGO authoring this study. "With experience in field operations all the way to the highest strategic political level, we work hard to make a real difference in the fight against subversive power, polarization, and extremism protecting those under siege." Justice for Prosperity investigates, exposes, and predicts targeted undermining of our democracies by the extreme-right & left, ultra-conservative, and populist parties undermining our open societies and democracies. (Justice for Prosperity n.d.)

KPI - Key Performance Indicator:
"Quantifiable measurements used to gauge a company's overall long-term performance. [...] They can also be used to judge progress or achievements against a set of benchmarks or past performance" (Investopedia 2024).

LGBTQI+:
An acronym for persons who identify as lesbian, gay, bisexual, transgender, queer or intersex. It is inclusive of individuals with diverse sexual orientations, gender identities, gender expressions and/or sex characteristics who use other terms or no terms to describe themselves. (European Commission n.d.)

Social media terms and tools used by the CSO in this study:
Reels is an Instagram feature that allows users to create short-form video content and share them on their profile in a section called Reels, but also on Stories, or Feed. Reels are created within the Instagram app, and users can add text, music, filters, and other creative features to help their Reels stand out and become more engaging.
A carousel post is a dynamic form of social media content that allows users to showcase multiple images, videos, or a combination of both in a single post. Unlike traditional single-image or video posts, carousels are interactive, encouraging viewers to swipe or scroll through a series of slides. Carousel posts are popular across many major platforms, including Instagram, TikTok, LinkedIn, Pinterest, and Facebook.
Content sprints are a dynamic approach to content creation and allow you to dedicate a focused and intensive period of time to creating content. During a sprint, the way you produce content is accelerated and the time you do it is compressed to help you meet specific goals or deadlines.
Thumb-Stopping is a social media term that describes exemplary content, typically viewed on a mobile device, that catches the attention of the user and causes them to stop scrolling. This usually refers to content displayed on social media platforms but can also refer to display ads.

SECTION 1: Hate Speech, who it targets and legal and non-legal measures to address it

1.What is Hate Speech?

Despite there being a common understanding of what hate speech is, there is no agreed binding definition of hate speech at the international level. Rather there are multiple different definitions proposed by academics and institutions (UNESC 2022). The UN in its Strategy and Plan of Action on Hate Speech from 2019 defines hate speech as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor." (United Nations 2019).

At the European level the Council of Europe Member States adopted in 2022 the Recommendation CM/Rec(2022)16 on combating hate speech. This Recommendation and its explanatory memorandum provide guidance for member States to implement a comprehensive and calibrated set of legal and non-legal measures. It builds on International human rights standards and relevant case-law of the European Court of Human Rights and pays special attention to the online environment in which most of today's hate speech can be found. The Recommendation contains a broad definition of hate speech (§2 of the Recommendation) and distinguishes within this definition different layers of hate speech (§3 of the Recommendation). This Recommendation was agreed upon by all the Council of Europe Member States.

The Recommendation CM/Rec(2022)16 defines have speech as: "all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as "race",[4] colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation."

International courts do not define the term hate speech, which makes it complicated to determine where, when and how emotions and statements can be considered to be hatred (Jääskeläinen 2020). The European Court of Human Rights (ECtHR) however, has built up a solid body of case law, and established that Articles 8 and 10 of the ECHR deserve equal respect. Member States have, under Article 8, a positive obligation to protect victims of hate speech which reaches a certain threshold of severity, through criminal law or should resort to civil and administrative law provisions in line with the scope of Article 10.2 and the principle of proportionality. The Court also emphasised the need for strong policies to combat racial discrimination as a basis for reducing hate speech and reinforcing democracy's vision of a society in which diversity is not perceived as a threat but as a source of enrichment (see Gündüz v. Turkey, 2003; Erbakan v. Turkey, 2006).

[4] Since all human beings belong to the same species, the Committee of Ministers rejects, as does the European Commission against Racism and Intolerance (ECRI), theories based on the existence of different "races". However, in this document, the term "race" is used in order to ensure that those persons who are generally and erroneously perceived as "belonging to another race" are not excluded from the protection provided for by the legislation and the implementation of policies to prevent and combat hate speech.

Indeed the European Court of Human Rights has established with Member States that the preparation and implementation of policies and legislation to prevent and combat hate speech require a careful balancing of the right to respect for private and family life (Article 8 of the Convention), the right to freedom of expression (Article 10 of the Convention), the right to be free from discrimination in respect of protected Convention rights (Article 14 of the Convention), and that some forms of hate speech may undermine the Convention and fall outside its protection (Article 17 of the Convention).

At the level of the European Union, hate speech is defined in the 2008 Article 1(1)(a) of the EU Framework Decision on combating certain forms of expressions of racism and xenophobia. It states, "publicly inciting to violence or hatred against a group of persons or a member of such a group defined by reference to race, colour, religion, descent, or national or ethnic origin". This definition is also referred to in other EU legal instruments; these instruments include the Audiovisual Media Services Directive, as well as hate speech related policy documents, such as the Code of Conduct on Countering Illegal Hate Speech and the recent proposal for a Council Recommendation on Roma equality, inclusion, and participation (European Commission 2021).

Considering the European scope of this report, this Study adopts the definition provided by the Council of Europe in its Recommendation CM/Rec(2022)16 on combating hate speech. The Recommendation was developed by the Council of Europe's Committee of Experts on Combating Hate Speech. This committee operated under the Steering Committee on Anti-Discrimination, Diversity, and Inclusion (CDADI), and the Steering Committee on Media and Information Society (CDMSI). The drafting of the Recommendation was done by experts from the Council of Europe, Council of Europe Member States, independent specialists, and representatives of other relevant Council of Europe bodies. Their work was informed by existing human rights standards, including the European Convention on Human Rights (ECHR) and jurisprudence from the European Court of Human Rights (ECtHR), which will be discussed further in Chapters 2 and 8 below, as well as consultations with civil society organisations (CSOs) and other stakeholders.

2. Hate Speech, Disinformation and Hate Crime

Council of Europe Recommendation CM/Rec(2022)16 on combating hate speech outlines that hate speech covers a range of hateful expressions which vary in their severity, the harm they cause, and their impact on members of particular groups in different contexts. The Recommendation makes a differentiation between:

a. i. hate speech that is prohibited under criminal law;
   ii. hate speech that does not attain the level of severity required for criminal liability, but is nevertheless subject to civil or administrative law;

b. offensive or harmful types of expression which are not sufficiently severe to be legitimately restricted under the European Convention on Human Rights, but nevertheless call for alternative responses, as set out below, such as: counter-speech and other countermeasures; measures fostering intercultural dialogue and understanding, including via the media and social media; and relevant educational, information-sharing, and awareness-raising activities.

The Court has developed a set of factors that should be applied for assessing the severity of hate speech. They are summarised in §4 of CM/Rec(2022)16: the content of the expression; the political and social context at the time of the expression; the intent of the speaker; the speaker's role and status in society; how the expression is disseminated or amplified; the capacity of the expression to lead to harmful consequences, including the imminence of such consequences; the nature and size of the audience, and the characteristics of the targeted group.

Recommendation CM/Rec(2022)16 on combating hate speech recognises that certain categories of hate speech, which constitute the most serious expressions of hatred, should be criminalised. They are therefore also covered by the hate crime definition of CM/Rec(2024)4. These categories of hate speech are outlined in paragraph 11 of CM/Rec(2022)16, and are as follows:

"Member States should specify and clearly define in their national criminal law which expressions of hate speech are subject to criminal liability, such as:

a. public incitement to commit genocide, crimes against humanity or war crimes;
b. public incitement to hatred, violence or discrimination;
c. racist, xenophobic, sexist and LGBTI-phobic threats;
d. racist, xenophobic, sexist and LGBTI-phobic public insults under conditions such as those set out specifically for online insults in the Additional Protocol to the Convention on Cybercrime concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems (ETS No. 189);
e. public denial, trivialisation and condoning of genocide, crimes against humanity or war crimes; and
f. intentional dissemination of material that contains such expressions of hate speech (listed in a-e above) including ideas based on racial superiority or hatred."

This list is derived from international treaties and conventions that have widely been ratified by the Member States and are outlined in paragraphs 55-63 of the Explanatory Memorandum of Recommendation CM/Rec(2022)16. This includes the EU Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law. Article 1(1)(c) of the Framework states: "publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes as defined in Articles 6, 7 and 8 of the Statute of the International Criminal Court", and 1(1)(d) states "publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes and in

Charter of the International Military Tribunal appended to the London Agreement of 8 August 1945" (European Commission 2021).

The Committee of Ministers of the Council of Europe adopted in May 2024, Recommendation CM/Rec(2024)4 of the Committee of Ministers to member States on combating hate crime. The Recommendation defines "hate crime" as a criminal offence committed with a hate element based on one or more actual or perceived personal characteristics or status, where:
a. "hate" includes bias, prejudice or contempt;
b. "personal characteristics or status" includes, but is not limited to, "race"[5], colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender, sexual orientation, gender identity and expression, and sex characteristics[6].

The definition in the Recommendation encompasses, but is broader than, "bias motivation," a term which has so far been used by a significant number of organisations and Member States as an operational framework for hate crime.

Building on existing case law of the European Court of Human Rights, Recommendation CM/Rec(2024)4 outlines that hate is manifested with different degrees of severity and acknowledges that the occurrence of hate crime can be a direct consequence of the escalation of hate speech. It also acknowledges that hate speech, such as verbal abuse may constitute a hate crime. As such, in combating hate speech, it is equally possible to contribute to preventing and combating hate crime and vice versa.

The following is an example to show the difference between hate crime and hate speech. An action would be categorised as a hate crime if a person physically assaults someone because of their race/religion/sexual orientation/gender or other protected characteristic of status (under national law) while shouting specific slurs directed towards the group they belong to. This qualifies as a hate crime as it involves a crime (assaulting someone) and a hate element, in this case an expression of bias towards gender/race… If the person calls for violence towards a minority without committing (physical) harm (themself) it would be criminalised as hate speech. However, in many European states this would also be classified as a hate crime because incitement to hatred and/or violence is a crime. If on the other hand the person posts a racist 'joke' about a minority group on a social media feed, this would be hate speech as there seems no intent to incite to violence. Depending on the severity of the hatred in the joke, the speaker's role and status in society, and how the expression is disseminated or amplified, there may be sufficient grounds to initiate administrative, civil or even in limited cases, criminal legal measures. In any case non-legal measures should be considered such as counter speech to call out the hatred and to stand with the victim.

[5] Since all human beings belong to the same species, the Committee of Ministers rejects, as does ECRI, theories based on the existence of different "races". However, in this document, the term "race" is used in order to ensure that those persons who are generally and erroneously perceived as "belonging to another race" are not excluded from the protection provided for by legislation and the implementation of policies to prevent and combat hate crime.
[6] In accordance with Article 10.2.c of the Rules of Procedure for the meetings of the Ministers' Deputies, the Republic of Bulgaria reserves the right of its government to comply or not with paragraph 2.b of the Appendix to Recommendation CM/Rec(2024)4 of the Committee of Ministers to member States on combating hate crime. Following Decision No. 13/2018 of the Constitutional Court, the term "gender identity" is incompatible with the legal order of the Republic of Bulgaria.

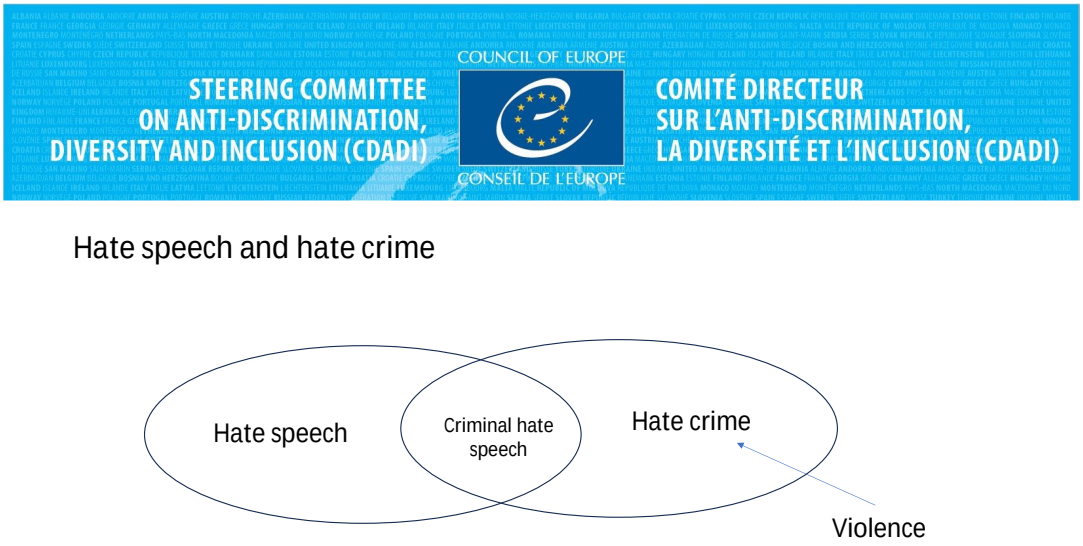


Figure 1: Intersection of hate speech and hate crime. (Source: CDADI)

The Council of Europe's Study on preventing and combating hate speech in times of crisis shows how "disinformation" is frequently intertwined with hate speech, or deliberately disseminated to trigger hate speech and hate crime. Disinformation is "Information that is false and deliberately created to harm a person, social group, organisation or country." (Council of Europe 2017: 20). It can be observed on social media and in the speeches of many politicians, for example stating false and negative information about migrants in order to support their political agenda (e.g. Trump stating that Haitians eat dogs). Another word related to hate speech is "misinformation," this is defined as "Information that is false, but not created with the intention of causing harm" (Ibid: 20). Although the information is not spread with the intent to manipulate or discriminate, it can indirectly do so by influencing people's thoughts based on false information. The third terminology here is "mal-information," this refers to "Information that is based on reality, used to inflict harm on a person, organisation or country" (Ibid: 20). In the definition of UNESCU its underlined that mal-information means "facts deployed out of context or in ways intended to manipulate or mislead" (UNESCO 2023: 17). Another term is "propaganda" which is defined as "false biassed information that is intended to deceive, manipulate or mislead" (Ibid: 17). This form of manipulated or false information can be used by governments to undermine and discriminate against minorities within or outside their countries. Lastly, the term "inciteful speech" refers to speech that "explicitly and deliberately aims to trigger discrimination, violence, terrorism or atrocity crimes" (Ibid: 21). According to binding and non-binding standards, incitement to hatred, violence, and discrimination should be criminalised.

To further add to the layers of complexity, we can also distinguish between offline hate speech and online hate speech. There is no explicit definition of what constitutes online hate speech, but Article 9 (2) of the EU Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law mandates that EU Member States must ensure jurisdiction extends to cases where the conduct is committed through an information system. The "Explanatory Memorandum of the Framework Decision" clarifies that these conducts can refer to public incitement to violence or hatred, and covers actions committed "by any means," including through information systems (European Commission 2021). Forms of online hate speech can vary from misinformation to trolls to images (in internet slang 'trolls' are people who make deliberately offensive posts).

In line with the Budapest Convention on Cybercrime, the Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems, and Court Case law, the Council of Europe Recommendation CM/Rec(2022)16 on combating hate speech, states that to limit the harm done by online hate speech, Member States should enact effective legislation for preventing its publication and for ensuring its removal where it has already been published. Given the amount of online hate speech and given the fact that quick action is needed to avoid its wide proliferation, Member States cannot alone ensure this task but need to ensure that internet intermediaries contribute to it.

The following example illustrates the diversity and range that hate speech can attain online. Three 'trolls' were convicted in a District Court in Finland, of systematic defamation against a journalist. The court rejected their arguments of exercising the right to freedom of speech because the trolls' attacks, made as online false accusations, continued systematically for more than three years. The primary motive for this was to undermine and destroy the journalist's professional credibility and reputation (Higgins 2018). The journalist received death threats, was mocked online as a subject of insulting memes, had her face photoshopped onto pornographic images, and had her address, medical records, and contact details published online (BBC News 2018. Jääskeläinen 2020). This case demonstrates how broad the concept of hate speech is online, as it can present itself through various forms such as "jokes," memes, photographs, harassment, and threats.

3. Spread of Hate Speech

New technologies have not only transformed the mediums through which hate speech is delivered, disseminated, and accessed, but also significantly altered its content and political impact (UNESCO 2023). “Hate speech can be communicated in a wide range of oral, written and visual forms: from the spoken and printed word in statements, speeches, news reports, blogs and texts through still and moving images, video memes and drawings to sounds, songs and more.” (UNESCO 2023: 18). Due to the lack of regulation on platforms and how globally accessible these are, hate speech is disseminated fast and globally. A simple hateful tweet, meme or song has the capacity to be retweeted, watched or listened to thousands of times, victimising and polarising an unprecedented number of people (Pavlovic 2022). As an example, in France the songs of the rapper ‘Freeze Coreleon’ went viral on the internet in 2023 for its antisemitic lyrics: “I arrive determined like Adolf in the 30s (...)”, “I have the propaganda techniques of Goebbels (...)”, “Every day I don't give a fuck about the Shoah”, So that my family can live like Jewish rentiers” (Le Figaro). The song has been deleted from certain streaming platforms like Deezer (French music streaming platform) but not from Spotify (Charts in France 2020).

Figure 2: Antisemitic and anti-immigrant/racist memes on Reddit. (Source: Mic 2016)

Figure 2 is a second example, whereby the images represent antisemitic memes spread on the Reddit platform posted by ‘r/The _ Donald’ which has now been banned for violating Reddit rules (Mic 2016). The image on the left implies that Jewish people were behind the 9/11 attack whilst the image on the right makes fun of Latino Americans (dressed stereotypically) for not being able to cross the border because of Trump’s policies.

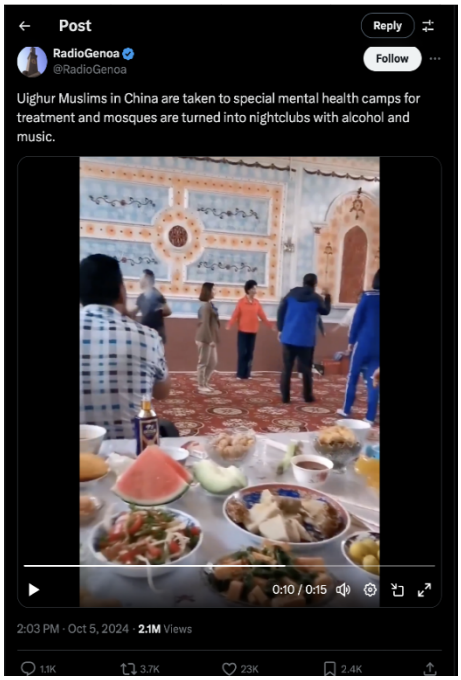


Figure 3: Islamophobic tweets praising the China Uyghur policy. (Source: RadioGenoa 2024)

Figure 3 showcases a 'tweet' which commends China's treatment of the Uyghurs. The Chinese Government's policies toward the predominantly Muslim ethnic minority in Xinjiang province has involved mass arbitrary detentions, forced labour and intrusive surveillance. The government justifies this as counter-terrorism measures. These actions have led to widespread human rights abuses, including cultural suppression and family separations, prompting international condemnation and accusations of crimes against humanity (Human Rights Watch 2024).

X, formerly known as Twitter, is a prime example of how hate speech can have such global reach on a social media platform. Hate speech content is being posted, liked and retweeted every day by accounts who have millions of followers. For instance @RadioGenoa is an account with more than one million followers which posts tweets in which it criticises immigrants with claims that they are violent and dangerous for Europe. On October 5th, 2024, @RadioGenoa posted a video where people are apparently celebrating in a Mosque, with the caption "Uyghurs Muslims in China are taken to special mental health camps for treatment and mosques are turned into nightclubs with alcohol and music".

Figure 4: Replies from Islamophobic Tweet. (Source: RadioGenoa 2024)

Although the captions from Figure 4 do not directly praise China's Uyghurs policy, considering the responses from the accounts, it appears to show that viewers interpreted with a positive perspective @RadioGenoa's post. Furthermore, the comments of the tweet speak for themselves with people praising such a policy.

These examples demonstrate how easily and quickly hate can spread online. The original tweet was posted on October 5th and on October 7th it had reached 2.1 million views, 3.7k retweets, 23k likes and 1.1k comments. Moreover, it is important to consider that this account posts similar tweets multiple times a day daily.

The Council of Europe in its Study on preventing and combating hate speech in times of crisis provides two case examples based on analyses of large data sets of social media posts. The first case looks at the COVID-19 pandemic which produced excessive amounts of information on the health crisis, including false and misleading information. This phenomenon became known as the "infodemic". The infodemic saw a huge increase of hate speech against individuals and groups (such as Chinese and people of Asian descent, migrants and refugees, national minorities), revamped antisemitism (via conspiracy theories), and was intertwined with numerous hateful narratives.

The second case relates to the full-scale war of aggression by the Russian Federation against Ukraine. This has fostered violent, dehumanising rhetoric, disinformation campaigns, and hate speech in individual countries and across Europe against "The West", and hatred against Ukraine, Ukrainian nationals, and refugees from Ukraine. Nationalistic hate speech has been used to trigger and fuel the conflict. Its circulation also represents a challenge to the media sector and to internet intermediaries, who are asked to disentangle hateful narratives and provide the public with objective information about the Russian Federation's aggression.

4. Contextualisation

4.1 What has happened in the past years?

A notable increase in hate speech and hate crime has been observed in the past decade at the highest level of public administration within certain EU Member States, according to a study conducted by the Policy Department of the European Parliament (European Commission 2021). The minorities suffering the most from those acts of hate speech being sexual minorities and migrants, in particular in the context of discriminatory political rhetoric (Combs et al. 2009).

The Council of Europe monitoring bodies, most notably the European Commission against Racism and Intolerance (ECRI), have noted in consecutive annual reports an increase in hate speech year on year. They report that hate speech has a detrimental effect on individuals or groups who are particularly vulnerable (migrants, national minorities including Roma and Travellers, women, LGBTQI+, and persons in certain professions e.g. female journalists and politicians), either because they are subjected to more and more severe abuse or because they face greater obstacles in obtaining justice. Many have insufficient understanding of their rights, and most are reluctant to report hate speech incidents as they lack trust in the justice and law enforcement institutions.

The rise of hate speech nowadays and its increasing presence online can be partially explained by the current political scenario globally, and the rise of the (extreme) right wing in Europe. Chapter 4.2 below explores how minorities from all groups have noted an increase in hate speech in the past 5 years. Prejudicial attitudes, and especially social dominance attitudes that drive hate speech are empirically linked to far-right political figures and movements (Cramer et al. 2020; Van Assche et al. 2019).

Hate speech can be exacerbated by the rise of far-right nationalism and xenophobia in election campaign speeches. Such campaigning legitimises and normalises hateful rhetoric because it is being pronounced by politicians or religious leaders. This also means it is being exponentially disseminated through contemporary communication channels. Piazza (2020) empirically observed that when politicians use hate speech, political violence increases. This happens because politicians' inflammatory language increases and legitimises already existing polarisations. This phenomenon is illustrated in the 1994 Rwandan genocide. Indeed, the genocide was fuelled through systematic and regular anti-Tutsi radio broadcasts that dehumanised the Tutsi, thereby fomenting widespread violence. The hateful targeting of minority groups is beneficial to politicians as it is a strategy to mobilise massive political support by dehumanising and alienating a part of the population (Piazza 2020). Consequently, the Council of Europe considers that public figures have an important responsibility, as their capacity to spread hateful messages is so much greater than the average citizen (Council of Europe 2022b and European Commission 2021).

It should be noted however that it is difficult to get a complete picture of the scale of hate speech in Europe. There is no data collection system on incidents of hate speech at the EU or Council of Europe levels. Such systems only exist at the national level where certain states collect hate speech data through different methods, while other states do not collect such data at all (European Commission 2021). ECRI notes in its reports that a lack of comprehensive and comparable data on the reporting of hate speech impacts effective monitoring of Council of Europe conventions as this data gap impedes having a full understanding of the prevalence of hate speech in both Council of Europe and EU Member States. This has been addressed by one

of the CSOs participating in this Study, which set a national-wide hate speech reporting portal in collaboration with national authorities. For this, the CSO won an 'Erster Trusted Flagger' award. The EU Agency for Fundamental Rights has conducted various research initiatives whereby they have collected data on groups targeted by hate speech, including immigrants and descendants of immigrants, LGBTQI+, and religious minorities. This type of research enables interested parties to understand the scope of hate speech beyond reported hate crimes. The existing data collection measures allow for an understanding of the scope of the problem but may be incomplete. This is because they do not necessarily report on the size and severity of the impact of hate speech on society and the various individuals and groups within it.

4.2 Contextual challenges

As previously mentioned, hate speech has not appeared by itself, but is part of a broader context which includes a decline in social cohesion, a decline in social trust in democratic institutions, a rise of authoritarianism, and a rise in support for political violence. These factors are all enhanced by new challenges faced in Europe, including the so called 'migration crisis', populism, disinformation, and the COVID 19 pandemic. All of these contribute to increased feelings of insecurity and make the future less certain (LIBE Committee 2020). This rise of hate speech can be observed both in the online and offline worlds - which are not separate but rather feed off and into each other.

The parliamentary assembly of the Council of Europe and ECRI have expressed their concern for the increasing scale of online hate speech (See ECRI annual reports, the 5th and 6th cycle country monitoring reports, resolutions of the No Hate Parliamentary Alliance, and Racism, Intolerance, Hate Speech). This rise can be witnessed in all social media platforms, for example Facebook, Instagram, LinkedIn, TikTok, and YouTube. In all these platforms there is a general increase in content flagged as hate speech and an increase in content which is being removed or actioned upon (European Commission 2021). Such expressions of hate can also be found in other youth-oriented spaces online, such as online gaming, meme-sharing sites, and video-based social media platforms (UNESCO 2023). The way the online world is set out creates the ideal environment for hate speech to spread at a global and accelerated rate. AI and algorithm decision making systems (ADM) have the capacity to embed biases in data labelling, data-based decision-making and content recommendations. AI and ADM systems also have the ability to generate hate-oriented information echo chambers (UNESCO 2023). Therefore, a single expression of hate online has amplified consequences as it can reach thousands of people simultaneously. Considering that algorithms reward engagement by increasing exposure, this can only favour the spread of polarised and hateful expressions, as their shock factor creates more views (UNESCO 2023).

The hate spread online does not remain confined to the screens but influences the attitudes and actions of people offline. In fact, the Council of Europe has determined that the growing availability of the internet and social media platforms has led to an increase in instances of sex-based hate speech both online and offline (see Council of Europe 'Factsheet on Combating Sexist Hate Speech' and European Commission, 2021). Online Hate speech, spread through emails, websites and social media platforms, permeates various forms of social interaction, including educational settings, family environments, social circles, public spaces, and workplaces (Council of Europe 2016c). This phenomenon applied to all xenophobic, racist, homophobic, and transphobic hate speech and can particularly arise during election campaigns (ECRI 2019). In Spain, during the 2019 electoral campaign, misinformation about migrants was exploited, WhatsApp reported that 25% of the disinformation shared included racist and hateful material. (Szakács and Bognár 2021)

An example of how the online world influences the offline world is the rise in conspiracy theories online which often scapegoat Jews, Muslims, and other minorities. When the COVID-19 pandemic started people of Asian descent (especially those perceived as Chinese), and other communities were blamed for the spread of the virus. This resulted in hate speech and hate crimes being committed offline (European Commission 2021 and Cramer et al. 2020 and Council of Europe 2023d).

Figure 5: Tweet about Asian hate speech affecting real life. (Source: EuroNews 2020)

Figure 5 presents a tweet from a French-Vietnamese woman who explains that “people are insulted and kicked off public transport because they are Asian. It is not just jokes/hatred on social media. Discrimination also happens in real life.” Such examples were also witnessed in Italy, where the Chinese community experienced a rise in hate speech attacks in 2020. In Florence where there is a large Asian community, 46 hate speech acts were reported at that time (CDADI 2023).

Another example of Asian hate speech and crime which is not related to COVID-19 is the case of a 49-year-old Chinese migrant who was beaten to death in 2016 by young people shouting racist slurs in a Paris suburb. They believed he had large amounts of cash with him, which is a recurring cliché about the Chinese community, but found he only had “sweets and cigarettes.” (EuroNews 2020). The above examples show the influence that online conspiracy theories have on offline hate speech and hate crimes.

Although it is clear that offline and online hate speech influence each other, it remains complicated to identify where hate speech is rooted. There is so much online hateful content, as well as many politicians who use hate speech-based narratives, that it is complicated to trace back to when certain narratives started spreading, what context sparked them, or who was the first to mention it. Politicians' speeches and political campaigns can cause a spike in a certain topic on social media, but the opposite also happens with online fringe narratives increasingly spread online until they influence mainstream political narratives. An example of how offline and online hate feed off each other is the development of the ‘Great Replacement’ conspiracy theory. Originally, it was a fringe and hateful ideology that claims that white Europeans are soon going to be replaced by non-white ethnic minorities due to immigration. In the past decade this conspiracy theory has spread online to the point that it has been normalised in many European countries and has made its way into official political speech. The Netherlands saw the usage of the “Omvolkingstheorie” (Great Replacement Theory) by the new government of Prime Minister Dick Schoff, where two members of the cabinet have referenced the Theory. This referencing by politicians resulted in a peak of online mentions of the Theory (WhoDis 2024).

5. Targets of Hate Speech

In Europe, there are several groups who suffer the most from hate crimes and hate speech. Women are targets of misogyny; people of colour, migrants, national, ethnic, religious and/or linguistic minorities, are victims of racism and/or xenophobia; and members of the LGBTQI+ community and/or gender non-conforming persons, are targets of homophobia and/or transphobia (United Nations n.d. and ECRI annual reports). It should be noted that individuals with intersecting identities (for example an LGBTQI+ person from a minority group) tend to face more frequent and concerted attacks.

5.1 Women

Women, in particular young women and girls, are often the targets of gender based and misogynist hate online which can escalate into hate crime offline (European Parliament and Council 2024). Recommendation CM/Rec(2019)1 of the Committee of Ministers to Member States on 'preventing and combating sexism', further elaborates on the online sexist hate speech as a continuum of sexism. "Acts of "everyday" sexism in the form of apparently inconsequential or minor sexist behaviour, comments and jokes are at one end of the continuum. However, these acts are often humiliating and contribute to a social climate where women are demeaned, their self-regard lowered and their activities and choices restricted, including at work, in the private, public or online sphere. Sexist behaviour such as, in particular, sexist hate speech, may escalate to or incite overtly offensive and threatening acts, including sexual abuse or violence, rape or potentially lethal action." (Council of Europe 2019: 4)

In Europe, 63% of girls and young women have reported that they have personally experienced some form of online harassment on social media platforms. This tendency seems to be worsening, as a 2021 study by the EU Commission observed that the increase in the scale of hate speech disproportionately targets women and girls, in particular women that are public figures. Some groups of women, including: human rights defenders, women in politics, journalists, bloggers, young women, women belonging to ethnic minorities and indigenous women, lesbian, bisexual and transgender women, women with disabilities, and women from marginalised groups and migrant women, are particularly targeted by technology-facilitated violence (UNGA 2018 and Council of Europe 2016c).

Indeed, women politicians find that once they become public figures, the hate they receive is multiplied. The same goes for journalists and women rights' defenders, who suffer more attacks than male human rights activists. There are cases where women have been killed because of their work (Council of Europe 2016c and UNGA 2018). According to a UNESCO Study in 2020 on Online Violence Against Women Journalists, 73% (of the 625 women surveyed), experienced online violence in the course of their work. However, it is complicated to assess the true extent of hate speech as a large part of it does not enter into the statistics with many targeted women not reporting it (Council of Europe 2016c). The Council of Europe Gender Equality Strategy 2024-2029 in paragraph 43 elaborates that "In the same way as with other forms of violence against women and girls, sexist hate speech remains underreported, but its impact, especially on girls and young women, be it emotional, psychological and/or physical, is devastating". Unfortunately, this is not surprising, as according to the UNESCO 2020 study, only 25% of the respondents reported incidents of online violence to their employers, with the most responses being "no response (10%) and advice like "grow a thicker skin" or "toughen up" (9%), while 2% said they were asked what they did to provoke the attack" (UNESCO 2020b: 3).

Figure 6 represents the most common types of online threats that women journalists experience according to UNESCO, figure 7 represents the most significant impacts of such online threats, and figure 8 describes how women journalists responded to such online attacks (UNESCO 2020b).

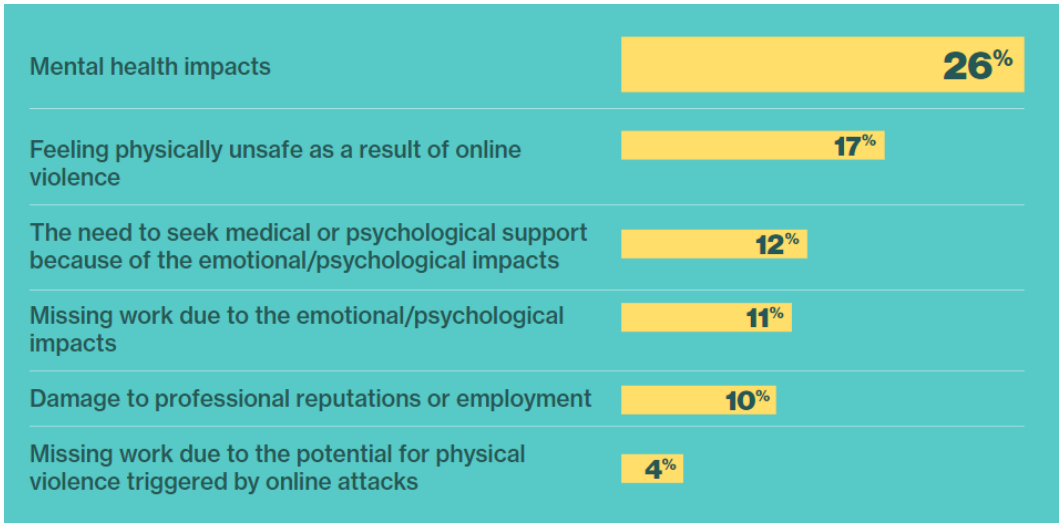


Figure 6: Most significant impacts of online women violence experienced by women journalists. (Source: UNESCO 2020b: 9)

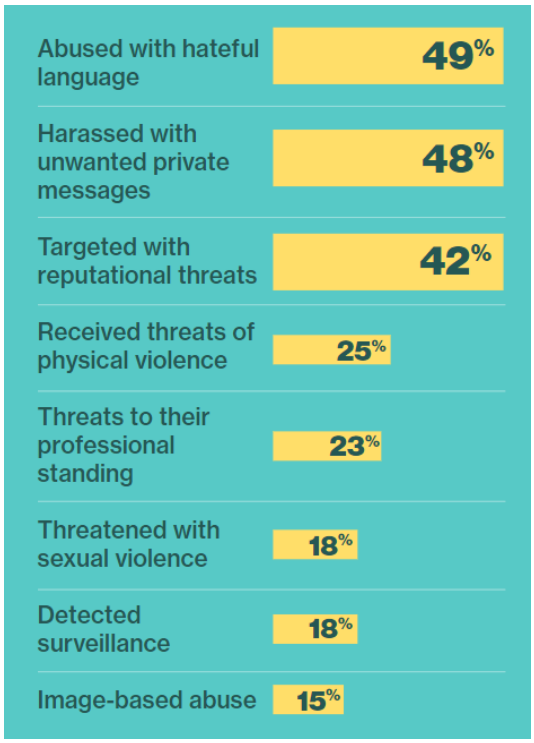


Figure 7: Types of online threats that journalists experience. (Source: UNESCO 2020b: 6)

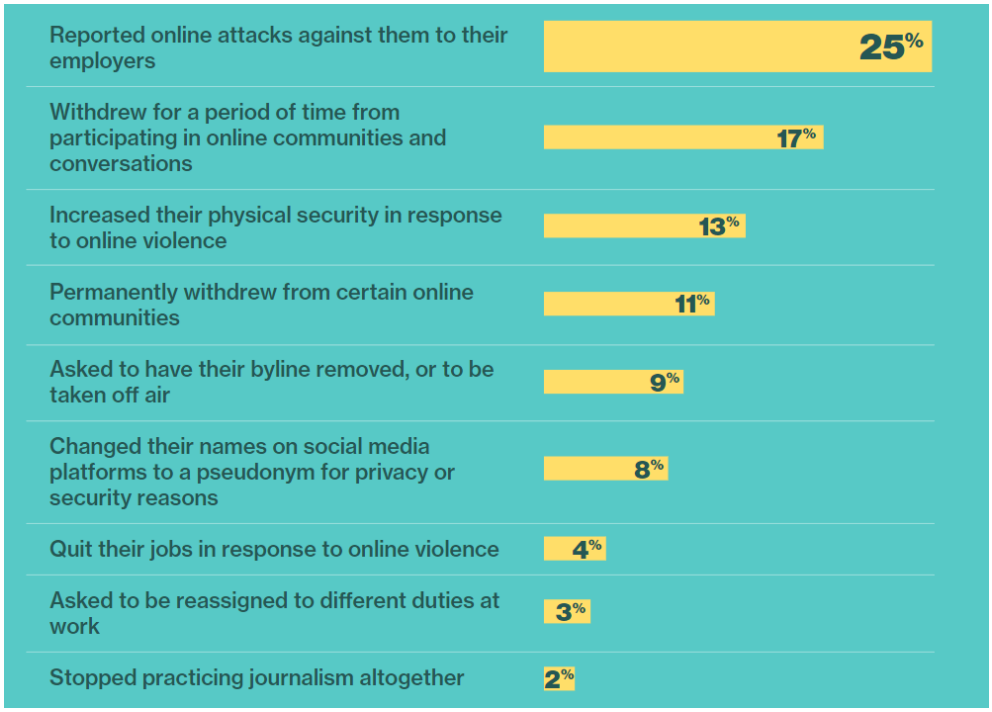


Figure 8: Main reactions from women journalists to online attacks. (Source: UNESCO 2020b: 11)

Gender-based hate speech is defined by the Council of Europe as "any supposition, belief, assertion, gesture or act that is aimed at expressing contempt towards a person, based on her or his sex or gender, or to consider that person as inferior or essentially reduced to her or his sexual dimension" (Council of Europe 2016c: 2). Gender-based hate speech can be encountered both online and offline through different shape and forms; namely victim blaming; slut-shaming; body-shaming; revenge porn; brutal and sexualised threats of death, rape and violence; offensive comments on appearance, sexuality, sexual orientation or gender roles; false compliments or supposed jokes using humour to humiliate and ridicule the target. The aim of gender-based hate speech is to humiliate or objectify women in order to depreciate their skills and opinions, destroy their reputation, make them feel vulnerable and fearful, and control and punish them for not following a certain behaviour (Council of Europe 2016c). The Council of Europe Group of Experts on Action against Violence against Women and Domestic Violence has in its General Policy Recommendation, General Recommendation No.1 on the digital dimension of violence against women (Council of Europe 2021b), that such acts constitute online sexual harassment and henceforth are prohibited under the Istanbul Convention on preventing and combating violence against women and domestic violence (Council of Europe 2011).

Gender-based hate speech dissemination is a consequence of the structure of patriarchal societies, in which "persistent and unequal power relations between women and men" permit and encourage the spread of degrading messages about women or girls (Council of Europe 2016c: 2). For instance, in 2015, women only made up 25% of people in the news, reinforcing gender stereotypes and bias by contributing to the invisibility of women in public discourse. Its consequences are particularly gendered, as "women and girls suffer from particular stigma in the context of structural inequality, discrimination and patriarchy" (UNGA 2018: 7), and the fact that its perpetrators often remain anonymous also contributes to increasing fears of violence and thus an increasing sense of insecurity and distress on the victim's side.

5.2 Migrants

Migrants and refugees tend to be particularly vulnerable to racism, discrimination and status-related intolerance (United Nations n.d.). They have been victims of increased hate speech and hate crime in the past year, caused by the anti-migrant and scapegoating rhetoric of politicians or public figures in the media. The Council of Europe Steering Committee on Anti-discrimination, Diversity and Inclusion (CDADI) on preventing and combating hate speech in times of crisis Study from 2023 quotes examples of how these hate narratives by public figures and in the media have risen in past years. The study adds that even in countries where there is a general positive perception on migrants and refugees, people are become more 'opinionated' and 'radicalised'.

According to the United Nations, increasingly, migrants and refugees are portrayed as "unable to adapt to local customs and life, and routinely associated with fears of violence and terrorism, while their positive contribution to societies is ignored" (United Nations n.d.). Furthermore, according to a FRA survey conducted in 2022 regarding discrimination and harassment experienced by migrants and descendants of migrants, 30% of them have experienced some form of harassment (FRA 2023). Migrant women in particular, are disproportionately affected by hate speech, as they face not only xenophobic and racist rhetoric but also gender-based violence and gender stereotypes (Recommendation CM/Rec (2022)17 of the Committee of Ministers to member States on protecting the rights of migrant, refugee and asylum- seeking women and girls, Council of Europe 2022c).

There is increasing concern regarding violence against migrants, as hate groups, politicians, and news agencies fuel hate speech against migrants to serve their own agendas and in consequence acts of intimidation and violence spike and disinformation intensifies (United Nations n.d.). This dynamic of acts of hate means that migrants become one of the main targets of hate crimes and are vulnerable to degrading treatment and social marginalisation. This amounts to a blatant violation of their fundamental rights.

5.3 Religious minorities

Regarding minorities, according to the Special Rapporteur on Minority Issues Thematic Report, 70% or more of those targeted by hate crimes or hate speech in social media are minorities. In addition to being the main targets of hate speech, the Thematic Report also claims that members of minority groups are also more likely to be affected by restrictions and/or removals by social media content moderation systems (United Nations, n.d. c).

Muslim and Jews are the largest European religious minorities in Europe, and both seem to be facing increasing hate.

Both groups continue to report threats, violence, harassment, intimidation and vandalism against their religious spaces and community centres. Since October 7th, 2023, the situation has been aggravated for both communities (European Commission n.d. a). The Council of Europe has conducted an initial consultation on the rise of hate against Muslims by consulting Muslim organisations. All the organisations noted an increase in anti-Muslim conspiracy theories circulating online. They have also noticed Islamophobic slogans spreading online such as: "le départ ou le cercueil" ("leave or you will end up in a coffin") in France. In the UK "just eradicate every single f****** Muslim" was circulating online. Fear is instilled against Muslims by claiming that they are "Islamising" Europe and the West, they want to "ban Christmas", they are referred to as "terrorists" or "wife beaters". Most of the Muslim organisations interviewed in a study in 2021 noted that

this hate is mainly encountered online and was most often committed by extreme right identitarian movements (Council of Europe 2021).

Also, in Europe, 80% of Jewish people have noted an increase in antisemitism in their country between 2019 and 2024. In 2023, 90% said that they encountered antisemitism online with 56% stating it was coming from people they know (European Agency for Fundamental Rights 2024a). It should be noted that this data was collected before October 7th, 2023. Data collected from 12 Jewish organisations post October 7th demonstrates that some organisations have reported an increase of 400% in antisemitic incidents. In past years, Jewish people have been blamed for things like the COVID-19 pandemic and for the rise in arrivals of migrants and refugees. These are depicted as a 'Zionist' plan to destabilise Europe (CDADI 2023). In recent years, the COVID-19 pandemic and possibly also the Russian Federation's full-scale invasion of Ukraine, have caused many antisemitic narratives and conspiracy theories to be recycled (CDADI 2023).

5.4 LGBTQI+

Lesbian, gay, bisexual, transgender, queer and intersex people, or those who do not conform to any gender, are continuously exposed to discrimination, stigma, hatred and abuse. This is sometimes due to the mere perception of their homosexuality or trans identity (Council of Europe 2023b). The abuses these communities face tend to be extremely severe, with some countries criminalising non-conforming sexual orientations and gender identities (United Nations n.d.). In doing so, these countries contribute to fostering the acceptance of intolerance, stigmatisation and violence towards the people from these communities.

The EU Agency for Fundamental Rights (FRA) conducted its third LGBTQI+ survey in 2023 (European Union Agency for Fundamental Rights 2024), revealing that trans individuals face disproportionately high levels of hate-motivated harassment and violence compared to other LGBTQI+ groups. The survey found that over two-thirds of trans and intersex respondents experienced hate-motivated harassment in the year preceding the survey - a significant increase from previous years. More than one in ten LGBTQI+ individuals reported being physically or sexually attacked in the previous five years, with trans and intersex people being the most targeted. In total, 59% of lesbian women, 52% of gay men, 77% of trans women and 72% of trans men, experienced some form of harassment in the 12 months before the survey. The same study concluded that most EU countries consider that violence, prejudice and intolerance have increased towards the LGBTQI+ population. On average in the EU, prejudice and intolerance increased by 36% in 2016 and by 53% in 2023, whereas violence increased from 43% to 59%. One of the German survey respondents stated "Being LGBTIQ is scary right now because of all the hate being created by politicians and other leaders." (European Union Agency for Fundamental Rights 2024: 64).

In addition, a record number of hate crimes were committed against transgender people in 2023 in England and Wales, even as racist and homophobic hate crimes recorded by police fell for the first time on record. In 2023, 4,732 hate crimes against transgender people were recorded, which is a rise of 11% compared to the previous year. The UK Home Office report speculates that comments by politicians and the media over the last year may have led to an increase in these offences, including the speech of Rishi Sunak in which he seemed to argue that transgender identities were not valid. "A man is a man and a woman is a woman. That's just common sense" (Goodier 2023).

Finally, certain religious and political leaders also blamed the LGBTQI+ community for COVID-19, giving rise to an increase in hate speech (ECRI 2020). For instance, in Ukraine, Patriarch Filaret, head of the Ukrainian

Orthodox Church – Kyiv Patriarchate, claimed that COVID-19 was, "God's punishment for the sins of men" - specifically linking it to same-sex marriage. This statement prompted a lawsuit from a Ukrainian LGBTQI+ group (Bacchi and Georgieva 2020). Similarly, in Belarus, some religious leaders attributed the pandemic to the presence of LGBTQI+ individuals, further inciting discriminatory rhetoric (Radio Free Europe Radio Liberty 2021a).

Some political leaders also contributed to the creation of an environment of intolerance and hate. In Hungary, Prime Minister Viktor Orbán's government intensified attacks on LGBTQI+ rights, including banning legal gender recognition for transgender and intersex individuals in May 2020 (Human Rights Watch 2020). Additionally, in December 2020, the Hungarian Parliament passed constitutional amendments further restricting LGBTQI+ rights (Amnesty International 2020). Similarly, in Poland, political leaders intensified anti-LGBTQI+ rhetoric throughout 2020-2021, with numerous municipalities declaring themselves "LGBT-free zones" - actions that were rationalised as defending public morality amid the health crisis (Radio Free Europe Radio Liberty 2021b). These measures fostered a climate of heightened discrimination against the LGBTQI+ community.

5.5 Ageism

Ageism "refers to the stereotypes (how we think), prejudice (how we feel), and discrimination (how we act) towards others or oneself based on age" (WHO 2021). Ageism can stem from internalised stereotypes about what a person of a certain age can be or do.

Ageism can affect the young people, and according to the World Health Organisations data, young people report more age discrimination in Europe than any other age group. An example of a stereotype about young people is that they do not want to work.

On average, half the world's population is ageist against older people. Despite it not being commonly mentioned, the elderly suffer from a lack of societal awareness of their existing discrimination. The global prevalence of elder abuse in community settings, where they suffer some form of a high level of psychological abuse is around 15.7% or one in six of the older population (European Commission 2021). However, this numbers varies according to world regions, for example in Asia it goes from 14% in India to 36.2% in China. In Europe the range varies from 2.2% in Ireland to 61.1% in Croatia. In the Americas it ranges from 10% in the USA to 79.7% in Peru (Yon et al. 2017).

Elder abuse, which encompasses both hate speech and hate crime directed towards older individuals, remains an underreported and largely hidden issue, despite its alarming prevalence across Europe (European Commission 2021b). This gap is further evidenced in this Study, as none of the CSOs involved in the research focused their campaigns on addressing hate against the elderly.

Elder abuse is acknowledged as a serious violation of human rights that requires prompt intervention. Abuse can present itself through various forms, namely physical abuse, such as causing bodily harm or restraining an adult against their will. Emotional/psychological abuse takes place in the form of addressing someone with hurtful words or threats. Financial abuse consists of misusing or stealing from an older person, such as taking their pension, using their social security benefits, using their credit card or withholding crucial information regarding their finances. There is neglecting abuse which consists of not providing the elder person with the needs they require. There is also sexual abuse (National Institute on Aging 2023). Older

women, in particular face heightened vulnerability to abuse. This is due to intersecting factors such as financial dependence and social isolation. Elder abuse can take place in a range of places, the elder person's home, a nursing home, a relative's, or friend's home. The mistreatment can come from relatives, strangers, friends, healthcare providers or caregivers.

Ageism can change how we view ourselves, erode solidarity between generations, devalue or limit our ability to benefit from what younger and older populations can contribute, and can impact an individuals' wellbeing (WHO 2019). It is also a severe public health issue with broad negative impacts on families and society as a whole, leading to considerable health consequences including heightened risks of morbidity, mortality (earlier deaths by 7.5 years), poor physical and mental health, slower recovery from disability in older age, institutionalisation, and hospital admission (WHO 2019, and Yon et al. 2017).

5.6 Disablist and ableism hate

In addition to ageism, disability hate crime is one of the least statutory recognised hate crimes globally. This is despite persons with disabilities being routinely harassed verbally, physically, and sexually in public spaces (Wood 2024). Unfortunately, it is not a rare hate crime but simply remains underreported and unnoticed by the police and criminal justice agencies compared to other hate crimes. According to Wood, "Disability is considered unique in comparison to other minority populations since it is more complex as not all individuals with impairments identify as disabled, alongside the perceived notion of vulnerability attached to disability labels" (Wood 2024: 5). Trolling and flaming (act of posting insults, often including profanity) are the two most common hate mechanisms employed online against people with disabilities. There is also ableism, which is based on the belief that typical abilities are superior. Additionally, many disability hate crimes involve the theft of money or valuables (Sherry 2010). Persons with disabilities often receive comments that reinforce their difficulty to find a romantic partner, maintain sexual relationships, or have children due to their disability (García-Prieto et al. 2023). Women with disabilities are particularly targeted, facing not only disablist hate but also misogynistic abuse, which often includes harmful stereotypes about their sexual and reproductive capabilities.

It must be noted that often the hate crimes are committed by insiders, notably friends, family or carers. This creates a unique perpetrator-victim relationship. For this reason, the new term 'mate crime' has arisen (Wood 2024). An example of this can be seen with Brent Martin's August 2007 case. This was a British disability hate crime committed by Miller, Hughes and Bonnellie. These three 'friends' of Martin beat and kicked him at least 18 times in the head before he died (Sherry 2010). In this story, we see the strange offender-victim relationship where Martin even apologised to the perpetrators on multiple occasions while they attacked him.

Unfortunately, 9 out of 10 victims do not report these crimes (García-Prieto et al. 2023).

5.7 Intersectionality of protected characteristics or status

Persons who are at the intersection of two or more protected characteristics or status may encounter specific challenges and substantial hate speech (OHCHR 2022). For example, "Violence against women and domestic violence can be exacerbated where it intersects with discrimination based on a combination of sex and any other ground or grounds of discrimination as referred to in Article 21 of the Charter, namely race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion,

membership of a national minority, property, birth, disability, age or sexual orientation ('intersectional discrimination')" (Directive (EU), 2024/1385, Article 6). An example is given by Amnesty International (2018), who claim that black women are 84% more likely to be mentioned in abusive or problematic tweets than white women, and "one in ten tweets mentioning black women was abusive or problematic, compared to one in fifteen for white women" (Amnesty International 2018). Additionally, migrant women frequently encounter multiple discrimination based on both gender and migration status, leading to cumulative disadvantages. The Council of Europe notes that such discrimination can create significant barriers, including limited access to employment and social services (Pandea, Grzemny and Keen 2019). The risk of suffering from hate speech is even higher when you have more than two intersecting identities. For example, according to the Center for Women's Policy Studies, "the fact of disability raises the chances that a woman will be the victim of a crime, and women of colour even more so" (Davis 2002).

This phenomenon is acknowledged in the Council of Europe's Recommendation CM/Rec(2022)16 on combating hate speech. It emphasises that individuals and groups can be targeted based on multiple, intersecting grounds - necessitating special protection measures. The Recommendation highlights the importance of understanding the various expressions and impacts of hate speech - including intersectional hate speech – in order to effectively prevent and combat it. It calls for comprehensive strategies which consider the compounded effects of intersectional discrimination. The Steering Committee on Anti-discrimination, Diversity and Inclusion is preparing a feasibility study and possible draft Recommendation on preventing and combating intersectional discrimination (for more information Intersectional Discrimination - CDADI)

According to a FRA survey, for women of all ages, whether they are immigrants or descendants of immigrants, whether they openly display their religion or keep it private, they remain more discriminated against than men (FRA 2023).

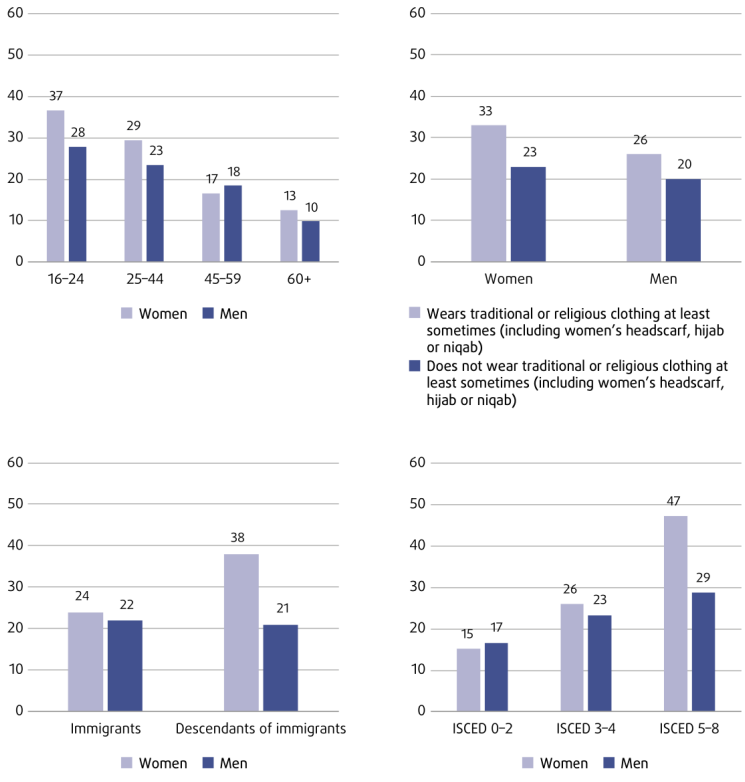


Figure 9. Prevalence of racist harassment against women and men by age, wearing of traditional or religious clothing, immigrant generation and education (%). (Source: FRA, 2023, p.62)

6. The Aim and Main Perpetrators of Hate Speech

6.1 The aim of hate speech

Hate speech is often conducted with the intention to harm the reputation of people belonging to a minority group characterised by particular aspects, such as disability, ethnicity, gender identity, race, religion, sexual orientation or other similar characteristics by depicting them as inferior (Jääskeläinen 2020). Hate speech often originates from judging someone based on the group they belong to. This happens through cognitive shortcuts caused by biases and exaggerated beliefs associated with the group this person belongs to (Cramer et al. 2020). These biases are complex to deconstruct as they are embedded in cultural differences, local, national and world history and contemporary socio-economic dynamics. These cultural, social, economic and historical factors can be caused and/or reinforced by a lack of appreciation of diversity and diverging opinions as well as a lack of education in managing such diversity (UNESCO 2023). Furthermore, biases are often instilled by societal influences such as family or friends (Martin et al. 2017). Lastly, hate motivated behaviour can be incentivised by other factors, such as attitudes stemming from social dominance.

Hate speech not only directly harms its targets but also indirectly curtails their freedom of expression. Victims of hate speech often feel unsafe or threatened in environments where hateful language is pervasive, making them hesitant to express their views openly (UNESCO 2023). This suppression of speech can be further compounded when hate speech is used as a pretext to silence human rights defenders and journalists, thereby restricting public discourse and accountability (UNESCO 2023). Moreover, hate speech serves to polarise and terrorise targeted groups, creating an atmosphere of fear and division (UNESCO 2023). Through dehumanisation and incitement to violence, hate speech fosters hostility toward those presented through the hate speech as a threat, further silencing the voices of marginalised communities (UNESCO 2023). This undermining of free speech erodes democratic dialogue, particularly for vulnerable groups already at risk of marginalisation.

6.2 Identifying the most common perpetrators of hate speech

The most common perpetrators of hate crimes and hate speech tend to be young people, most of whom are male, white, people with high emotional instability, and people with a lack of exposure to diversity (Cramer et al. 2020). Furthermore, hate speech often emerges from large-scale intergroup dynamics, particularly in contexts of social segregation or perceived competition for resources in times of crises (for example COVID-19, inflation, etc.) (Cramer et al. 2020). Such dynamics fuel resentment, and the fear of losing status or power can lead to hate-motivated behaviour, with individuals or groups attempting to dehumanise, victimise, and marginalise others. Understanding these dynamics is critical in addressing hate speech, as it often stems from these broader socio-political tensions - creating a fertile ground for discriminatory rhetoric that justifies exclusion or violence against the out-group (Bukowski et al. 2016). Additionally, hate speech that persists over an extended period of time can be the cause of acts of violence against the groups targeted by the hateful words. The spread and normalisation of hateful narratives instils and encourages the dehumanisation of targeted groups (UNESCO 2023). This explains, according to Tsesis (1999), why historical atrocities such as the Holocaust, the displacement of Native Americans, and the enslavement of Black people, were facilitated by the proliferation of hate propaganda. Over time, this propaganda built a conceptual framework which normalised intolerance, discrimination, and violence - laying the groundwork for oppressive and racist policies (Jääskeläinen 2020). While some scholars, such as Desai (2003) and Heinze (2016) argue that there

is insufficient legal or scientific evidence to directly attribute these events to hate speech, the role of repeated hate propaganda in fostering environments conducive to large-scale human rights violations should not be ignored (Jääskeläinen 2020).

In addition, certain types of notable events can also be associated with spikes of hate crime (Cramer et al. 2020). These events include terrorist attacks, sudden increases in immigration numbers, and the expansion of civil rights for minority groups.

Finally, in recent years, social media platforms have been found to play a significant role in exposing young individuals to far-right ideologies. These platforms facilitate the rapid dissemination of hateful content - often under the guise of humour or countercultural narratives - making such ideologies more palatable to the young people. Algorithms on platforms like YouTube can inadvertently lead users towards increasingly hateful content by recommending videos that align with previously viewed material, a phenomenon known as the "alt-right pipeline" (Ribeiro et al. 2019). Such social media processes foster 'echo chambers' where extreme views are normalised and reinforced. Additionally, far-right groups exploit social media to recruit members by creating a sense of community and belonging - appealing to young individuals seeking identity and purpose. Young people with a less stable sense of identity, belonging and community, might be more vulnerable to such targeted messaging (National Counterterrorism Center (n.d.). This can manifest in cases when such individuals feel disenfranchised and/or deprived of opportunities - a situation for which they perceive a need to attribute blame to others for their circumstances (Harpviken 2020). The opportunity of remaining (relatively) anonymous, along with the vast and rapid reach offered by social media platforms, can further embolden users to express and spread hateful views without suffering immediate repercussions (The Role of the Internet and Social Media on Radicalization, Office of Justice programmes n.d.). Consequently, social media not only facilitates the spread of far-right ideologies but also accelerates the radicalisation process among susceptible youth.

7. Consequences and Impact of Hate Speech

Hate speech, is not only directly harmful, but also has the potential to create conditions that exacerbate harm in the long-term, leading to deeper consequences such as human rights violations, discrimination, emotional and psychological damage, and even acts of violence. The "domino effect" of hate speech is often observed as disempowerment, marginalisation, suppression of voices, and the silencing of minority groups, all of which further exacerbates systemic inequalities. This harm extends beyond individual victims, fostering division and the polarisation of communities (Jääskeläinen 2020).

One of the most notable impacts of hate speech is the psychological distress it causes to the stigmatised individuals. The constant exposure to hate-filled narratives against one's community can cause deep emotional and mental harm. Hate speech also erodes social cohesion by creating environments of fear and hostility. These erode trust and solidarity within communities which in turn worsens mental health in individuals (UNESCO 2023).

In educational settings, the influence of hate speech is particularly detrimental. Evidence suggests that hate speech, especially when shaped by gender norms, can contribute to bullying and harassment in schools. This, in turn, has a detrimental impact on students' academic performance, leading to decreased motivation, concentration difficulties, and, in some cases, students dropping out of school altogether. The long-term consequences are significant: lower academic achievement, reduced access to higher education, and fewer employment opportunities for affected students. This in turn further perpetuates cycles of disadvantage and exclusion (UNESCO 2023).

The impact of hate speech is not confined to national borders; it has significant cross-border implications, particularly in the digital age. According to a study by the European Commission (2021), 93% of stakeholders surveyed, believed that hate speech perpetrated online could have spillover effects across national boundaries. This is largely due to the widespread access to similar online content across Europe, which allows hate speech to transcend local contexts and influence audiences in other States. Online platforms and media outlets play a crucial role in amplifying the dissemination of hate speech, emboldening hate groups by expanding their reach and broadening their audience (European Commission 2021). As such, the cross-border spread of hate speech poses a serious challenge to efforts aimed at curbing its harmful effects, both within and beyond national jurisdictions. The Council of Europe Recommendation CM/Rec(2022)16 on combating hate speech includes in its list of recommendations, national coordination and internation co-operation as an essential aspect for any comprehensive approach to combating hate speech covering legal and non-legal measures.

8. Legal Measures on Hate Speech

This section explores how countering hate speech is framed by international and European legislation, and looks at how EU policies attempt to tackle hate speech. The first part of this chapter focuses on the international legal guidance provided to states, as well as on the limits of this guidance. The second part focuses on European Law. Finally, the third section focuses on European policies and recommendations which aim to tackle hate speech and hate crime.

8.1 United Nations

International law dictates that not all forms of hate speech should be considered acceptable within the boundaries of freedom of speech. Hate speech should be criminalised when it is linked to incitement of violence, hostility, discrimination or racism (UNESCO 2022). These provisions stem from two important international guidelines: the International Covenant on Civil and Political Rights (ICCPR) and the International Convention on the Elimination of all Forms of Racial Discrimination (CERD). The enforcement of these international mechanisms is limited by the nature of international law and individual states' incorporation of these provisions into their domestic legal systems. The descriptions related to hate speech in the ICCPR and CERD have helped lawmakers assess what forms of hatred fall beyond the scope of freedom of expression and qualify as hate speech (UNESCO 2022). The definitions of hate speech within these articles, however, are broad, leaving room for interpretation by courts and policy makers (O'Flaherty 2012).

The UN framework categorises hate speech into three levels - offering distinct responses for each. The top level comprises hate speech which fulfils all the criteria of the "six-part threshold test". This was developed by the United Nations Office on Genocide Prevention and the Responsibility to Protect (UNOGPRP). This test is designed to assess when hate speech reaches the threshold of incitement to discrimination, hostility, or violence as prohibited under Article 20(2) of the ICCPR and Article 4 of the CERD. This threshold test is used by the UN to distinguish between hate speech that, while harmful, is protected under freedom of expression, and hate speech that constitutes incitement to violence or discrimination which states are obliged to prohibit. The six criteria are as follows:

- Context - The social, political, economic, and cultural conditions in which the speech occurs. For example, speech in a tense or conflict-prone environment might be more likely to incite violence.
- Speaker - The authority, influence, or credibility of the individual or group delivering the speech. Leaders, public figures, or those with large platforms have more capacity to incite action.
- Intent - Whether the speaker intended to incite discrimination, hostility, or violence. This requires careful analysis of the speaker's goals and the likely consequences of their speech.
- Content and Form - The specific language, imagery, and tone of the speech. This includes assessing whether the speech includes explicit calls for violence or dehumanisation of a group.
- Extent and Magnitude - How widely the speech is disseminated and its reach or audience. For instance, hate speech broadcast to a large audience has a higher potential for harm.
- Likelihood of Harm - The probability that the speech will lead to harm, including violence or discrimination. This involves considering the vulnerability of the targeted group and the historical context.

The intermediate level includes speech that may be prohibited under international law if the speech violates the respect of the rights or reputation of others or it violates the protection of national security, public order or of public health or morals.

Finally, the least severe form of hate speech may include expressions which are offensive; denial of historical events; blasphemous speech; disinformation; misinformation; and mal-information. These forms of hate speech are generally protected under free speech norms and cannot be legally restricted under international law.

Level	Definition and examples	Legal response
Top level	Hate speech that constitutes incitement to discrimination, hostility or violence and fulfils all the criteria of the six-part threshold test, such as: • Incitement to genocide and other violations of international law • Incitement to discrimination, hostility or violence • Incitement to racial discrimination	Must be prohibited under international law
Intermediate level	Hate speech that does not reach the threshold of incitement, such as: • Threats of violence • Harassment motivated by bias	May only be restricted if it fulfils the three-part test of Article 19 in the International Covenant on Civil and Political Rights
Bottom level	The least severe forms of hate speech, such as: • Expressions that are offensive, shocking or disturbing • Condonation or denial of historical events • Blasphemous speech • Disinformation, misinformation and malinformation	Should not be prohibited, even if offensive, but that should still be addressed through non-legal measures

Figure 10. The Three Levels of Hate Speech Under the United Nations Strategy and Plan of Action according to the UN. (Source: UNESCO 2023: 24)

Indeed, Article 19(3) of the ICCPR provides that freedom of speech may be limited "if provided by law, and if necessary and proportionate to respect the rights or reputation of others or to protect national security, public order or public health or morals" (UNESCO 2023: 21). Yet, these are broad terms which can be circumvented and/or construed in an expansive manner by the courts asked to interpret and apply them. This lack of clarity in delineating the boundaries between 'freedom of speech' and 'hate speech' might provide space for hate speech and manipulation to occur.

Article 20 of ICCPR, furthermore, prohibits any advocacy or hatred which constitutes incitement to discrimination, hostility or violence (UNESCO 2023). However, what constitutes "incitement or advocacy" of "hatred" can be interpreted differently (OHCHR 2010). An additional problem is the fact that the language of Article 20 of the ICCPR is seldom incorporated into domestic legislation. The lack of direct reference to "incitement" in domestic legislation indicates that states may either be reluctant to adopt the terminology of Article 20 of the ICCPR or are simply unaware of it. Yet, despite all these regulations, large sections of the Internet are 'de facto governed by private corporations', and therefore, state control over user content is limited - giving rise to 'ineffective enforcement of state jurisdiction in cyberspace' (Sjöholm 2024).

On the other hand, the CERD is more detailed and leaves less room for interpretation. Article 4(a) specifies the need for restrictions on expressions which share ideas on the "superiority or inferiority" of people distinguished by race (UNESCO 2023). Additionally, this article does not require evidence of intent or the "advocacy of hatred" and includes dissemination in the list of punishable practices (UNESCO 2023).
Prohibitions of hate speech under certain conventions even provide a more clear delineation of specific instances and consequences. For instance, Article 3 of the Convention on the Prevention and Punishment of the Crime of Genocide prohibits "conspiracy to commit genocide" and "direct and public incitement to commit genocide" - both of which can be linked to hate speech.

8.2 European Union on hate speech and hate crimes

8.2.1 EU Directives

The EU's Framework Decision 2008/913/JHA, criminalises hate speech or hate crime based on race, colour, religion, descent, or national or ethnic origin. It is the only EU criminal legal provision which harmonises the definition and criminal sanctions for some specific forms of hate speech and hate crimes. The Framework Decision, nevertheless, does not include sexual orientation nor gender identity as grounds for being targeted - consequently leaving the criminalisation of such acts to the authorities of EU's Member States. The objective of the Framework Decision is to guarantee that severe cases of racism and xenophobia face justifiable, reasonable, and deterrent criminal penalties across the entire EU. The Framework Decision requires EU Member States to take the appropriate steps to make public incitement to violence or hatred illegal on the grounds mentioned above. EU Member States were obliged to incorporate the requirements set forth in the Framework Decision into their national law by 2010 - requiring them to have national laws which criminalise hate speech and hate crimes on the grounds of race, colour, religion, descent or national or ethnic origin. While the Framework Decision aims to establish a common approach and standards across EU Member States, they do not provide guidance on what the sanctions or punishment for the perpetration of the acts they criminalise should entail (Peršak 2022).

Consequently, the EU Member States had to implement laws which made public incitement to violence or hatred illegal on the grounds of racist and xenophobic hate speech, the Member States had discretion in deciding which sanctions such crimes should carry. This resulted in discrepancies between certain Member States - some of which imposed stricter punishments than others. This is attributed to the fact that the EU has very limited criminal law scope and is restricted to specific areas which mostly include serious cross-border crimes or matters essential for the functioning of the Union, as specified in Article 83(1) of the Lisbon Treaty. Hate speech and hate crimes do not fall within the "Eurocrimes" listed there (Peršak 2022). Because of the negative effects brought about by the EU's lack of scope and precision in sanctioning hate speech and hate crimes, the Commission and the European Parliament have been calling for the amendment of Article 83(1). Such an amendment will permit a deeper harmonisation across EU Member States legislation by adding hate speech and hate crimes to the list of Eurocrimes. This will expand hate crimes and hate speech to include targeting on the grounds of race, religion, gender or sexuality. Such an amendment, nevertheless, is complicated to achieve. An amendment of this kind must be passed by a unanimous decision of all EU Member States and with the approval of the European Parliament.

Another important Directive in connection to (illegal) hate speech and hate crime calls is the Directive 2012/29/EU which establishes minimum standards for the implementation of rights and provision of support and protection for all victims of crime. The Directive mandates that victims should be acknowledged and

treated with respect, sensitivity, professionalism, and without discrimination by all relevant actors they encounter (European Union 2012: Article 1). It also requires that victims have access to specialised support services tailored to their individual needs, particularly ensuring that those most vulnerable, such as victims of hate crimes, receive special protection measures (European Union 2012: Article 8(3)). To further enhance the enforcement of the Victims' Rights Directive, the European Commission published the EU Strategy on Victims' Rights (2020-2025) in 2020. This strategy applies to all crime victims, but gives particular attention to the most vulnerable, among whom are victims of hate crimes. The Directive emphasises the need for targeted and integrated support for these victims, through close collaboration with relevant communities (European Commission 2021). It should be noted that in various Member States, severe expressions of hate speech and/or certain forms of hate are covered by criminal law. Victims of illegal hate speech therefore fall under the scope of this Directive.

8.2.2 EU Code of Conduct on illegal hate speech

The European Commission considers the fight against hate crime as a policy priority (European Commission 2021). Although the EU cannot impose common sanctions for crimes concerning hate speech and hate crimes, it does have the authority to impose minimum requirements within that area upon its 27 Member States.

In 2016, the European Commission adopted the Code of Conduct on Countering Illegal Hate Speech Online. It was signed as an agreement between the Commission, Google (YouTube), Facebook, X (formerly Twitter) and Microsoft in 2018. This Code of Conduct was subsequently also signed by Instagram, Google+, Dailymotion, Snap and Jeuxvideo.com in 2019. By 2019, the Code covered 96% of the EU market share of online platforms that could be affected by hateful content (European Commission 2019). This agreement significantly streamlined the review and removal of hate speech content. In 2016, 28% of such content was removed and by 2019 increased to 72%. While 40% of notices were reviewed within 24 hours in 2016, that percentage rose to 89% in 2019. The Code also contributed to increasing trust and cooperation between IT Companies, CSOs and Member States authorities. This took the form of a structured process of mutual learning and exchange of knowledge. The Code of Conduct also complemented the effective enforcement of the Council's Framework Decision 2008/913/JHA, which prohibits racist and xenophobic hate crimes and hate speech. It additionally supported efforts performed by national authorities to investigate and prosecute hate-motivated offences, both offline and online (European Commission 2019).

On 20 January 2025, the Code of conduct on countering illegal hate speech online was revised and integrated into the regulatory framework of the Digital Services Act (DSA) following a positive assessment from the Commission and the European Board for Digital Services (the Board). The revised Code of conduct (referred to as the 'Code of conduct+') strengthens the way online platforms deal with content deemed illegal hate speech according to EU law and Member States' laws. It facilitates compliance with and the effective enforcement of the DSA in this specific area. Following its integration, adherence to the Code of conduct+ may be considered as an appropriate risk mitigation measure for signatories designated as Very Large Online Platforms (VLOPs) and Search Engines (VLOSEs) under the DSA.

Focusing on countering racism, in 2019, the Commission published a report on "Countering Racism and Xenophobia in the EU: Fostering a Society Where Pluralism, Tolerance and Non-Discrimination Prevail', which provided an overview of the main areas of policy actions in the fight against racism and xenophobia in the EU. Following the report, the Commission set out the EU Anti-Racism Action Plan covering 2020 to 2025,

which defined the main principles of combating racism, both at the EU and national levels (European Commission 2021). These guiding pillars include measures which aim to address hate speech online, and also include the DSA.

8.3 The Council of Europe

The Council of Europe specifically has adopted a large range of Conventions and Recommendations providing policy guidance to its Member States on hate speech.

In 2016, the European Commission against Racism and Intolerance (ECRI) published General Policy Recommendation No.15 on Combating Hate Speech (ECRI 2016). The Recommendation outlined the need for rapid responses to hate speech from public figures, which are expected to both condemn it, as well as take action by reinforcing the values which a specific instance of hate speech has threatened (ECRI 2016). The General Policy Recommendation encourages media self-regulation and highlights the importance of raising awareness about the harmful effects of hate speech. Additionally, it suggests withdrawing financial and other support from political parties which actively engage in hate speech and criminalising its most extreme forms. The ECRI recommendation highlights that anti-hate speech measures should be well-grounded, proportionate, non-discriminatory, and must not be misused to limit freedom of expression, assembly, or to suppress criticism of government policies, political opposition, and religious beliefs.

The importance of data collection is underlined and followed by advice on how to engage in effective data gathering. The document also recommends imposing penalties which efficiently respond to the offence yet remain proportionate (ECRI 2016). Furthermore, it is important to note that the recommendation includes, for the first time, “sex” and “gender”, as grounds for being targeted by hate speech (Council of Europe 2016c).

The work of the Council of Europe on hate speech covers the online dimension and is also covered by the Internet Governance Strategy for 2016-2019 (CM(2016)10) and the Council of Europe Digital Agenda 2022-2025, Protecting human rights, democracy and the rule of law in the digital environment. . The Digital Agenda includes objectives aimed at establishing common policies on internet governance, specifically targeting network and information security and safeguarding vulnerable populations, particularly women and children. Such measures would create a secure digital environment which promotes accountability and protection for those most at risk of online abuse, reducing the prevalence of hate-driven content.

8.3.1. Committee of Ministers Recommendation CM/Rec(2022)16 on combating hate speech

In 2022, the Council of Europe’s Committee of Ministers adopted a Recommendation CM/Rec(2022)16 on combating hate speech. This Recommendation outlined a comprehensive approach to addressing hate speech by using a Human Rights framework, and included both legal and non-legal measures. It provided a definition of hate speech, factors for assessing the level of severity of hate speech, and guidance for developing appropriate responses for different layers of hate speech. Within this Recommendation, the Council of Europe specifically mentions CANs as a non-legal method through which to address hate speech, as well as other general awareness-raising, education and training methods (Council of Europe 2022b).

This Recommendation and its explanatory memorandum provide guidance for member states to implement a comprehensive and calibrated set of legal and non-legal measures. It builds on international human rights

standards and relevant case-law of the European Court of Human Rights and pays special attention to the online environment. The Recommendation also addresses other key actors, including public officials, elected bodies and political parties, internet intermediaries, media, and civil society organisations.

The Recommendation contains a broad definition of hate speech (see §2 of the Recommendation) and distinguishes within this definition different layers of hate speech. Namely it distinguishes between hate speech liable under criminal law and that under administrative and civil law for expressions which are not sufficiently severe to be legitimately restricted under the European Convention on Human Rights but nevertheless call for alternative responses (see §3 of the Recommendation). It furthermore provides factors for assessing the level of severity of hate speech and guidance for developing appropriate and proportionate responses for those different layers of hate speech (§§4 et seq. of the Recommendation). The Recommendation then outlines the necessary legal frameworks, including expressions of hate speech which are subject to criminal liability in line with international standards and court case law, this includes online hate speech.

The Recommendation pursues a comprehensive approach to preventing and combating hate speech. Therefore, it not only deals with the necessary legal framework for combating hate speech but also contains important guidance for addressing the root causes of hate speech through non-legal means. This is particularly through recommendations made in Chapter 4 of the document covering awareness-raising, education, training and the use of counter- and alternative speech. The different constitutional and legal orders and the varying situations in the Member States make it necessary to explore various avenues for implementing this Recommendation.

An Explanatory Memorandum accompanies the Recommendation. It outlines the reasoning behind the provided recommendations, making reference to the case law of the European Court of Human Rights, International treaties and other relevant standards. The Explanatory Memorandum provides further examples and considerations to assist legal professionals and practitioners who are working with the Recommendation in their respective fields.

8.4 Legal measures against hate speech targeting specific groups

8.4.1 Gender-based hate speech

In 2011, the Council set out a legal precedent by adopting the Convention on Preventing and Combating Violence against Women and Domestic Violence (Istanbul Convention). This is an extremely comprehensive legally-binding treaty addressing root causes of violence against women and calling for greater equality between women and men (Council of Europe 2016c). Articles 33-40 of the Convention specifically address the issue of hate crime, by requesting state parties to criminalise forms of violence which relate to gender based hate speech, namely stalking and sexual harassment (Council of Europe 2011). This Convention also establishes a specific monitoring mechanism, the 'Group of Experts on Action against Violence against Women and Domestic Violence' (GREVIO). This group was established to ensure effective implementation of its provisions (Council of Europe 2024). In monitoring the implementation of the Istanbul Convention, GREVIO has identified that the digital dimension of violence against women is often being overlooked in domestic laws and policies. In its General Recommendation No.1 on the digital dimension of violence against women, it outlines the problem of both gender-based violence against women committed online and facilitated by technology. Based on the four pillars of the Istanbul Convention - Prevention, Protection,

Prosecution and Coordinated Policies - the Recommendation proposes specific actions to be taken. It coins the term "the digital dimension of violence against women" as comprehensive enough to comprise both online acts of violence and those perpetrated through technology, including technology yet to be developed.

In 2024, the EU ratified the Council of Europe's Istanbul Convention, thus ensuring it applies not only to the EU Member States which have ratified it, but also binds the decision-making institutions of the EU. These include the Parliament, Commission, Council, as well as the EU Court of Justice - which may interpret how the Convention applies within the EU legal order. This EU-wide ratification potentially opens a path to create a legal framework at an EU level to protect women against all forms of violence.

In 2013, the Council of Europe's Recommendation CM/Rec(2013)1 of the Committee of Ministers to Member States on gender equality and media outlined specific guidelines to ensure gender equality and for combating gender stereotyping in the media. The Recommendation encourages the establishment of new mechanisms for media accountability and civic responsibility, including for a public debate. It calls for the establishment of online and offline platforms to facilitate direct exchanges between citizens (Council of Europe 2013 and Council of Europe 2016c). Complimentary to this, and in response to the increasing violence that female journalists and other female media actors had been facing, such as sexist, misogynist and degrading abuse, threats, intimidation, harassment, and sexual aggression and violence, the Council of Europe set out Recommendation CM/Rec(2016)4 on the protection of journalism and safety of journalists and other media actors in 2016.

In 2014, the Council of Europe's Gender Equality Strategy 2014-2017 explicitly prioritised combating sexism as a form of hate speech under its first strategic objective, which aimed to challenge gender stereotypes and sexism (Council of Europe 2016c). By framing sexism as a form of hate speech, this strategy advanced efforts to dismantle deeply entrenched biases, to promote gender equality, and reduce the prevalence of hate speech targeting women.

The follow-up strategy for 2018-2023 built upon the previous strategy and extended its focus another six years. It also incorporated objectives which tackle gender-based hate speech in expanded contexts, the rights of migrants, refugees, and asylum-seeking women and girls. The strategy continued to address gender-based hate speech as a form of sexism, analysing and monitoring its impact in cooperation with other relevant sectors of the Council of Europe. It also drafted a recommendation to prevent and combat sexism, which included guidelines for addressing gender-based hate speech both online and offline. These guidelines covered new forms of sexism affecting individuals in private and public spaces and included sexist language (drawing on the Committee of Ministers Recommendation No. R(90)4 on the elimination of sexism from language), gender-based hate speech, and sexism in media and advertising.

The current 2024-2029 strategy aims to support Member States in implementing campaigns to prevent and combat sexism, and target gender-based hate speech both online and offline. These efforts span various sectors including education, justice, culture, sport, STEM fields, and the private sector (such as social media and social networks). The strategy also seeks to promote the implementation of other Council of Europe instruments addressing human rights violations rooted in prejudices, customs, and traditions based on stereotyped gender roles. Among those instruments is the Committee of Ministers Recommendation on combating hate speech.

The onset of AI systems brings new challenges and exacerbates existing problems. For example, content moderations increasing reliance on AI risks not identifying hate speech on some grounds or mis-labelling content which is part of counter narrative strategies. The Council of Europe study on the impact of artificial intelligence systems, their potential for promoting equality, including gender equality, and the risks they may cause in relation to non-discrimination gives the following example: "sexist and other forms of online hate speech have been highlighted as contingent on the rising use of social media platforms. At the same time, content moderation particularly affects minority groups, who are at risk of being silenced while at the same time subjected to hate campaigns. For example, the stereotypical association of words associated with the lesbian community (e.g., 'lesbian') with pornographic content often results in so-called 'shadow-bans' that limit the reach of social media posts, or in the outright impossibility to use certain words in account names and handles. The silencing effects of content moderation have a severe negative impact on the visibility and reach of organisations, activities and events aimed at countering hateful and discriminatory narratives targeting such minority communities." (Council of Europe 2023c: 27-28).

At the level of the European Union, the EU Gender Equality Strategy of the period 2020-2025, includes the proposal for a directive on combating violence against women and domestic violence. Within it the EU committed to do "all it can to prevent and combat gender-based violence, support and protect victims of such crimes, and hold perpetrators accountable for their abusive behaviour" (European Commission 2020: 3). It also wishes to extend ''gender-based violence' as a 'Eurocrime' under Article 83(1) TFEU (Treaty on the Functioning of the European Union). They also committed to propose a Digital Services Act, to clarify online platforms responsibilities concerning user-disseminated content (European Commission 2020: 5). This Act was adopted in 2022.

In May 2024, the EU proposed a new Directive on Violence Against Women, which "introduces definitions of relevant criminal offences and penalties, the protection of victims and access to justice, victims support, enhanced data collection, prevention, coordination and cooperation" (European Union 2024: 1). This directive seeks to criminalise several (online) offences, namely "non-consensual sharing of intimate or manipulated material, cyber stalking, cyber harassment, cyber flashing, and cyber incitement to violence or hatred" (European Union 2024: 9). This Directive additionally urged EU Member States to provide the possibility to "submit complaints online or through other accessible and secure Information and communication technology (ICT) for the reporting of violence against women or domestic violence," (European Union 2024: 25).

At the International level, the UN Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) provides guidelines which could be linked to regulating hate speech against women. This Convention, ratified by 186 countries, suggests its state-parties to include the elimination of gender-based hate speech in their domestic legal systems (ACLU 2010 and Sjöholm 2024). The United States is one of only seven countries (Iran, Sudan, Somalia, Nauru, Palau and Tonga) which have not ratified CEDAW (ACLU 2010).

CEDAW provides a framework to achieve equality between women and men by requiring women's and men's equal access to and opportunities in political and public life. It affirms the reproductive rights of women and highlights the link between discrimination and the reproductive role of women. It requires its state-parties to guarantee a woman's right to freedom and to provide women with the necessary education and means to exercise these rights (UNHCR n.d.). Nonetheless, its effectiveness is limited, in particular since the CEDAW Committee has no enforcement authority and can therefore only provide recommendations in particular areas where more progress could be achieved (ACLU 2010). In May 2023 the Committee recommended

eliminating gender bias in AI to effectively detect and regulate gender stereotypes, including hate speech (CEDAW 2023).

In various UN bodies, gender-based hate speech has come up numerous times over the years. In 2020, the UN mentioned gender as a 'protected ground' in their Strategy and Plan of Action on Hate Speech. In the same year, the Special Rapporteur on freedom of opinion and expression encouraged the inclusion of gender-based hate speech in Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR). and provided briefings on the effects of gender-based hate speech on women's freedom of expression including women's access to the public sphere (Sjöholm 2024). Gender-based hate speech was also briefly addressed in relation to gender-based violence in 2021, whereby the Special Rapporteur on violence against women and girls obliged its state-parties to regulate online against violence against women (Sjöholm 2024).

8.4.2 Hate speech targeting LGBTQI+

The Council of Europe aims to ensure the successful enforcement of legal, policy, and practical strategies to prevent and address discrimination and violence based on the grounds of sexual orientation and gender identity. The Sexual Orientation, Gender Identity and Expression and Sex Characteristics Unit (SOGIESC) also works on protecting LGBTQI+ people from hate speech and hate crimes. To do so, it supports Member States in implementing Recommendation CM/Rec(2010)5 on measures to combat discrimination on grounds of sexual orientation or gender identity. The Committee of Experts on Sexual Orientation, Gender Identity and Expression, and Sex Characteristics (ADI-SOGIESC) commenced its mandate in 2024. Operating as a subordinate body of the Steering Committee on Anti-discrimination, Diversity and Inclusion (CDADI), which oversees the Council of Europe's efforts to advance equality and foster inclusive societies, ADI-SOGIESC is entrusted with supporting Member States in the development and implementation of effective policies related to SOGIESC. It is specifically mandated by the Committee of Ministers to produce key deliverables, including drafting a Recommendation on the equality of rights of intersex persons. It is responsible for devising a Council of Europe strategy to combat discrimination and promote the rights of LGBTQI+ individuals. It is also tasked with conducting a comprehensive and thematic review of the Council of Europe Recommendation CM/Rec(2010)5. The Council of Europe and the EU are currently working on a joint project on the prevention of Anti-LGBTIQ+ Hate Speech and Violence. The SOGIESC Unit runs other cooperation programmes and has numerous resources on various themes that can provide guidance and build the capacity of member state authorities and civil society.

The European Commission launched on 12 November 2020 its LGBTQI+ Equality Strategy 2020-2025, which serves as a comprehensive framework to promote equality and fight discrimination against LGBTIQ+ people across the EU. The strategy outlines targeted actions across four main pillars: tackling discrimination, ensuring safety, building inclusive societies, and leading the call for LGBTIQ+ equality worldwide (European Commission n.d. f). So far, the Framework has accomplished major benchmarks, including proposals to extend the list of 'EU crimes' to cover hate speech and hate crime based on sexual orientation, gender identity, or sex characteristics. The Commission has also worked with EU Member States to strengthen the legal and policy framework protecting LGBTIQ+ rights and provide funding for civil society organisations which promote LGBTIQ+ equality. Despite progress, challenges remain, particularly in EU Member States with significant opposition to LGBTIQ+ rights (Implementation of the LGBTIQ Equality Strategy 2020-2025).

8.4.3 Hate speech targeting migrants and religious minorities

The Council of Europe has established a comprehensive framework to address hate speech, particularly against migrants and religious minorities. At the core of this framework is the European Convention on Human Rights, specifically Article 10, which safeguards freedom of expression but permits restrictions to protect public safety and the rights of others. Article 14 complements this by prohibiting discrimination based on factors such as religion and national origin (Hate Speech - Freedom of Expression). The European Court of Human Rights has interpreted these provisions in cases like Erbakan v. Turkey (2006) - emphasising that expressions inciting hatred are incompatible with democratic values.

Additionally, the European Commission against Racism and Intolerance (ECRI) provides policy recommendations and monitors hate speech incidents. It offers guidance on preventing and responding to such offenses (Council of Europe on hate speech). ECRI's General Policy Recommendation No. 15 provides comprehensive guidelines for preventing and responding to hate speech - urging Member States to adopt legal measures, support victims, and promote counter-speech initiatives. This recommendation specifically highlights the need to protect vulnerable groups, including migrants and religious minorities, from hate speech.

In Recommendation 1805(2007), titled "Blasphemy, religious insults and hate speech against persons on grounds of their religion", the Council of Europe's Parliamentary Assembly addressed the issue of hate speech directed at individuals based on their religious beliefs. The recommendation emphasises the importance of balancing freedom of expression with respect for religious diversity, and calls on Member States to enact legislation that penalises expressions inciting hatred, discrimination or violence against individuals due to their religion.

Finally, the Council of Europe's Committee of Ministers' Recommendation CM/Rec(2022)16 on combating hate speech provides guidance for Member States to implement a calibrated set of legal and non-legal measures to combat hate speech, with particular attention to protecting migrants and religious minorities. Additionally, recognising that migrants, refugees and asylum-seeking women and girls are particularly targeted by hate speech, Recommendation CM/Rec(2022)17 of the Committee of Ministers recommends that Member States take measures to protect them from hate speech and sexism.

8.4.4 Ageism and hate speech targeting children

Through its Strategy for the Rights of the Child (2016-2021), the Council of Europe also addresses hate speech affecting children. This strategy connects two of its priority areas to combating hate speech; ensuring that all children live free from violence (priority area 3); and protecting children's rights in the digital environment (priority area 5). These two priority areas contribute to reducing hate speech by creating safer online spaces and ensuring that children are protected from hate-based harm, especially in digital contexts.

The Council of Europe currently lacks a policy recommendation or legislation specifically aimed at protecting the elderly from hate speech. Despite this, the Council of Europe is active in defending their rights. For instance, it funded the study "Against Ageism and Towards Active Social Citizenship for Older Persons" which provided recommendations on how to enhance elder participation in society (ECHR Press Unit 2023).

The European Court of Human Rights has been active in protecting elders from abuse, as witnessed in the case of 'Dodov v Bulgaria' which involved the disappearance of the applicant's mother, who had Alzheimer's

disease, from a state-operated nursing home for the elderly. The Court found a violation of Article 2 (right to life) of the European Court of Human Rights, thus establishing a direct connection between the homes personnel failure to supervise the woman, despite instructions to never leave her unattended. Sadly, her disappearance ultimately caused her death (Quinn and Doron 2021).

8.4.5 Ableism hate speech

A pivotal component within the Council of Europe's regulatory framework on combating hate speech against persons with disabilities is the Recommendation CM/Rec(2022)16. This Recommendation defines hate speech as encompassing expressions that denigrate individuals based on personal characteristics - including disability. It urges Member States to implement calibrated legal and non-legal measures to prevent and combat hate speech, ensuring protection for vulnerable groups such as persons with disabilities.

Complementing this, ECRI issued General Policy Recommendation No. 15. This recommendation provides comprehensive guidelines for preventing and responding to hate speech, emphasising the necessity of protecting vulnerable groups - including persons with disabilities - from such expressions.

In December 2017, the Council of Europe organised a seminar on Hate Speech in Copenhagen which focused on the impact of hate speech on persons with disabilities. This event facilitated the exchange of experiences and effective awareness-raising strategies among Member States, civil society, and various sectors. Its aim was to address and mitigate hate speech targeting individuals with disabilities (Seminar on Hate Speech - Rights of Persons with Disabilities).

Finally, the Council of Europe's Disability Strategy 2017-2023 underscores the importance of protecting persons with disabilities from hate speech and discrimination. The strategy advocates for the implementation of comprehensive measures to ensure equal opportunities and safeguard the dignity of individuals with disabilities (Rights of Persons with Disabilities).

9. Non-Legal Measures to Combat Hate Motivated Behaviour and Hate Speech

Nowadays, there exists an extensive and diverse amount of approaches to combat hate motivated behaviour beyond legal measures. This Study will now briefly mention the following, policing solutions, psychological and other health-related strategies, technological strategies, and countering strategies through CANs.

9.1. Community-based policing

Community based policing can be seen as an effective method of enforcement, as it seeks to reduce intergroup tensions and bias-based violence. Community based policing is different from the traditional surveillance method of policing because it implements a strategy that puts the citizens at the core of its attention. It merges the concepts of security, safety and care, by building trust between the police and the communities targeted by hate crimes. This approach has the intention of increasing hate crime reporting. This method has been openly supported by the Council of Europe, who have worked with local police through the Intercultural Cities Programme to define and develop this concept of community policing.

The Council of Europe has developed two key manuals to promote inclusive policing within the framework of community policing. The aim is to enhance trust and cooperation between law enforcement and diverse communities. The Intercultural Cities Manual on Community Policing (Council of Europe 2019) guides local police, including high-ranking managers and public safety directors, in implementing community policing principles tailored to diverse societies. It emphasises building mutual trust and engaging citizens in defining safety solutions, thereby fostering peaceful coexistence.

The Policing Hate Crime against LGBTIQ+ Persons: Training for a Professional Police Response (Council of Europe 2025) is intended to be used by police trainers, investigators, managers, and frontline officers. This manual provides tools and information for training law enforcement on addressing hate crimes against LGBTIQ+ individuals. It aligns with Council of Europe standards, including the principles enshrined within the European Convention on Human Rights, to ensure a professional and respectful police response.

The EU, through the European Crime Prevention Network (EUCPN) and the EU Agency for Law Enforcement Training (CEPOL), have produced the Community-Oriented Policing in the European Union Today (EUCPN 2019) toolbox. It presents recent good practices in community-oriented policing across EU Member States. It serves as a practical resource for law enforcement agencies aiming to implement effective community policing strategies.

All three of the resources aim to equip law enforcement agencies with the knowledge and strategies necessary for effective community engagement - fostering an environment where diversity is respected and protected.

9.2 Education and training strategies

9.2.1 Education and training programmes, and their methodological approaches

As outlined in Recommendation CM/Rec(2022)16 on combating hate speech and numerous other European and international policy documents on hate speech, education and training are seen as an effective way of addressing the root causes of hate speech by sensitising learners of any group about the consequences or harmful rhetoric online and offline.

The competencies to be developed through education for human rights and democratic citizenship are outlined in CM/Rec(2010)7 on the Council of Europe Charter on Education for Democratic Citizenship and Human Rights Education. They include the development of knowledge, personal and social skills, critical thinking and understanding that reduce conflict, increased appreciation and understanding of the differences between faith and ethnic groups, building mutual respect for human dignity and shared values, and dialogue and non-violence for the resolution of problems and disputes. It is understood that such competences are essential to equip persons to recognise hate speech, the risk it poses to a democratic society, and to be able to take initiatives to address it. Tools for implementing education for human rights and democratic citizenship have been developed by the Council of Europe's Youth Department and Education Department. They include the educational manuals Compass, Compasito, Mirrors, and the manuals on Democratic citizenship and intercultural dialogue. Positive experiences addressing hate speech through human rights education and counter speech, have been gained with the use of the educational manuals Bookmarks and We CAN!. These were developed for the Council of Europe's Youth Departments No Hate Speech Movement campaign. Another useful tool is the UN Faith for Rights Framework and Toolkit, 2019, which uses a peer-to-peer learning methodology. Illustrative examples of educational civil society initiatives are provided in this study.

In the field of education, research connected methodologies exist, many of which are explored in the book "Hate-motivated behaviour: Impacts, risk factors, and interventions" by Cramer et al 2020. The methods include Measurement Development, Perspective Taking, Counterfactual Thinking, Intergroup Contact Approaches, and Social-Emotional Learning. These methods go beyond teaching why hate speech is harmful by addressing various root causes of hate speech.

As an example, the Perspective Taking method aims at deconstructing cultural messaging and stereotypes which are used to spread hate in political speech, social media and education (UNESCO 2023). It addresses inequalities by facilitating conversation on the topic of inequality, exploitation and privilege, and educates learners and educational staff against hate speech.

Education and training approaches in order to be effective must include a strategy to improve the social inclusiveness and diversity in society. It should also improve social and emotional skills to enable learners to identify their strengths and manage negative emotions. This approach supports people in building their self-confidence and improving their critical thinking skills. It can potentially reduce hate speech by permitting individuals to regulate their emotions, control their impulses, and engage in safe, ethical and responsible behaviours. It can help to cultivate perspective-taking, supporting them in negotiating conflicts constructively, and encouraging them to acknowledge strengths in others and work with them to solve problems (UNESCO 2023).

Bullying intervention programmes, which often incorporate peer support, counselling, and social-emotional learning techniques, aim to equip young people with crucial skills such as emotional regulation, goal setting,

and social awareness. These interventions have been shown to reduce both overt and subtle forms of prejudice among young people, potentially fostering more inclusive and empathetic environments (Cramer et al. 2020). Self-regulation skills developed through such programmes can be particularly effective in mitigating prejudiced attitudes and contributing to a more tolerant social climate. However, the effectiveness of these interventions is often hindered by a lack of adequate training for educators. Many teachers and school staff are insufficiently prepared to address complex issues related to hate, violence, and bullying prevention (UNESCO 2023). This gap in training can result in inconsistent applications of anti-bullying measures, potentially limiting the overall impact of such programmes and leaving some forms of bias and aggression unaddressed.

Similarly, psychological intergroup contact-based interventions consist of creating direct contact between the stigmatised group and the offenders. This has been empirically demonstrated to be effective, as prejudice and stereotypes stem from ignorance and not knowing each other. Yet, it could appear as quite confrontational and mentally hard for the stigmatised group.

Media and information literacy is an essential element of education to reduce hate speech, since hate speech is often disseminated in the media as outlined in §47 of CM/Rec(2022)16 and numerous other Council of Europe and international standards and policies outlined earlier in the Study. The Council of Europe's Digital Citizenship Education Handbook and the Internet Literacy Handbook may be particularly useful in that regard. Therefore, providing critical thinking skills adapted to internet communication challenges is crucial for tackling the hate speech problem caused by misinformation, biassed algorithms, and conspiracy theories (UNESCO 2022).

Furthermore, it is argued that public awareness campaigns or more generally public health programming can be effective. However, there is no formal evaluation of the efficacy of these strategies (Cramer et al. 2020). Arts, arts education and other cultural activities can offer constructive methods for deterring hate speech (Jääskeläinen 2020 and §49 of CM/Rec(2022)16). They can "help prevent hate speech from happening in the case where hate speech does not present an immediate threat" (Jääskeläinen 2020: 345-349).

9.3 Countering hate speech

9.3.1 Counter and alternative narratives

The Council of Europe Recommendation CM/Rec(2022)16 on combating hate speech in §194 of the explanatory memorandum has defined counter and alternative speech as expressions of counter and alternative narratives. These narratives are designed to combat hate speech by discrediting, deconstructing and condemning the narratives on which hate speech is based by reinforcing the values that hate speech threatens, such as human rights and democracy. It also clarifies that counter and alternative narratives to hate speech promote openness, respect for difference, freedom, and equality. While counter speech is a short and direct reaction to hateful messages, alternative speech usually does not challenge or directly refer to hate speech but instead changes the frame of the discussion (see Council of Europe manual We CAN!).

The use of counter and alternative speech forms are particularly important for addressing hate speech that does not reach the severity level for being addressed via criminal, civil or administrative procedures (see §§3 and 4 of Recommendation CM/Rec(2022)16). Research argues that counter speech exposes hate, deceit, abuse, and stereotypes through clarification and providing a counter narrative as well as advancing counter

values such as sharing experiences and uniting communities (Jääskeläinen 2020). The aim of counter narratives is to dispute and contradict commonly held beliefs and stereotypes related to minorities by sharing a different point of view which is based on human rights and democratic values, such as respect of plurality, freedom and equality (Jääskeläinen 2020). It does not function by directly discrediting hateful beliefs but by deconstructing the narrative which has helped in the dissemination of those hateful beliefs. It provides alternative ways of approaching and thinking about the issues. These alternative perspectives are based on accurate information against hate speech propaganda and may use humour to appeal to emotions (Jääskeläinen 2020). The UN guide for countering hate speech recommends the promotion of positive narratives and the fostering of user engagement, and empowerment towards these narratives, in order to spread counter narratives more effectively (United Nations 2023).

9.3.2 Human rights-based CANs

Human rights-based narratives are a solid structure to create a CAN to combat hate speech. They permit building an alternative way of thinking based on human rights or democratic values to the perpetrators hate speech. Human rights-based narratives that counter hate speech must be formulated through the universality of human rights language, and therefore cannot use scapegoating or outline divisions between "us" and "them" or distinguish between superior and inferior persons or arguments. These narratives must promote human rights by using concepts of "acceptance, dialogue, cooperation, non-discrimination, empowering for unity, empowering solidarity" (Step 4. Define the human rights-based narrative - Toolkit for human rights speech and Council of Europe n.d. c: 8). Equally, these messages can be communicated in various forms, namely by supporting, empowering or celebrating human rights, by informing someone of their rights, by empowering duty-bearers to fulfil their obligations or by empowering right-holders to speak up. These messages can also be spread by calling for action, such as calling for respect for human rights, for non-discrimination, or for justice, etc. (Step 5. Develop the message based on your human rights-based narrative - Toolkit for human rights speech and Council of Europe n.d. c: 10).

9.4 Digital tools

9.4.1 Technologies

Social media companies content moderators should carry an important responsibility for balancing human rights principles (Council of Europe 2023d). Social media can be a powerful tool to challenge hate speech through the dissemination of alternative narratives and counter-speech. Yet, the UN Special Rapporteur on Freedom of Opinion and Expression has warned against using excessively strict penalties for users or excessively intrusive technology (Sjöholm 2024). Interestingly, in the campaign undertaken by Jugendstiftung Baden-Wurttemberg from Germany - one of the CSOs that participated in the Study - when respondents who reported instances of hate speech online were asked whether they prefer freedom of speech or protection from harm, the responses were extremely polarised, with respondents favouring the extremes. Nonetheless, around 75% of respondents leaned towards preferring a 'protection from harm' approach. "The support for regulating free speech, particularly in the context of violence and hateful content, appears strong, while views on the limits of free speech in non-violent contexts are more varied". For more information on this campaign, refer to Section 3 of this Study. In this regard The Gender Equality Strategy 2024-2029 notes that "Freedom of expression is often abused as the pretext put forward to avoid accountability for unacceptable and offensive behaviour." As outlined in previous section on the Human Rights Court, balance of rights must be made. The exercise of freedom of expression by some individuals can result in, among other things, the

silencing of others. Therefore, it is crucial to strike a balance, ensuring that the right to freedom of expression is upheld for everyone, without infringing upon the rights and voices of others.

The Digital Services Act (European Commission n.d. b) aims at regulating different spheres of the online world, namely marketplaces, social networks, content-sharing sites, app stores, and online travel and accommodation services. Its primary objective is to curb illegal and harmful activities, as well as the dissemination of disinformation. It also promotes user safety, safeguards fundamental rights, and fosters a fair and transparent online platform ecosystem. The Digital Services Act is composed of different provisions, each tackling a different problem. The Code of Conduct+ is one of these provisions, and it aims at countering illegal hate speech online. The enforcement of the Code of Conduct is assessed through regular monitoring conducted in collaboration with a network of organisations across various EU countries. These organisations, following a commonly agreed methodology, evaluate how social media companies are fulfilling their commitments outlined in the Code (European Commission n.d. g).

The new developments in AI and Data Analysis have permitted the development of AI tools to detect the spread of hate speech and polarising narratives online. This is the case of the WhoDis AI detection tool which combines political analysis and technological innovation to support the detection of the spread of hate speech patterns online and offline. The detection of such patterns by the AI tool is enabled by Justice for Prosperity's lexicon and taxonomy of keywords, activities and strategic language used by anti-democratic actors. The usage of terms within the lexicon and taxonomy are risk indicators of the increase of antidemocratic or hate speech trends online. The AI tool detects any increase or decrease of these risk indicators and has the capacity to link their usage to events, narratives or the people who are behind the peaks or decreases in their usage. The capacity of the tool to detect toxic language is enabled by its use of Natural Language Processing instead of Large Language Models. The latter cannot process toxic language, making it complicated for it to process hate speech. Previous platforms that worked on hate speech detection failed to identify implicit hate speech messages whereas the WhoDis tool can detect this implicit language.

9.4.2 Content moderation and hate speech detection

However, because a tech companies main aim is commercial rather than the defence of democratic and human rights values, it is essential to have effective mechanisms to influence company policy regarding content moderation (Gorwa 2024). Lobbying and other forms of influencing need to be supported by research and the exposing of the main actors behind the propagation of hate speech. The WhoDis AI detection tool is a crucial piece of research equipment for researchers, journalists, and policymakers. It can provide them with the capacity, within minutes, to identify the causes of the spread of online and offline hate effectively. Some of the CSOs involved in this Study are using this tool. Furthermore, it can be a key tool for online content moderators who face the challenge of effectively detecting online hate speech in its various forms, for example when it shows up as idioms and lexical nuances, which vary widely across cultures, languages and regions (Council of Europe 2023d). The WhoDis tool can help tackle these issues as it is able to operate in 48 different languages that use the Latin alphabet. It is currently being developed to use other languages that do not use the Latin script, such as Cyrillic, Arabic or Hanzi (Chinese characters).

9.4.3 Exposing foreign influence actors who aim to destabilise democracy

It is important to gain a broad understanding of where hate speech is coming from. This includes recognising that there are foreign influence actors who aim to destabilise democratic regions. To learn more about such

threats, there are many studies that can be followed including the study conducted by the European External Action Service of the European Union "Foreign Information Manipulation and Interference Threats". The study identifies the topics that are found to be often manipulated by foreign actors, including Russia's full-scale invasion of Ukraine. It also mentions who the main countries are behind such destabilisation, namely Russia and China. The report additionally identifies different strategies adopted by foreign influences, such as impersonating famous newspapers by copying their style and the translation of content in 30 plus languages. The report finds that this increase in information manipulation and interference as well as hate speech and incitement to genocide have been crucial in justifying the invasion (EEAS 2023).

A second interesting study which exposes foreign influence actors whose aim is to destabilise democracies is the WhoDis Report (Justice for Prosperity 2023). It explores how different actors and groups undermine democracy and human rights in Europe. These actors are often European networks of individuals, conservative organisations or populist politicians who are supported and funded by their foreign homologues whose aim is to destabilise Europe by undermining democratic principles. They operate via various tactics, namely through the usage of transnational conservative organisational networks, the transformation of far right narratives into mainstream discourse, the spread of polarising narratives online, targeted disinformation campaigns against minorities, offering trainings via workshops or online events to teach methods on conservative agenda execution, the infiltration of democratic spaces to gain legitimacy, and most importantly the funding obtained through multiple income streams (private corporate, religious and political donors…) (Justice for Prosperity 2023).

10. Placing this study in previous (CAN) research

The Council of Europe has produced various research papers on the issue of combating hate speech. In 2016, the Council of Europe produced the report "Combating Sexist Hate Speech" (Council of Europe n.d. d), which provides regulatory recommendations for Member States, companies and CSOs on how to combat gender-based hate speech. Yet, it does not mention 'Counter' or 'Alternative Narratives' (Council of Europe n.d. d).

In 2023, the Council of Europe's Committee on Anti-Discrimination, Diversity and Inclusion produced a 'Study on preventing and combating hate speech in times of crisis'. This study highlights that in certain cases, counter narrative initiatives have been effective (Council of Europe 2023d)). The report presents three cases. One of the examples explains that a CAN campaign, which was led by the Scottish government in 2020, was designed to respond to a rise in hate crime caused by the pandemic. The campaign consisted of a sequence of letters appearing in advertisement spots in Scotland which started with "your hate has no home here", and were signed "Yours, Scotland". This was set up in this way to encourage those who saw the advertisements to report hate speech and hate crime if they witnessed it (Council of Europe 2023d). The impact evaluation of the 2020 campaign indicated that individuals were willing to act and report after experiencing or witnessing hate crime and hate speech if they felt part of a common initiative. Despite these case studies providing clarification on the effectiveness of CANs to combat hate speech, they do not provide a systematic data analysis methodology to analyse the effectiveness of CANs, something that this Study does in Section 3.

The EU institutions have also contributed to research on the deterring of hate speech. In 2020, the European Parliaments Policy Department for Citizens' Rights and Constitutional Affairs commissioned the report 'Hate speech and hate crime in the EU' (European parliament 2020). It performed an evaluation of online content regulation approaches. The study suggests that counter-speech, or "talking back against hate speech," is ineffective without institutional backing through appropriate functioning of the administrative and law enforcement procedure. While members of the majority and those directly targeted by hate speech may engage in counter speech, the severe impact of hate speech, discrimination, and other biases, undermine the dignity of the targeted individuals. It takes exceptional strength and bravery for those affected to respond, and does not necessarily promote their empowerment.

List of References Section 1:

Amnesty International (2018). Crowdsourced Twitter study reveals shocking scale of online abuse against women. https://www.amnesty.org/en/latest/press-release/2018/12/crowdsourced-twitter-study-reveals-shocking-scale-of-online-abuse-against-women/

Amnesty International (2020), Hungary: Dark day for LGBTI community as homophobic discriminatory bill and constitutional amendments are passed, Amnesty International, available at https://www.amnesty.org/en/latest/news/2020/12/hungary-dark-day-for-lgbti-community-as-homophobic-discriminatory-bill-and-constitutional-amendments-are-passed-2/?utm_source=chatgpt.com

Bacchi, U and Georgieva, M. (2020), LGBT+ group sues Ukraine religious figure linking coronavirus to gay marriage, Thomson Reuters Foundation News, available at https://news.trust.org/item/20200413191406-79wt9?utm_source=chatgpt.com

BBC News (2018) Jessikka Aro: Finn jailed over pro-Russia hate campaign against journalist. BBC News, 18 October. [online] Available at: www.bbc.co.uk/news/world-europe-45902496 (accessed 6th December 2024).

Bonotti M (2017) Book review: Hate speech and democratic citizenship. Social & Legal Studies26(2), 276–280. DOI: 10.1177/0964663917704734a.

Bukowski, M., De Lemus, S., Rodriguez-Bailón, R., & B. Willis, G. (2016). Who's to blame? Causal attributions of the economic crisis and personal control. Social and Personality Psychology Compass, 20(6).

Calderwood, I. (2018, September 27). "Dear Haters": Scotland's brilliant new ad campaign tackles discrimination and hate crimes. Retrieved from https://www.globalcitizen.org/en/content/scotland-ad-campaign-hate-crime-abuse/

Charts in France (2020). Freeze Corleone : Spotify refuse de retirer ses titres après la polémique.chartsinfrance.net. https://www.chartsinfrance.net/Freeze-corleone/news-115572.html

Combs,Turner, Whittle, – 2009 – Transphobic hate crime in the EU – ILGA-Europe. https://www.ilga-europe.org/sites/default/files/transphobic_hate_crime_in_the_european_union_0.pdf

Council of Europe (n.d. a). A guide to the interpretation and meaning of Article 10 of the European Convention on Human Rights. Council of Europe. Centre for Law and Democracy.

Council of Europe. (n.d. b). Intersectionality and multiple discrimination. Retrieved December 5, 2024, from https://www.coe.int/en/web/gender-matters/intersectionality-and-multiple-discrimination

Council of Europe (n.d. c) Toolkit for developing human rights-based narratives: Step-by-step guide. Retrieved from https://rm.coe.int/guide-to-developing-human-rights-based-narratives/1680a20c60

Council of Europe (n.d. d) Combating Sexist Hate Speech, Council of Europe Gender Equality Strategy, available at https://rm.coe.int/1680651592

Council of Europe (1997) - Council of Europe, Recommendation No.R(97)20 - 1680505d5b (coe.int)

Council of Europe (2011) Convention on preventing and combating violence against women and domestic violence (CETS No. 210) https://rm.coe.int/168008482e and https://www.coe.int/en/web/gender-matters/council-of-europe-convention-on-preventing-and-combating-violence-against-women-and-domestic-violence

Council of Europe (2013) Recommendation CM/Rec(2013)1 of the Committee of Ministers to member States on gender equality and media https://search.coe.int/cm?i=09000016805c7c7e

Council of Europe (2015), Council of Europe Gender Equality Strategy 2014-2017, Retrieved from https://rm.coe.int/1680590174

Council of Europe (2016a), Bookmarks, A Manual for Combating Hate Speech Online Through Human Rights Education. Rev. ed. Budapest: Council of Europe Publishing. Available at: rm.coe.int/168065dac7 (accessed 6th of December 2025).

Council of Europe (2016b) CM(2015)175 – Strategy for the Rights of the Child for 2016-2021: https://search.coe.int/cm#{%22CoEIdentifier%22:[%2209000016805c1d08%22],%22sort%22:[%22CoEValidationDate%20Descending%22]} - DOCUMENT URL: https://search.coe.int/cm?i=09000016805c1d08

Council of Europe (2016d)CM/REC 4 - Recommendation CM/Rec(2016)4[1] of the Committee of Ministers to member States on the protection of journalism and safety of journalists and other media actors - (Adopted by the Committee of Ministers on 13 April 2016 at the 1253rd meeting of the Ministers' Deputies): https://search.coe.int/cm?i=09000016806415d9

Council of Europe(2016e) CM10 – Internet Governance Strategy for 2016-2019- https://search.coe.int/cm?i=09000016805c1b60

Council of Europe (2016f) Recommendation (CM/REC(2016)4) of the Committee of Ministers on the Protection of Journalism and Safety of Journalists and Other Media Actors in 2016 https://search.coe.int/cm?i=09000016806415d9

Council of Europe (2017) Information disorder: Toward an interdisciplinary framework for research and policy making http://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77

Council of Europe (2019) Recommendation CM/Rec(2019)1 of the Committee of Ministers to member States on preventing and combating sexism, available at https://rm.coe.int/168093b26a

Council of Europe (2021a) Online hate speech is a growing and dangerous trend, Initial results of a consultation of Muslim organizations. Special Representative on antisemitic, anti-Muslim and other forms religious intolerance and hate crimes.

Council of Europe (2021b) GREVIO General Recommendation No. 1 on the digital dimension of violence against women https://rm.coe.int/grevio-rec-no-on-digital-violence-against-women/1680a49147

Council of Europe (2022a) General Policy Recommendation N°15. (Portal) https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/recommendation-no.15

Council of Europe (2022b), Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech and Explanatory Memorandum http://rm.coe.int/prems-083822-gbr-2018-recommendation-on-combating-hate-speech-memorand/1680a70b12

Council of Europe (2022c) Recommendation CM/Rec(2022)17 of the Committee of Ministers to member States on protecting the rights of migrant, refugee and asylum- seeking women and girls.

Council of Europe (2022d), Digital Agenda 2022-2025, Protecting human rights, democracy and the rule of law in the digital environment, Council of Europe Committee of Ministers, available at https://rm.coe.int/coe-digital-agenda-2022-2025-pro-eng-web/1680aa3e1b

Council of Europe (2023a) European Union accession to the European Convention on Human Rights - Questions and Answers. (2023, October 7). https://www.coe.int/en/web/portal/eu-accession-echr-questions-and-answers#:~:text=What%20does%20EU%20accession%20to,the%20EU%20itself%20is%20not

Council of Europe (2023b). Newsroom Stop intolerance and discrimination against LGBTI persons: new recommendation published - Portal. Retrieved from https://www.coe.int/en/web/portal/-/stop-intolerance-and-discrimination-against-lgbti-persons-new-recommendation-published?utm_source=chatgpt.com

Council of Europe (2023c), Study on the impact of artificial intelligence systems, their potential for promoting equality, including gender equality, and the risks they may cause in relation to non-discrimination https://rm.coe.int/study-on-the-impact-of-artificial-intelligence-systems-their-potential/1680ac99e3

Council of Europe (2023d), Study on preventing and combating hate speech in times of crisis, Steering Committee on Anti-Discrimination, Diversity and Inclusion, Strasbourg, available at https://rm.coe.int/-study-on-preventing-and-combating-hate-speech-in-times-of-crisis/1680ad393b

Council of Europe. (2024). The European Union deposited the instrument of approval of the "Istanbul Convention." Portal. Retrieved from https://www.coe.int

Council of Europe. (2024a). Recommendation CM/Rec(2024)4 of the Committee of Ministers to member States on combating hate crime and Explanatory Memorandum https://rm.coe.int/combating-hate-crime/1680b08c6b

Council of Europe (2024b). Council of Europe Gender Equality Strategy 2024-2029 https://rm.coe.int/prems-073024-gbr-2573-gender-equality-strategy-2024-29-txt-web-a5-2756/1680afc66a

Council of Foundations. (n.d.). Definitions of hate and extremism. Retrieved December 5, 2024, from https://cof.org/page/definitions-hate-and-extremism

Cramer, R. J., Fording, R. C., Gerstenfeld, P., Kehn, A., Marsden, J., Deitle, C., King, A., Smart, S., & Nobles, M. R. (2020). Hate-motivated behavior: Impacts, risk factors, and interventions. Health Affairs.

Davis, L. J. (2002). Bending over backwards: Disability, dismodernism, and other difficult positions. New York University Press. https://doi.org/10.18574/nyu/9781479820108.003.0012

DeMorgen (2023) "Meerdere figuren met wortels in de neonaziscene groeperen zich onder Action Radar Europe en kwamen samen in Brussel na aanslag"

Desai A (2003) Attacking Brandenburg with history: Does the long-term harm of biased speech justify a criminal statute suppressing it? Federal Communications Law Journal 55(2), 352–394. [online] Available at: www.repository.law.indiana.edu/fclj/vol55/iss2/8

European Agency for Fundamental Rights (2023)" Being Black in the EU. Experiences of people of African descent" Publications Office of the European Union

European Agency for Fundamental Rights (2024a) "Jewish People's Experiences and Perceptions of Antisemitism" Publications Office of the European Union

European Agency for Fundamental Rights (2024b) "LGBTIQ Equality at a crossword. Progress and Challenges"Publications Office of the European Union

ECHR Press Unit (2023), Older people and the European Convention on Human Rights. (2023, July). https://www.echr.coe.int Last Consulted October 8th 2024.

ECHR. (2003). Case of Gündüz v. Turkey (Application no. 35071/97). Retrieved December 5, 2024, from https://hudoc.echr.coe.int/eng

ECHR. (2006). Case of Erbakan v. Turkey (Application no. 59405/00). Retrieved December 5, 2024, from https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-76234%22]}

ECRI (2016a) General Policy Recommendation N°15. On combating hate speech. December 2 https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/recommendation-no.15

ECRI (2016b). General Policy Recommendation No. 16 on safeguarding irregularly present migrants from discrimination. Adopted on 16 March 2016. Retrieved December 5, 2024, from https://rm.coe.int/ecri-general-policy-recommendation-no-16-on-safeguarding-irregularly-p/16808b5b0b

ECRI (2019). Annual report on ECRI's activities covering the period from 1 January to 31 December 2019 https://rm.coe.int/ecri-annual-report-2019/16809ca3e1

ECRI (2020). Annual report on ECRI's activities covering the period from 1 January to 31 December 2020 https://rm.coe.int/annual-report-on-ecri-s-activities-for-2020/1680a1cd59

EEAS. (2023). Stratcom's responses to foreign information manipulation and interference (FIMI). Retrieved December 5, 2024, from [source URL https://www.bing.com/ck/a?!&&p=9187e504c1dceedd26889f3c3f3f14c1fb1697f13cf769125d6277bf7c73b0e2JmltdHM9MTc0MzEyMDAwMA&ptn=3&ver=2&hsh=4&fclid=2a42b7dd-233a-6109-28ac-a4fe22246004&psq=Stratcom%e2%80%99s+responses+to+foreign+information+manipulation+and+interference+(FIMI)&u=a1aHR0cHM6Ly93d3cuZWVhcy5IdXJvcGEuZXUvc2l0ZXMvZGVmYXVsdC9maWxlcy9kb2N1bWVudHMvMjAyNC9FRUFTJTIwU3RyYXRjb20lMjBBbm51YWwlMjBSZXBvcnQlMjAyMDIzLnBkZg&ntb=1].

EHRC (2021). Article 10: Freedom of Expression. https://www.equalityhumanrights.com/human-rights/human-rights-act/article-10-freedom-expression#:~:text=Everyone%20has%20the%20right%20to,authority%20and%20regardless%20of%20frontiers.

EuroNews (2020). Coronavirus: France faces 'epidemic' of anti-Asian racism. Coronavirus: France faces 'epidemic' of anti-Asian racism | Euronews

European Commission (n.d.) EUR-lex https://eur-lex.europa.eu/EN/legal-content/glossary/civil-society-organisation.html

European Commission (n.d. a) "Combating anti-Muslim hatred" https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/combating-anti-muslim-hatred_en. Last Consulted, October 7th, 2024.

European Commission (n.d. b), The Digital Services Act. Ensuring a safe and accountable online environment, Retrieved, October 7th, 2024 https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

European Commission (n.d. c), The impact of the Digital Services Act on digital platforms, European Commission, available at https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms

European Commission (n.d. d), European Board for Digital Services, European Commission, available at https://digital-strategy.ec.europa.eu/en/policies/dsa-board

European Commission (n.d. e), DSA: Very large online platforms and search engines, European Commission, available at https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops

European Commission (n.d. f), LGBTIQ Equality Strategy 2020-2025, European Commission, available at https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/lesbian-gay-bi-trans-and-intersex-equality/lgbtiq-equality-strategy-2020-2025_en#lgbtiq-equality-strategy-2020-2025

European Commission (n.d. g), The EU Code of conduct on countering illegal hate speech online, European Commission, available at https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

European Commission (2008) Framework Decision on combating certain forms of expressions of racism and xenophobia https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178

European Commission (2016) European Commission and IT Companies announce Code of Conduct on illegal online hate speech. (2016, May 31). Available at https://ec.europa.eu/commission/presscorner/detail/en/ip_16_1937

European Commission (2018). EU High Level Group on combating racism, xenophobia and other forms of intolerance. Guidance note on the practical application of council framework decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law.

European Commission. (2019). Assessment of the Code of Conduct on countering illegal hate speech online: State of play. Information note to the Permanent Representatives Committee and the Council. Retrieved December 5, 2024, from https://commission.europa.eu/system/files/2020-03/assessment_of_the_code_of_conduct_on_hate_speech_on_line_-_state_of_play__0.pdf

European Commission (2020), Communication from the commission to the European Parliament, the council, the European economic and social committee and the committee of the regions, a Union of Equality: Gender Equality Strategy 2020-2025, available at file:///C:/Users/User/Downloads/gender_equality_strategy_2020_2025_en_77C86437-0983-F10D-E0FF41E71D577EE0_68222.pdf

European Commission (2021a), Study to support the preparation of the European Commission's initiative to extend the list of EU crimes in Article 83 of the Treaty on the Functioning of the EU to hate speech and hate crime. Final Report. Spark Legal Network & Justice And Consumers.

European Commission (2021b), Feedback from: AGE Platform Europe, available at https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12872-Hate-speech-hate-crime-inclusion-on-list-of-EU-crimes/F2232017_en

European Commission (2021c), Extending EU crimes to hate speech and hate crime, European Commission, available at https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/extending-eu-crimes-hate-speech-and-hate-crime_en#:~:text=On%209%20December%202021%2C%20the,down%20in%20Art%2083%20TFEU.

European Commission (2025), The Code of conduct on countering illegal hate speech online +, European Commission, available at https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online

European Parliament and Council. (2012). Directive 2012/29/EU of the European Parliament and of the Council of 25 October 2012 establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA. Official Journal of the European Union, L 315, 57–73. Retrieved December 5, 2024, from https://eur-lex.europa.eu/eli/dir/2012/29/oj

European Parliament (2018) Resolution – on the rise of neo-fascist violence in Europe (2018/2869(RSP)) - Texts adopted - Rise of neo-fascist violence in Europe - Thursday, 25 October 2018 (europa.eu)

European Parliament. (2020). The impact of disinformation on democratic processes and human rights in the world. Retrieved from https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf

European Parliament (2020a), Hate speech and hate crime in the EU and the evaluation of online content regulation approaches, Policy Department for Citizens' Rights and Constitutional Affairs, Directorate-General for Internal Policies, available at https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf

European Parliament (2024), Extending the list of EU crimes to hate speech and hate crime, European Parliament resolution of 18 January 2024 on extending the list of EU crimes to hate speech and hate crime (2023/2068(INI)), European Parliament, Strasbourg, available at https://www.europarl.europa.eu/doceo/document/TA-9-2024-0044_EN.html

European Parliament and Council. (2024). Directive (EU) 2024/1385 of the European Parliament and of the Council of 14 May 2024 on combating violence against women and domestic violence. Retrieved December 5, 2024, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401385#:~:text=The%20purpose%20of%20this%20Directive,domestic%20violence%20throughout%20the%20Union.

European Union (2012), Directive 2012/29/EU of the European Parliament and of the Council of 25 October 2012 establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA, European Union, available at https://eur-lex.europa.eu/eli/dir/2012/29/oj

European Union (2024), Directive (EU) 2024/1385 of the European Parliament and of the Council on combating violence against women and domestic violence, available at https://eur-lex.europa.eu/eli/dir/2024/1385/oj/eng

European Union Agency for Fundamental Rights (2024), LGBTIQ equality at a crossroads - progress and challenges, Agency for Fundamental Rights, available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2024-lgbtiq-equality_en.pdf

Explanatory Memorandum to Proposal for a Council Framework Decision on combating racism and xenophobia (COM/2001/0664) - EUR-Lex - 52001PC0664 - EN - EUR-Lex (europa.eu)

Feldman SM (2013) Review essay: Hate speech and democracy. Criminal Justice Ethics 32(1), 78–90. DOI: 10.1080/0731129X.2013.777254.

García-Prieto, V., Bonilla-del-Río, M., & Figuereo-Benítez, J. C. (2024). Discapacidad, discursos deodio y redes sociales: Video-respuestas a los haters en TikTok. Revista Latina de Comunicación Social, 82, 1–21. https://doi.org/10.4185/rlcs-2024-22581

Global Justice Journal. (2021). "Hate speech and international criminal law" https://globaljustice.queenslaw.ca/news/hate-speech-and-international-criminal-law. Last Check 2024, October 6th

Goodier, M. (2023, November 21). Hate crimes against transgender people hit record high in England and Wales. The Guardian. Retrieved from https://www.theguardian.com

Gorwa, R (2024), 'Explaining Government Intervention in Content Moderation', The Politics of Platform Regulation: How Governments Shape Online Content Moderation, Oxford Studies in Digital Politics (New York, 2024; online edn, Oxford Academic, 23 May 2024), https://doi.org/10.1093/oso/9780197692851.003.0004, accessed 6 Dec. 2024.

Harpviken A N, (2020), Psychological Vulnerabilities and Extremism Among Western Youth: A Literature Review. Adolescent Res Rev 5, 1–26, available at https://doi.org/10.1007/s40894-019-00108-y

Healy, J.. (2019). Thinking outside the box: Title continuation if applicable. Retrieved December 5, 2024, from https://www.britsoccrim.org/wp-content/uploads/2019/12/Thinking-outside-the-box-PBCC19.pdf

Heinze E (2016) Hate Speech and Democratic Citizenship. New York: Oxford University Press.

Higgins A (2018) Three internet trolls convicted of systematic defamation against journalist in Finland. The New York Times, 18 October. [online] Available at: www.nytimes.com/2018/10/19/world/europe/finland-internet-trolls-defamation.html (accessed 31 October 2018).

Human Rights Watch (2020), Hungary: Intensified Attack on LGBT People, Human Rights Watch, available at https://www.hrw.org/news/2020/11/18/hungary-intensified-attack-lgbt-people?utm_source=chatgpt.com

Jääskeläinen, T. (2020). Countering hate speech through arts and arts education: Addressing intersections and policy implications. Policy futures in education.

Judit Szakács and Éva Bognár (2021), 'The impact of disinformation campaigns about migrants and minority groups in the EU', European Parliament, Policy Department for External Relations, Directorate General for External Policies of the Union – PE 653.641 – June 2021, available at: https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-

Koltay A (2016) Book review: Hate speech and democratic citizenship. Journal of Media Law 8(2), 302–306. DOI: 10.1080/17577632.2016.1209318.

Le Figaro (2020). Ce que révèle le Rap antisémite de Freeze Corleone. https://www.lefigaro.fr/vox/culture/ce-que-revele-le-rap-antisemite-de-freeze-corleone-20200918

LIBE Committee (2020). Hate speech and hate crime in the EU and the evaluation of online content regulation approaches. Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies PE 655.135 - July 2020

Linh-Lan Dao, @linhlandao (January,2020) "Oui oui. Des gens se font insulter et expulser des transport parcequ'ils sont asiatiques. Y a pas que les blagues/ de la haine sur les réseaux sociaux. La vrai discrimination se passe aussi dans la vrai vie" Twitter

Martin, D., Cunningham, S. J., Hutchison, J., Slessor, G., & Smith, K. (2017). How societal stereotypes might form and evolve via cumulative cultural evolution. Social and Personality Psychology Compass, 11(9). https://doi.org/10.1111/spc3.12338

McQueen, A., & Kreuter, M. W. (2010). Women's cognitive and affective reactions to breast cancer survivor stories: A structural equation analysis. Patient Education and Counseling, 81, S15–S21. doi: 10.1016/j.pec.2010.08.015.

Mendel T (2012) Does international law provide for consistent rules on hate speech? In: Herz ME, Molnar P (eds) The Content and Context of Hate Speech: Rethinking Regulation and Responses. New York: Cambridge University Press, pp.417–429.

Mic (2016,). How conservative trolls turned the rare Pepe meme into a virulent racist. https://www.mic.com/articles/143778/is-pepe-meme-racist?utm_source=policymicTWTR&utm_medium=main&utm_campaign=socia

Molnar P (2012) Responding to "hate speech" with art, education, and the imminent danger test. In: Herz ME and Molnar P (eds) The Content and Context of Hate

National Counterterrorism Center (n.d.), Understanding and Mitigating Youth Vulnerabilities to Extremist Messaging: a guide for community authority figures and Bystanders, United States Government, available at https://www.safeguardrisksolutions.com/wp-content/uploads/2022/08/Youth-vulnerabilties-to-extremist-messaging.pdf?utm

National Institute on Aging (2023) "Elder abuse". July 21st https://www.nia.nih.gov/health/elder-abuse/elder-abuse Last Consulted, October 7th 2024

Niederdeppe, J., Roh, S., & Shapiro, M. A. (2015). Acknowledging individual responsi- bility while emphasizing social determinants in narratives to promote obesity-reducing public policy: A randomized experiment. PLOS One, 10, e0117565. doi: 10.1371/jour- nal.pone.0117565.

OCHR. (n.d.). Guidance note on intersectionality. Retrieved December 5, 2024, from https://www.ohchr.org/sites/default/files/documents/issues/minorities/30th-anniversary/2022-09-22/GuidanceNoteonIntersectionality.pdf

OHCHR (2010) Towards an interpretation of Article 20 of the ICCPR: Thresholds for the Prohibition of incitement to hatred Work in progress. (2010). In Office of the High Comissioner for Human Rights. Office of the High Comissioner for Human Right. Retrieved October 3, 2024, from https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/ICCPR/Vienna/CRP7Callamard.pdf

O'Flaherty, M. (2012). Freedom of Expression: Article 19 of the International Covenant on Civil and Political Rights and the Human Rights Committee's General Comment No 34. Human Rights Law Review, 12(4), 627–654. https://doi.org/10.1093/hrlr/ngs030

Office of Justice programmes, US Department of Justice, n.d. The Role of the Internet and Social Media on Radicalization, https://www.ojp.gov/ncjrs/virtual-library/abstracts/role-internet-and-social-media-radicalization-what-research

Pandea A, Grzemny D, and Keen E, (2019), GENDER MATTERS A manual on addressing gender-based violence affecting young people, Council of Europe, available at https://www.coe.int/en/web/gender-matters

Pavlovic, Z. (2022). Hate speech, violence and disseminating of false news on the global social networking highway. Institut Za Kriminološka I Sociološka Istraživanja, 155.

Piazza, J. (2020). Politician hate speech and domestic terrorism. International Interactions, Empirical and Theoretical Research in International Relations. 46 (3), pp431-453. https://doi.org/10.1080/03050629.2020.1739033

Peršak, N. (2022). Criminalizing hate crime and hate speech at EU level: extending the list of eurocrimes under Article 83(1) TFEU.Criminal Law Forum, 33(2), 85–119. https://doi.org/10.1007/s10609-022-09440-w

Quinn, G., & Doron, I. (2021a). Against Ageism and Towards Active Social Citizenship for Older Persons. The Current Use and Future Potential of the European Social Charter. Retrieved from https://rm.coe.int/against-ageism-and-towards-active-social-citizenship-for-older-persons/1680a3f5da

Radio Free Europe Radio Liberty (2021a), Homophobic Hate Speech On The Rise In Europe, Says New Report, Radio Free Europe Radio Liberty, available at https://www.rferl.org/a/homophobic-hate-speech-rise-europe-report/31106026.html?utm_source=chatgpt.com

Radio Free Europe Radio Liberty (2021b), The Worrying Regression Of LGBT Rights In Eastern Europe, Radio Free Europe Radio Liberty, available at https://www.rferl.org/a/lgbt-rights-eastern-europe-backsliding/31622890.html?utm_source=chatgpt.com

RadioGenoa, @RadioGenoa. (October, 2024). Uighur Muslims in China are taken to special mental health camps for treatment and mosques are turned into nightclubs with alcohol and music." Twitter. https://x.com/RadioGenoa/status/1842535977828434294

Reagle J (2015) Counter speech. In: Joseph Reagle. [online] Available at: reagle.org/joseph/2015/07/counter-speech.html (accessed 31 October 2018).

Ribeiro M H, et al. (2019), Auditing Radicalization Pathways on YouTube, Cornell University, available at https://arxiv.org/abs/1908.08313?utm_source=chatgpt.com

Sherry, M. (2016). Disability hate crime: Does anyone really hate disabled people? Routledge. https://doi.org/10.4324/9781315577371

Silva, C. (2023). Fighting Against Hate Speech: A Case for Harnessing Interactive Digital Counter-Narratives. In: Holloway-Attaway, L., Murray, J.T. (eds) Interactive Storytelling. ICIDS 2023. Lecture Notes in Computer Science, vol 14383. Springer, Cham. https://doi.org/10.1007/978-3-031-47655-6_10

Sjöholm, M. (2024). Regulation of online gender-based hate speech and international human rights law: Current status and challenges. QIL Zoom in.

The Conversation (2020). When politicians use hate speech, political violence increases. Piazza https://theconversation.com/when-politicians-use-hate-speech-political-violence-increases-146640

Tsesis A (1999) The empirical shortcomings of first amendment jurisprudence: A historical perspective on the power of hate speech. Santa Clara Law Review 40, 728– 786. [online] Available at: digitalcommons.law.scu.edu/lawreview/vol40/iss3/3

UNESCO (2020a). School-related gender-based violence (SRGBV) - A human rights violation and a threat to inclusive and equitable quality education for all. Constanza Ginestra. https://unesdoc.unesco.org/in/documentViewer.xhtml?v=2.1.196&id=p::usmarcdef_0000374509&file=/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_88edf536-9d80-47e8-8a0a-c83ac392f045%3F_%3D374509eng.pdf&updateUrl=updateUrl3488&ark=/ark:/48223/pf0000374509/PDF/374509eng.pdf.multi&fullScreen=true&locale=en#%5B%7B%22num%22%3A187%2C%22gen%22%3A0%7D%2C%7B%22name%22%3A%22XYZ%22%7D%2C69%2C576%2C0%5D

UNESCO (2020b) Online violence Against Women Journalists: A Global Snapshot of Incidence and Impacts, available at https://unesdoc.unesco.org/ark:/48223/pf0000375136

UNESCO (2022). How to address online #HateSpeech with a human rights-based approach? https://www.youtube.com/watch?v=JirA4suOdXI

UNESCO (2023) Addressing hate speech through education. A guide for policy-makers. https://unesdoc.unesco.org/in/documentViewer.xhtml?v=2.1.196&id=p::usmarcdef_0000384872&file=/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_26004d6b-ae0f-48cc-b394-06109eaaef51%3F_%3D384872eng.pdf&updateUrl=updateUrl8567&ark=/ark:/48223/pf0000384872/PDF/384872eng.pdf.multi&fullScreen=true&locale=en#1200_22_hate_speech.indd%3A.86035%3A635

United Nations (2023) Countering and Addressing Online Hate Speech: A Guide for policy makers and practitioners. (2023, July). United Nations. https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf

United Nations. (n.d. a). What is hate speech?| United Nations. Retrieved from https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech#:~:text=In%20common%20language%2C%20%E2%80%9Chate%20speech,that%20may%20threaten%20social%20peace. Last concuslted, 2024 Octobe

United Nations (n.d. b). Targets of Hate. Hate Speech. United Nations https://www.un.org/en/hate-speech/impact-and-prevention/targets-of-hate

United Nations (n.d. c), Special Rapporteur on minority issues, available at https://www.ohchr.org/en/special-procedures/sr-minority-issues

UNGA (2018). Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective. Retrieved December 5, 2024, from https://documents.un.org/doc/undoc/gen/g18/184/58/pdf/g1818458.pdf

UNHCR (n.d) Convention on the Elimination of All Forms of Discrimination against Women New York, 18 December 1979. (n.d.). Retrieved October 8, 2014, from https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-elimination-all-forms-discrimination-against-women

Van Assche et al....The Social-Psychological Bases of Far-Right Support in Europe and the United States. Journal of Community & Applied Social Psychology. (2019)

Waldron J (2012) The Harm in Hate Speech. Cambridge, Massachusetts: Harvard University Press.

WHO (2021), Ageing and ageism, available at https://www.who.int/news-room/questions-and-answers/item/ageing-ageism, retrieved December 5, 2024.

Wood, S. C. (2024). Disablist hate crime: A scoping review of current research and understandings. Retrieved December 5, 2024, from https://www.diva-portal.org/smash/get/diva2:1871916/FULLTEXT01.pdf

Yon et al., (2017) - elder abuse prevalence in community settings: a systematic review and meta-analysis - https://www.thelancet.com/pdfs/journals/langlo/PIIS2214-109X%2817%2930006-2.pdf

SECTION 2: The Impact of Counter and Alternative Narratives to Combat Hate Speech

Executive Summary

The first chapter of this section overviews the persuasive nature of narratives, and explains the aim, origin and definitions of CANs (Counter and Alternative Narratives). A narrative is a synonym of a story with a plot and characters that often indulges in ideas and values of how society is deemed to function. Adding multiple narratives together makes one general narrative. This general narrative becomes dangerous when it transmits hateful or discriminatory ideals against certain minorities, and undermines democratic and human rights values. Narratives can be particularly dangerous because of their persuasive effects, which are explained in detail in Chapter 2 of this section. Yet, when narratives are employed in a CAN format, they can also be particularly effective in countering the hateful and discriminatory ideals.

Chapter 1 explores the origin of CANs which can be traced back to public health campaigns. These campaigns aimed to destigmatise certain illnesses. In a political context it has been used for countering violent extremisms, for example 'Anti-Salafi-Jihadist Terrorism'. After explaining the origins, Chapter 1 delves into the definitions of counter-narratives and alternative narratives. The former refers to narratives which are designed to directly oppose a dominant or mainstream narrative. They do this by focusing on identifying and dismantling harmful, misleading, or oppressive narratives, exposing their flaws, biases, or inaccuracies. Whereas the latter promote positive, inclusive, and constructive ideas, aiming to engage the entire population, including those who may be producing the hate speech. Both are complementary. Counter-narratives target those who sympathise or hold hateful views in order to change their existing opinions. Alternative narratives constitute strategic communication efforts that aim to undercut hateful narratives by focusing on "what we are for" rather than "what we are against".

The second chapter delves into the potential for using CANs to combat hate speech. According to numerous reports and publications, CANs have been deemed effective in combating hate speech. This is because they are narratives countering other narratives, which are more effective than factual arguments (non-narratives). When faced with opposite opinions, individuals tend to react negatively towards non-narratives, regardless of the solidity of the argumentation. This can be explained by something called reactance theory – which is discussed extensively. Additionally, the chapter also explains the potential of narratives in using persuasion communication theories to undermine the occurrence of negative reactions. Other related theories that will be explored in the chapter include transportation theory, identification theory, parasocial interaction theory, and processing fluency theory. Additional theories, including perceived realism, utopia and grand narratives of significance theories, are also presented. The chapter overviews other methods and variables that need to be considered when developing a CAN. These include entertainment, education, the medium used to transmit the story, the characters, and the interactivity of a story. Finally, the chapter highlights how the use of emotions in narratives can also have a persuasive appeal on the target audience, as these emotions – empathy, anticipated regret, fear and humour – reduce possible negative reactions from the target audience by increasing the narrative's capacity and persuasiveness.

The third chapter overviews the risks which can take place when conducting CANs. For example, if the research part of the CAN is not conducted appropriately in terms of the situation and people involved, it can make the CAN backfire or be ineffective. Another risk is to conduct a poor-quality evaluation which can lead to a misunderstanding about the impact of the CAN. This chapter explores the challenges that come with CAN campaigns, especially those conducted online. This topic area includes looking into issues such as having a lack of reach towards non-social media users or experiencing a loss of control of the campaign at the hands of spreaders of hateful messages. Finally, it explains the risks behind campaigns which have a limited amount of financial and human resources.

The fourth chapter overviews a case study which explores the use of empathy in a CAN campaign against hate speech.

Chapter 5 provides guidance on the creation of effective CAN campaigns for combating hate speech. This guidance is based on the information and topics covered in Chapters 3 and 4. It promotes the conducting of appropriate audience targeting by making the target audience as narrow as possible and distinguishing between age groups, ethnicities, location and levels of radicalisation. It also advises on the use of an adequate messenger to whom the extremist audience can relate to. Moreover, it advises conducting proper evaluation, which measures both reach and behavioural impact, through surveys, focus groups, sentiment analysis, and individual interviews. Lastly, it advises CSOs to conduct CAN campaigns through partnerships and collaboration, to improve the impact of their respective campaigns.

Methodology of Section 2

For this section, a desk research was conducted that involved a thorough review of academic literature across several fields. These include communication science – particularly with a focus on persuasive communication and marketing as well as CANs, public health communication papers assessing the effectiveness of CANs, anti-terrorism and counter-extremism research within political science, and security studies centred on CANs aimed at countering jihadist extremism. Additionally, papers discussing the use of CANs to combat hate speech specifically were reviewed. The research also included institutional studies by the European Union and the Council of Europe, as well as other security-focused institutes and think tanks. A total of 190 sources were used. Papers focusing on public health communication, counterterrorism, extremism, and persuasive communication science were included because of the lack of research regarding CANs in connection to hate speech specifically.

The literature reviewed came from a wide range of fields, the period of publishing was also wide ranging. Communication and public health papers typically ranged from the 1970s to the 2020s, while anti-terrorism and institutional research mostly dated from the 2010s to the 2020s. Studies on anti-hate speech CANs were the most recent, with the oldest one being from 2017. This research has integrated insights from diverse disciplines and time periods. The synthesised findings build a comprehensive understanding of effective approaches to combating hate speech.

However, this study has limitations due to restrictions within the existing literature. There is a notable lack of research explicitly addressing how communication theories might enhance CANs against hate speech. Consequently, the persuasive theories referenced, drawn primarily from marketing, public health, and anti-jihadist terrorism research, may not translate directly to countering hate speech, but apply more directly to violent extremism and radicalisation. Despite these findings being considerably relatable or generalisable to those that would apply for hate speech, the lack of literature specifically on hate speech underscores the need for further research specifically investigating the applicability of these theories to address hate speech through CANs.

1. The Aim, Origin and Definition of Counter and Alternative Narratives

1.1 The persuasive nature of narratives

1.1.1 What are narratives?

A narrative is a formal way to refer to a "story", something that provides an account of events or experiences, whether true or fictional, or a combination of both. A non-narrative provides information in statistical or didactic formats or uses evidence and logical reasoning instead of stories or examples (Bullock et al 2021). Narratives involve characters and a plot and often indulge prevailing ideas and values of how society is deemed to function, as well as what ought to be normal or not (Council of Europe 2017).

Although stories can have different plots and characters, the underlying meaning or values they connect to are often similar. For instance, there are multiple love stories with different plots and characters. Yet traditionally, most love stories depict mainly heterosexual couples, reflecting a false reality that only heterosexual couples existing, or only heterosexual couples are acceptable. Stories which transmit similar values and understandings of society make a big general narrative. This is presented as the only viable narrative and denies other narratives, even on occasions inciting violence against other narratives (Council of Europe 2017).

General narratives can undermine the pluralistic and diverse society necessary for a democracy. They become increasingly problematic when they lead to or are composed of multiple hateful narratives towards minorities. They can be dehumanising, target a particular group, and promote prejudice and intolerance, by having a message that portrays these minorities in a negative perspective. To do this they use stereotypes, labels or particular terms, inaccurate facts, and undermine their human rights (Council of Europe 2023).

General narratives are not a story written by one person or a group of persons, they appear over long periods of time as a result of conversations, debates and exchanges of ideas, both offline and online. They are shared by different persons and groups of persons who have similar political wills or beliefs and aim to influence public opinion in a particular direction. General narratives are not always transmitted as a story, they can also be transmitted through misinformation, debates, as facts that are exaggerated or taken out of context or using a specific example to scapegoat an entire community. They are often used by extremist, radical or anti-establishment groups. For example, narratives have been used by the Salafi-Jihadists movement to radicalise people to incite them to commit terrorist attacks. This is achieved by depicting western culture in a negative way, characterising western US and NATO forces as oppressing Islam and Muslims, therefore highlighting the necessity to fight against them (Rusi 2011).

1.1.2 Why are narratives effective?

Narratives are particularly effective as they can increase persuasion. Persuasion communication science demonstrates that narratives have an advantage over non-narrative forms of communication. The narrative's information is transferred in a manner which renders its content and meaning more structured and imaginable, as opposed to non-narrative forms (Glaser et al 2009 and Bullock et al 2021). Narratives are dangerous because of the emotions they can cause, and they can reduce reactance and rationality from

individuals. This allows them to convene dangerous, discriminatory and anti-democratic ideas, regardless of their lack of rationality and logic. Although studies that compare narratives to non-narratives remain uncommon, meta-analytic research has demonstrated that when both types of persuasion are compared, narratives have both more effective short-term and long-term persuasive effects (Shen et al., 2015; Oschatz and Marker, 2020; Bullock et al 2021). These persuasion theories - Transportation Theory, Identification Theory, and Parasocial Theory - apply to any form of narrative and CANs. They will be detailed in Chapter 2 of this section which defines the potential of persuasion theories to create effective CAN campaigns.

1.2 What are CANs?

1.2.1 The origin – CANs as method to undermine narratives

CANs aim to combat the dangerous narratives which produce radicalisation, stigmatisation or violence against particular groups of people. However, CANs are not unique to political narratives. They were originally used to destigmatise people suffering from medical conditions which were negatively perceived by the public, such as AIDS or autism (Slater & Rouner 2002). They have also been heavily used in a political context during times of multiple jihad terrorist attacks in Europe as a way to de-radicalise Islamists (RUSI, 2011; Lombardi et al 2020; Van Eerten et al, 2017; RAN, 2015). Most literature analysing the effectiveness of CANs is based on public health literature, communication science literature, or Anti-Terrorist and security literature. Only recently have several institutions, CSOs and governments started using CANs to combat the rise of far-right extremism, hate speech and hate crime (Council of Europe 2017).

CANs are either counter narratives or alternative narratives although some literature additionally distinguishes between government and institutional narrative categories (Van Eerten et al 2017; RAN 2025). However, this study focuses solely on counter-narratives and alternative-narratives which will be defined below.

1.2.2 Defining counter-narratives

Ferguson (2016) and Van Eerten (2017) underlines that counter-narratives are used in multiple academic and policy circles and as such have multiple definitions across different disciplines. Yet, there tends to be a consensus that counter-narratives should contradict the underlying logic of dominant general narratives (Carthy et al 2020).

Grossman (2015) defines counter-narratives as a “variety positional or relational discourse, at once overtly constructed and implicitly normative, that seeks to disrupt, dismantle, or speak back to other narrative trajectories that exert discursive power”. In other words, he defines counter-narratives as a discourse that aims at undermining the general narratives that shape society and societies values, identities, and worldview. Briggs and Fever (2013) state that counter narratives “directly or indirectly challenge extremist narratives either through ideology, logic, fact or humour” and target groups of people who are either already radicalised or are on the path to becoming radicalised. Hussain & Saltman (2014) argue that counter-narratives should actively dismantle the messages promoted by individuals or groups advocating violent extremist ideologies (see also Carthy et al, 2020).

Building on persuasion and communication theories, Carthy et al (2020) and Braddock and Horgan (2015) conceptualise counter-narratives as "narratives composed of content that challenges the themes intrinsic to other narratives" (p386). These authors additionally highlight that the persuasive function of counter-narratives should be tailored towards targeting individuals vulnerable to radicalisation. McDowell-Smith, Speckhard, and Yayla (2017) agree, and add that counter-narratives should aim to persuade by enhancing and surpassing the narrative through the emotional appeal of their content.

The Council of Europe 2017 manual We CAN! describes counter-narratives as concise responses to hateful content, aimed at deconstructing and discrediting hateful messages. They are used to block or challenge hate, particularly in two scenarios: 1) reacting swiftly to hate targeting minorities after specific events, and 2) challenging entrenched hateful narratives linked to societal prejudice or political dominance. The goal is to reclaim public spaces through actions like creating online memes or offline "love speech" - such as subverting hate graffiti. These tools empower activists, educators, and youth workers to effectively counter hate (Council of Europe 2017: 80).

An example of a counter narrative is a campaign in Germany #IchBinHier (translated as #IAmHere) from 2017. A few hours after the truck terrorist attack in Stockholm of that year, the Facebook page of a prominent German TV news outlet began filling with comments condemning refugees. In reaction to that, the private Facebook group called #IchBinHier coordinated an effort to counteract the online trolls, leading to other users appearing on the page to also challenge these negative responses. They questioned the assumption that the attacker was a refugee and whether religious motivations were behind it, and urged patience until all facts were known (Chase 2017).

1.2.3 Defining alternative narratives

Alternative narratives are often part of the counter-narrative definition for many experts. For instance, Van Eerten et al (2017) considers that counter-narratives are composed of three different approaches 1) counter-messaging (i.e. activities that challenge extremist narratives head on); 2) alternative messaging (i.e. activities that aim to provide a positive alternative to extremist narratives); and 3) strategic communication by governments (i.e. activities that provide insight into what the government is doing). The manual We CAN! also considers counter-narratives and alternative-narratives to be part of the counter-narrative definition and hence refers to them as Counter and Alternative Narratives (CANs).

Alternative narratives constitute strategic communication efforts that aim to undercut hateful narratives by focusing on “what we are for” rather than “what we are against” (Schmid, 2014; Van Eerten et al 2017). They promote positive, inclusive, and constructive ideas, aiming to engage the entire population, including those who may currently produce hate speech but could be exposed to new perspectives.

The primary goal of alternative narratives according to Briggs and Feve (2013) is “to influence those who might be sympathetic towards (but not actively supportive of) extremist causes, or help to unite the silent majority against extremism by emphasising solidarity, common causes and shared values”.

Unlike counter narratives, which directly respond to and challenge (often isolated) harmful narratives, alternative narratives emphasise presenting positive (different) alternatives. The aim of which is to foster a broader shift in mindset and support lasting societal change, without reinforcing or validating the harmful

narratives they oppose. Alternative narratives therefore offer a constructive vision that moves beyond merely opposing negative content (Council of Europe 2017).

An example of an alternative narrative campaign is the "Dare to be Grey" campaign, which encourages people to embrace the complexity of opinions and identities that lie in the grey middle ground. It seeks to promote understanding, tolerance, and dialogue rather than confrontation or exclusion. By encouraging open discussion and challenging binary thinking, the campaign indirectly undermines the divisive rhetoric of hate groups (Dare to be Grey n.d).

1.2.4 The complementarity of counter narratives and alternative narratives

Counter and alternative narratives are complementary, and both are needed to combat hateful narratives. Those who intend to spark changes in narratives will need to go from direct (short-term) reactions (counter narratives), to developing alternative stories to support those reactions and create change (Council of Europe 2017).

	Counter Narrative	Alternative Narrative
How?	Directly confronting an oppressive narrative	Aiming at creating an alternative vision of society
What?	Undermine authority and myths that oppression relies on	Offer a "what we are for" as a different perspective to look at the issue from
Where and when?	Small scale, shorter period of time	Wide project, long-term
For example?	• Debunking of discriminatory myths about a certain group in society through a public information campaign. • Former haters testimonies about the negative impacts of extremist movements on their lives. • Painting a mural celebrating diversity over racist comments on walls.	• All Different – All Equal campaign, a campaign promoting human rights • Reports on inter-faith dialogue youth meetings • Documentaries about the lives of refugees depincting them as human beings and not as criminals • Series of posters showing how fathers can also enjoy paternity leave and take care of children (a role often taken by mothers).

Figure 11: The complementarity of Counter and Alternative Narratives. (Source: Council of Europe 2017: 82)

This study uses the Council of Europe's CAN definition from the Recommendation CM/Rec(2022)16, Combating Hate Speech: "narratives that are designed to combat hate speech by discrediting, deconstructing and condemning the narratives on which hate speech is based by reinforcing the values that hate speech threatens, such as human rights and democracy. Counter and alternative narratives to hate speech also promote openness, respect for difference, freedom, and equality. [...] The use of counter and alternative speech forms is particularly important for addressing hate speech that does not reach the severity level for being addressed via criminal, civil or administrative procedures." (Council of Europe 2022: 42-43).

2. The Potential of CANs to Combat Hate Speech

The potential of CANs to combat hate speech can be explained through a number of theories and methodologies, such as Social Judgement Theory, narrative engagement methods, and persuasion communication theories, etc. The majority of sources on the use of CANs are linked to combating (hateful) radical violent extremism, and whilst these findings can be transferable to combating hate speech, they are not fully compatible. Additionally, it must also be noted that if used by malicious actors, the potential of CANs can also be used to create more hateful ways of reasoning, transforming CANs from being constructive to being a risk.

2.1 Social Judgement Theory

Social Judgment Theory suggests that an individuals' receptiveness to persuasive messages largely depends on how these messages align with their pre-existing attitudes (Sherif & Hovland 1961; Sherif, Sherif, & Nebergall 1965; Van Eerten et al 2017). Rather than assessing the content solely based on its merits, individuals gauge the messages position relative to their own stance and then determine their willingness to accept it (O'Keefe 2015). This evaluation process incorporates three key judgmental "latitudes": the latitude of acceptance, the latitude of rejection, and the latitude of non-commitment.

The latitude of acceptance encompasses positions individuals find agreeable, while the latitude of rejection includes viewpoints they oppose, and the latitude of non-commitment consists of positions they are indifferent toward. When messages align with one's latitude of acceptance or non-commitment, persuasion is likely, potentially fostering an attitude shift. Conversely, messages that fall within the latitude of rejection may be outright dismissed or, through a "boomerang effect," may result in reinforcing opposition to the position (Perloff 2010).

The concept of the "contrast effect" and "assimilation effect" further explain Social Justice Theory's influence process. Messages significantly different from one's stance (i.e. within the latitude of rejection) may be exaggeratedly perceived as further removed (contrast effect), whereas messages somewhat aligned may seem closer than they actually are (assimilation effect). This evaluative process becomes more polarised in individuals with high "ego-involvement," or when the issue is central to their identity, often limiting the latitude of acceptance while broadening the latitude of rejection (O'Keefe 2015; Perloff 2010; Van Eerten et al. 2017).

Thus, highly invested individuals are especially resistant to persuasion and often engage in black-or-white thinking, particularly within radicalised contexts where extremist ideologies become absolute. For CAN efforts, this suggests that advocating for moderately discrepant but non-rejectable positions is more effective than direct confrontations, as suggested by the Social Justice Theory's core principles (de Wolf & Doosje 2010; Van Eerten et al 2017).

2.2 Should (non-narrative) argumentation and logic be irrefutably used?

Creating CAN content which challenges narratives and arguments stemming from radical sources has been deemed as a viable approach to counter extremism. Van Eerten et al (2017) add that (non-narrative) argumentation and logic can be useful for those who are at the initial phase of radicalisation. This was

observed by Schmidt (2015: 3) who conducted a study on ISIS radicalisation and declared that "while the fanatical extremists of ISIS might no longer be open to rational, persuasive arguments, many of those not yet fully radicalised might still have open minds. They can be confronted with facts and rational reasoning and might then be able to see ISIS for what it is…".

"Message sidedness" refers to the extent to which a message takes into consideration differing perspectives. There are one sided messages and two-sided messages. The former refers to messages that do not consider the opposite arguments or perspective and only offer the perspective of the messenger. Whereas the latter considers arguments from both sides. Meta-analysis was done on studies comparing one-sided and two-sided messages, conclusions showed that two-sided messages tend to be more persuasive than one-sided ones, particularly when they counter opposing viewpoints (Allen 1998; O'Keefe 1999). This persuasive advantage is attributed to an increase in the source's credibility and the presentation of strong arguments that challenge opposing perspectives (Perloff 2010). Nevertheless, much depends on the target audience, if the audience has no prior knowledge regarding a particular topic, providing a one-sided message has been shown to be more effective (Keller and Lehmann 2008).

Brock (1967) noticed that underlining inconsistencies of arguments in a (hate) narrative was successful in reducing the shift of beliefs held by recipients in certain cases. The Inoculation Theory explains this cognitive reaction as a type of vaccine. Explaining the weak arguments for a cause that the targeted audience may support could prevent them from starting to believe stronger hateful arguments. In a meta-analysis of 54 cases, Inoculation Theory was found to be an effective form of creating resistance to persuasive messages (McGuire 1961; Banas & Rains, 2010; Carthy et al 2020).

2.3 Narrative engagement methods and persuasion theories

2.3.1 Persuasion communication theories

In the context of radicalisation and violent extremism, Schlegel (2021) argued that there is no current evidence which proves that current or past CAN campaigns have been effective at reducing the violent threats of radicalisation. One possible reason for this is that "counter-propaganda videos are sarcastic, emotionless and do not focus on creating a compelling narrative" (Gerstel 2016: 7). Regardless, communication literature claims that narratives which are emotionally compelling are more effective. Although (non-narrative) arguments have proven to be useful, academics have often found that narratives tend to be more persuasive. This can be explained by different theories which demonstrate that humans, when faced with ideas opposite to their beliefs, tend to reject them no matter what the facts or logic are. Narrative messages sidestep potential reactance by lowering the audience's motivation or capacity to critically assess the message (Moyer-Gusé 2008; Sun & Radcliff, 2020).

Compelling narratives might reduce resistance to beliefs or thoughts. To persuade individuals more effectively, it is preferable to make them leave their thoughtful and focused state, which can be realised most effectively by avoiding logical arguments and facts (non-narratives). Narratives have the capacity to disable the condition of systematic processing of information, permitting counter arguments to develop. As opposed to factual arguments (non-narratives), emotions, evoked by narratives, reduce forms of resistance. CANs can be effective due to the power of persuasion they have through their simplicity and emotional appeal, as opposed to facts and numbers (non-narratives). Responding to simple messages through emotions, art and

other forms of simple messaging has a greater chance of working (Moyer-Gusé, 2008; RAN, 2015; Ratcliff and Sun 2020; Jääskeläinen, 2019).

It is crucial to avoid forms of reactance when spreading CANs or providing arguments. CANs must follow narrative engagement persuasion communication methods to reduce reactance's such as Transportation Theory, Character Identification Theory, and Parasocial Interaction Theory (these are explained in detail below). To this day, not many CAN campaigns countering hate speech use communication persuasive strategies, rather they use more education based strategies. Yet, despite CANs combating hate speech not having persuasive communication tools as their primary objective, they often include an emotional element as a persuasive factor, thus engaging with such tools.

An interesting case study to showcase this is presented in Chapter 4 of this section. Communication persuasion strategies have been used more systematically and explicitly in CAN campaigns aiming at countering (ISIS) radicalism. Hence, more studies should be conducted on the effectiveness of communication theories on CANs combating hate speech specifically.

Transportation Theory

Transportation Theory describes the process in which narratives influence people into changing their attention from the real world to the world constructed in the narrative (Kero and Schmitt 2023). It is defined as "an engrossing temporary experience" that shifts the audiences focus away from the critical evaluation of the message (van Laer et al., 2014, p.800; Green & Brock 2000; Moyer-Gusé & Nabi 2010; Slater & Rouner 2002; Ratcliff & Sun 2020). Transportation requires the construction of a story world which involves characters and a plot, yet is not strictly reserved to fictional narratives. It can also occur in non-fictional narratives such as advertisements or even Instagram selfies (Lien & Chen 2013; Farace et al 2017; Schlegel 2021).

Psychological immersion enables individuals to temporarily set aside their real-world beliefs and knowledge, reducing both their motivation and capacity to critically assess the message, and increasing their susceptibility to persuasion (Ratcliff and Sun 2020). 'Transportation' is seen as a key mechanism for narrative persuasion to occur. High levels of transportation are positively correlated with persuasion, as participants experience rich imagery of the story events and take over the perspective of the protagonists (Green 2004; Tukachinsky & Tokunaga 2013; Schlegel 2021).

A mechanism of transportation is the usage of emotional shifts and intense emotions. Nabi and Green (2015) observe that emotional shifts throughout a narrative, effect persuasion before, during, and after story exposure. They describe emotional shifts as shifts experienced by the audience when being subject to a narrative, "from negative to positive (i.e. fear to relief), from positive to negative (i.e. happiness to sadness), and even from one negative or positive emotional state to another of a similar valence (i.e. fear to anger or happiness to pride)"

In fictional narratives, it is necessary to build a believable world that the characters inhabit. In order to develop a convincing narrative, it is necessary to invest thought into world building and construct a coherent story world. This may entail depicting the characters' environments, location or relationships. The audience needs to grasp and picture the characters' surroundings and the unfolding events. The characters' personalities, attitudes, identities and relationships amongst each other must fit the story world logic of their surroundings (Turn 2019; Schneider 2001; Schlegel 2021).

Identification Theory

Identification Theory, another type of persuasion communication theory, refers to the character's viewpoint being transferred to the reader or viewer. This is different to Transportation Theory where the reader changes their attention to the narrative's world, but does not necessarily feel like the main character in that world (Bullock et al 2021; Winkler et al 2022).

Identification Theory plays a crucial role in fostering emotional engagement with a narrative which can enhance its persuasive impact and generate both cognitive and behavioural responses. The greater the audience's identification with a character, the stronger the emotional resonance of the story, leading to increased persuasive effectiveness. Factors like psychological proximity, character likability, and emotional relatability enhance this identification, which increases narrative involvement and persuasive impact (De Graaf et al. 2009; Igartua 2010; Igartua & Betsch et al. 2011; Barrios 2012; Keer et al. 2013; Murphy et al., 2011; Hoeken & Sinkeldam, 2014; Schlegel 2021).

This identification applies in narratives where the media figure is perceived as being similar to the audience (Kero and Schmitt 2023). In audiovisual formats, this could be seen through a 'Point of View' shot, where the camera is positioned to represent what a specific character is seeing, giving the audience the feeling that they are looking through the characters' own eyes (Schlegel 2021).

This mechanism is typical of conservative or far-right content influencers who share their everyday life on social media through personal stories which can allow viewers to feel identification with them. An illustration of this is the rise of the trad-wife movement on social media, in which women record themselves being housewives and explain how becoming a housewife has increased their quality of life. The narrative in this particular example is that women don't have to be stressed because of their jobs and women are not meant to hold these types of stress. Women who watch this content are often stressed because of their studies or work and are therefore able to identify with the narrative and be influenced by it. The increased trust provided by the identification narratives increases the credibility of the story. This leads to a dismissal from the audience of the negative sides that come with being a housewife (i.e. no financial independence) (The Guardian 2024; The New Yorker 2024).

It should be noted that characters do not have to be similar to the audience physically but rather must promote identification through emotional or social aspects. For instance, non-human characters in fictional narratives, such as robots or aliens, can also foster identification by addressing sensitive topics. This has some advantages, by offering a psychologically safe distance and reducing potential resistance as the audience does not feel like the story is directed to them. By minimising perceived social differences, fictional characters may encourage viewers to relate to diverse experiences without triggering group biases, allowing audiences to reconsider ideological positions (Slater et al. 2006; Schlegel 2021).

Parasocial Interaction Theory

Parasocial Interaction Theory is another type of persuasion communication theory. It explores "the psychological process in which viewers/readers like and/or trust a media figure so much that they feel as if they are connected to them" (Bradock 2020; Kero and Schmitt 2023: 5). It can be described as a "quasi-social, one-direction interaction" in which individuals develop a one-sided connection with a character within a narrative (Tukachinsky & Tokunaga 2013; Ratcliff and Sun 2020). This type of engagement unfolds as

audiences consume media content where the character or narrator directly addresses them, as seen with TV hosts, talk show personalities, and influencers.

This is different to Identification Theory because the reader does not need to feel like the main character, but they do need to feel very connected to them. It is also different to the Transportation Theory because in that theory the reader can feel anger or fear towards the main character rather than connecting with them. When this Parasocial process takes place over a longer period through multiple "interactions" between the audience and the media figure, it becomes a parasocial relationship (Kero and Schmitt 2023).

A Parasocial interaction or relationship with a media figure reduces the user's reactance to the content presented and facilitates the adoption of beliefs, attitudes and behaviours (Bradock 2020; Kero and Schmitt 2023). This can be observed in far-right content creators who post on a daily basis to their millions of followers and receive likes and shares from their fan base. By directly addressing their audience, influencers can increase their parasocial relationship with their followers, and increase their authenticity, reliability and approachability. This then increases the feeling of trust which in turn makes it less likely that people will check the truth of the information they are receiving. People are then influenced further into the belief of disinformation or conspiracy theories they are engaged in (Kero and Schmitt 2023). Parasocial interactions with a primary character in a story can therefore decrease audience resistance, provided the message does not overtly reveal its persuasive intent (Moyer-Gusé & Nabi 2010; Ratcliff and Sun 2020).

Processing Fluency

Processing fluency refers to subjective feelings of ease or difficulty that occur while processing new information. Narratives are easier to read and remember than non-narratives, and therefore information transferred through narratives is perceived more positively than information contained in non-narratives. The ease of understanding a narrative is called "processing fluency" (Bullock et al 2021).
Several studies have connected processing fluency to narrative persuasion showing that narratives may be easier to process, which facilitates persuasion (see among others Graesser et al. 1980; Bullock et al 2021; Adaval and Wyer 1998; Bullock et al 2021).

Perceived Realism

For fictional narratives to be persuasive, they must maintain internal coherence and realism that aligns with the story's constructed world. There are two types of realism in narratives: external and internal. External realism refers to the narrative's alignment with real-world expectations, while internal realism concerns the logical consistency within the story itself (Busselle & Greenberg 2000; Schlegel, 2021; Shapiro & Chock 2003; Busselle & Bilandzic 2008; Schlegel, 2021). Notably, narratives with lower external realism can still be persuasive if they are "real enough" to appear plausible, allowing for audience engagement through transportation and emotional involvement (Cho et al. 2014; Bilandzic & Busselle 2011; Schlegel 2021). Hamby et al (2018) termed this verisimilitude, emphasising that characters - even if non-human - must act in ways that resonate with social norms familiar to the audience.

Most current CAN initiatives employ high external realism by presenting factual, non-fictional messages. However, when fictional elements are used, CAN campaigns aiming for authenticity must reflect the target audiences' experiences accurately, often requiring direct insights from the audience for credibility. Fictional narratives with low external realism may yield comparable persuasive effects without needing such direct audience connection. Low external realism also enables CAN campaigns to utilise "comic exaggerations,"

similar to political cartoons, highlighting key themes without risking the perception of mockery that high levels of realism depictions might evoke (Lippe and Reidinger 2020; Schlegel 2021).

Furthermore, narratives with low external realism may sidestep audience preconceptions, as their distance from reality makes them less prone to being dismissed as unrealistic. For instance, discussing discrimination through alien characters rather than humans allows the audience to consider the topic without the interference of existing beliefs about real-world discrimination. Since the narrative is perceived as "just a story," audiences may be more open to the story logic and less likely to dispute the content. Particularly for sensitive issues, low external realism and fictional characters provide CAN efforts the flexibility to address complex subjects with a reduced risk of backlash, enhancing engagement with challenging content (Schlegel 2021).

Utopian Narratives

A core part of right-wing extremists' communication messaging includes a utopian narrative strategy portraying visions of an ideal society. This aligns with Benford and Snow's (2000) theory on collective action frames, which emphasises the importance of providing not only a problem diagnosis, but also a prognosis - positive change resulting from proposed solutions. Extremist diagnostic frames often include dystopian depictions, such as the "The Great Replacement" theory, while utopian narratives tend to envision ethnically homogenous societies with strong collective identities (Davey & Ebner, 2019; Baldauf et al., 2017; Schlegel, 2021). This dual narrative strategy can be highly motivational. Utopian narratives also often feature themes of personal significance and heroism, urging individuals to take bold actions against societal structures (Kruglanski et al. 2014, 2019; Winter 2015; Schlegel 2021).

In contrast, CAN campaigns rarely employ utopian elements, as their focus on external realism precludes the fictional storytelling that would allow such visions. As Berger (2016) observes, pro-establishment messaging lacks the appeal of revolutionary change, which may inadvertently support extremist narratives appeal. RAN (2019) noted that extremist messages often provide "awe-inspiring" solutions, adventure, and success guarantees that CANs cannot match by defending the status quo. When CAN campaigns uphold the present societal structure, they risk alienating audiences who may see it as corrupt or unjust, leaving them without a desirable alternative. Even if CANs included utopian elements, Western societies' lack of grand narratives for societal improvement make such efforts challenging. The absence of a compelling, transformative visions that can rival extremist narratives limits any CAN's effectiveness (Webb 2013; Schlegel 2021).

Integrating fictional elements could help CAN campaigns tap into heroic and transformative aspirations and move beyond merely defending the status quo. Fictional storytelling does not require strict adherence to reality and can provide alternative visions of societal progress. Such narratives that address societal issues like racism or discrimination within fictional utopian contexts, offer a critique of the present, but also potential solutions. Fictional characters also play a key role in shaping real-world self-perceptions and aspirations. For example, female heroes in fiction, such as Hermione Granger (Harry Potter story), have had profound impact on young female viewers by offering an empowering role model (Quiroga 2018; Schlegel 2021). Fictional CAN campaigns might inspire alternative, non-extremist pathways for individuals to find meaning and significance, effectively offering narratives of heroism without sacrificing external realism (Schlegel 2021).

2.4 Emotions and transportation

Returning to the Transportation Theory, recommendations in counter-extremism and communication health literature often claim that CAN messages should focus on arousing emotions, since emotions increase transportation, and consequently persuasion. Indeed, people tend to remember messages better when stories have been presented with high emotional stakes than when presented in a neutral way (Van Eerten et al 2017; Civettini & Redlawsk 2009). Persuasion literature provides insight into fear, appeal, humour and regret - these will be explained in detail below. It should be noted that this section will not specifically explore how emotional narratives impact the effectiveness of CAN messages due to a lack of research in this area. However, there are lessons to be learnt through examining some of the different emotional responses that can be found utilised in narratives.

2.4.1 Emotional shifts

Broader literature states that when narratives are conveyed through emotional flows rather than constant emotions, they can be more persuasive (Nabi 2015). However, Schmidt et al (2023) found that continuously positive stories are more persuasive than those with a negative middle part, yet both stories with and without emotional shifts remain persuasive. They also suggest that the lack of effectiveness stemming from emotional shifts might be caused by the length of the story. In consistently positive stories, recipients are likely to allocate fewer cognitive resources towards critically processing the story's message, as they are motivated to avoid information, such as counterarguments, that could disrupt their emotional (i.e. happy) state. In contrast, the story with emotional shifts, moving between positive and negative emotions, may enhance message processing by relatively increasing cognitive resource allocation compared to the always-positive story. Thus, the persuasive effect might be influenced by a participants' levels of narrative transportation. Future research should investigate the mechanisms behind these findings further (Petty and Briñol, 2015; Clayton et al. 2021).

2.4.2 Fear

Van Eerten et al (2017) suggest that fear appeals might be successful in inducing change in peoples' behaviour if conducted properly. Ruiter et al (2014), considers that effective fear appeals must present a threat and a prospect to prevent this threat. Fear presents a danger to which the person might be susceptible, such as lung cancer provoked by smoking. The prospect, which must reinforce the belief that individuals can carry it out, must suggest protective actions to mitigate the threat. When individuals see the recommended response to address fear as both effective and achievable, they are more likely to adopt it. However, if individuals doubt their ability to avert the threat, this can lead to defensive or counterproductive reactions. Although fear appeals are widely used across various fields, and some research supports their effectiveness, their usefulness remains debated. Studies in health promotion (Peters et al. 2013 & 2014; Van Eerten et al. 2017) and crime prevention (Turpin-Petrosino & Buehler 2003; Van Eerten et al 2017) indicate that fear appeals can be ineffective and may even produce adverse effects.

However, fear appeals lack usage and reviews in CAN hate speech literature. Jacobson (2010) explains that fear might be effective to counter radicalisation through making individuals understand the dangers of engaging in violent actions. But others, such as Beutel et al (2016), provide that since this strategy lacks certainty, and in the context of violence and extremism, it is best to be avoided. This might be different in the context of undermining hateful and discriminatory languages; however, it requires further research to be confirmed.

2.4.3 Humour

Humour-based messaging has been widely applied across fields such as commercial advertising, entertainment, and health campaigns. Despite its extensive use, research on the persuasive effectiveness of humour appeals remains mixed. Some studies suggest that humour can enhance psychological states conducive to persuasion (Nabi 2016; Van Eerten et al 2017). However, there are notable concerns regarding its impact. For example, a meta-analysis in advertising found that while humour can increase attention, positive effect, brand attitudes, and purchase intentions, it does not necessarily influence positive or negative cognition or affect liking toward the advertiser (Eisend 2009; Van Eerten et al 2017). Humour can even diminish the perceived credibility of the message source (Van Eerten et al 2017). Additionally, research on humour in entertainment and educational programming reveals that, although humour may reduce counter-arguing and support persuasion, it can also lead to message trivialisation and discounting (Moyer-Gusé, Mahood, and Brookes 2011; Nabi, Moyer-Gusé, & Byrne, 2007; Van Eerten et al 2017). Further, humour may cause individuals to remember the humour rather than the message content itself, rendering the content ineffective (Konijn 2008; Van Eerten et al, 2017).

In counter-messaging, humour has received limited focus. Bartlett, Birdwell, and King (2010) argue that satire could help diminish the allure of extremist groups, though they caution against governmental use of satire. Historically, satire has contributed to undermining extremist movements like the Ku Klux Klan. A study by the Centre for the Analysis of Social Media at the Demos Institute found that humorous tones in counter-speech against extremism on Facebook generated significant engagement (e.g., likes, comments, shares). This was especially so when parodying extremist language. Although humour's potential risks cannot be overlooked, its ability to foster message sharing suggests it may serve as a valuable tool in certain contexts. Tuck and Silverman (2016) observed instances where satire and humour were used to undermine the claims made online by haters and Gemmerli (2016) even suggested making fun of extremists' ideological claims.

However, other scholars emphasise that humour carries risks; for example, Goodall et al (2012) warn that humour can be divisive and may hinder policy goals if misapplied. Beutel et al (2016) highlight the importance of understanding the humour's target and intent: humour aimed at discrediting terrorist leaders could reduce their appeal, but ridicule directed at potential recruits may provoke resentment and defiance or retaliatory ridicule.

Most studies on humour and persuasion do not differentiate between humour types. Nonetheless, scholars propose that different forms, such as slapstick, irony, and satire, may influence audiences differently. Due to this, humour types should be carefully considered, as certain types may counteract the message's effectiveness (Van Eerten et al 2017).

2.4.4. Regret

Persuasive appeals can emphasise the expected positive or negative emotions that may arise if a recommendation is (not) followed. Messages may thus guide individuals to imagine the feelings they might experience should they follow - or disregard - a suggested action, thereby shaping their intentions and behaviours (O'Keefe 2015; Van Eerten et al. 2017).

Research has examined the role of anticipated regret. This negative emotion occurs when individuals realise or imagine that a present outcome could have been better if they had chosen differently. The Theory posits that individuals typically aim to avoid regret from their decisions, especially when the consequences of risky behaviour are severe and irreversible. Accordingly, evoking anticipated regret could discourage risky actions (Sandberg & Conner 2008; van der Pligt & Vliek 2016; Van Eerten et al 2017).

The potential impact of anticipated regret has been demonstrated in various studies, particularly within preventive health behaviours. Evidence indicates that anticipated regret influences health and safety practices such as substance use, safe sex, vaccination uptake, organ donation, and adherence to road safety regulations (Koch 2014). A meta-analysis by Brewer, DeFrank, and Gilkey (2016) suggested that anticipated regret is more influential when related to high-risk activities with severe consequences. Their analysis also highlighted that anticipated regret associated with inaction tends to have a more significant impact than regret linked to action (Van Eerten et al. 2017).

Despite this evidence, the use of anticipated regret strategies in counter hate speech efforts remains largely unexplored. In the context of radicalisation and violent extremism, de Wolf and Doosje (2010) propose that encouraging individuals drawn to radical ideologies to consider the potential negative emotional repercussions of their actions could have a preventive potential. However, this research is limited to counter-terrorism radicalisation in which it can be easier to leverage the emotion of regret, since terrorism involves violent actions. Leveraging anticipated regret regarding hate speech might be more complicated. More research should be conducted on the topic.

2.4.5 Empathy

Empathy can be a powerful emotion to deter the perpetration of hate speech, and a motivation to take part in CANs. The literature defines empathy as "the ability to understand and share the feelings of others" (Batson 2009; Wachs et al 2024: 4). Having empathy makes people who are not victims of hate speech understand the deep consequences of the psychological and sociological impact that hate speech can have.

Studies have observed with consistency a negative correlation between empathy and aggressive online behaviour (Barlett et al. 2023; Zych et al. 2019; Wachs et al 2024). Hate speech research specifically has reported that adolescents who empathised with the pain and exclusion that hate speech can cause were less inclined to perpetrate hate speech themselves (Soral et al. 2022; Wachs et al. 2024). Furthermore, it can encourage young people to stand up against hate speech by using counter-speech, or engage in alternative narrative by promoting a positive discourse. Additionally, studies have demonstrated that empathy plays a crucial role in encouraging young people to defend victims of bullying online (Deng et al. 2021; Zych et al. 2019; Wachs et al 2024).

Other research shows that empathy can be negatively related to prejudice, often considered a root cause of hate speech (Pettigrew & Tropp 2008; Ballaschk et al. 2021, 2022; Castellanos, et al 2022; Wachs et al. 2024).

2.5 Other methods to consider

2.5.1 Entertainment education vs advocacy

Narrative stimuli can be broadly classified as having either a more preeminent persuasive appeal (i.e. public service announcements and commercial ads) or a subtler appeal embedded within entertainment (e.g., full-length films and TV shows). Such genre differences can be classified into advocacy and education entertainment messaging. In terms of genre, Ratcliff and Sun (2020) observed a larger effect for education entertainment than for advocacy messages. This aligns with the position that narrative effects vary depending on the explicitness of the persuasive effort. Entertainment has a unique ability to mitigate perceived persuasive intent and other forms of resistance by engaging audiences. Compared to education entertainment, advocacy narratives (i.e. advertisements, policy advocacy) may be perceived as more prominent (Moyer-Gusé 2008; Ratcliff & Sun 2020).

2.5.2 The medium and length

The medium of a narrative - text or video - can influence how audiences engage with it. Text-based stories are often believed to promote active mental imagery generation, leading to potentially greater audience transportation through self-created imagery. Whereas watching a video is considered to be a more passive experience (Green et al. 2008; Ratcliff and Sun 2020). However, audiovisual formats provide a “sensory richness” that can enhance engagement. Prior meta-analyses have shown mixed results; Braddock and Dillard (2016) found no difference in the persuasive effects of narratives based on medium, while Shen et al (2015) reported that audio and video messages were more persuasive than print.

Narrative length is also recognised as a crucial factor in assessing effectiveness. Longer narratives may allow for deeper audience immersion or more time for character development, which strengthen audience identification. Igartua (2010) suggests that extended exposure enhances character identification through increased imagined experiences with the main characters. Mid-length narratives (e.g. 15-20 minute TV segments) tend to produce a larger effect size than very short (e.g. 30-second public service announcements or single-page stories) or very long (e.g. 30-40 minute programmes) messages. Short narratives may not allow sufficient time for engagement, while longer ones risk inducing fatigue, which can detract from the persuasive impact (Ratcliff and Sun 2020).

However, both in terms of medium and length, research is limited. Hence, further studies should be conducted in order to analyse to what extent these variables matter, as well as to study them in the scope of CANs.

2.5.3 The characters

Narratives also vary based on the number of primary "character units" featured. Some narratives centre on a single character (e.g. one woman benefiting from a community health programme), while others involve multiple main characters (e.g. several women sharing breast cancer survival stories). This variation can affect audience engagement levels with the character(s). For example, single-character narratives may foster deeper character-based engagement by focusing attention on the one person, potentially enhancing immersion and identification. Narratives featuring only one main character may also reduce measurement inconsistencies in studies by minimising participants' confusion over which character they relate to or critique

(Ratcliff and Sun, 2020, see also McQueen & Kreuter 2010; Niederdeppe, Roh, & Shapiro 2015; Dahlstrom et al. 2017).

2.5.4 Interactivity

Interactive narratives have been deemed effective by communication persuasion literature. Traditional forms of narratives such as newspaper articles, TV shows, movies, and novels, etc. have a limited capacity for efficiently communicating the complexity of certain topics. To fully grasp such topics, perception of the different sides, factors, situations and how they interplay with each other, is necessary. Although certain forms of traditional media attempt to provide different perspectives they do not actually enable us to understand how a series of decisions are taken, or how certain decisions lead to one specific outcome and not another. This is because it is the author who decides on the content and sequencing (Koenitz et al 2022).

Digital media facilitates the representation of complexity, as it has the capacity of holding limitless amounts of data, dynamically interconnecting them with each other, and the audience can participate and influence its dynamics. They can choose between different viewpoints, add their own perspectives or discuss their insights with others, all within the same Interactive Digital Narrative (IDN). An IDN has the potential for enhancing the understanding of complex issues as it empowers audiences to be interactors, requiring them to make decisions by picking perspectives. IDNs permit users to revisit their decisions, take a different route and explore other perspectives (Koenitz et al 2022).

The agency provided by IDNs can have great potential in reducing resistance to persuasion. One potential reason for this can be the control the user has over the narrative. The increased control can reduce the possibility of inciting reactance to persuasion and can decrease the feeling of someone attempting to influence their choices. Additionally, it can also increase the feeling of identification, as the audience feels like the character who is suffering the consequences of a certain situation, thereby reducing resistance. Interactivity can also increase transportation, since the audience's interactive role makes the recipient dive into the story and be "transported" by it (Koenitz et al 2022). However, there is limited research which analyses the use of interactivity on CANs to combat hate speech, hence, further research should be conducted.

2.5.5 Education

Education is recognised as a critical tool for deradicalisation and combating hate speech. UNESCO (2023) emphasises that "Education is a powerful tool to build learners' resilience to violent extremism and mitigate the drivers of this phenomenon." This is because it reinforces students' commitment to peace by challenging and offering alternatives to hateful or violent narratives. Research by Chetty and Alathur (2018) and Blaya (2019) supports the role of education in raising public awareness and empowering young people to generate counter-speech. Wachs et al (2024) demonstrate that specific school modules focused on hate speech are effective in countering its proliferation online.

Moreover, engaging students as agents of change within their school communities has shown promise in reducing hate speech by fostering a culture of positive messaging. By creating and sharing educational materials - such as flyers, exhibitions, podcasts, and videos - students actively participate in developing a school ethos that opposes hate speech. This peer-to-peer educational model enhances the programmes reach and efficacy, promoting norms of respectful communication and counter-speech practices across the

school community. The collaborative approaches not only reduce instances of hate speech and victimisation in the immediate school environment but also strengthens students' resilience against such behaviours in broader social contexts (Wachs et al 2024).

2.5.6 Working with technology and social media companies

Technology and social media companies, along with marketing and advertising agencies, possess a wealth of expertise in developing and promoting ideas across digital and traditional media, making them potentially valuable allies in CAN campaigns against hate speech. Their experience in shaping public attitudes and perceptions could be instrumental in designing impactful campaigns to address radical ideologies (RAN 2015). Yet, several limitations affect their engagement in this area. Firstly, in the context of violent extremism, companies generally lack specialised knowledge on extremist movements and radicalisation processes, which limits their ability to develop informed counter-narratives. This limitation however could be surpassed in the more neutral context of hate speech. Secondly, their primary objective often centres on their core business and delivering returns to shareholders. This may reduce their investment in initiatives that fall outside these goals, such as counter-narrative campaigns. Yet, collaborating with non-governmental organisations on advancing and protecting human rights could also garner them positive publicity through being seen to be engaging in social corporate responsibility. Thirdly, the highly sensitive nature of anti-extremism work may create hesitations, as companies might fear reputational risks associated with entering a contested domain or aligning too closely with government efforts to combat extremism. Concerns over potential brand impact may limit their willingness to contribute fully to anti-extremism campaigns, despite their technical capabilities (Ibid). Yet, this again could be overcome in the polarising yet not extremist context of hate speech, as it covers a range of different severities of hate. Also, it might signal that technology and social media companies might be more willing to join civil society on alternative-narrative efforts, rather than counter-narrative ones.

2.6 Case studies on potentials of using CANs

2.6.1 Art-based case studies

Counter-narratives can be effectively promoted through arts education, offering both a creative means to address hate and a pathway for students to explore empathy and social issues. Artists have historically embedded counter-narratives in their work, which can provide valuable templates for arts educators.

Ana Teresa Fernández's project, *Borrando la Frontera* ("Erasing the Border"), was carried out in 2016 along parts of the U.S.A-Mexico border. By painting sections of the border fence sky blue, Fernández visually "erased" the barrier, symbolising unity and challenging divisive, physical borders (Taylor 2016; Jääskeläinen 2020).

The activist artist Irmela Mensah-Schramm, also known as the "Hate Destroyer," has dedicated decades to eradicating racist and anti-Semitic graffiti in Berlin, illustrating how counter-narrative art can directly confront hate by removing harmful symbols (Caruso 2017; Jääskeläinen 2020).

Arts-based counter-narratives also reach broader audiences through public installations.

The "Wall of Hope" was created by EGS and Jani Leinonen at Finland's World Village Festival. It displayed pieces of hate speech that were transformed into hopeful messages aligning with the persuasion

communication theory that continuously promoting positive stories is more persuasive (see Chapter 2.4.1). It additionally symbolises an individuals' responsibilities in countering hate, here aligning with the notion that giving individuals agency has potential in reducing resistance to persuasion (see Chapter 2.5.4). (Amnesty International Finland 2017)

These examples underline how arts education can enable individuals to actively engage with CANs and foster an inclusive environment (Jääskeläinen 2020). Although these reach the population, there is no knowledge on whether they generate behavioural impact on the audience. More studies focusing on the behavioural impact of art CAN campaigns needs to be conducted.

2.6.2 Interactive activities case studies

As previously mentioned in Chapter 2.5.4, interactivity has been described as useful for the target audience to fully grasp certain topics. It provides perception of the different sides, factors, situations and how they interplay with each other (Koenitz et al 2022). Examples include narrative-focused digital games, interactive documentaries, and journalistic stories.

In Portugal, Isca et al (2020) speak about a game called "Enredo" ('Plot' in English). The creators defined it as a "gamified counter-narrative". Its goal as a CAN tool is to incite "changes in behaviour" in the specific context of Online Hate speech.

The game was used in schools, and its rationale is as follows:
Players enter a cabin and decipher the reason behind a website's feed discontinuation. Subsequently, they explore the system of the discontinued website and encounter several hate messages, making them understand the shutdown. The competitive nature of the game encourages dialogue (and interactivity) amongst the players and allows situations to occur where they can feel insulted at certain points. The insults can be on sexual orientation, gender, harassment, or attacks on minorities.

This gamified CAN allows players to be invested in the game (transportation theory) and feel identification with the victims of hate speech (identification theory).

"Brasileiras Não se Calam" ("Brazilian Women Do Not Stay Quiet") has interactive pages on both Instagram and Facebook. This project has been identified as a case of, "cyberfeminism, aiming to combat coloniality and long-standing stereotypes of Latin women, particularly Brazilian women who are often subjected to sexual objectification and stigmatisation" (Silva 2023). The campaign began in 2020 on Instagram as a way for women to anonymously report situations of harassment, discrimination, and prejudice experienced by Brazilian immigrant women in Portugal.

The project has transformed into an emotional support network for Brazilian women who struggle with prejudice and discrimination. They hold online and physical meetings, where women can share stories and seek advice and support (promoting continuity in campaigns, which will be addressed in Chapter 3.3.4). This example represents a project transitioning from being an interactive direct (short-term) response to specific discriminatory events, to becoming a long-term network counter-narrative project.

The App "No More Haters", developed in Spain, is designed to engage users and confront them with particular imaginary situations. In these situations, they have to make decisions on how they are going to

confront hate speech against different minorities (building on the Identification Theory). The game sets them in scenarios such as “you read a viral Tweet and… what do you do?” The App aims to debunk disinformation which targets certain minorities, as well as seeks to teach what constitutes hate speech.

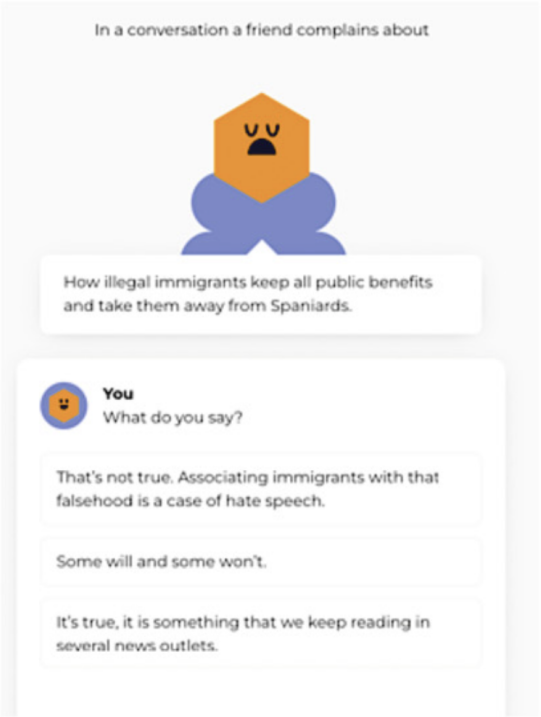


Figure 12 and 13: Screenshots of the App “No More Haters”. (Source: Silva 2023)

The problem with the above-mentioned examples is that they fail to measure whether hate speech prevalence has been reduced. As will be highlighted in Chapter 3.3.3, this is a historical issue. Campaign developers are not measuring whether their initiatives, despite having good intentions, are truly impactful and leading to behavioural change, or not. Therefore, they might be concentrating their efforts on initiatives that actually do not bear impactful real-life results.

3. Possible Risks of Using CANs to Combat Hate Speech

Most academics and experts analysing the level of effectiveness of CANs come to similar conclusions, this being that the lack of research on the issue undermines the potential to determine CAN's effectiveness. It is thus crucial to understand what the objective is, who is the target audience, and what communication method is most effective for the particular issue at hand (Van Eerten et al 2017). Each variable can change the reception of the message. If the message is received negatively, it can either result in the target audience ignoring or rejecting the message or could even result in it backfiring, with the result of increased hatred from the target audience. In this chapter, the reactance and cognitive dissonance theories are explained. It will also look at the need to use theories of persuasion and employing emotions with caution, and explore other possible risks that one can encounter which can lead to your CAN backfiring.

3.1 Reactance and cognitive theories: why do people react negatively to CANs?

3.1.1 Reactance Theory

Resistance is "a reaction against change in response to some perceived pressure for change" (Moyer-Guse 2008: 414) or "an individual's motivated response, triggered by the perceived persuasive attempt and enacted to disregard the intent and/or the content of persuasion" (Ratcliff and Sun 2020: 415). This tends to occur when individuals hold a strong belief and are faced with a logical argument or fact which undermines their core belief. Instead of agreeing with the counterargument, they are resistant to it, even if the (non-narrative) argument makes sense, or they internally agree with it (Seyle and Besaw 2020). In communication science, "the operationalisation of the notion of resistance can be summarised into four categories: counterarguing, perceived threat to freedom, message derogation, and anger" (Ratcliff and Sun 2020: 415-416).

Counterarguing considers the "generation of thoughts that dispute or are inconsistent with the persuasive argument" (Slater and Rouner 2002: 180). The target audience will often create resistance by coming up with counter arguments to the initial arguments provided, to persuade them to change opinion (Ratcliff & Sun 2020).

Perceived freedom threat refers to when the audience targeted by counter arguments may perceive their freedom to be threatened. In other words, they might believe that their freedom to think, feel and act how they want is threatened by a persuasive message. The Theory suggests that when people sense an external entity limiting this autonomy, they experience a state of "reactance." This reactance manifests as an uncomfortable motivational state, characterised by feelings of anger, uneasiness and counteractive thoughts, which propel individuals to engage in cognitive and behavioural actions aimed at reaffirming their autonomy. For example, an individual declared: "I am uncomfortable being told how to feel about tobacco use" (Ratcliff and Sun 2020: 416). This negative feeling may present a risk, discouraging target audiences from listening and following the message. Once reactance happens, individuals may attempt to regain their freedom, which may cause a 'boomerang effect', driving individuals to reinforce their beliefs or ignore the message advocated by the counter argument (Van Eerten et al 2017).

Message derogation refers to a negative response that ignores the message of the counterarguments, regardless of the critical assessment of the persuasive arguments. It rejects the persuasive message regardless of its content, unlike the counterargument reactance, which refutes "specific points in the message". It can be manifested with adverse responses to the story's characters or events (i.e. "the event is disgusting") or with negative evaluations of the overall message, irrelevant to the persuasive content itself (i.e "the video was too long") (Zhou & Shapiro 2017: 1301; McQueen & Kreuter 2010: 19; Ratcliff & Sun, 2020).

Anger has been examined both as an aspect of psychological reactance and as a potent form of resistance. Persuasive messages can trigger this emotional response, if audiences resist and resent directives.

3.1.2 Cognitive Dissonance Theory

The Cognitive Dissonance Theory holds that an individuals' cognitions need to be consistent. In other words, elements of knowledge, perception, attitudes, beliefs and feelings relevant to each other, must be consistent (Harmon-Jones and Mills 1999; van Eerten et al. 2017). When this does not happen, individuals will experience a negative mental tension referred to as cognitive dissonance. This results in individuals aiming to rescue the dissonance by looking for information that counter argues the information that produces the dissonance (Van Eerten et al 2017). For instance, Festinger (1957) explains that smoking is detrimental for your health. If a smoker comes across information that supports this fact, it is likely to cause dissonance to them. Since it causes a negative feeling, the smoker will seek to reduce the dissonance by avoiding situations and information which are likely to increase it. Similarly, smokers could seek information that highlights the positive sides of smoking (i.e. produces joy, social benefit, relaxing, etc.), to reduce the dissonance (in Van Eerten et al. 2017).

In the world of video games, the concept of "ludonarrative dissonance" is not uncommon. It refers to "a temporal dissonance between the player's interpretation of the ludonarrative process and the interpretation of the overarching narrative". This implies that if players are provided too much agency, it might undermine the process of transportation and identification and interrupt the flow of the narrative. Accordingly, the audience might get confused or frustrated by the choices (Roth and Koenitz 2016). Therefore, interactivity of CAN digital narratives should remain limited, to prevent counter narrative messages from being hindered. However, more research needs to be conducted in the topic of CANs and Digital Narratives.

3.2 Using the theories of persuasion and emotions with caution

Communication literature and counter-terrorist literature focused on CANs show that using persuasion narratives theories such as Transportation Theory, Parasocial Theory and Identification Theory can be useful to persuade and change people's behaviour (as explored in Chapter 2). However, literature also emphasises certain risks that these forms of communication can lead to. Bullock et al (2021) emphasise that all narratives are not equally persuasive, as they depend on the interaction between message content and audience specific characteristics, including their knowledge or preexisting beliefs (Yeshurun et al. 2017; Huskey et al. 2017; Bullock et al.2021).

The counter-violent extremism literature suggests that counter-narratives should evoke emotions. Yet, there is little consensus on which emotions are most effective. Studies indicate that fear appeals are generally

ineffective, and humour, while potentially impactful, requires cautious application. Anticipated regret shows promise in guiding behaviour away from radicalisation, although research evidence is limited. Working with younger audiences may further complicate strategies, as developmental factors - such as increased aggressiveness, sensation-seeking, and risk-taking - can undermine the effectiveness of emotional appeals, with severe messaging sometimes trivialised as 'humour' by this demographic (Van Eerten et al 2017).

3.3 Other issues to consider

3.3.1 Lack of credibility from the messenger

There is not a specific theory which explains the typology of the ideal messenger to transmit CANs as the credibility of the messenger depends on the target audience (Lombardi, Lucini and Maiolino 2020). One particular person or group might be perceived as credible to a certain audience, whereas to another audience they might be perceived oppositely. Taking an example from the countering violent extremism literature, if the intended receptor of your narrative is a group of radicalised Islamists, a priest is probably not a credible messenger for this group. Yet, a priest might be a credible messenger for a far-right Christian target group. When preparing a CAN, it is important to conduct extensive research on the CAN target audience, as this allows the campaigner to understand who is considered a credible messenger for the audience.

Government communications generally lack credibility. This is especially true with CANs as government messages, they can either lack credibility (due to the "say-do-gap") or be perceived negatively by the target audience (Renes et al. 2011; Van Eerten et al. 2017; Briggs & Feve 2013; RAN 2015; Lombardi et al. 2017). The "say-do-gap" refers to discrepancies between the governments' actions and their communication or differences between the government's policies and the realities and perceptions of the target audience (Aistrope 2016; ISD 2014; Van Eerten et al. 2017). This lack of credibility from the messenger can result in the CANs effectiveness being undermined or in the message backfiring with malign actors leveraging the say-do-gap (Ingram and Reed 2016; Van Eerten et al. 2017). In the context of violent extremism, the field where the majority of CAN research is conducted, certain governmental objectives of challenging Salafi-Jihadi extremist narratives backfired. For example, after a governmental campaign, radical narratives reached broader audiences, resulting in dignifying Salafi-Jihadi extremist movements (Van Eerten et al 2017). Although studies focusing on combating hate speech do not address or demonstrate that governmental CAN campaigns can have similar effects, considering the anti-governmental stance that many far-right groups have, it is possible that governmental campaigns could have a similar lack of credibility among this target audience.

Similarly, CSOs or institutions that support particular causes can lack credibility with certain hate groups since their aim is to undermine them. For instance, a CSO that works against racism might be perceived negatively by a white supremacist audience. Moreover, often CSOs and governments collaborate which can lead to even more reduced credibility.

3.3.2 Risk of not segmenting the target audience correctly

Accurately segmenting a target audience is extremely complicated as within one target group, members might have various profiles. Hate group members can have different reasons to engage in hate-motivated behaviour, they can have been influenced through different methods or have different psychological and sociological backgrounds. Within the target audience, variables such as age, gender, race or ethnicity, social class, religion, and geographic location can differ and can have an impact on the reception of the message.

For instance, even if an adult and a teenager have the same ideology, teenagers are generally more emotionally sensitive, which can lead them to interpret certain CAN messages differently to adults (Van Eerten et al 2017). Young people are generally more likely to engage in aggressive behaviours and therefore are more prone to sensation seeking and risk behaviours, meaning that emotional appeals or warnings about dangers may easily backfire (Lombardi et al 2017). On certain occasions, warnings about violence may even appear humorous to such an audience (Van Eerten et al. 2017).

Campaign designers can create a message that they consider impactful yet is not impactful for their target audience. Additionally, messages can be interpreted differently by the messenger and the target audience. For instance, if a CAN depicts violence in a scary way, the hate group target audience might not necessarily perceive this violence as scary or negative (O'Keefe 2015; Van Eerten et al. 2017).

Using humour or mocking might be particularly risky if the audience is not segmented appropriately, as segmenting incorrectly can result in the audience feeling ridiculed. In the context of violent extremism, if a message makes fun of potential recruits and the audience is composed of potential recruits, they might feel offended and disregard the message. Yet, if the message makes fun of the leader recruiting them, then the target audience is less likely to feel offended by the message as it does not make fun of them directly (Beutel et al 2016). It is thus crucial to adapt the CAN to each audience, so the message does not result in the audience feeling offended by the message or associating negative feelings to it.

3.3.3 Lack of effective evaluation in CAN campaigns.

It is complicated to measure a CAN's impact and effectiveness, or create accurate evaluation methods as the qualitative side of CANs, which includes changes in beliefs, behaviours or opinions of people, is hard to measure due to the ambiguity and dynamicity of such aspects (Lombardi et al 2017; RAN 2021). This leads campaigners to measure the effects of campaigns through quantitative methods, including measuring the reach, numbers of views, likes, shares, comments, replies, participants, costs, and geographic and public reach of the campaign (RAN 2021). Yet, quantitative measurements do not provide much understanding of the actual impact the CAN campaign has had on people's behaviour and beliefs, (this will be further explored in Section 3 of this study). Many campaigns consider the reach objectives as the overall goal and hence consider that their campaign was successful if it reached a certain level of interaction, shares, and likes, etc. However, this does not provide insights into people's changes of behaviours or beliefs.

CAN campaigns must evaluate their campaigns more systematically through accurate methods - more information is provided in chapter five on effective evaluation methods that be used to assess a campaign's impact.

3.3.4 Limited resources and lack of continuity in campaigns

Lack of time, human resources, and financial resources, can be detrimental when conducting a CAN campaign, resulting in a lack of continuity and limited impact. Lack of resources lead to inadequate research and analysis of the target audience and issue at hand, which can lead a campaign in the wrong direction or reduce its long-term perspective (CDADI 2024). A lack of resources can also increase complexity in finding credible messengers, convincing credible messengers to participate in your CAN campaign, or accessing expertise in marketing.

Resource shortage can also slow responses to any campaign criticisms on social media. Fast responses to CAN campaigns are necessary for the message to avoid backlash (see Chapter 2).

3.4 Challenges with social media CAN campaigns

The challenge with a digital CAN campaign is the global reach that it might have. The broadness of the campaigns reach can be risky as the posts might land in the hands of people who should not be targeted by the campaign, defeating the purpose of properly segmenting a campaign to a specific target audience. A campaign whose reach is too broad can lead to the potential loss of control over the message through reposts, comments, and other forms of user engagement (Heldman et al. 2013; Van Eerten et al. 2017). This risk is particularly relevant for CAN initiatives, which may provoke adverse reactions, including hostile commentary or hate speech (Ernst et al., 2017; Van Eerten et al., 2017). Indeed, entire campaigns may be used by adversarial groups and repurposed for opposing messages, as seen in the "Say No To Terror" initiative which was rebranded by opposing factions as "Say Yes to Jihad," resulting in counter-content that outnumbered the original material (Briggs & Frennet 2014; Van Eerten et al 2017).

Additionally, online CAN campaigns might struggle to find their entire target audience on social media, as part of their target audience might use traditional media or might be more easily reached through interpersonal communication (Ingram & Reed 2016; Levac & O'Sullivan 2010; Stevens 2010; Van Eerten et al. 2017; Zeiger 2016).

A last challenge lies in assessing the value and impact of social media campaigns. It is complicated to measure the impact of the goals of such a CAN campaign because it is hard to interact with the recipients of the campaign (Adewuyi and Adefemi 2016; Heldman, Schindelar & Weaver 2013; Taylor 2012; Van Eerten et al. 2017).

4. Applied Case Study: The HateLess Programme

Below we present an education CAN case study which has employed the majority of the potentials of CANs highlighted in Chapter 2. It has also considered nearly all the risks possible that have been explained in Chapter 3. This Case Study is the perfect example to showcase how, if one considers the potentials and risks of CANs thoroughly, it can lead to positive and effective outcomes. It is also an appropriate case study to apply all the previously mentioned research into practice.

The 'HateLess. Together Against Hatred' programme aimed, through its prevention programme, to reduce hate speech among young people (particularly 7th and 8th graders – aged 12 to 13 years) in schools in Germany. "In five modules, the students jointly learnt what makes hate speech so dangerous, where it comes from, and what harm it does, so they can fight it with the right strategy and free their school from hatred and hate speech" (University of Potsdam n.d.). Its five-module education programme aims at two things, first, increasing empathy towards victims of hate speech (in-line with research findings on transportation of narratives using empathy as described in 2.4.5), and secondly, increasing the belief of young people in their capacity to effectively stand up against hate (providing agency to youth, reducing their resistance to persuasion) (Wachs et al 2024).

The programme aimed at teaching the following skills:

- Increasing knowledge about the nature of hate speech;
- Building competencies in areas such as implementing effective counter-speech;
- Developing self-efficacy for dealing with hate speech;
- Increasing emotional competencies, namely improving empathy and increasing moral engagement;
- Building social competencies;
- Promoting cooperation;
- Building methodological competencies such as ethical media use.

Wachs et al 2024.

The programme mixed different types of methods to reduce hate speech, namely knowledge-based interventions (e.g. educating adolescents about the phenomenon of hate speech or how information and communication technologies may influence their online behaviour), intergroup contact interventions (e.g. reading stories and watching a movie about members of other social groups), and individual skill development (e.g. empathy and self-efficacy training) (Wachs et al 2024).

In this programme, each module has its own main characters: Anura, Bennet, Carla, Hamza, and Laura. The protagonists are of the same age as the students, so they can become confidants and students may identify with them (University of Potsdam n.d.). Here the employment of the Identification Theory can be seen as part of the persuasion communication group of theories. It also shows how the programme developers thought about the characters by ensuring that each module had a single-character narrative, and thus potentially enhancing immersion and identification, as was explained in Chapter 2.5.3. Furthermore, these five characters all learn in one class and are "as diverse as is typical of heterogeneously composed classes today" (Wachs et al 2024). This decision demonstrates that the programme developers carefully thought about segmenting the target audience and recognised that within one target group (7th and 8th graders in particular), there are still heterogeneous profiles, as was explained in Chapter 3.3.2. Additionally, with characters being presented as 'credible,' it enables empathy and reflection on the diverse life situations that

individuals can have. This also demonstrates that the campaign developers have thought about and addressed elements including 'risk of lack of credibility of the messenger' (Chapter 3.3.1). Moreover, the characters help the children "on their journey to a school without hatred" (University of Potsdam n.d.). This also aligns with the potential of emotions and transportation explained in Chapter 2.4.1, which argues that continuously positive stories are more persuasive than those with a negative middle part.

Programme Modules (University of Potsdam n.d.)
1 - Students learn to distinguish between hate speech, verbal abuse, and bullying. 2 - Students search for the underlying reasons and motives. 3 - Students are encouraged to think about the consequences for society. 4 - Students are directed in how best to deal with hatred. 5 - Students are encouraged to set up an interest group at school.

Throughout the entire programme, it is perceived that integrating fictional characters could help the CAN campaign tap into heroic and transformative aspirations amongst the 7th and 8th graders, especially when during Module 5 when the fictional character encourages them to set up an interest group at school. Additionally, for Module 2 specifically, there are enhanced experiences of empathy, as students engage in a role-playing exercise, whereby they experience what it is like to be excluded. As highlighted in Chapter 2.4.5, empathy can be a powerful emotion to deter hate speech, and a motivation to take part in a CAN. Furthermore, we can also see the Transportation Theory being used, "by changing perspectives and putting themselves in someone else's shoes they understand how much words can hurt" (University of Potsdam n.d.).

Furthermore, to support teachers in the process, they are given a detailed manual with instructions on how to navigate sensitive subjects, as well as didactic tools, illustrative PowerPoint presentations, animated videos, and a short film. Here, we see that the programme developers have also considered the 'resources and continuity in campaigns' aspect, ensuring that there are no resource shortages, and that those who have to implement the campaign are fully supported.

Despite this project not having finished, and therefore the final results from the analysis not being public, there are possible avenues to measure behavioural changes amongst the young people in the 7th and 8th grade and the schools following this programme. For example, by recording hate speech incidents in school before, during, and after the programme, and seeing whether after the programme the incidents have decreased. According to initial results, the involvement of classes in the HateLess programme has been linked to a subsequent reduction in both online hate speech and victimisation. It has also shown an increase in the adolescent's engagement in countering online hate speech in the month following their participation. Another alternative to measure behaviour change would be to see how many of the young people decide to set up an interest group, and for how long this interest group lasts (days, weeks, months or years). These are some possible ways of measuring behavioural change through impact achievement.

Overall, this project represents a substantial contribution to existing research, offering the first empirical evidence of HateLess as an effective intervention for equipping adolescents to confront and manage online hate speech. These findings are consistent with prior research demonstrating the programmes positive impact on adolescent's offline counter-speech behaviours within school settings (Wachs et al 2023).

5. Practical Guide: How to Create Effective CAN Campaigns

5.1 Initial campaign phase

Communication science and political science literature emphasise that research is critical in the initial phase of any campaign (Van Eerten et al 2017). This research stage requires:

- a deep understanding of the subject.
- a clear definition of the problem and target audience.
    - insights into effective programme strategies adapted to such an audience.
    - consideration of risks due to the specific problem and target audience.
        - including the education and literacy levels of the target audience as it can determine factors in the reception of the message.
        - goal to reach a narrow and as homogenous as possible target audience.
- outlining the campaign's structure.
    - consideration of how to combine offline and online components.
- creating a broad plan for its design - which includes:
    - the application of theory (including communication persuasion theories).
    - defining and understanding the audience.
    - examining past hateful narratives.
    - assessing available resources.
    - setting specific goals, objectives and clear criteria for success - among which the evaluation after the campaign will be based.
- Other important aspects to consider.
    - how information reaches and flows through communities.
    - how the credibility of information is determined.
    - whether formal or informal power structures are of greater importance in disseminating the message and shaping its interpretation.
    - which individuals or groups have the greatest potential to influence the behaviour of the campaign target group - messengers that the target audience relate to.

5.2 Communication persuasion theories, emotions and education - impactful CANs

A fundamental principle from communication research is that applying theory or multiple theories is essential to designing effective communication strategies. Effective communication efforts are distinguished by their "careful application of theoretical principles," in contrast to less effective strategies that rely on “intuitive, ad-hoc approaches” (Perloff 2010). This would mean being able to avoid any negative reactions, explained in Reactance Theory, in which individuals reject the message because the CAN does not follow persuasion rules.

When making narratives it is crucial to reflect on and use different theories, some of which may include:

- Transportation
- Parasocial interaction
- Character identification

- Easy processing fluency
- Perceived realism
- Other methods including entertainment advocacy, a correct medium and length, characters (single or multiple), interactivity, and interactions with technology and/or social media companies.

Communication theories also hold that specific emotions might be effective to transport the audience and hence persuade them more effectively. However, when it comes to (highly) sensitive topics, such as political and violent extremism, such emotions need to be transmitted with care, since they can easily backfire.

Furthermore, education seems to be an effective tool to reduce and counter hate speech (online). It has been used in multiple experiments and seems to have fostered positive effects. Additionally, education through arts has also been a method used regularly. Yet, no evaluation has been conducted to assess whether these campaigns have been effective.

Finally, it is crucial that campaigns are clear and precise on the message they are spreading in order to avoid any miscommunications between the campaigners and the target audience (Manevska 2024). Avoiding such miscommunication is important as it can result in backlash from the public and/or hateful movements themselves.

5.3 Human rights-based CANs

It is advisable that CANs use human rights values, principles and characteristics, to project a positive storyline which highlights values including democracy, freedom, rule of law, equality, and respect for human rights (Dafnos 2014). These narratives must not contain hate, violence or discrimination, neither as a means nor an aim. “They shall encourage equality, respect and solidarity among everyone and should promote the equal dignity of all human beings,” as well as promote critical thinking, fair dialogue and correct information (Van Eerten et al. 2017).

5.4 Audience targeting

Developing a theory of change is key to a successful campaign, therefore it is important to ask the following three questions:

- Who do you want to influence?
- What influences them? (e.g. facts, emotions, satire, credible voices)
- Where do they congregate? (RAN 2015).

The answer to these questions will determine the specific variables that need to be considered when constructing CAN campaigns (adequate messenger, adequate evaluation, appropriate dissemination channels), in order to conduct an effective campaign (Noar 2011; Slater 1996; Van Eerten et al. 2017).

In the context of violent extremism, the optimal approach is to obtain such information through the former target audience, meaning through deradicalised extremists who understand the psychology and other crucial characteristics of the radicalised audience (RAN 2015). In the context of reducing hateful speech, a similar approach could be explored. For example, consulting with individuals who formerly spread hate (online) in order to understand what drove them away from such practices.

The more homogenous the audience, the more likely the message will have an influence on the audience. No single theory exists to systematically conduct effective audience segmentation as it depends on the particular variable of the particular issue. Yet, the audience could be subdivided by:

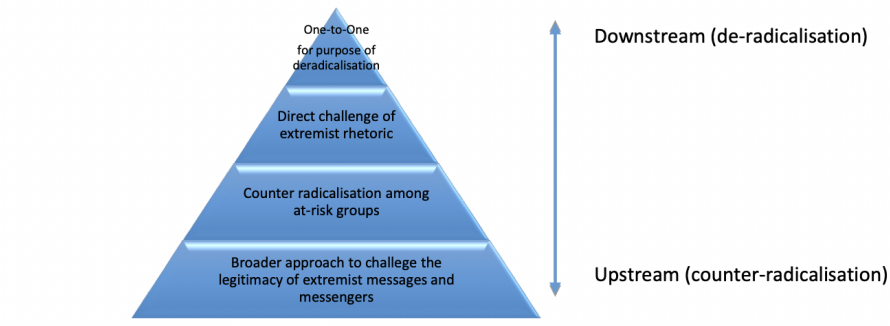
- Demographic variables (e.g., age, gender, race or ethnicity, social class, religion).
- Geographic location (e.g., city, region).
- Cultural psychographic, attitudinal and behavioural variables.
- With whom are the members of the target group in contact with.
- Level of education or professional experience the target audience has.
- (Former) interests.
- Sources the target audience uses to find their information.
- Combination of all or several of the above.

(Boslaugh et al. 2005; Egner 2009; Van Ginkel 2015: Van Eerten et al. 2017).

It is recommended that campaigns avoid broad targeting and instead direct efforts towards more specific, narrowly defined groups, particularly individuals displaying curiosity or sympathy for hateful messaging (Van Eerten 2017).

Furthermore, another step to understand the target audience (especially for violent extremism CANs) is to distinguish between people on the verge of radicalisation (upstream) and people who are already deeply radicalised (downstream). Preventing radicalisation at both stages is necessary to effectively counter narratives.

- When countering narratives upstream, it is recommended to adopt reactive strategic communication and one to one messages. This requires tailored and focused strategies, frequently involving direct, interpersonal engagement.
- When countering a downstream audience, it may be more adequate to tailor broad messages or to spread alternative messages which stand for democratic and human rights values. (RAN 2015).



Examples of counter-narrative campaigns – including where they sit on the radicalisation spectrum – are included in the Annex to this paper.

Figure 14: Examples of counter-narrative campaigns. (Source: RAN 2015)

5.5 Adequate messenger

Literature mainly states that even if a CAN message is perfectly shaped and targets the correct audience, if its messenger is not adequate (credible) to the target audience, the message loses its value (Van Eerten et al. 2017; Briggs & Feve 2013; Davies et al. 2016; Fink & Barclay 2013; Braddock & Horgan 2016; van Ginkel 2015; Weimann 2015; Zeiger 2016).

Several factors can impact the credibility of the source:

- Trustworthiness.
- Expertise.
- Attractiveness.
- Likability, similarity and familiarity (Atkin and Rice 2012).

As a general rule, the most effective message spreaders are found (Hedayah and ICCT 2014):

- At the grassroot level.
- People of the same community.
- Victims, as they can induce empathy from the target audience, representing a powerful emotion to influence change into people (Beutel et al. 2016; Briggs & Feve 2013; Hedayah and ICCT 2014; Schmid 2012; van Ginkel 2015; Zeiger 2016; Van Eerten 2017).
- Former radicalised members (Van Eerten 2017; Hu & Sundar 2009).
    - They can empathise with the target audience since they understand their psychological mechanism to induce behavioural and opinion change. They can probably spark more empathy and willingness to listen.
        - These actors can also be effective in assisting in the finding of other credible messengers or building a credible message themselves.
- It might be complicated to find someone willing to oppose their former companions (Hedayah and ICCT 2014; RAN 2015).
    - They might be ashamed of their past and not want this part of themselves to affect their current life.
    - They might fear being perceived as a betrayer of their group and become a target of the radical group.
    - It is essential to provide security to former radicals involved in CANs.
- Choose a messenger with whom the target audience can relate (Van Eerten 2017; Hedayah & ICCT 2014; RAN, 2015; Richardson 2013).
    - If a CAN targets young people, it is important to choose a messenger which mirrors them, namely someone young or who appeals to a young audience. An older person might lack credibility with them.
- Include a "face" for your campaign, this has the capacity of making the campaign more persuasive by connecting more with the audience (Manevska 2024).
    - This has to be conducted with precaution as the person representing the campaign risks facing hate and security threats. When such threats take place, it is the responsibility of the organisation to absorb the threats and protect the "face" of the campaign.
    - An alternative to avoid such threats is to create a fictional character which can become the face of the campaign.

In the context of social media, everyone can inadvertently become a messenger and a receiver through liking a post, retweeting or sending it to a friend or family (Hu and Sundar 2009; Van Eerten 2017). A message is more trustworthy if it has been sent by a person they trust, this is therefore considered a positive side of using social media to spread CANs.

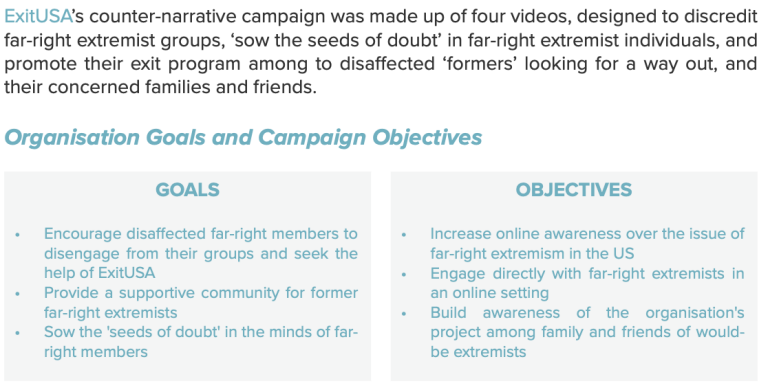
5.6 Adequate evaluation

Effectively measuring the impact and outcomes of a CAN campaign (online) is difficult. For example, the correlation between the CAN campaign and a decrease in hateful rhetoric in the same time period does not necessarily point out a causation between the two. There can have been many other variables that might have influenced that relationship.

An effective evaluation requires (RAN 2015; RAN 2021; ISD 2016):

- Forward planning.
- Clear criteria for success, among which the evaluation will be later based.
- Fixed objectives and goals.
    - Objectives here refer to the, 'what needs to be achieved in order to reach the goals of the campaign'.
        - The objectives of a campaign also need to be evaluated.
    - Goals refer to the expected impact the campaign has on the behaviour and the beliefs of the target audience.
        - To correctly identify what the needed goals are, it is necessary to identify the reason behind the groups/individuals hate-motivated behaviour. Once this is identified, it is important to understand the necessary changes the individual needs in order to replace the "reason" with something more positive. This approach is complicated and not always possible as it requires close contact with the target audience in order to be able to identify this type of information.

Combining the two types of quantitative (metrics such as number or shares, views and likes) and qualitative (behaviour change metrics) evaluations measuring objectives and goals is crucial.


ExitUSA's counter-narrative campaign was made up of four videos, designed to discredit far-right extremist groups, 'sow the seeds of doubt' in far-right extremist individuals, and promote their exit program among to disaffected 'formers' looking for a way out, and their concerned families and friends.

Organisation Goals and Campaign Objectives

GOALS	OBJECTIVES
• Encourage disaffected far-right members to disengage from their groups and seek the help of ExitUSA • Provide a supportive community for former far-right extremists • Sow the 'seeds of doubt' in the minds of far-right members	• Increase online awareness over the issue of far-right extremism in the US • Engage directly with far-right extremists in an online setting • Build awareness of the organisation's project among family and friends of would-be extremists

Figure 15: Exit USA campaign's differentiating between objectives and goals. (Source: ISD 2016: 16)

Stricter methodological designs would permit addressing the issue of determining a causality between the campaign and its impact on the goals of the campaign (randomisation, control groups or a more comprehensive insight into the behavioural foundations of messaging strategies). Implementing evaluation methods which permit analysing people's behaviour or mindset is necessary (ISD 2016), such as:

- Sentiment analysis: using data mining and natural language processing to gather a sample of text and analyse its meaning through an automated process, categorising it into different categories: far-right, racist, hate speech, homophobic, etc. This permits it to detect general decreases or increases in hateful tendencies on the internet.

An example of such a tool is the "WhoDis" AI detection tool, which combines political analysis and technological innovation to support the detection of the spread of hate speech patterns online and offline. The detection of such patterns by the AI tool is enabled by Justice for Prosperity Foundation lexicon and taxonomy of keywords, activities and strategic language used by anti-democratic actors. The usage of terms within the lexicon and taxonomy are risk indicators of the increase of anti-democratic or hate speech trends online.

The AI tool detects the increase or decrease of these risk indicators and has the capacity to link their usage to events, narratives or the people who are behind the peaks or decreases in their usage. The capacity of the tool to detect toxic language is enabled by its use of Natural Language Processing innovation, instead of Large Language Models who refuse to process toxic language making it complicated for them to process hate speech.

Previous platforms that worked on hate speech detection failed to be able to identify 'implicit' hate speech messages, which is something that the WhoDis tool can do. This data can be gathered throughout different platforms such as X, YouTube, and Instagram. This is an interesting method because it does not require direct interaction with the target audience, which could result in biassed responses, and avoids the complication of reaching out to the target audience when campaigns are conducted online.

- Online surveys: These can collect a greater depth of the sentiment and impact of the CAN campaign. They can be a source of open answers and provide quantitative impact. Nevertheless, surveys can hold biases or survey fatigue, which leads to inaccurate responses to some questions and reduced data quality. Therefore, surveys should be used complementarily to other evaluation methods. Distributing the surveys can be done through links added to the content of the campaign, so people who have just visualised it can follow it or add the social media profile of the campaign. Nevertheless, it cannot be guaranteed that the audience has seen the whole campaign, which might also bias the results. Another possibility would be to send it to individuals who have visualised, liked, commented or reposted the full content of the campaign.
- One to one in-depth interviews: These allow one to ask accurate questions and build an understanding of the extent to which that person has changed their mind or behaviour, as well as what made them do so, and how likely they are to change their minds again. Interviews also allow evaluators to directly talk through the strengths, weaknesses and impact of their CAN campaign and identify improvements for future CAN campaign design. Yet, interviews are complicated to establish as the target audience might be reluctant or uncomfortable to answer. It might also be complicated to establish contact with individuals, though this depends on how the campaign was set up and how numerous they are.
- Focus groups: These can provide the same type of accuracy as in-depth interviews, although group dynamics need to be taken into consideration by moderators. However, focus groups might be complicated to organise as some individuals might prefer anonymity when it comes to their past or current (hate-based) opinions.

5.7 Appropriate dissemination channels and partnerships

As was explained in the audience sub chapter 3.3.2, the goal should not be to reach the largest audience possible, but rather to reach a narrow and as homogenous as possible audience. It is essential to find the right communication channels of such an audience and subsequently spread the message in those particular channels (Van Eerten et al 2017; Ingram & Reed 2016; Noar 2011; Stevens 2010). Indeed, it is crucial to "tell the message where the audience already is" in order to make it as effortless as possible for the target audience to access the message (Manevska 2024).

It is crucial to collect data on which channels are mainly used by the target audience (Van Eerten et al 2017).

- Targeted digital advertising to spread the message to the identified target group can be an interesting avenue to explore.
- Spread the message using particular hashtags which are regularly used by the target audience or on popular accounts they use.
- Partner up with an influential social media account to spread the message.
    - This is especially useful when the CSO involved in the campaign lacks the necessary resources to conduct appropriate digital marketing analysis.
- Coordinating different campaign efforts is important in order to increase the power of the communication effort.
- Accurate timing: Even the most effective initiatives could result in failure if launched at an inopportune moment.
    - For instance, choosing the time when the target audience comes back from work or when they have breaks.
- Combining online and offline action is considered to be the most effective approach, especially considering that online campaigns do not always reach the target audience. Face to face communication is more effective in reaching and influencing individuals. This is particularly true for contexts where CAN campaigns target those more prone to violent extremism.

5.8 Flexibility and adaptability in the social media landscape

Considering the global reach and speed at which content circulates on social media, campaigns require flexibility and rapid adaptability. The posts themselves can become the target of negative responses, including hateful comments or receive comments from people who are interested in learning more (Taylor 2012; Hedayah and ICCT 2014).

- Campaigns need to have the right resources to be able to react to such instances.
    - Organisational procedures, such as multi-tier approval processes, may impede timely responses to audience interactions.
    - Organisations must also consider their legal accountability for all content, including user-generated comments, highlighting the need for sufficient human resources to manage these activities effectively (Freeman et al 2015).

6. Summary of Main Findings and Concluding Remarks

Combating hate speech through counter and alternative narratives has shown some effectiveness, yet it comes with inherent challenges and risks. CAN campaigns need to consider a range of variables to reach their intended impact. Literature suggests that countering harmful narratives with narratives, rather than with direct (non-narrative) arguments, is generally more effective as (non-narrative) arguments alone often lack persuasive strength and have risks. But, CANs that employ narrative persuasion theories are less likely to elicit negative responses, as they create more engagement and immersion.

Much of the research presented in this study that is advocating for communication-based persuasive theories in CAN campaigns originates from anti-terrorism and public health literature, as this is currently the only available research. Yet, audience receptiveness to such strategies may differ in the context of hate speech, therefore applying these theories to combating hate speech must be approached with caution. Further targeted research is needed to understand how communication theories can enhance CANs aimed at combating hate speech specifically.

Although the use of emotions in narratives has generally been deemed to increase persuasion, further research needs to be conducted, particularly in the scope of regret, humour, fear and emotional shifts. Based on the available literature, empathy seems to be the emotion that can be used with the least associated risks and can lead to effective behavioural changes. Furthermore, other variables such as the medium, number of characters in the narrative and the length of the narrative could potentially play a role in the effectiveness of a CAN campaign. Yet, again, further research is required in the field of combating hate speech.

Specific research focusing on counter hate speech and CANs, demonstrates that education can have an important impact in reducing hate speech amongst young people, as shown by the practical case study on The HateLess Programme. Considering that education entertainment has also been effective in communication science, it might be advisable to combine education and education entertainment to create compelling educative narratives to combat hate speech.

Finally, CAN campaigns for combating hate speech face several persistent challenges. Primarily, a lack of targeted research and limited funding. These limitations can lead to errors such as misidentifying the intended audience (both online and offline), choosing ineffective messengers, and failing to evaluate the campaign's true impact. Evaluation remains a core issue, as many campaigns rely on self-assessment measures like views, likes, or shares to gauge success. While these metrics reflect reach, they do not necessarily indicate any lasting behavioural change within the target audience. To increase effectiveness, CAN campaigns addressing hate speech need to be supported by more comprehensive research and funding. This support would facilitate well-informed campaign designs and rigorous evaluation. Recognising that CSOs often lack the resources to fully fund such initiatives, a collaborative approach is recommended. CSOs can partner with other organisations or institutions and engage with influencers or other digital media companies to expand their reach and overcome resource constraints. This approach would allow for more accurate audience research, the selection of credible and influential messengers, and the implementation of more robust evaluation methods.

These improvements are essential for CAN campaigns to move beyond broad reach metrics and achieve tangible changes in attitudes and behaviours concerning hate speech.

SECTION 3: Observations on impact from participating Civil Society Organisations using CANs as a methodology to combat hate speech

Introduction

This study provides insights to help practitioners and civil society organisations (CSOs) to understand how CANs can effectively combat hate speech online. This Section details the observations from six CSOs (from Greece, Ireland, Italy, the Netherlands, and Germany) who used CANs as a methodology to develop campaigns combating hate speech. The organisations were identified in an open call launched by the Council of Europe.

This Section details the methodology used by the CSOs for their campaigns. This includes an overview of the step-by-step process used to produce the findings which are outlined in this Section. It additionally mentions all the tools they used (available as additional material), the details of the training provided to CSOs, and the informal support provided. At the end of the methodology chapter, there is a brief comment on the methodology of this Study, including its limitations and opportunities.

The first chapter of the section introduces the CSOs and provides an overview of the working methods they used to develop their campaigns. This includes the campaign's characteristics such as duration, geographic coverage, objective, topics, and target audiences. There is also a shortened version of the quantitative raw data from each CSO based on their inputs to the Google Spreadsheet monitoring tool. The main challenges and the CSO-specific recommendations and conclusions are presented, including conclusions on what to keep or change for future campaigns. This is followed by an aggregated analysis of all the CSOs quantitative and qualitative data and provides conclusions on platform specific audience engagement. Lastly, some of the aggregated results on the CSOs general satisfaction levels towards their campaigns are presented. These include perceptions of the campaign's effectiveness, and some of the most recurring challenges faced when designing the campaigns.

The second chapter delves into further observations from the Study by answering the predefined research questions on the effectiveness and impact that CANs can have on combating hate speech. This chapter contains conclusions on the approach (educating instead of countering), campaign format, language and timing, platform-specific considerations, target audience, risk mitigation, and meaningful evaluation of the CAN campaigns. It is built on knowledge acquired by the researchers through individual meetings with the CSOs, as well as the written contributions at the Study's final evaluation report. This chapter additionally presents observations on behavioural changes, strategic partnerships, reacting to criticisms towards the organisation or campaign, and dealing with the mainstream media.

The third chapter presents some key take-aways and recommendations for online and offline campaigns, as well as recommendations for online campaigning. Furthermore, it has recommendations for future research based on the study's findings, as well as, concluding remarks.

Research objectives

The main aim of the study is to provide insights for practitioners on how CANs can be effective in combating hate speech online. The secondary research objective is to provide insight for CSOs and practitioners into good practices from peers and other fields and to strengthen the understanding of the impact of CANs. This study was conducted as a desk research (Section 2 of this Study), as well as by gathering best practices from the observed CSOs.

Methodology of Section 3

This Study has obtained unique insights into the work of six different CSOs operating across five countries. It has done so through a 'Direct Observational Methodology', whereby it collected mainly quantitative but also qualitative data on the CAN campaigns of these CSOs. The primary focus of a direct observation method is to first-hand and systematically examine all phases of the campaigning, including 'Developing and Designing a Campaign', 'Online Deployment of a Campaign', 'Audience Engagement in a Campaign´, 'Offline Activities' performed by the CSOs as part of their campaign, 'Behavioural Change Metrics', the CSOs exposure to 'Hate Speech Incidents' whilst campaigning, and whether CSOs received any 'Criticisms' of their campaigns.

The figure below presents a description of the different phases of the Study.

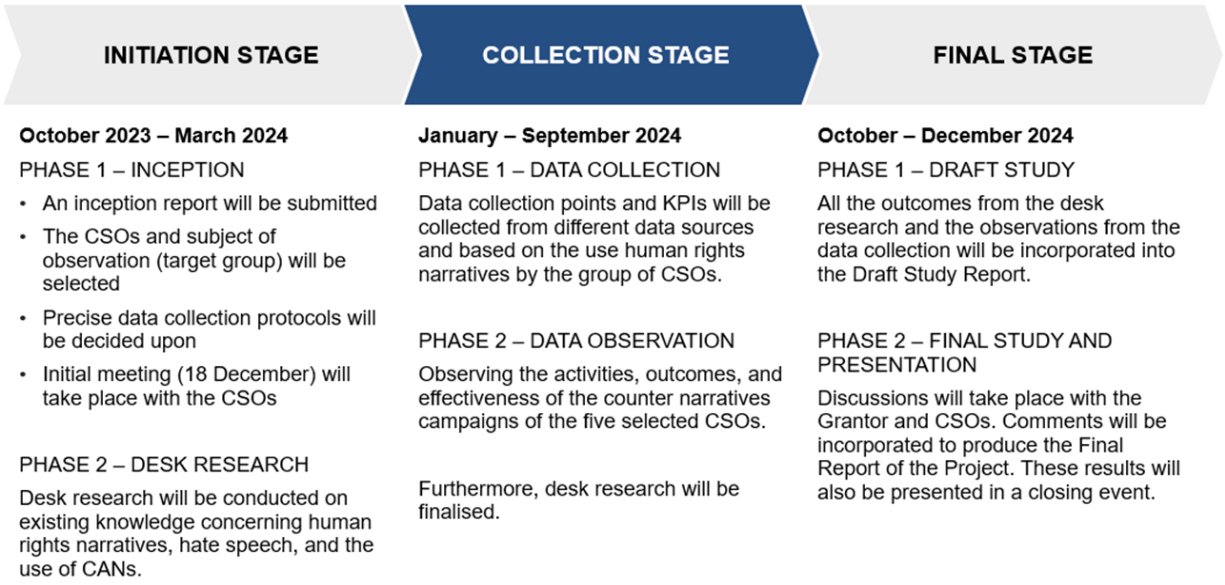


Figure 16: Project Phases for the Study "Effectiveness, risks and potentials of using CAN in combating hate speech".

Initiation Phase

Step 1: Inception Report

The research team produced an Inception Report. The report consisted of the 'Project Synopsis', 'Background' to the Study, the 'Beneficiaries and Parties', a 'Research Plan', and the 'Implementation Arrangements' including the project deliverables.

Step 2: Selection Process of CSOs

The Council of Europe launched a 'Call for CSOs to participate in a Study on the effectiveness, risks and potential of using CANs to address hate speech'. The call defined the eligible profiles of CSOs that could participate in the study as the following:

1. Based and operating in any of the 27 EU member states or be working on the European level, in at least 8 different EU member states.
2. Experience with running communication activities or campaigns based on human rights narratives.

3. Past campaign experiences with messages that are in line with the Council of Europe and EU values and principles.
4. Having clearly defined campaign initiatives in their work plan for 2024.

Following the call, a selection committee consisting of the Council of Europe and Justice for Prosperity selected six CSOs, four who were working with the Hate speech, hate crime and AI Unit, and two who were working with the SOGIESC Unit. These six CSOs operate in Italy, Greece, Ireland, Germany and the Netherlands. They all address different topics and target audiences with their campaigns. Further information on each CSO and their individual areas of work can be found in Chapter 1 of this Section.

Step 3: CSO Analysis Template

Justice for Prosperity created a CSO Analysis Template in order to get to know each CSO and to better understand where and how they operate. The completed template contains information about each CSOs campaign topics, target audiences, platforms of diffusion, working methods, tools for collecting data, and what they would consider as successes and failures.

Step 4: Creation of a systematic monitoring tool to evaluate campaign impact.

Justice for Prosperity conducted research into different methodologies that measure campaign impact and effectiveness. They also explored different criteria that make campaigns impactful according to academia and practitioners. Based on this research they developed 13 Key Performance Indicators (KPIs).

The 13 KPIs established were:

1. Audience engagement.
2. (non-) Homogeneity of audience.
3. Audience reach.
4. Website and social media analytics.
5. Partnership effectiveness.
6. Media exposure.
7. Sentiment.
8. Behaviour change.
9. Effectiveness perception.
10. Increase in awareness and understanding.
11. Incident reporting.
12. Timely response to hate speech incidents.
13. Response to CAN criticism.

A detailed explanation of what they each entail is included in Appendix 3. With these in mind, Justice for Prosperity developed a user-friendly form to collect the necessary data. Through the Spreadsheet[7], Justice for Prosperity systematically examined all the processes involved in campaigning. The observations were fine-tuned based on each CSOs campaign and target audience. For some, due to their modi operandi, certain elements were not relevant, for example, if the CSO did not use video as a format of campaigning, the data input on 'Viewer retention' was not relevant. Other CSOs however had quite elaborate offline activities. For these, the Spreadsheet was expanded to include further offline campaigning criteria.

[7] If you wish to receive a copy of such a monitoring tool to implement it within your organisation, please email info@justiceforprosperity.org.

The diagram below presents the processes the tool measured:

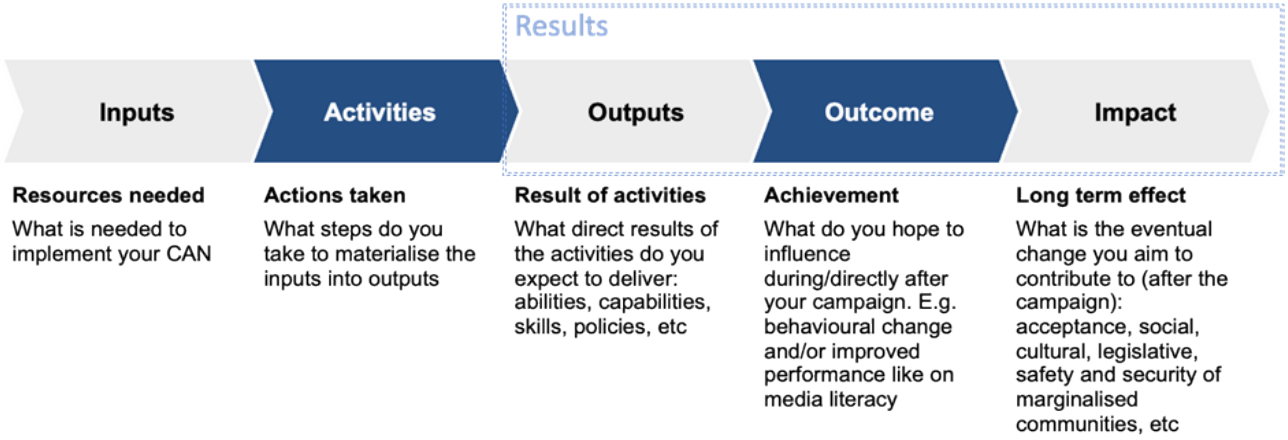


Figure 17: Steps measured from CSOs campaigning.

The steps from Figure 17 are measured through seven different chapters, comprising seven different pages, in the Spreadsheet (Appendix 3).

Step 5 Introducing the CSOs to the study

Justice for Prosperity developed a 'Data Collection Guidelines' document for the CSOs. It organised an online session with them to introduce them to the Study, and its methodology - including an overview of the data collection tool they were expected to use. They also provided a two-day training on the tool at the European Youth Centre in Strasbourg.

Data Collection Phase

Step 6: Data collection process

The data collection process for all the CSOs started in March 2024. Each CSO had different operational timelines and campaigning structures. For this reason, some CSOs data was recorded over a six-month period and four different campaigns, and for others, the data was recorded for one month based on a single campaign – this has limited the standardisation of the data collection.

Ongoing Support to the CSOs

Throughout the entire process, Justice for Prosperity kept in close contact with each of the CSOs liaising with them and solving any issues that the CSOs encountered whilst participating in this Study. During the process over 30 bilateral meetings were held along with more than 10 meetings that included the Council of Europe's secretariat.

Step 7: Study evaluation report

Once the CSOs finished their campaigns and had filled in all the mainly quantitative data in the Spreadsheet tool, Justice for Prosperity developed a qualitative End of Study Evaluation Report. This Report was meant to complement the mainly quantitative data from the Spreadsheet, by getting the CSOs to reflect on and be critical about their campaign results, outline lessons learnt and conclusions. The End of Study Evaluation Reports different sections included 'Qualitative scaling on campaign satisfaction', 'Campaign effectiveness', 'Use of the Spreadsheet Monitoring Tool', 'Most appropriate timing and format for campaigns', 'Engaging

with hate towards your campaign', 'Future campaigns and lessons learnt', and 'Messages for the audience of the Study'.

Final Phase

Step 8: Data observation and analysis

Data observation and analysis was conducted by the senior researcher. It included bilateral calls with some of the CSOs to clarify or elaborate on their data and gather more information on possible lessons learnt. In September 2024, Justice for Prosperity organised a plenary 'End of Study Get-Together', where even more feedback was collected for the study through a series of discussion questions.

This analysis is presented in this Chapter.

Once the final phase was completed the research team compiled and analysed all the data collected by the CSOs and produced this report.

Strengths and Limitations of this Study

The main strengths of the Study:

- Justice for Prosperity has created two systematic analysis tools (the Spreadsheet and the Study Evaluation Report), to monitor and evaluate the effectiveness of online and offline CAN campaigns. These tools can contribute to conducting similar evaluations by both academia and practitioners.
    - According to the participating CSOs, the biggest potential from this methodology was its capacity to foster critical analysis about their campaigns. An additional benefit of these tools is that they provide the option to look at platform specific differences, behavioural changes, and even differences across post engagement. All the CSOs said they would continue to use the tools in the future.
- These monitoring tools can be used to guide a collective review of the work, stimulate conversation, and build on or improve the campaigns.
- The CSOs welcomed Justice for Prosperity's flexibility and openness to propose adaptations to and tailoring of the data collection and working methods to the needs of each CSO. Justice for Prosperity also accepted additional documents and sources of information from the CSOs.
- This study has proven useful for gathering best practices, lessons learnt, and interesting recommendations based on the experiences and results of the observed CSOs.

The Study also provides: a definition of hate speech, lists the main perpetrators, lists the major risk groups, describes the impact and consequences of hate speech, and the legislative system and policies put in place to address it both at a European level and internationally (Section 1). It also includes good practices, lessons learnt in using CANs, communication theories, an applied case study, and insights from communication, public health, security and political science fields (Section 2).

The main limitations of the study:

- Because there is no one size fits all approach, it is extremely difficult to accurately measure impact, effectiveness, behavioural changes, and levels of success.

Each CSO has different target groups, objectives, and modus operandi. As a result, some tabs from the Spreadsheet were not used by some of the CSOs, as they were not applicable to them. Each CSO also operates very differently from the others, which makes it complicated to compare their work. Additionally, as this study was conducted with the experience of only six organisations, the findings and conclusions of it cannot be generalised for all organisations and should not be considered as the absolute truth.

- During the data collection tasks, one problem for the CSOs was the amount of time and technical knowledge required to fill in the data collection spreadsheet (with over 130 data inputs). The CSOs have financial and human resource constraints, which makes filling in an extensive data collection spreadsheet burdensome. Therefore, the amount or details of data provided was different from one case to another, creating challenges in comparing them. As a result, some sub-sections of Chapter 1 are longer for some CSOs than for others. Similarly, including gender and age-disaggregated perspectives in Section 3 has been difficult for similar reasons.
- Data analysis was further complicated due to errors in adding up aggregated figures in the spreadsheet. Therefore, Justice for Prosperity's researchers had to manually check all the statistics. However, it might also be the case that some CSOs did not input the data 100% correctly, therefore decreasing the validity of the research findings. Furthermore, no reliability or inter-coded reliability tests were conducted, which could also decrease data quality.

1. The Civil Society Organisations' Campaigns and Results

This Chapter provides an in-depth overview of each of the CSOs campaigns and results. Each CSO is presented with an introductory box, specifying the campaign's format, duration, geographic coverage, objectives, topics that it addresses, target audience(s), platform(s) where it will be deployed, the working methods[8], and what they consider would be any successes and failures.

These boxes have been produced with information from the CSO Analysis Template which the CSOs filled in at the start of the project. They also contain information compiled through bilateral meetings and more broader data. Further details about the CSOs and copies of some of their campaign visuals are provided to help better understand and conceptualise their work.

This chapter also delves into the data collected, prioritising three or four main points per CSO, the challenges or risks encountered, and specific conclusions from each organisations campaign.

The CSOs and their campaigns are presented in the following order:
CSO 1. Transgender Equality Network, Ireland: Another Way is Possible
CSO 2. Rutgers, the Netherlands: Spring Fever Week
CSO 3. ICEI, Italy: Deconstructing Stereotypes and Discrimination
CSO 4. Thessaloniki Pride, Greece: LGBTQI+ Rights in Greece (three campaigns)
CSO 5. Jugendstiftung Baden-Wurttemberg, Germany: Meldestelle REspect!
CSO 6. APICE, Italy: Tackling Hate Speech (seven campaigns)

In addition, this chapter presents the aggregated perspective of the performance of different CANs and an overview of the CSOs satisfaction with their own campaigns including their effectiveness and lessons learned.

1.1 Spreadsheet - CSOs case by case

The following sub-section outlines the main findings for each organisation within each chapter of the Spreadsheet including a table with the aggregated and average results in the end. An outline of the Key Performance Indicators[9] as used in the Spreadsheet is provided in Annex.

[8] Among the social media tools used: Reels is an Instagram feature that allows users to create short-form video content and share them on their profile in a section called Reels, but also on Stories, or Feed. Reels are created within the Instagram app, and users can add text, music, filters, and other creative features to help their Reels stand out and become more engaging.
A carousel post is a dynamic form of social media content that allows users to showcase multiple images, videos, or a combination of both in a single post. Unlike traditional single-image or video posts, carousels are interactive, encouraging viewers to swipe or scroll through a series of slides. Carousel posts are popular across many major platforms, including Instagram, TikTok, LinkedIn, Pinterest, and Facebook.
Content sprints are a dynamic approach to content creation and allow you to dedicate a focused and intensive period of time to creating content. During a sprint, the way you produce content is accelerated and the time you do it is compressed to help you meet specific goals or deadlines.
Thumb-Stopping is a social media term that describes exemplary content, typically viewed on a mobile device, that catches the attention of the user and causes them to stop scrolling. This usually refers to content displayed on social media platforms but can also refer to display ads.
[9] Description of the Key Performance Indicators can be found in the annex.

For each CSO, the researchers have provided an overview of the campaign planning process, results for online and offline deployment, results for behavioural changes, ways they responded to the hate speech incidents, and main success factors. They also included challenges and not-so-successful aspects of their campaigns. Finally, the CSO specific conclusions and future campaign avenues are presented.

The results of the online deployment from each CSO are presented in a Summary Results Table. This table includes the aggregated analysis for each campaign, including data on the impressions, reach, likes, saved content, shares, reactions/comments, and clicks on links for their entire campaign(s). It also includes an average of the audience's engagement results for each CSO, for example, the average 'likes' per post.

It is important to clarify the differences between some of the indicators being measured:

- Reach - entails the total number of people who see the content.
- Impressions - the total number of times the content is displayed, thus also accounting for users who see the content two or more times.
- Clicks - represent the number of times a user clicked on a particular link within the post (i.e. leading the user to the CSOs website or to fill in a form, etc).
- Saved content - the number of times a user has saved the content on their device(s) (usually done on Instagram).
- Shares - the number of times a user has shared the content with another user within the same or another platform.
- Reactions or comments - the number of times that a user reacts with an emoji to the content or writes a comment to it.
- Re-plays - the number of times that a video or reel has been played more than once by users.

CSO 1. Transgender Equality Network, Ireland: 'Another Way is Possible'

Table 1

Another Way is Possible	
Format	Social media graphics and website landing page
Duration	July - October 2024
Geographic coverage	Ireland
Objective	To inform the public about the reality of trans healthcare in Ireland (explaining core concepts and problems of trans healthcare in Ireland)
Topic that campaign addresses	Awareness and understanding about gender affirming health care for trans people in Ireland. More specifically: 1. Highlight the reality that trans people receive some of the worst healthcare in the EU in Ireland. 2. Raise awareness that the World Health Organisation (WHO) does not agree that being trans is a mental health issue, yet in Ireland trans people are treated as if they have a mental health issue.

	3. Raise awareness that the medication trans people take is neither novel nor only taken by trans patients, with many of the drugs being prescribed to cisgender patients without controversy or media fanfare by their local GP. 4. Provide a reasonable 'Another Way' approach to Trans Healthcare in Ireland, with links to models of care being delivered in both the UK and the Netherlands.
Target audience	1. General Public - the moveable middle who does not have all the information about trans healthcare yet are being informed by media stories and anti-trans groups. 2. Policy Makers and Politicians.
Platforms	Online: • Instagram • Facebook • CSO Website
Working methods	The organisation used polling data to develop their campaign and better understand the moveable middle's perspectives on the trans community and trans healthcare specific to Ireland. They worked through a messaging design process, followed by a creative design process. The development of their campaign required 300 hours of staff time, including the work of three board members, two CSO and three agency staff members. They used a marketing company to produce social media graphics that prepared the public for the key campaign messages. Within these graphics, users are directed to a new website landing page with information about solutions for gender affirming care in Ireland. They used 'Content Sprints' as a campaign format and spread complex messages through the 'Scroll Stopper' or 'Thumb-Stopping' strategy. • Content Sprints consist of a single post with catchy sentences featured in the post followed by an explanatory description in the post bio. • The Scroll Stopper/Thumb Stopper strategy, normally used in advertisements, entails capturing and engaging users enough so that they stop scrolling and pay attention to the content.

	Their posts had a single catchy sentence aimed to educate readers through the comment on the post bio. Their messaging was short, blunt and catchy, and their bio used eight sentences maximum. Tools they used to collect data were: • Instagram • Facebook • Backroom data centre from their website
Successes and failures	Campaign successes: • In qualitative terms, the content of the responses and engagement with their campaign were overall positive. Campaign failures: • In quantitative numbers, they received less than 75% of the usual engagement on their campaign posts.

Results of online deployment

They made 16 posts for their 'Another Way is Possible' campaign. All the posts (graphic with text) were deployed on Instagram and Facebook. The aggregated findings for their posts and the total averages for them can be found in Table 2 below.

Table 2. Aggregated Results for the 'Another Way is Possible' campaign

	Impression & Frequency	Reach (Volume per target group)	Likes	Reactions or comments	Saved content	Shares	Clicks on link
Aggregated online deployment	IG: 85,391 FB: 9,793	IG: 75,283 FB: 9,429	IG: 3,328 FB: 491	IG: 17 FB: 72	IG: 180 FB: 0	IG: 470 FB: 159	IG: 90 FB: 56

(IG – Instagram, FB – Facebook)

In terms of average, the mean 'impressions' were 5,336.94 per post on Instagram, and 612.06 per post on Facebook. The average 'reach' per post on Instagram was 4 705 and 589.31 per post on Facebook. There was an average of 208 'likes' per post on Instagram and 30.67 per post on Facebook. Regarding 'comments' there was an average 1.06 per post on Instagram and 4.5 per post on Facebook. Based on the average (mean) calculations, Instagram had approximately 771.96% more 'impressions' than Facebook, 698.39% more 'reach' than Facebook, and 578.19% more 'likes' than Facebook. Whereas Facebook has 324.53% more 'comments' than Instagram.

Audience engagement

Measurements on saved content, shares, comments or clicks on links per post are appropriate (quantitative) indicators to measure audience engagement with the campaign and its content. On average, on Instagram there were 11.25 'saved content' per post, 29.34 'shares' per post, 1.06 'comments' per post, and 5.63 'clicks on link' per post. Whereas on Facebook, there were on average 9.94 'shares' per post, 4.5 'comments' per post, and 3.5 'clicks on link' per post – there was no 'saved content' on Facebook.

As visible by these figures, despite Instagram having higher engagement in all other indicators, Facebook has a higher average of comments per post, highlighting that perhaps there is more audience visualisation on Instagram, but more audience engagement on Facebook. The organisation considered that the overall interaction with the content was hugely positive with some trans people even sharing their experiences.

Concerning the comments from users, these mainly compared other countries' healthcare systems, mentioned the Irish GPs treating trans patients poorly, argued about the validity of trans healthcare, called for change to best practices, or called the community to action.

Their top performing post was 'Being Trans' (Figure 18), with 14 196 impressions on Instagram, 398 likes, and 5 comments, on Facebook there were 723 impressions and 44 likes. Their worst performing post was 'In Correct Care 2.0' (Figure 19), this had an average of 393 impressions, 62 likes and 0 comments on Instagram, and 160 impressions, 14 likes and 14 comments on Facebook.

Figure 18: Being Trans Campaign Post. (2024)

Figure 19: In Correct Care 2.0 Campaign Post. (2024)

Results of offline deployment

The Transgender Equality Network only performed one offline deployment as part of their campaign: a presentation to their board. The purpose of this meeting was to provide information about the campaign progress and inspire the strategic communications direction of their organisation in the future. After reviewing the work, the board decided to engage on social networks at more strategic times in the year, and to find funders to resource their work.

Results of behavioural changes

With this campaign, the organisation aimed to raise awareness rather than drive a particular set of actions, which makes it complicated to measure behavioural change. Yet, the audience did appear more engaged and more aware, and there was a modest amount of commenting. From 2023 to 2024, there was an average increase of 617.24% of likes on Instagram and 91.69% on Facebook. In that regard, this campaign was successful for increasing the reach of their awareness building, especially on Instagram.

How this CSO responded to hate speech incidents

The organisation shared that they did not have to deal with any substantial incidents, possibly because their content was exposed to allies. However, they believe that by using paid ads to further promote their content, they may face more criticism from audiences who do not agree with their work.

The only criticism they came across was one from a troll who said that children should not receive trans healthcare. The organisation did not directly reply, but rather a user tackled the comment. The Transgender Equality Network was surprised to see how the audience countered the troll, telling him he was "full of sh*t". The organisation removed the comment because they have a 'Zero Tolerance Policy' to non-constructive speech.

CSOs main success factors

Within the organisations team, there are three experts on social media: the current CEO is a digital marketer, and two other board members are a graphic designer and communications expert. Furthermore, the Transgender Equality Network worked with an agency that provides expertise in creative socials to help with messaging. The organisation trained a designated staff member on social media, thereby contributing to building institutional capacity. Following the training, the quality of the creatives (with vivid colours and strong messaging) stood out, and the campaign work was more impactful. Additionally, the funding was used for a staff member's time to attend a free META training, and another one on Instagram and Facebook 101. This was included training on posting, scheduling, moderating and reporting. The staff member was also responsible for filling in the Spreadsheet, therefore also improving the organisation's monitoring and Evaluation Analysis.

Finally, the organisation believes their campaign garnered a positive sentiment from the audience towards them, as the campaign was mentioned by other activists.

Main challenges experienced by this CSO

Among their main challenges were how to convey such complex messages in a simple way, and how to communicate nuances around the difficulties of a trans person's access to healthcare. The organisation decided to take a 'bare minimum' approach and communicate to the audience the essential information that they had, in order to inform about the issue but not overload their social media. They spent a lot of time with the expert agency analysing drafts and figuring out the best way to communicate with their intended audience.

Furthermore, they also faced (1) risks associated with the sensitivity of the topic – the reaction of cis-people to the provision of healthcare to trans people, and (2) risks that their CSO would be attacked by 'bad actors'. They believed that risk (1) will be present whether it is a CAN, or another methodology and they would mitigate risk (2) by promoting their content to an audience who would not be outright hostile. Added to this,

they ensured that they actively monitored posts and comments and that their terms of engagement were clearly laid out on their platform.

As part of their campaign, the organisation would make one post, and re-post a version 2.0 about that same topic few weeks later, as a way to remind users of that content and encourage them re-engage. This appeared not to be that successful, as often the audience had already seen the original post, and thus did not engage extensively with the reminder/refresher content.

Additionally, despite significant improvement from the previous year's campaign on audience visualisation and engagement, their figures represent only a portion of the population, notably users who are positive towards their work. The Transgender Equality Network believes that with more resources, they could use paid Ads to reach more people (also outside their supporter's bubble) and influence the 'moveable middle'.

Furthermore, they were disappointed to receive hundreds of likes but barely any comments. This might be due to the lack of call to action in their posts. Finally, the Transgender Equality Network also realised that when they posted about what their organisation was doing, posts had less engagement.

CSOs conclusions and future campaigns

The organisation overall felt inspired by the effect of their work and believe they are in a better position for next year's campaign. They intend to continue upskilling their team and would use paid adverts on social media. From the analysis, it can be concluded that the Transgender Equality Network was successful with its campaign. It reached its objectives and met its designated campaign success criteria with having a majority of positive responses and engagement. The organisation considered it would have been a failure if the posts had less than 75% of their usual engagement. Yet, their campaign did not fail, rather it exceeded expectations. There was an improvement of awareness and interest among the target audience, with a 14 times increase in engagement with their content on Instagram (from an average of 366 impressions, to a new average of 5 337 with their campaign content).

This organisation used the small additional funding for training a staff member, enhancing their skillset in communication, developing of messages, and using social platforms. They also reflected that "if it's worth doing, it's worth measuring".

If they continue to post these 2.0 versions in the next campaign, they will do so through paid content in order to reach more people. Also, they want to change the colour and/or wording of the post, so that people who have already seen it do not automatically realise and hopefully will engage with it again, thereby positively influencing their algorithms.[10] With additional knowledge and budget they would also explore multimedia content.

Whilst recognising that tackling the topic of trans children is a very complex field that often receives a lot of controversy, the Transgender Equality Network feels better prepared and might start addressing this more complex issue in their next campaign. They also want to consider a more robust call to action approach to be able to better measure behavioural change and activation amongst the audience. For example, encouraging people to sign up to their mailing list or asking questions such as "Do you think it is okay that trans people have to wait for 7 years to receive access to healthcare?" in order to evoke a response from the audience.

[10] For more information on how the algorithms works click here.

The organisation wants to run three campaigns at Pride, Trans and Intersex Pride, and Trans Day of Remembrance in 2025. They feel this would maximise the attention given to their community by broader society and keep the energy and enthusiasm within their team.

Take-away 1:

Creativity and simplicity of messages are crucial in grabbing an audience. This made their campaign so much more successful than previous experiences that used their usual content. As a case study from this organisation, it can be concluded that more resources and funding should be made available for improving quality and higher standards of social media presence.

Take-away 2:

"If it's worth doing, it's worth measuring" - having to go back and engage with the statistics made the organisation a lot more aware of what was working and what was not. It also gave them a sense of a baseline of their social media presence, as well as a better understanding of the people reached on each platform.

Take-away 3:

"This can be a really powerful tool for us in our drive to build awareness of the lives that trans people in Ireland lead, what we need from our state, and how that can come about. I'm really excited at the thought of getting to do this again - but with the resources to go broader."

CSO 2. Rutgers, the Netherlands: 'Spring Fever Week'

Table 3

Spring Fever Week	
Format	Yearly campaign - offline and online education on sexual and reproductive rights
Duration	21 February - 15 March 2024
Geographic coverage	The Netherlands
Objective	1. Create awareness of the importance of sexuality education at primary schools (make relational and sexuality education among elementary school children more discussable). 2. Reduce polarisation, so more people recognise the positive effects of age appropriate relational and sexuality education for children. 3. Online resilience for pupils - making children more resilient online and also better informed about relational and sexual development at each stage of life.

Topic that campaign addresses	The organisation is running campaigns once a year which are concentrated into a week. The aim is to create awareness of the importance of age-appropriate sexuality education in primary schools and facilitate sexuality education lessons at primary schools. The topic of focus for 2025 was 'awareness for online safety and resilience for pupils'.
Target audience	Different target audiences (all in The Netherlands): • Primary schools and teachers. • Parents of children in primary school between 20-50 years old. • Primary school children aged 4-12 years. • Professionals (GGD'en (Dutch Public Health Services) - including Sexuality Education Trainers). • General Public.
Platforms	Offline platforms: • Campaign materials for schools. • Traditional media. Online platforms: • Posts on social media (Instagram, LinkedIn, X). • Create traffic to the CSOs Website.
Working methods	Rutgers developed their campaign according to the different target audiences including: • An explanatory website for parents, to guide them in the sexual education of their child (children). • Social media content for young people about different topics (like online resilience). • Information sessions for Members of Parliament. • Newsletters (with tips, additional information and referrals) for schools and partner organisations. To develop this campaign, 699 hours of staff time went into it. The CSO had to design the campaign, organise their media campaign and train their staff in online resilience training, Human Rights and flowcharts. To collect data, the organisation used a social (listening) monitoring system 'Obi4wan', to track traditional media and online platforms.

	Through the above-mentioned platform Rutgers knew where the online conversation and narrative was going and could 'jump in' and intervene as necessary. They also recorded the emails and the telephone logfile from their reception when they received questions from concerned parents.
Successes and failures	Campaign success: • Multiple primary schools participating. Campaign failure: • Online hate and disinformation appearing as a result of the campaign. A failure in this context would be if parents 'buy into the mis- and disinformation' and there is a reticent position to use the CSOs materials due to online hate and disinformation.

Results of online deployment

On the 4th of March 2024, they made three posts for their 'Spring Fever Week Campaign', one on each of the following platforms: Instagram, LinkedIn and X. For all the actions, they used a carousel post format containing a graphic with text. On Instagram, they additionally used stories to support their posts. The aggregated findings for their posts and the total averages for them can be found in Table 2 below.

Table 4. Aggregated findings for the Spring Fever Week campaign

	Impression & Frequency	Reach (Volume per target group)	Likes	Saved content	Shares	Reactions or comments	Clicks on link
Aggregated online deployment	IG: 9,250 LI: 14,368 X: 9,437	N.A.	IG: 238 LI: 244 X: 101	IG: 26 LI: 0 X: 2	IG: 0 LI: 24 X: 38	IG: 23 LI: 6 X: 19	IG: 27 LI: 428

(IG – Instagram, LI – LinkedIn, X – X (formerly Twitter))

Overall, their best platform was LinkedIn with 14,368 'impressions', 244 'likes', and 428 'clicks on link'. Despite Instagram and X having very similar results, their worst performing platform was X, with 9,437 'impressions', 38 'shares' and 19 'reactions or comments'. As visible by these figures, there appears to be more audience visualisations and engagement on LinkedIn, compared to X and Instagram. LinkedIn is also particularly noteworthy in terms of 'clicks on links', which might signal that if the organisation wants their

website or any other product to be professionally visualised, LinkedIn would be a good place to promote it in a professional context.

Audience engagement

Regarding comments from users, they vary according to platform. Instagram was not a positive place, with the majority of comments coming from religious conservative groups and supporters of right-wing accounts. Facebook was even more negative. Due to this, Rutgers decided to no longer generate organic content on that platform, which is why it is not featured in the statistics. They did however continue to use paid advertisements to specifically target certain groups on Facebook. Interestingly, Facebook is a place that generates a lot of discussion (as will be seen according to other CSOs experiences). On X, there were a lot of 'fast, unkind, thoughtless and harsh' messages. There were many hard discussions about the campaign and a lot of polarised opinions (either completely for or against), but little space for nuance. In general, it was an unfriendly discussion on the platform, interestingly though compared to 2023, more users stood up for Rutgers. Positively, on LinkedIn, there was not much negativity regarding their campaign, critical questions were asked, but overall LinkedIn had a positive sentiment, probably due to the more professional profile of the platform.

Interestingly, their content was mentioned by several third parties; there was one mention of their website, four mentions of their Instagram, five mentions of their Facebook and eight mentions of their LinkedIn. Furthermore, their two campaign hashtags were used 865 and 184,472 times respectively.

Results of offline deployment

As part of their campaign, this organisation did multiple offline deployments. The first one was a series of information sessions for the members of the Tweede Kamer der Staten-Generaal (the House of Representatives of the Netherlands). The purpose was to answer questions and concerns of the members in the Tweede Kamer (following the unrest of the 2023 campaign), and prepare them for possible questions from colleagues, electorate and even family members. The activity was launched in February, a week before their campaign started. It was well received and there was not much opposition within the meeting.

Secondly, Rutgers also organised an official launch of the Spring Fever Week at two different schools in the country on the first day of their campaign. The purpose of these activities was to make a celebration for the schools and the children, encouraging them to be involved in the topic and promote the message; 'we are in this together, it's a party to learn about this'. Among the participants of this kick off, were two classes of children, the press, teachers, school principals, and Mariëlle Paul, the Dutch Minister for Primary and Secondary Education and Equal Opportunities.

Results on behavioural changes

The 2024 campaign seemed to garner less behavioural change compared to the previous year when a media storm happened after a deliberate online hate attack (Kro NCRV, 2023). In March 2023, there were 31,500 posts online about their campaign, this was the result of a deliberate disinformation campaign on the topic of sexuality education. Yet, in March 2024, posts online referring to their campaign week were down by nearly 50%, with only 15,000 posts, which means less attention - even if online hate remained to some extent. In April, there were only 670 posts on the topic. These figures show that there appears to have been less online discussion in 2024 than in 2023.

The Spring Week Fever campaign started in 2005, and until 2023, it had not received any online hate. In the year 2019,599 schools participated. In 2020, during COVID, the schools participating increased to 657 and in 2021 the campaign did not happen. In 2022, when Rutgers re-started the campaign, it had 739 participating schools and reached its peak in 2023, with 936 participating schools. However, in 2024 there was a reduction in the number of participating schools with only 611 schools participating. This represented a 34.7% decrease. This could have been a result of the disinformation against the campaign in 2023.

Despite the reduced participation in the number of schools Rutgers considers that the 2024 campaign was more successful due to the reduced "amount of online hate and… more allies and people 'standing up for them' in 2024." Rutgers used a strategic media approach resulting in another narrative in the mainstream media containing factually correct information and exposing the disinformation campaign of 2023. Some advice from Rutgers media strategy experts on how to engage with the media is available here in Chapter 2.5.2 of this section. This Study is unable to determine why less schools participated and cannot attribute the positive feedback about the campaign to the online presence of the campaign, where the comments were in general quite negative. The positive aspect appears to be more related to their offline influence as the organisation had 98 media appearances, 72 newspaper articles, and featured in 26 television programmes. They also held presentations and created a specialised website to directly address the concerns of their main stakeholders – the parents. This supports one of the main take-aways from this Study which is that we must not forget about the potentials of offline campaigning - "Through offline activities and human interactions, discussions are fostered, and through meaningful and sustained discussions, opinions are changed."

How this CSO responded to hate speech incidents

Despite their efforts to prevent disinformation being spread about their campaign, Rutgers faced criticism around several topics. Disinformation was spread using the frame of 'sexualizing kids'. This was spread mainly on the platform X and was detected by Rutgers using a social media analytics tool. Overall, they registered 961 separate content writers spreading this disinformation frame. A second disinformation frame was that of 'indoctrinating children', a third was that they were 'spreading gender ideology' and 'forcing it' upon children. There were over 108 separate posts about these two frames. The fourth frame was the accusation of Rutgers 'being woke', which appeared in 412 posts. The fifth and final frame was that they were paedophiles. They took the decision to not respond to any of the disinformation frames. For more information on how different CSOs address disinformation, hate and criticism, please refer here – Chapter 2.4 of this Section.

Unfortunately, this CSO also experienced real threats revolving around their campaign. Rutgers is equipped with a Crisis Management Team which came about as a result of the deliberate online attacks in 2023. According to the level of threat and the incident, the Crisis Management Team either refers attackers to victim services or reports the incident to the police.

Most of threats were in the months of February (the month before their campaign) and March (the month of the campaign). The threats had different characteristics. The majority were people threatening to go to the office, or that they would 'have no staff left alive', or that someone would #firebomb their office, or that thousands would come to their office, or people would burn the organisations books in front of the Central Station. Most of the threats came through X and Facebook, however they also received one threat through their personal website contact form. There was also a physical demonstration against the CSO. On one occasion the threat was taken extremely seriously, and the staff were asked to work from home.

On one occasion, there was an actual physical 'threat' when a woman came physically to their office. Up until this incident, nobody had actually come to their office. A staff member sent her away and despite the woman never coming back, she protested at the Parliament multiple times complaining about Rutgers. This incident was a wake-up call for the CSO.

The month before the campaign is when they received the most threats. The majority were online or on the phone. Despite this, Rutgers is planning to prioritise the physical protection of their staff and office before and during all future campaigns. This Study recommends that other CSOs should follow the advice of Rutgers and take note of Take Away 5 of this Study.

CSO main success factors

Apart from campaigning, Rutgers also conduct quantitative and qualitative research and provide training and advice to other professionals. They develop training courses, workshops, online courses and intervisions. Furthermore, they work on several different themes including, contraception, safe abortion, prevention of sexual and gender-based violence, sexuality education, and sexual health (CSO Website, n.d.).

Additionally, their offline work represents a big support for their campaigns. They actively engage and interact with their target groups: parents, teachers and children. These actions are important because they provide opportunities to tackle possible backlash and counter any disinformation being spread about their campaign.

Main challenges experienced by this CSO

Among the main challenges Rutgers foresaw when planning the campaign, were a lot of opposition, disinformation, misinformation, demonstrations, and political games. They took all of these into consideration when designing their campaign and included the information sessions with Tweede Kamer der Staten-Generaal members, the introduction of a new website for parents - which included sex-education information, and media interviews prior to the campaign starting.

The not so successful aspects included the misinformation, disinformation, hate-mongers, journalists who were trying to make a name for themselves, very concerned and influential parents, and time and money constraints.

CSOs conclusions and future campaigns

From the data Justice for Prosperity has received, the 2024 campaign was less successful than the previous year when it comes to the number of schools participating. However, online hate was less prevalent than in 2023, which can be considered a success. Additionally, many primary schools did participate, meeting Rutgers internal criteria of success - which was having 'multiple primary schools participating'.

Nonetheless, they still faced several misinformation and disinformation campaigns against them which they considered as a failure.

For 2025, we recommend that Rutgers focuses more on its offline deployments and pushes to recruit more schools. It could also set up even stronger mechanisms to address and prevent disinformation and misinformation being spread about their campaign. A very positive aspect that Rutgers should keep is their Crisis Management Team, which has proven extremely successful in addressing different types of threats and crises throughout the year.

Take-away 1:

"Through offline activities and human interactions, discussions are fostered, and through meaningful and sustained discussions, opinions are changed" - remember to also focus on ways to combat hate speech offline. This campaign was able to reach the parents of the students at participating schools through offline methods, hence increasing the likelihood that the parents will continue discussions about sexuality outside of school.

Take-away 2:

Sometimes, it may be better to respond to misinformation and disinformation. The reason for the decrease in schools that participated in this campaign could have been the misinformation spread about it. Had they responded to this, they could have avoided this decline in participants.

Take-away 3:

Extra help can never hurt. Because of the misinformation spread about and the threats against Rutgers, they created a Crisis Management Team. This team worked to keep the CSO safe, allowing them to stop worrying about the threats they received and put more focus and energy into their campaign.

CSO 3. ICEI, Italy: Deconstructing Stereotypes and Discrimination

Table 5

Deconstructing Stereotypes and Discrimination	
Format	Offline components: • Focus groups • Workshops in local cities • One National workshop • National survey Online components: The campaign consisted of 9 carousels and 1 video. • Carousels - explained different types of prejudice and discrimination; sexual orientation, religion, fatphobia, racism, disability, ageism and gender identity. • Video teaser - in the first 30 seconds the actors show the mistake of judging at first glance. In the last 30 seconds they change perspectives and learn to see people without preconceptions and instead see them with curiosity and respect.
Duration	Offline component: March - October 2024. Online component: September - October 2024.
Geographic coverage	Italy: • Cities in the regions of Piemonte, Lombardia, Emilia Romagna, Toscana, Abruzzo.

Objective	Offline objective: Activate young people to become change makers - show them the contrast between discrimination and the deconstruction of stereotypes. Online objective: (aimed at Generation Z) a. Objective 1: Promote critical thinking regarding the use of words in everyday life. b. Objective 2: Raise awareness about discrimination and stereotypes from an intersectional perspective. c. Objective 3: Counter hate speech online and offline among young people.
Topic that campaign addresses	The campaigns topics were defined with the support of the young people participating in the project DiversaMente - Youth Against Discrimination. They involved them directly in the designing and defining of the campaigns concept and message. Overall, the campaign addressed: • Discrimination and stereotype deconstruction - based on the principles of intercultural and anti-discriminatory approaches and the Anti-Rumours methodology. • The role of 'change makers'. How young activists can become the actors of change in their local territories and at a national level. This project empowered young people thanks to the promotion of inclusion and critical thinking development, the Anti-Rumours methodology, and advocacy capacity-building - as part of the 'Intercultural Cities Programme'. Through this platform, different cities are connected to each other and share best practices. The Anti-Rumours Handbook and Toolkit for Anti-Rumours Dialogue are part of this Strategy.
Target audience	Primary target audience: • Young people (particularly those born from late 1990s to early 2010s - Generation Z) and their communities. • This generation is characterised as being digital natives, growing up surrounded by technology, internet and social media - this all shapes their communication styles and social interactions. Secondary target audience: • Youth workers and youth organisations - through these, they mainly engaged young people in activities (especially offline awareness raising initiatives and trainings). Tertiary target audience: • Municipalities and policymakers.

Platforms	Online: • Social media channels (Instagram, Facebook, LinkedIn, other CSO Socials and YouTube). • Third-parties online media engagement, notably digital creators/ influencers and online magazines. • Digital PR. Offline: • Focus groups • Workshops • Survey
Working methods	The working method was a bottom-up approach and was based on user-generated content. The campaign directly involved young people from youth centres and youth organisations. Together, they defined the messages and the narratives. The online campaign concept was developed and designed, based on the young peoples' insights and by ICEIs partner communication agency. Additionally, ICEI worked with eight influencers. The campaign used the language of Generation Z to encourage reflection. Additionally, as a narrative device, they used the "POV" (Point of View) strategy, which is very popular on social media. It invites viewers to put themselves in a specific situation and reverse their perspective. The POV is a tool Generation Z uses to express itself online: a protective shield that employs irony, allowing them to talk about themselves, without revealing too much. The campaign required 410 hours of staff time and cost 29.000€. This included creating the digital content for social media, the influencers, the press office and the creation of a specific landing page. ICEI used the following tools to collect data: • Google Analytics • Social Media Analytics • Feedback and surveys during offline events
Successes and failures	Campaign success: • Involved and engaged the target audiences in the offline and online activities. Campaign failures: • Some messages were not relevant and could not reach their primary or secondary targets.

Results of online deployment

The ICEI campaign lasted from 2 September to 16 October 2024. Their online campaign had two components, a video campaign and a graphic campaign. They deployed a total of 14 posts: one video teaser, four video interviews (video campaign), and 9 carousels (graphic campaign). Their graphic campaign was published on different platforms including, Instagram, Facebook, LinkedIn and other profiles of CSOs on social media and networking websites. The video campaign was published on YouTube. The organisation distributed a press release to the media, presenting the project and the communication campaign. Seven press articles were published, but unfortunately there are no quantitative results available for this channel as the press did not provide any data for ICEI. The aggregated findings for their posts (video and graphic campaigns combined) and the total averages can be found in Table 6. In the case of this organisation, the 'other CSO Socials' indicator includes the combined data results from 8 accounts of social networks influencers.

Table 6. Aggregated Results for the Deconstructing stereotypes and discrimination campaign

	Impression & Frequency	Reach (Volume per target group)	Likes	Saved content	Shares	Reactions or comments	Clicks on link
Aggregated Online Campaign	IG: 21,840 LI: 5,086 FB: 2,253 CSO: 189,592 YT: 98,631	IG: 16 ,707 LI: 3,212 FB: 1,474 CSO: N.A. YT: 80,232	IG: 504 LI: 37 FB: 83 CSO: 2,232 YT: 1,511	IG: 44 LI: 0 FB: 0 CSO: 75 YT: 0	IG: 69 LI: 0 FB: 2 CSO: 180 YT: 0	IG: 17 LI: 3 FB: 3 CSO: N.A. YT: 2	IG: no link LI: 759 FB: 2 CSO: 15,600 YT: 0

(IG – Instagram, LI – LinkedIn, FB – Facebook, CSO – Other CSO Socials, YT – YouTube)

Based on the data results, it can be seen that the video campaign had the most impressions on 'other CSO Socials' with the 5 videos averaging 37,918.4 impressions per video. This is followed by YouTube and Instagram which had an average of 19,726.2 and 4,368 impressions per post respectively.

The high results on 'Other CSO Socials' and 'YouTube' could have been driven by the campaigns media plan which included some sponsored content on these channels. Facebook had a considerably lower performance and was the worst performing channel overall. According to ICEI, Facebook is becoming more and more a secondary channel when engaging with young people as a target.

For the graphic campaign there was 1 campaign overview carousel, 7 carousels addressing different types of discrimination and stereotypes, and 1 campaign claim carousel. In this case, the top performer channel was by far Instagram, followed by LinkedIn, and the lowest performance channel again was Facebook. In the past two years, ICEI has given more focus to their LinkedIn channels' development which is starting to yield results. They received 759 'Click on links' for this campaign.

Overall, based on the data provided, the video campaign garnered more attention and higher results. This finding is in line with results seen from all the CSOs; reels garner more views and engagement.

Audience engagement

ICEI ran the campaign with a positive message and tone and wanted to stimulate young people to understand the narrative through critical thinking.

They contracted eight influencers to develop the videos of their campaign, where they addressed seven types of discrimination: sexual orientation, religion, fatphobia, racism, disability, ageism, and gender identity. Each influencer produced content which described the type of discrimination they were working on and a solution or a reality of it. They each fully engaged in the acting required for the campaign, and even helped refine the video script. Apart from the campaign discrimination video, a second (inspirational) video interview was produced for each topic. Here the influencers shared why they decided to advocate for their chosen topic and addressed how they approached social media to become a spokesperson on this topic. This shows that working with influencers can foster better results for a campaign.

The Videos accounted for 99% of the views/impressions and 98% of the likes. The Video Interviews with influencers were not sponsored and within the campaign framework and strategy, served primarily as supplementary materials for the main campaign videos. The top performing carousel post was that on the topic of fatphobia, which had 5,271 aggregated impressions, 76 likes, and 8 shares across Instagram, Facebook, and LinkedIn. The worst performing post was the one that addressed ageism, with only 771 impressions, 26 likes and 3 reshares across the three platforms.

Figure 20: Carousel 2 – Fatphobia. (2024)

Figure 21: Carousel 6 – Ageism. (2024)

Interestingly, ICEI provided the following Figure 22 detailing from highest to lowest level, the ranking of views and impressions vs likes on the different carousel topics and across the different platforms. On the Instagram carousels the highest impressions were for the posts on Fatphobia, Disability, and the campaign Claim, whereas on Facebook these were mainly the campaign Claim, Overview (of the campaign), and Religion. On

LinkedIn they were Overview, Disability and Racism. Generally, it can be seen that the topics of fatphobia and disability created more engagement, this could be interesting to explore in future studies.

Top 3 topics by platforms by views/impressions		
Instagram	Facebook	LinkedIn
Fatphobia	Campaign claims	Campaign overview – introduction
Disability	Campaign overview – introduction	Disability
Campaign claims	Religion	Racism
Top 3 topics by platform by likes		
Instagram	Facebook	LinkedIn
Fatphobia	Campaign overview – introduction	Campaign overview – introduction
Campaign claims	Campaign claims	Racism
Disability	Religion	Disability

Figure 22: Ranking highest to lowest of user's specific engagement with each discrimination type according to each platform. (Source: CSO3)

It is interesting to note that with this CAN campaign, followers increased by 125 on Instagram, 16 on Facebook, 250 on LinkedIn and 31 on YouTube. The ironic and humorous element used by the CSO resonated well on social media.

Results of offline deployment

ICEI was quite engaged offline with their project 'DiversaMente'. Through their in-person activities: they gathered information on how young people feel, and what type of discrimination they tend to be exposed to. They then used this information to develop and strengthen their online campaign and used their offline events to assess and qualitatively evaluate the impact of their online campaign. In this case, the offline deployments were crucial in order to develop an effective and ground-based online campaign.

Additionally, the organisation conducted focus groups which were effective for measuring behavioural changes. They also conducted workshops which they used to gather inputs and materials to elaborate their co-designed campaign. Finally, they conducted a national survey to find out how young people feel about the discrimination they face as part of their daily lives.

Table 7. Explanation of offline activities

Offline Activity	Number
Local focus groups	5 (in four cities) • Pontedera, Reggio Emilia, Torino, and Milano
Control focus groups (Local)	4 (2 pre-campaign, 2 post-campaign) - with 2 groups in total. • ICEI was only loosely involved in the selection of young people who participated and did not participate in the control group activity, in order not to influence its results.
National focus group	1

Workshops (Local)	7 (in four cities) • Pontedera, Reggio Emilia, Torino, and Milano
National survey	1

All the focus group took place between February and September of 2024. The aim of the focus groups was to evaluate the target's attitude towards rumours, discriminations and other related drivers. Particularly regarding the beneficiaries of the project, ICEI was interested in studying how their attitudes and perceptions changed from the beginning of the project to its end.

The local focus groups involved young people who were already engaged in the DiversaMente project, and who were invited by their youth workers. The control focus groups that happened pre and post campaign involved young people who were not related to the project. They established these by reaching out to some volunteers from organisations connected to the AVIS network and a vocational school in the province of Turin. Within all of the groups there was high engagement and a strong desire to share their opinions, including feelings of optimism, positivity, and hope for building a future more open to differences.

The national focus group was conducted at a Turin Youth Camp which gathered participants who were highly engaged with the campaign and enthusiastically shared their thoughts. During the session, they used the wheel of emotions to facilitate a deeper and more intimate recognition of their feelings. This tool allowed everyone to explore and verbalise sentiments that might otherwise remain unexpressed, creating an environment of open sharing and mutual support. Some of the reactions that the campaign evoked included empowerment, optimism for the future, respect for others, freedom to express their identity without fear, and not judging without knowing.

In addition, the key messages from the national focus group were:

1. Knowing to prevent hate and discrimination: The campaign highlighted the importance of knowledge and education as essential tools to prevent hate and discrimination.
2. Not judging without knowing: The campaign developed an awareness of the importance of approaching others with openness and curiosity.
3. Respect for others: The campaign emphasised the commitment to recognising and valuing differences among people. This approach allowed ICEI to build stronger and more meaningful relationships which contributed to a community where everyone felt welcomed and valued.

In March of 2024 ICEI also conducted several local workshops in the cities they operate in. The purpose of these was to reflect on the key messages, keywords, emotions, communication channels and targets related to the project's topics. Similarly, a national workshop was held in April in Reggio Emilia. This workshop was organised by the communication agency Puntozero.

Lastly, ICEI conducted a national survey from March to October 2024, for which they obtained a total of 266 responses. The majority of these responses came from the 4 cities where the organisation operates. The survey involved young people participating in the project's activities in its design and was shared through the communication channels of the involved youth centres and by the youth workers.

Among the key findings, a majority of the young people (73% of respondents) recognised that prejudice and discrimination are present in their life, mostly against foreigners, LGBTIQ+ persons, and women. They mainly

encountered such discrimination in school, with friends, or at an institutional level. They also came across it on social platforms, including TikTok, Instagram, and WhatsApp. Finally, among the respondents, over 80% had witnessed or been victims of discrimination.

Regarding how participants dealt with discrimination, they primarily seek help from friends or members of their families. They saw institutions as mostly ineffective or simply having an awareness-raising role. As will be seen later, this is different to the trust in institutions in Germany, which according to CSO5 survey respondents, is considerably higher than in Italy. According to 10% of respondents, schools are not doing anything to prevent episodes of discrimination which is worrying data. Overall, respondents asked for more intercultural and dialogue activities, youth participation, and awareness raising campaigns.

Results on behavioural changes

For this organisation, both the offline and the online campaigns mutually complemented each other, as opposed to other CSOs, where campaigning was more focussed on either the online or offline elements.
Despite it being difficult to measure behavioural changes as there was no specific call to action, higher levels of engagement and interaction with the content of the campaign have been seen. The online deployments saw over 250 000 video views, there were high levels of participation and engagement in all the focus groups and workshops with over 300 young people from diverse backgrounds taking part, and there was a high number of respondents to the national survey.

In the offline activities, because ICEI conducted the control focus groups before and after the campaign, they were able to measure the differences in reactions of the young people and thereby measure any behavioural changes. Additionally, behavioural and attitudinal change is also visible in these groups, as there was a stark difference between the first round and the second round - once they had been exposed to the campaign. According to ICEI, in the first round, participants expressed ignorance, injustice or rage regarding the topics, whereas in the second round, they expressed more positive hopeful and optimistic attitudes. This serves to demonstrate that the positive-based narrative of organisation's campaign managed to foster attitudinal change and hope within the population which could lead them to actions against hate speech and discrimination.

One of the key findings from the local focus groups was that the participants saw the increase in their awareness and skills in recognising rumours or when they were hearing of incidents of discrimination and hate speech which in turn supported them in coping with such incidents within their own families and friendships. This demonstrates that the campaign fostered an impact, meaning that behaviours and ways of thinking changed. The focus group confirmed a positive impact on young people also in terms of greater mobilisation through locally organised activities and actions. The groups evaluated the experience positively and pointed out a growth in their own knowledge, awareness, and a change in attitudes - exactly what is needed to be able to determine that behaviour change occurred.

How this CSOs responded to hate speech incidents

ICEI came across two incidents during their campaign, one on YouTube and another one on Instagram. They do not have specific guidelines which they follow.

CSOs main success factors

ICEI was especially focussed on targeting the campaign message to their target audience. The organisation used a participatory co-designed process to develop the content of the campaign with the young people involved in the offline activities. Additionally, ICEI also thought extensively about the target audience and who would appear as a credible messenger for their Gen Z audience, notably young influencers. They invested in building strong and meaningful relationships, contributing to creating a community where everyone felt welcomed and valued.

Furthermore, the ICEI was also successful in their strategy to use a positive narrative based on similar emotions and empathy, fostering less backlash from the audience during the campaign. They followed the recommendations from Section 2 on utopian narratives, which consider it not only important to state the problem and its diagnosis but also show its prognosis or solution. This was done with the influencers showing how, despite them having suffered from discrimination, they are still empowered in their differences and specificities.

ICEI was continuously asking for feedback from the target audience throughout the campaign, this was through surveys, polls and focus groups. This helped them to understand what was resonating and what was not resonating with young people in general and particularly Generation Z. This was also positive as it signals that there was a close and tight relationship with the target audience.

Moreover, ICEI had different media partners who contributed to promoting their material and increasing the outreach of their campaign. These included the influencers themselves, and used press releases and articles online. All this contributed to the high reach of their campaign.

The organisation was effective in engaging young people with the campaign content and critically reflecting about it. It fostered behavioural and mentality change, as the young people went from rage and anger to feelings of hope and optimism about future prospects after the focus groups. Additionally, their use of "pop" language seems to also have resonated with Generation Z. Furthermore, using a message of empowerment played a vital role in framing the campaign, rather than a didactic lesson, making the content more relatable and engaging for younger audiences.

Main challenges experienced by this CSO

The challenges ICEI foresaw whilst planning the campaign were:

1. Scepticism and Distrust - Many young people are wary of corporate messaging and may view awareness campaigns as insincere, leading to disengagement, scepticism and a lack of trust.
   They mitigated it by building trust through authenticity and transparency. Collaborating with trusted influencers and community leaders also gave them credibility. They also used personal stories and real-life experiences to create a genuine connection with the audience.
2. Information Overload - Gen Z is constantly overwhelmed with information from multiple sources, making it challenging for any campaign to stand out.
   They mitigated it by simplifying the messaging and making it clear, concise, and relatable. They highlighted the campaign's purpose and significance. They used engaging visuals, infographics, and interactive content to capture attention and enhance understanding.
3. Resistance to Change - There can still be resistance among some individuals who may feel defensive when confronted with discussions about their own biases or behaviours.

They mitigated it by fostering an open dialogue with conversations around biases and discrimination without judgement, they ensured spaces for dialogue were safe spaces with open sharing of perspectives and the asking of questions.

4. Digital Divide - Not all young people have equal access to digital platforms. Economic disparities and varying levels of tech literacy can prevent some from engaging fully with online campaigns.
   They mitigated it by ensuring accessibility and providing offline resources and materials. They partnered with community centres to reach those without internet access.
5. Short Attention Spans - Capturing and maintaining interest long enough to convey meaningful messages can be a significant challenge.
   They mitigated it by applying current trends and popular formats on social media, such as memes and short videos, to resonate with Gen Z interests and encourage interactions.
6. The Role of Algorithms - Social media platforms often use algorithms that prioritise certain types of content over others.
   They mitigated it by leveraging social media algorithms by creating shareable and engaging content that encouraged likes and comments. They used trending hashtags and collaborated with influencers.
7. Cultural Differences - Generation Z is incredibly diverse, and cultural nuances can affect how messages are received. A one-size-fits-all approach may not resonate with all subgroups.
   They mitigated it by embracing diversity through culturally relevant content and diverse voices in the campaign to ensure that various experiences and perspectives were represented.

CSOs conclusions and future campaigns

ICEI is very satisfied with the strategy of their campaign, which primarily focused on engaging third parties - notably influencers and media partners. Although this approach required greater effort during the planning, organisation, management, and measurement phases, they exceeded the campaign's initial goals.

Additionally, the organisation reached its success criteria which was to involve and engage the target audiences in both the offline and online activities. In the future they want to build on the strategy of collaborating with third party media to gain further experience and enhance their impact and reach.

The organisation intends to implement these learning outcomes in two future projects. One on strengthening the role of Italian cities in developing local anti-discrimination strategies and the other is a European project on engaging young people and youth workers in a campaign on inclusion and diversity. Both projects will include offline awareness-raising and training activities and a communication campaign targeting both institutional stakeholders and citizens (especially young people).

Take-away 1:

The participative construction of the communication campaign including focus groups, workshops and small group feedback sessions helped tailor high impact messages at the campaign level. It also improved the performance and the effectiveness of the content and overall results. Moreover, it increased the young people's sense of ownership and their capacity to use the underlying narratives in engaging other actors, such as local policy makers.

Take-away 2:

The involvement of young digital creators/influencers in advocating online against discrimination was particularly effective in reaching and engaging young people. This was both in the development of the campaign and in the media plan which created a viral effect in social media and fostered the dissemination of the content, while optimising the media budget effort.

Take-away 3:

The use of a comprehensive measurement spreadsheet, including detailed online and offline indicators, proved to be challenging at times, but improved the quality of insights. As a result of this, they were able to adjust the activities' implementation and daily operations when needed. It also offered a space for deeper and more thoughtful reflections.

CSO 4. Thessaloniki Pride, Greece: LGBTQI+ Rights in Greece

Table 8

LGBTQI+ Rights in Greece – three campaigns	
Format	Various offline and online components
Duration	21 - 29 June 2024 (nine-day event)
Geographic coverage	Thessaloniki, Greece.
Objective	Promote the European LGBTQI+ Festival 2024 1. To amplify the community's voice through the festival. 2. To reach out to many people and invite them to attend the festival. 3. Make as many people aware of the issues and problems that concern LGBTQI+ people.
Topic that campaign addresses	The campaign was mainly focussed on the festival. They wanted to boost awareness about the events taking place in Thessaloniki in June 2024.
Target audience	Members and allies of the LGBTQI+ community who are interested in and advocating for their rights. More specifically, European and Balkan countries, and citizens aged 18-60 from all sexes.
Platforms	Online platforms: • Facebook • Instagram
Working methods	Thessaloniki Pride developed their campaign with their volunteers. The campaign consisted of: • Launch of the Festival Official Song with a positive and upbeat tone. It was produced by the CSO, and the songwriter and singer were both their volunteers. The songs contained

	impactful lyrics, and the song production required 10 hours of staff time and 2 000 euros. • Announcement of the organisation's Human rights conference which required 10 hours of staff time and 500 euros. • UberTalks – six Instagram Reels with members of the CSO's team and recorded inside Uber taxis. Each was interviewed about their experience at the EuroPride, being queer in Thessaloniki, and their motive behind joining Thessaloniki Pride. This collaboration required 20 hours of staff time and 500 euros. They published posts with images for each of their events and posts inviting people to attend and register for the Human Rights conference. Tools to collect data: • Google Analytics
Successes and failures	Campaign success: • Human rights conference Campaign failure: • Lack of consistency

Results of online deployment

Thessaloniki Pride deployed nine posts in total during their participation in this Study. For each of the three campaigns, they deployed the song, one announcement about the Human rights conference, and seven UberTalk Interviews. Nearly all posts had different formats. Their festival song was shared with a post presenting the official EuroPride song video and encouraged people to listen to it and get inspired by its lyrics. Announcements about the Human Rights conference were in a social media post format containing a graphic with text. For the UberTalks, they did a post promoting the collaboration, followed by a Reel format containing the series of interviews. Each Reel was an interview with a different member of their CSO and was deployed on Instagram. Only three posts were deployed on Facebook. The aggregated findings for their posts (video and graphic campaign combined) can be found in Table 9 below.

Table 9. Aggregated results for the LGBTIQ+ Rights in Greece campaign

	Impression & Frequency	Reach (Volume per target group)	Likes	Saved content	Shares	Reactions or comments	Clicks on link
Aggregated Online Deployment	IG: 1,645,042 FB: 1,161,449	IG: 6,441 FB: 6, 893	IG: 6,173 FB: 269	N.A.	IG: 36 FB: 26	IG: 271 FB: 274	N.A.

(IG – Instagram, FB – Facebook)

Interestingly, other CSOs in the study had a higher number of 'Likes' in Facebook, this campaign seems to break that pattern. This CSO had more Reels and videos which could indicate that this style of content is more successful on Facebook than carousels.

Their most successful campaign was the UberTalks collaboration. It must be noted however that this campaign had seven posts, as opposed to the other two campaigns which only had one each. So naturally, audience impressions were higher.

Audience Engagement

Their top performing post out of all posts made on Instagram, was 'Post 8: UberTalks Interview 5' (Figure 23), with 101 thousand impressions on Instagram. Their worst performing post was 'Post 1: EuroPride Official Song' (Figure 24), which received substantially less numbers in all the categories. This reduced engagement in comparison to other posts might be due to releasing this post too late. They reflected that they released the song only 2 months before the Festival started, which for them meant they lost a significant number of listeners. They believe that if they had released the song more in advance, it would have been listened to more.

Figure 23: UberTalks Interview 5. (2024)

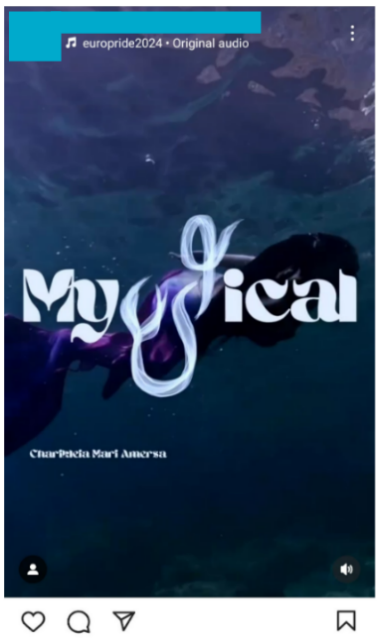


Figure 24: EuroPride Song Campaign Post. (2024)

Results of offline deployment

Thessaloniki Pride completed two offline deployments as part of their campaign: they held a Human Rights conference and a Parade. The Human Rights conference took place over three days in June 2024. It included panels concerning various social and political issues. The aim of the Conference was to act as an amplifier of the voice of the LGBTQI+ community. They selected topics about the challenges and hostile environments being faced by trans people and LGBTQI+ people in general. These included the legislation of same sex marriage. The location, the Olympion Theatre, was chosen purposefully, as it is the oldest cinema theatre in Thessaloniki. By choosing this location, they wanted to spread a strong and important message that LGBTQI+ people also have the right to be present and visible to everyone. Furthermore, some months previously, there was an attack by a mob against two queer people at the same location. They therefore wanted to reclaim the area and make it clear to mobs that they will not hide or be afraid. Thessaloniki Pride organised the conference towards the final days of EuroPride in order to make use of the period when most people would be in the town. Throughout the three days, more than 800 people attended their conference and shared positive feedback. The attendees found the conference informative and reassuring for the future of LGBTQI+ people.

A Parade was organised on the final day of their Festival, during which the organisation members invited all LGBTQI+ people and their allies to walk in the city centre with music and dance. The Parade took place in the city centre of Thessaloniki, claiming the right to be present and open in every part of their city. It took place on a Saturday afternoon, and more than 27 000 people joined the parade. Attendance was more than double their expectations. For Thessaloniki Pride this level of attendance for the parade gave them strength and determination to continue their fight for LGBTQI+ rights. They were especially pleased when 'random people' spontaneously joined the parade, and many joined the festivities from their balconies, including older aged citizens.

Results of behavioural change

Thessaloniki Pride have a clear example of behavioural change and increased engagement from individuals through their offline components. They had 800 people attend the Human Rights conference and around 27 000 people joined the Parade. Based on the data provided by the CSO, both events motivated reflection and gave the attendees hope and a sense of empowerment.

The participants who joined the Human Rights conference found the experience of being able to discuss the issues openly extremely reassuring. This is an example of the power of conversations and human interactions, showing how important it is to have constructive discussions and debates to achieve change. These discussions were felt to be more fruitful because they were done in person. Whereas it is more complicated to achieve the same sort of in-depth conversations online.

How this CSO responded to hate speech incidents

According to Thessaloniki Pride, during their entire campaign they received only 14 negative reactions on Facebook, and seven negative comments on Instagram. However, they did not react to these, as they believe that it is not worth, 'wasting your time' with people who are already radicalised.

CSO main success factors

Hosting such a huge event and promoting it throughout the UberTalks interviews contributed to promoting the organisation's team and volunteers.

The light spirit of the song and Reels, as well as the up-beat rhythm of the song, helped promote their reach and message. According to the organisation, people do not tend to engage in something they consider too serious or distressing. This strategy was helpful to amplify the positive message that "each of us is a wonderful human being exactly as they are and are worthy of love and respect".

One of their volunteers (pro-bono) wrote and composed the song, enabling the CSO to use some limited funding on hiring a professional recording studio.

Thessaloniki Pride invested in ads on both Facebook and Instagram to improve the reach of their content. Furthermore, they also went to Brussels to raise awareness about the Festival which could have also contributed to increased reach.

Key emotions identified by attendees of the conference and the parade were 'optimism' and 'empowerment'. Others talked about their realisation that equality is an issue that needs to be addressed and fought for by everyone. The parade contributed to giving agency and empowerment to the target audience.

Main challenges experienced by this CSO

The organisation feared before the campaign that their song might receive hate speech and push back from homophobic people. However, they did not let this fear affect their course of action.

After this campaign, they shared that they struggled to engage people not actively involved in LGBTQI+ issues whilst staying true to their community.

Time management was a significant problem as they were severely understaffed. This resulted in them facing difficulties keeping the appropriate timing for the deployment of the posts. They released the song only two months before the Festival started, which in their opinion resulted in them losing a significant amount of listeners.

Unfortunately, Thessaloniki Pride did not provide all the indicators for this study, such as the data on the amounts of 'clicks on links' that their Song received or the number of applications they received for the Human Rights conference. This has made it difficult to evaluate the impact of the online campaign on the turnout regarding the offline campaign. They also did not provide many results from Facebook which made it impossible to establish a cross-platform comparison. Because of this, many tabs in their Findings Table have been left blank or 'N.A.'.

CSOs conclusions and future campaigns

Thessaloniki Pride was very proud to be able to organise this big European Pride Festival, which is done once a year in a different European city. Thessaloniki was selected in 2020 to host this festival, but due to COVID it was postponed. They were able to finally host the event in 2024.

This CSO had the best performing content in terms of audience impressions out of all the participating organisations. This could be due to multiple reasons, one being that their content really appealed emotionally to their audience, through lively songs and testimonies with their UberTalks campaign. However, it could also be that they have higher audience engagement as they were organising a Europe-wide event. It could also be due to their paid ads on Instagram and Facebook. A future study could look at another situation when

they are not organising a Europe-wide event or using paid ads, but rather a national or even local event with organic reach, if they have the same amount of audience impressions, or whether these considerably go down.

The organisation had intended to reach at least 1 000 people and aimed for lower awareness yet more sustained engagement. In the end they had 800 attendees at their conference and 27 000 attending the Parade. This exceeded the expectations of their success criteria (explained in Box 3) and shows that their campaign was highly successful.

For next year, they intend to be more organised with planning the time and execution of their posts.

Take-away 1:

Importance of timing the launch and implementation of the campaign. The same post when communicated at the optimal time will have much greater impact and effectiveness.

Take-away 2:

Incorporate various formats in the campaign. A single post with a picture does not have the same impact as one that is enriched with a short Reel that can capture the audience's interest much more.

Take-away 3:

Always be true to your authentic identity and values. The audience responds much more to authenticity and honesty than anything else.

CSO 5. Jugendstiftung Baden-Wurttemberg, Germany: Meldestelle REspect!

Table 10

Meldestelle REspect! (REspect Reporting Centre)	
Format	Online reporting portal
Duration	March - September 2024 Analysed campaign results are from April - August 2024.
Geographic Coverage	Germany
Objective	Overall goal: • Promote hate speech reporting and the filing of criminal complaints. It had three objectives: 1) To encourage people to file a criminal complaint to the Federal Police Department of Germany (more direct call to action). 2) To communicate about the negative aspects of hate speech (more awareness-raising). 3) To promote respectful online behaviour.

Topic that campaign addresses	Main focus of the campaign: • How to recognise and report hate speech - according to the organisation: 40% is about awareness raising, 60% activating people to report online hate speech. The campaign covered general content regarding hate speech, responded to current developments and trends regarding hate speech, and explored specific (hate speech) topics in greater depth. The campaign posts were clustered into four groups, notably: 1. General promotion of the Reporting Office - Meldestelle REspect! 2. Political engagement/EU elections. 3. Cooperation posts (with other CSOs and spreading across all their platforms). 4. Knowledge about hate speech.
Target audience	All internet users, mainly young people aged 12-27.
Platforms	• CSO Websites. • Ongoing Instagram and LinkedIn accounts. • Partner organisation's X account. • Partner organisation websites including, 'Federal Police Department of Germany', 'Ministry of Justice Bavaria', and 'State Police Department of Bavaria'.
Working methods	The organisation creates campaigning posts continuously throughout the year with both pedagogical and awareness-raising objectives. The campaigning materials are developed by their social media team. Jugendstiftung Baden-Wurttemberg has a part-time staff member who spends 25% of their time on social media campaigning and they spent a total of €16,000 on campaigning. Systematic analysis tools: Jugendstiftung Baden-Wurttemberg expected up to 10,000 reports of hate speech in the months of March to September 2024. The reporting of hate speech is broken down for analysis in the following ways: • Number of reports received (as a measure of engagement against hate speech). • Number of criminal charges based on the German law. • Offences of the criminal charges. • Groups affected.

	• Phenomenon areas. • Number of reports by platform. • Comparison of the data by relevant events and to the figures from previous years.
Successes and failures	Campaign success: • Over 10,000 reports per year of which, at least 2,000 resulted in criminal charges. Campaign failure: • Less than 8,000 reports per year, of which under 1,000 resulted in criminal charges.

Results of online deployment

The organisation is running a continuous campaign on social media using Instagram for more educational purposes, LinkedIn for network purposes, and X as a platform to support users and who want the CSO as a witness of specific hate speech incidents. Out of the 40 posts deployed on Instagram, only 16 were also deployed on LinkedIn. Due to hate speech directed against the CSO, their director decided to shut down the LinkedIn channel completely. Therefore, they can no longer access the analytic date from LinkedIn. The posts for their 'Meldestelle REspect!' campaign were all deployed from April to August 2024. They were mostly in the carousel format or the basic post format which contained a graphic with text. Exceptionally, they also deployed two reels during this period. Overall, there tended to be a substantially higher visualisation and engagement with the content on Instagram than on LinkedIn.

The aggregated findings for their posts can be found in Table 11 below.

Table 11. Aggregated Results for the Meldestelle REspect! campaign

	Impression & Frequency	Reach	Likes	Saved content	Shares	Reactions or comments	Clicks on link
Aggregated online deployment	IG: 77,843 (without CAN 3) LI: 689	IG: 55,843 (without CAN 3) LI:	IG: 3,839 LI: 504	IG: 180 LI:	IG: 0 LI: 46	IG: 179 LI: 16	IG: na LI: na

(IG – Instagram, LI – LinkedIn)

Audience Engagement

Among the comments from users on their posts, there were several critical comments, as well as intense discussions about antisemitism. Depending on the comment, Jugendstiftung Baden-Wurttemberg would choose whether to answer it or not.

Through the following graphs, the total number of followers on their Instagram grew by 334 from January-August 2024. However, this Study is unable to determine if the followers have grown due to the success of

their online campaigns or because the Jugendstiftung Baden-Wurttemberg became more popular through other means.

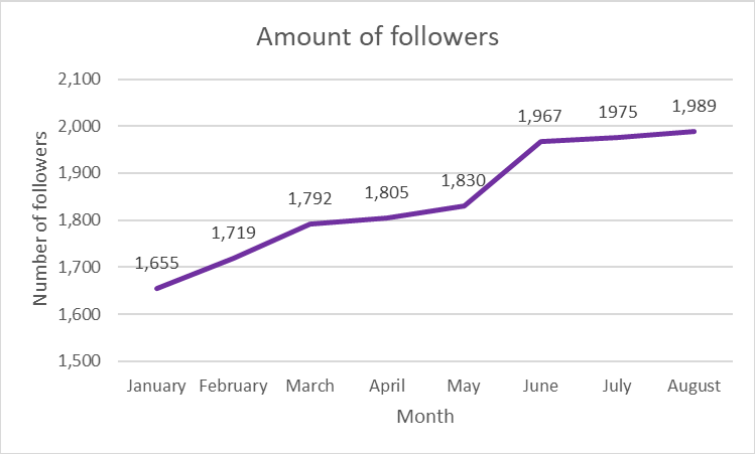


Figure 25: Distribution of CSO followers in absolute numbers (Year 2024).

The top performing post (Figure 26) was a counter-narrative which addressed the general promotion of their organisation. This seems to contradict the experience of the organisation working on Trans Healthcare Rights in Ireland, who received lower engagements on the posts where they mentioned their organisation. Jugendstiftung Baden-Wurttemberg's worst performing posts were those on cooperation and those on knowledge about hate speech.

The top performing post received 51,394 impressions, 1,133 likes, and 23 comments on Instagram (the reel was not deployed on LinkedIn). Outside of the data collection period, Jugendstiftung deployed their 'Trusted Flagger Deutschlands Award' (Figure 27), which they obtained for their excellent work in October 2024. This post had 134 likes and 91 comments on Instagram, and 69 likes, 18 comments, numerous emoji reactions, and 6 reposts on LinkedIn. Their worst performing post (Figure 28) had only 15 likes and 2 comments on Instagram.

Figure 26: Post 7. CAN 1. (Reel) (2024)

Figure 27: Trusted Flagger Post. (1 Oct 2024)

Figure 28: Post 23. CAN 3. (2024)

Results of offline deployment

Jugendstiftung Baden-Wurttemberg invested its efforts mainly in offline deployments and its reporting portal. They engaged in local events and had posters for the reporting portal in every police station in Bavaria. One of the reasons for this was to express their close cooperation with law enforcement and to help make the Police appear more accessible to the public.

As part of their campaign, Jugendstiftung Baden-Wurttemberg also issued a 'Press Release' and a 'Specialist Event' with the Bavarian Ministries during the No Hate Speech Week. The intention of the press release and event was to promote the reporting portal as well as their cooperation with law enforcement.

Jugendstiftung Baden-Wurttemberg had two additional interviews with the press in May 2024 which gave them even more visibility. The first one was with the Bavarian Ministry of Justice, where the organisation called for an expansion of the investigative jurisdiction to all areas of digital and analogue hate crime in the Bavarian justice system. The second one was with the Bavarian State Centre for New Media, to promote cooperation with their reporting portal.

The organisation was featured in 41 media articles, three regarding their reporting portal and 35 general articles. These featured their cooperation with the public authorities, information about criminalised hate speech, and the nationwide day of action against hate speech.

Results of behavioural changes

Jugendstiftung Baden-Wurttemberg was the first organisation to set up a hate speech reporting office in Germany. Anyone can report hate speech using this portal, including non-German citizens. The organisation started with approximately 10 000 reports in 2022, and already in the first half of 2024 they had received more than 20 000.

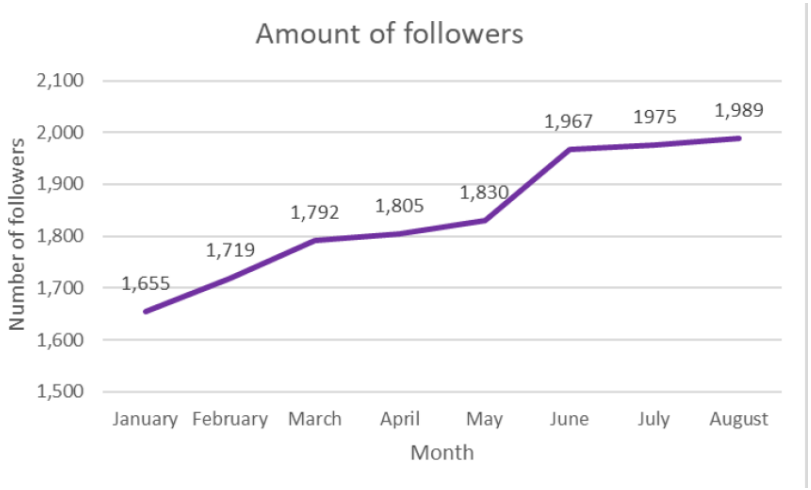


Figure 29: Distribution of CSO followers in absolute numbers. (2024)

As is visible in Figure 29, in the last two quarters (April-June and July-September), the period when Jugendstiftung Baden-Wurttemberg was collecting data for this Study, there was an increase from the first quarter. This could imply that behaviour change is happening because people are taking direct action when they come across hate speech incidents, and actively choosing to report them. Hate speech reports were consistently growing since the start of 2024; therefore, it appears that there is a positive audience engagement and behavioural change from their target group. Yet, this Study cannot conclude that this is

necessarily due to the online campaign, as the online component does not seem to garner high enough audience engagement with the content. Measuring what aspect of the campaign or external influence has led to such behavioural changes is a very complicated process. In this case, as the organisation within their campaigning did not make calls to action, this Study cannot attribute the increased reporting to the online campaign.

According to Figure 30, the peak in reporting was in the months of June, July and August. This could have something to do with the fact that at the end of May, Jugendstiftung released their press release and appeared in various media articles and posts, increasing their public visibility. However, the data from their online campaign does not show a direct link with the higher reporting. In the months of June, July and August, with exception of one post which received 1,130 likes on Instagram, none of the other posts had significantly higher likes or engagement. Furthermore, in their online campaigns, Jugendstiftung Baden-Wurttemberg spread awareness about responsible online consumption and hate speech more generally, instead of promoting their reporting portal.

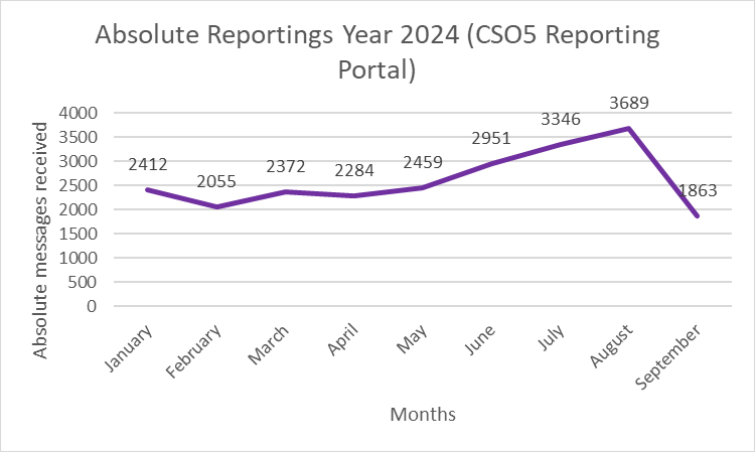


Figure 30: Number of users who report instances of hate speech on the CSO Reporting Portal. (2024)

Out of all the cases reported in 2024, approximately 11.89% were cases of antisemitism and within these, 47.54% came under criminal law. This implies that the users of such platforms have considerable knowledge of criminalised hate speech, as they relatively accurately report relevant cases under criminal law. Further studies could also look into how the users gained such knowledge.

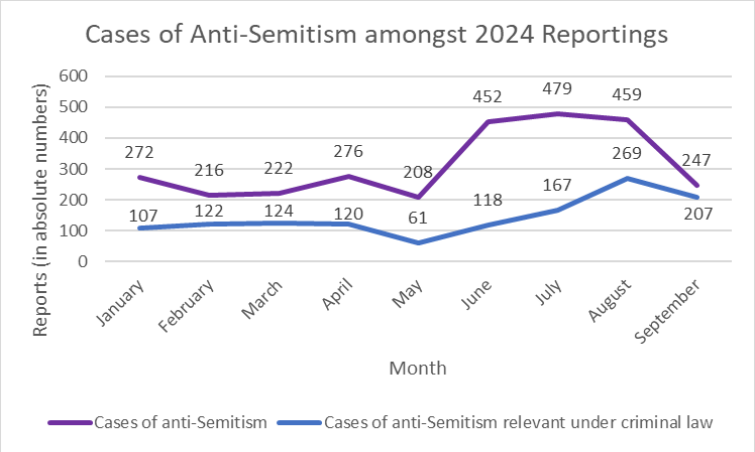


Figure 31: Cases of Anti-Semitism (and cases relevant under criminal law). (2024)

Furthermore, as part of this Study, Jugendstiftung Baden-Wurttemberg obtained the necessary funds to conduct a survey on the reporters. This first-of-a-kind survey was voluntary, and reporters could choose whether or not to fill it in once they finished reporting a hate speech incident through the portal. This survey was done in partnership with Technical University Munich and aimed to gather insights into the characteristics and motivations behind reporters using this reporting tool.

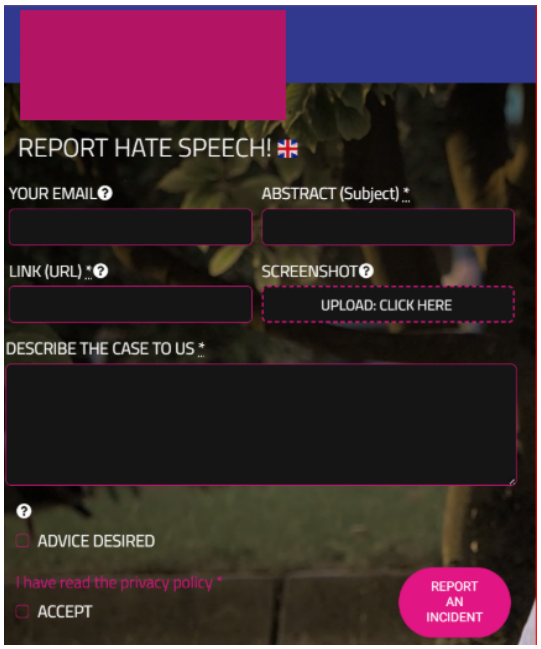


Figure 32: Form that users fill in to submit a report through the reporting website.

For this study the following are some relevant conclusions:

1. Most reporters come from Bavaria, a region where the organisation cooperates with local authorities and police stations.
2. The majority of users who use this reporting portal actively searched for such a portal through Google. Many of these people have heard about Jugendstiftung Baden-Wurttemberg, and therefore when searching are looking for them specifically. However, it is not because of the campaigning that they know of the organisation.
3. Motivation to report: according to the study, the majority of users do not report hate speech because they are directly affected or targeted by it. The majority of users are bystanders who report hate speech as a part of a civil reaction in order to raise awareness for third parties.

How this CSO responded to hate speech incidents

In terms of campaign criticism, Jugendstiftung Baden-Wurttemberg noted that they suffered considerably fewer direct threats and insults than in the early days of the reporting portal. They shared that during the first years, they received threats via social media, the post, telephone, and even physically. The latest incident was the publishing of the bio and private life information of one of its managers. This appears to have happened after the awarding of the 'Trusted Flagger' status they received. It was then that right-wing extremists used false information and made personal racist attacks on the director of the reporting office. As this was outside the period of data collection, this Study did not analyse the organisation's response to these attacks. There are potential consequences of such attacks which can include donors withdrawing their funding, the work against hate speech being jeopardised, and other employees facing personal attacks and threats.

CSO's main success factors

Jugendstiftung Baden-Wurttemberg is the only CSO working at the top federal level with the police in Germany. Their posters about their reporting portal appear in every police department in the country. They are also partners with the Ministry of Justice and Interior, and various other Bavarian authorities. These partners are considered as the 'perfect network' according to the CSO, as the organisation delivers information and files for criminal complaints, and the public authorities promote the cooperation and good practices of the CSOs with press releases.

Jugendstiftung Baden-Wurttemberg has a 'legal team', which checks if the hate speech reported is criminalised or not, and whether it should be sent to the police. Of the cases that they consider to be 'criminal', there is a 98% rate of acceptance by the police. Furthermore, once they submit a report to the police, the police provide feedback for them on the report. This contributes to their success by garnering increased credibility from the general public. Over the years, the barometers of the criminalisation of hate speech in German jurisdiction have changed. According to Jugendstiftung Baden-Wurttemberg, in 2007, only 10-12% of hate speech was considered criminal under German law, whereas in 2024, around 35% of hate speech was considered criminal under German law.

The Study is unable to determine the extent to which the online campaigns have contributed to the increased numbers of reports, however, it can be concluded that much of their visibility is due to their strategic partnerships and offline efforts. It is recommended that they consider using social media to reflect the strength of the reporting portal, and having an online campaign which would serve as a compliment to the tool. Having this as the topic for the online campaign could be more effective than trying to spread awareness on general hate speech topics. Alternatively, Jugendstiftung Baden-Wurttemberg could focus on exploiting further its offline capacities and spreading awareness through offline mechanisms.

Jugendstiftung Baden-Wurttemberg garnered extremely useful evidence-based data on the characteristics of their reporters and their motivations behind their reporting. This data will be extremely useful to tailor future campaigns, and to continue to increase audience engagement.

Main challenges experienced by this CSO

The main challenge for the campaign was to get in touch with young people as a target group. In the campaign, they do not need to advertise themselves as a stakeholder against hate speech as this is already recognised. However, they do recognise that they have a communication gap with young people in raising awareness about hate speech more generally.

With regards to their website, most referrals to the page came from search engines, followed by Facebook, www.bayern-gegen-rechts.de, X (with no activity of their own on the channel) and other websites. This shows that their Instagram and LinkedIn campaigns were therefore not very successful in gathering attention towards the website.

CSO's conclusions and future campaigns

Jugendstiftung Baden-Wurttemberg has been successful regarding their offline campaign and reporting portal. Initially, they considered a campaign to be a success if it gained over 10,000 reports per year, and yet just in the first half of 2024, they managed to double that amount. They should be proud of the results that they are achieving, as they have been successful in increasing the amount of reports by users, mobilising society and fostering behavioural change. Nonetheless, the months with the highest reporting on Instagram

do not necessarily have substantially higher audience engagement because of their social media campaign. Additionally, Instagram, where their social media campaign is majorly deployed, appears as the seventh source of referrals to their website, indicating that it is not raising visibility to the reporting portal.

For the next year, they will work more on the analysis of their campaigns and take measures to improve them. They plan to engage in strategic planning for their online social media presence with a focus on reaching young people.

Take-away 1:

An accompanying analysis of any campaign against hate speech is crucial and has to be implemented into any project from the beginning. It should be systemically included in projects against hate speech, following specific quality standards.

Take-away 2:

There is big potential in using the behavioural change metrics and the described correlation of data for the specialised task of reporting portals. This can be applied within the different reporting offices of Jugendstiftung Baden-Wurttemberg, and having country-comparable results on a standardised level.

Take-away 3:

Organisations which are experts in working against hate speech often do not have the capacity for marketing, but if they are planning a campaign, they should collaborate with professionals for the marketing.

CSO 6. APICE, Italy: Tackling Hate Speech

Table 12

Tackling Hate Speech	
Format	Online and offline: • Seven Mini campaigns (online). o 21 posts. o 102 Stories. • Three articles (Online). • Four Events (Offline). o One was at the National network against hate speech conference, where they shared their experiences and discussed ways of addressing future challenges.
Duration	5 February - 30 September 2024
Geographic coverage	Italy
Objective	These are the objectives for all the campaigns: 1. Raising Awareness on the importance of being safe online.

	2. Addressing the war in Gaza to talk about women's rights and intersectional feminism, as well as tackling double standards. 3. Raise awareness about the structural discrimination of Roma people. 4. Raising awareness of LGBTQI+ rights in Italy in the last year and on the meaning of IDAHOBIT[11]. 5. Deconstructing the narrative that hate speech is freedom of expression. 6. Countering the success of fascist mottos among young people in Italy. 7. Raise awareness about the forms of religious hate speech in Italy.
Topic that campaign addresses	This study analysed seven different campaigns on different topics: 1. Safer Internet Day Campaign. o Addressed the need to have a zero-hate internet and included a practical guide on how to deal with hate speech. 2. International Women's Day. o Addressed how to recognise human rights violations during wars, it spread knowledge on the sexist violence endured by Palestinian women and intersectional violence. 3. International Roma Day. o Portrayed stories of Italian Roma victims who died from tragic deaths and were ignored by the media and public. 4. 17 May International Day Against Homo-bi-transphobia. o Used data and humour whilst highlighting worrying discourse and political decisions amplifying homophobia and transphobia. 5. International Day Against Hate Speech. o Provided reliable sources on the right to freedom of expression and tackled the misconception that hate speech is free speech. 6. Reject Fascism, Embrace Caring.

[11] International Day Against Homophobia, Transphobia, and Biphobia

	o Countered the idea of strength being associated with violence and instead portrayed strength as a form of care. 7. Countering Religious Hate Speech. o Deconstructed the main stereotypes surrounding Islam and Muslims.
Target audience	Young people aged 18-35 based in Italy or of Italian origin (Italian speakers).
Platforms	• Instagram • Facebook
Working methods	The campaigns were all based on human rights and adopted the Council of Europe's materials for applying inclusive and intersectional approaches. The campaigns were based on a "dialogue approach", never provocative nor aggressive, but rather open to questions and respectfully confronted hateful narratives. APICE has been operating with the structure of producing one campaign per month since 2020. They engage with Italian volunteer activists aged 18-35 and co-create the campaign with them (nine activists on rotation). These volunteers take responsibility of certain aspects of the campaign design, deployment and monitoring. Online posts are released according to the calendar of international days of commemoration or awareness raising. Tools to collect data which were used in this Study include: • Instagram • Facebook • META Business Suite Tools • Feedback from activists through chat and this internal evaluation form
Successes and failures	Failure for this CSO would be to spread irrelevant messaging to young people. Successes vary for each campaign: 1. More young people being aware of the concept of online safety and the online environment. 2. Improved awareness on women's rights violations during conflicts.

	3. Develop more solidarity towards Roma communities in Italy. 4. Further awareness about the increase in homophobic and transphobic hate in Italy. 5. Counter the idea of hate speech as freedom of expression. 6. Offer young people a valid alternative to the violence of fascist ideology. 7. Raise awareness among young people so they can distinguish what is considered hate speech and what is not.

Results of online deployment

Within their seven campaigns, they deployed 27 posts in total. The majority of their content was in a carousel format, each post containing a graphic with text. APICE also published reels. The posts were deployed on Instagram and Facebook, although not the same format or the amount of posts were deployed equally in both platforms (27 posts were deployed on Instagram and 19 on Facebook). Their Instagram audience consisted of 72.8% women, 27.2% men, and on Facebook, 67.7% were women and 32.3% men. The aggregated findings for their posts (Seven campaigns combined) can be found in Table 13 below.

Table 13. Aggregated Results for the Tackling Hate Speech campaigns.

	Impression & Frequency	Reach (Volume per target group)	Likes	Saved content	Shares	Reactions or comments	Clicks on link
Aggregated Online Deployment	IG: 11,367 FB: 6,356	IG: 8,153 FB: 5,224	IG: 534 FB: 277	IG: 41	IG: 178 FB: 65	IG: 13 FB: 92	IG: 2 FB: 24

(IG – Instagram, FB – Facebook)

Overall, Instagram had on average 25.84% more impressions than Facebook 35.66% more likes than Facebook, and 92.69% more reposts than Facebook. However, there were more comments on Facebook than on Instagram. This confirms the trend that in terms of audience impressions, Instagram tends to surpass. But when it comes to audience engagement and users interacting with content, Facebook produces the higher numbers.

Their highest performing campaign was 'Campaign 5: Pensala come ti pare', and the worst performing campaign was 'Campaign 4: + Diritti = - Discriminazio'.

Their top Instagram performing post according to META Business Analysis was 'Post 2: Pensala come ti pare Reel 1'. This top-performing post was a Reel which interestingly only managed to get 31 'likes' between both platforms. Despite there being many audience 'impressions' (approximately 2,000 between the two platforms), it did not directly result in the audience engaging with the content.

According to the CSO, their online communications were less visible due to the EU elections and some national political developments, such as a new law being adopted, which are events of high visibility and engagement. These developments also impacted the motivation of the activists the last post of the campaign, even if it was previously prepared.

Their Instagram post, ´Pensala come ti pare Carousel 1´ had less reach but nearly double the amount of 'likes'. This supports the theory that for APICE, the audience visualises the reels but engages more with carousel content.

Figure 33: META Business analytics of the CSO best performing posts by reach on Instagram. (2024)

On Facebook, APICE's top performing post was 'Post 1: Diritti Bombardati Carousel 1'. Interestingly, according to APICE, this post received so much attention due to the comments from a user, which attracted more supporters and engagement.

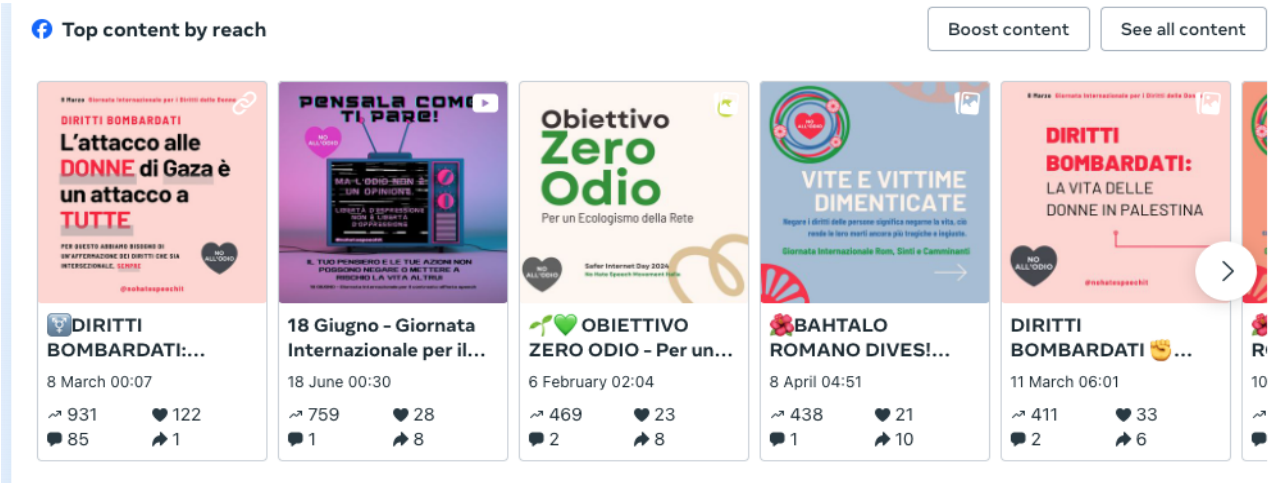


Figure 34: META Business analysis of CSO best performing posts by reach on Facebook. (2024)

Audience Engagement

In previous years the organisation would reach 150 likes per post. However, this year, their most successful post reached only 41 likes. This reflects a 72.67% reduction in likes from 2023. The APICE team think that this decline is because of the algorithm because often when they try to promote a post, it gets rejected, with Instagram classifying it as having 'political connotations' or being 'controversial'. Therefore, their campaigns

lacked the capacity to deeply engage and mobilise audiences. It also opened questions about how the algorithms promote or hide content.

An interesting observation is that for nearly every campaign, carousels had more likes, shares and comments than Reels. APICE clearly achieves better audience engagement with carousels and should therefore focus their efforts on this format. This seems to contradict the experiences found by other participating CSOs in this study, who have higher engagement with Reels.

The comments from users on their posts are generally very positive, finding the APICE's tools useful, reacting with positive emojis, resharing their posts, and praising the quality of the content.
During their campaigns, APICE gained 84 followers overall: 66 on Instagram and 18 on Facebook. The distribution of follower increase can be seen in Table 14 below.

Table 14. Increase in APICEs followers after each campaign.

Campaign 1 - Obiettivo Zero Odio	+10 followers (Instagram)
Campaign 2 - Diritti Bombardati	+10 followers (3 Instagram 7 Facebook)
Campaign 3 - Vite e Vittime Dimenticate	+11 followers (9 Instagram 2 Facebook)
Campaign 4 - +Diritti = - Discriminazione Aggregated	+13 followers (Instagram)
Campaign 5 - Pensala come ti pare	+18 followers (12 Instagram 6 Facebook)
Campaign 6 - Reject Fascism Embrace Caring	+7 followers (5 Instagram 2 Facebook)
Campaign 7 - Contro l'odio sull base della religione e del credo	+15 followers (14 Instagram 1 Facebook)

APICE does not share content on X but, after analysis, they have discovered that for issues such as gender, X as a platform provides space for more in-depth discussion.

As part of this Study, APICE was able to better understand the context of their campaigns and experiment with social media content. This experimentation included using the Sentiment Analysis within the WhoDis tool developed by Justice for Prosperity. They started using it in June 2024 for monitoring the campaign concerning the "Safer Internet Day". The WhoDis tool helped them to understand the spread, dissemination and reaction of specific keywords and hashtags within their campaigns. This was important for understanding how the audience engages with the content. The analysis also revealed platform-specific trends.[12]
The organisation found that using monitoring tools is beneficial for helping them to further understand the prevalence and spread of hate speech in social media, as well as to analyse the ethical implications for content moderation systems. In their experience, the social media monitoring tool SproutSocial does not comply with GDPR regulations nor employs a user-centred approach. Below are the main summaries of the sentiment analysis results.

[12] If any CSO wishes to receive more information about the WhoDis tool for AI detection and visualisation, please send an email to info@justiceforprosperity.org

- "Discorsi d'odio" (hate speech) as a term, often appeared in a negative context in accounts promoting a hate agenda. Whereas the hashtag "#noallodio" (no hate) was linked to positive, anti-hate campaign messages. This demonstrated that there is public support for hate speech countermeasures. This differentiation illustrates how specific language can shape the effectiveness and public perception of CAN campaigns.
- "Politically Correct" vs. "Hate Speech": The term "politically correct" generated more significant results than "hate speech". This suggests that public conversations on acceptable language and societal norms are more prevalent than direct discussions on hate speech itself. Peaks in mentions of the term "politically correct" typically aligned with controversies, often reflecting polarised views around free expression and inclusivity. Sentiment was split between those advocating for respectful language and those critical of perceived over-correction or "cancel culture."
- Mentions of "fascism" peaked during politically charged events, such as elections or controversies involving public figures linked to authoritarianism. The term was predominantly used critically, often as an accusation which reflects polarised political views within Italian social media.
- Keywords related to gender issues, such as "sexism" and "patriarchy," saw peaks in alignment with events like International Women's Day or newsworthy gender rights discussions. These terms showed a strong ideological divide: positive mentions were linked to feminist advocacy, while negative sentiments came from those opposing progressive gender norms. According to the tool's analysis, Instagram is particularly active in visual advocacy, whereas X provides a platform for in-depth discussions on gender politics.
- Peaks for LGBTQI+ terms corresponded with key dates, such as Pride Month and IDAHOBIT. This reflects public support or conversely, opposition. Positive sentiment around these keywords often signified solidarity and visibility efforts, while negative sentiment indicated resistance, especially from conservative groups. Instagram generally hosted community-focused support, while X included both advocacy and opposition, driven by influencers and public figures commenting on LGBTQI+ topics.
- The analysis also revealed platform-specific trends. Instagram's engagement favoured community-driven support and advocacy, with a younger demographic engaging in visual activism around gender and LGBTQI+ issues. Meanwhile, X featured more polarised analytical discourse, where ideologically driven discussions and political critiques were more prevalent. This platform distinction helps Justice for Prosperity and its partner CSOs tailor their outreach and counter-narrative strategies to maximise engagement and impact.

Results of offline deployment

APICE performed two offline deployments as part of their campaigns. They gave a presentation during a panel discussion for the European project "Stand by Me", led by Amnesty International Italy. The purpose of the activity was to discuss ways in which to prevent and react to online gender-based violence and hate speech. Around 40 people attended the activity including young people, professionals, educators, youth workers, activists, and representatives of organisations from Italy, Hungary, Poland and Slovenia. Amnesty International was in charge of promoting this presentation and the audience impressions were very positive and enthusiastic. During this Presentation, the organisation also discussed the META policies, as they perceive META to be uneasy about promoting politically linked content. According to this organisation's experience, CANs are internally flagged as political content by META's policy, reaching less people and often even mislabelled as hate speech towards public authorities.

The second offline activity was a Workshop on CANs based on Human Rights Education, called 'Discovering the Power of Human Rights Education in Contrasting Hate Speech'. This workshop took place in Strasbourg during the 'No Hate Speech Week'. By producing human rights-based CANs, the workshop introduced participants to the basic approaches of Human Rights Education and its role in countering hate speech. During the discussions, many participants expressed their anxiety and discouragement at the attitudes of social media platforms which they felt were amplifying hate speech during the ongoing global crises and wars. The participants expressed a general worry for the future and shared their doubts about how they could counter hate speech without proper political support. One positive aspect was that they appreciated the methodology and approaches for building CANs.

Results of behavioural changes

The organisation had an example of behaviour changes from one user. This person interacted through a direct message on Instagram on the topic of sexism and changed their tone from violent to talkative.

It is difficult to measure behavioural changes based purely on social media visualisations. From what can be seen from the data, their audience was less engaged on social media than in the previous year. It can be concluded that the organisation was not very successful in managing to change behaviours in the online spaces.

When combining all of their campaigns together, they gained 84 followers on Instagram and Facebook. However, their followers represent only a small percentage of the population, and the APICE is quite demotivated by how little they have managed to expand their sphere of social media influence.

How this CSO responded to hate speech incidents

The organisation shared that they do not often deal with substantial hate speech towards them or their campaign. During the study, they recorded two hate speech incidents.

Within the Campaign 2 - Diritti Bombardati they received a comment from a troll about 'women having no rights and being opportunistic, taking advantage of men'. The killing of a young woman, Giulia Cecchetin, in Italy, sparked great attention to issues of sexism and male violence against women. The troll's comment appeared publicly on one of the posts and was based on sources with links to articles and YouTube videos that shared fake news, stereotypes and manipulations. APICE engaged in a conversation with this troll through Instagram's direct messages.

During Campaign 6 - Reject Fascism Embrace Caring, the organisation published a screen shot from a documentary on 'Far Right Youth Groups Connected with the Government', which addressed far-right, extremism, xenophobia, racism, and antisemitism. This documentary mentioned offensive words, including very common fascist chants, and racist and antisemitic slurs. It additionally portrayed how these far-right youth groups often call for explicit threats or violence, including for the burning of black people, and the use of violence as a form of strength and resistance. They used this documentary as it is very widespread, with more than 500k views on YouTube and screenings in many Italian TV programmes. The reaction from the public towards this documentary has been quite polarised, with some criticising and condemning the journalists for simply revealing these movements rather than outright condemning it. As a response, APICE shaped their campaign to respond to the criticism and these issues.

CSO's main success factors

The high quality of their content, as well as the horizontal participatory process in the developing of their campaign with activists is a main strength and potential advantage of this organisation. Furthermore, they were successful in delivering offline deployments, such as training courses and presentations which received positive feedback from attendees.

Main challenges experienced by this CSO

APICE's staff found it difficult to follow the flow of events while doing research, collecting data and materials, producing graphics, attending calls, and other tasks. Additional frustration from the worsening of global and national political developments added a great toll on the general resilience of the activists group. They struggled to find the necessary energy and inspiration to commit to the entire duration of each campaign.

APICE outlined the changes of the past four years. For them, the expertise on hate speech at the national level in Italy has increased, which has resulted in the publication of more and better-quality literature on the topic. The appearance of more initiatives countering hate speech at different levels, for example, the National Network Against Hate Speech and Hate Phenomena, makes it more difficult for them to stand out.

They have not systematically reported the replays of Reels as Instagram stopped producing this data in 2024. Therefore, from July onwards they could not access the exact number of times some Reels were re-played.

APICE has noticed a significant decline in audience engagement in their 2024 campaigns, and they are quite concerned. They have linked some of this decline to the changes of Meta's publication policies for social and political topics.

For them, the extremely polarised societal context makes it difficult for their non-provocative human rights-based approaches to provoke. They conclude this because their two most successful posts had a more provocative style. However, too much provocation can lead to backlash and countering their own narrative. APICE expected that their featuring in the National Network against Hate newsletter in June would produce results and increase their audience engagement. However, this did not materialise.

CSOs conclusions and future campaigns

The strength of this organisation is in their in-person work, such as delivering trainings and presentations. For these, the feedback is positive, outlining them as valid and of high-quality. Potentially, APIOCE may need to focus more on their offline activities. This seems especially pertinent with the reduced audience engagement and lack of successful online campaigning. This will also lift the general atmosphere, addressing the need for motivating their activists.

Despite all the challenges, APICE was proud to have grown in this process by learning more about success indicators regarding CAN campaigns. They also increased their skills in using specific data collection tools, including the Spreadsheet and WhoDis Tool. The organisation is more aware of their strengths and weaknesses, and the need for a long-term, wider strategy on combating hate speech at the national level, based on offline activities and cross-sectoral cooperation. They also reflected on how social media's terms, conditions, and policies can influence their work and mental well-being. Finally, they built a wider network of organisation in Italy and Europe involved in this Study.

Take-away 1:

Divide the tasks for monitoring better and do not leave them for the end.

Take-away 2:

Do not be too demoralised by external factors. Even when it feels too much, try to find a way. Even though times are rapidly changing for the worse, look at the positive things happening around.

Take-away 3:

Trust the group and the process. When you have a participatory process, the magic keeps happening and the group can produce things everyone can be very proud of.

1.2 Spreadsheet - CSO aggregated engagements compared

Despite all CSOs operating in very distinct manners with different focus and target groups, they all combat hate speech online directly or indirectly. Table 15 below gives a comparison of their average engagements across the different platforms. For some organisations such as the Transgender Equality Network, the difference in engagement between Instagram and Facebook is quite stark, whereas for other CSOs such as APICE, this difference is not significant. Certain CSOs have more audience visualisations and engagements than others. Despite Instagram being quite a common social networking platform, not all organisations had the biggest audience impressions there.

In Table 15, it is clear that Instagram tends to be a platform where there are a lot of audience visualisations. Yet, these visualisations do not necessarily equal audience engagement. There tends to be considerably more comments on Facebook, at least for half of the organisations. For example, for APICE, there are 900% more comments on Facebook than on Instagram. Additionally, there also tends to be more sharing, reposting, and clicks on links on platforms like Facebook and LinkedIn than on Instagram.

This implies that Instagram is more useful as a platform for content to be seen and for the obtaining of impressions. But, if you want users to engage with your content or follow a specific call to action, your efforts must be deployed in a different platform, for example, Facebook. According to the analysis with the WhoDis tool conducted by Justice for Prosperity, there tends to be more in-depth political debate and discussion on Facebook and LinkedIn. On these platforms there is also more professional commenting, a high number of clicks on links, and conversations of more serious content (The Jugendstiftung Baden-Wurttemberg reporting portal in Germany had a diverging experience).

Regarding visualisations on Facebook, the highest number was reached by Thessaloniki Pride. What distinguishes Thessaloniki Pride is that nearly all their content is deployed in a Reel, song or video format. Based on this, it can be concluded that this type of audio-visual content achieves better impression results on Facebook than the posts with carousel graphics and text format. Additional research is needed to test this hypothesis.

Particularly useful here are the findings from the Jugendstiftung Baden-Wurttemberg respondent survey, which showed that TikTok and YouTube are the most popular platforms for daily use (Quint and Theocharis 2024 p.30). It would be interesting to further study if this is because these platforms have the most audio-visual content.

Table 15. Average CSO Engagement per post comparing across all CSOs performances.

	CSO1 Transgender Equality Network	CSO2 Rutgers	CSO3 ICEI	CSO4 Thessaloniki Pride	CSO5 Jugendstiftung Baden-Wurttemberg	CSO6 APICE
Impressions (Average per post)	IG: 5,336.94 FB: 612.06	IG: 9,250 LI: 14,368 X: 9,437	IG: 1,560 LI: 363.28 FB: 160.93 CSO: 37,918.4 YT: 19,726.2	IG: 182,782.44 FB: 387,149.66	IG: 2,289 LI: 43.06	IG: 421 FB: 334.54
Reach (Average per post)	IG: 4,705 FB: 589.31	IG: 2,950 LI: 14,368 X: 9,437	IG: 1,193.36 LI: 229.43 FB: 105.29 CSO: N.A. YT: 16,046.4	IG: N.A. FB: N.A.	IG: 1,642 LI: N.A.	IG: 301.97 FB: 274.95
Likes (Average per post)	IG: 208 FB: 30.67	IG: 238 LI: 244 X: 101	IG: 30.4 YT: 100.73 LI: 4.93 OSM: 222.73 CSO: 1.2 Press Release: 57.53	IG: 685.89 FB: 89.66	IG: 95.58 LI: 31.5	IG: 19.78 FB: 14.58
Comment (Average per post)	IG: 1.06 FB: 4.5	IG: 23 LI: 6 X: 19	IG: 1.21 LI: 0.21 FB: 0.21 CSO: N.A. from third parties YT: 0.4	IG: 30.11 FB: 91.33	IG: 4.46 LI: 1	IG: 0.48 FB: 4.84

Shares (Average per post)	IG: 29.34 FB: 9.94	IG: 0 LI: 24 X: 38	IG: 4.93 LI: 0 FB: 0.14 OSM: 36 YT: 0	IG: 6 FB: 13	IG: 0 LI: 2.88	IG: 0.48 FB: 4.84
Saved content (Average per post)	IG: 11.25	IG: 26 LI: 0 X: 2	IG: 3 LI: 0 FB: 0 CSO: 15 YT: 0	IG: N.A. FB: N.A.	IG: N.A. LI: N.A.	IG: 1.52
Clicks link (Average per post)	IG: 5.63 FB: 3.5	IG: 27 LI: 428	IG: No link LI: 54.21 18% average CTR FB: 0.14 CSO: 3,120 YT: 0	IG: N.A. FB: N.A.	IG: N.A. LI: N.A.	IG: 0.07 FB: 1.26

(IG – Instagram, FB – Facebook, LI – LinkedIn, X – X (formerly Twitter), YT – YouTube, CSO - Other CSO Socials, OSM – Other Social Media).

1.3 End of Study Evaluation - CSO aggregated results

1.3.1 CSOs Campaign Satisfaction and effectiveness

Figure 35 visualises the level of satisfaction amongst the CSOs regarding their campaigns. On average, considering 1 is very unsatisfied, and 5 is extremely satisfied, the CSOs are mid-satisfied with their campaign planning (3.25), campaign execution (3.75) and campaign results (3.00). This shows that the organisations acknowledge that the results are not negative, but not very positive either, which implies that changes would be welcomed for their next campaigns.

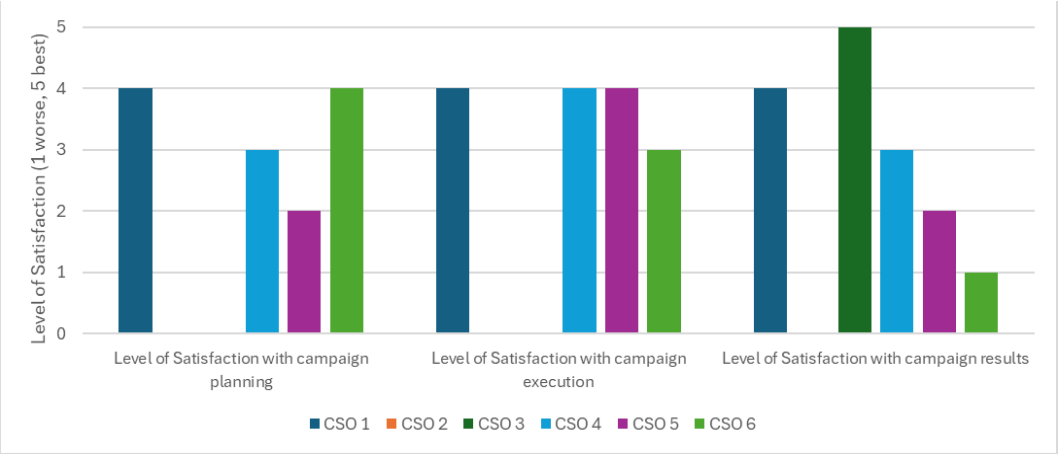


Figure 35: Level of Satisfaction Amongst CSOs with Campaign Planning, Execution and Results.

The most satisfied with campaign planning was CSO1: Transgender Equality Network, Ireland and CSO6: APICE, Italy. CSO1 believes that despite there always being room for improvement, they set out their goals clearly and built the campaigns towards these goals. CSO6 was content with their plan which was based on well-established practice. The least satisfied was CSO5: Jugendstiftung Baden-Wurttemberg, Germany, because their campaign was not built on a strategy, but consisted of "loose fragments not quite connected to each other".

Regarding the implementation of the campaign, all the CSOs answered between 3 and 4, in the middle of the ladder of satisfaction. The most variations amongst the CSOs' impressions are campaign results. CSO3: ICEI, Italy is very satisfied because their campaign played a crucial role in sparking a meaningful dialogue among young people about the language expressions used in everyday lives. Oppositely, CSO6 was extremely unsatisfied with the results of their online campaign.

Summary of factors contributing to the levels of satisfaction

When it comes to the level of satisfaction with campaign planning, the CSOs that were satisfied were the ones that had clear set goals which were then implemented throughout the campaign. Satisfaction can also be seen when a realistic plan was designed and based on established practices.

There was less satisfaction where there was no strategy, but rather different plan fragments connected to each other. When it comes to the satisfaction with campaign results, the organisations which experienced meaningful dialogue that was initiated in the campaign were the most satisfied.

As a conclusion, the following aspects contributed to higher perceived effectiveness:

- Creativity and simplicity of content. For example, using one sentence instead of a paragraph of text.
- Staff training and quality project management.
- Collaboration with a communication agency (professionals).
- Collaborating, engaging and integrating the target audience into the creation of the campaign.
- Collaborating with influencers.
- Non-divisive communication driven by hope.
- Engagement with the general public.
- Cooperation with law enforcement authorities and other public institutions.

Aspects contributing to lower perceived effectiveness:

- Not releasing the campaign at an appropriate time.
- No partnerships for dissemination of the content.
- Disconnected campaign elements, not having a solid strategy.

2. Lessons learnt from this study: answering critical questions

This chapter builds on the knowledge acquired by Justice for Prosperity through individual one-on-one consultation meetings with members of each CSO. It is also based on the qualitative self-critical evaluation reports submitted by the CSOs. This chapter provides insights on effective campaigning, such as choosing the campaign format, timing, platform specific considerations, proper communication channels, target audience, risk mitigation, and effectively measuring campaign impact and behavioural change when developing counter-narratives. Further insights include information on how CSOs can respond to the hate targeting their campaigns. It also advises on how to best engage with traditional media. The lessons learnt therefore address questions such as:

- What approaches and conditions should be met to make a CAN effective?
- How can a CAN be effectively promoted, used and managed?
- How can we adequately assess and mitigate risks by specific organisations or stakeholders before implementing a CAN?

Each CSO had a different profile target audience, campaign content, and modi operandi. This made any behavioural changes difficult to assess. Therefore, these research findings cannot give a universally clear answer to what extent CANs are effective in combating hate speech. Nonetheless, their results and experiences are valuable as lessons learnt and potential avenues for exploring the use of counter-narratives as a measure to combat hate speech.

2.1 Campaign effectiveness

CANs are effective when they mobilise the target audience to adopt the desired behaviour promoted by the message in the narrative. Below are some specific reflections on the effectiveness of CANs.

2.1.1 Educate, do not counter

From the experiences it seems that the successful CSOs are the ones that not necessarily counter hate speech, but rather educate through alternative narratives around the same topic or issue. For example, CSO1 which worked on trans healthcare rights in Ireland recommends that, “if you want people to be behind you, their literacy is important”. Similarly, CSO5 which worked on reporting hate in Germany did not campaign to increase the publics use of their reporting portal, but to raise awareness around the issues of hate speech and motivate civic responsibility. Perhaps one of the most notable examples was CSO3 who worked on deconstructing stereotypes and discrimination in Italy. Through their offline focus groups and interactions with young people, they managed to foster behavioural changes among them.

Focus groups provided the most marked differences in identifiable behaviours. Before being exposed to any campaign content, the members of the different focus groups shared feelings such as ignorance, injustice and rage. Whereas, after being exposed to the campaign content during the second focus group meetings, they expressed hope and optimism. After these focus groups, respondents even asked for more intercultural and dialogue activities, youth participation, and awareness raising campaigns. This was a clear message that the focus groups were a way to successfully engage the target audience.

2.1.2 Campaign format

Offline campaigns were especially effective due to the close contact they had with the target audience which created higher levels of trust. This enabled closer monitoring of any behavioural changes, whereas online campaigns were effective mainly as a complement to the offline campaign.

ICEI from Italy, Thessaloniki Pride from Greece, and Jugendstiftung Baden-Wurttemberg from Germany all argued that the best approach for campaigning was a combination of online posts and Reels, and offline activities such as presence in schools, and events with politicians and other stakeholders. As the online space is not reserved only for human rights narratives, the offline activities were where the organisation could engage with attendees, often resulting with the same or more impactful effects.

Regarding online campaigning, the CSOs agreed that the most effective format is Reels, as these garner higher reach and engagement rates. However, not all the CSOs could afford this format, which is why many used graphic posts instead. Nonetheless, there are disagreements as to whether carousels or content sprints are better. The Transgender Equality Network in Ireland picked Content Sprints for their campaign(.). Content Sprints cost much less to produce and work well when it comes to delivering complex messages. Alternatively, APICE from Italy chose carousels which they filled with information as they believed this was the best way to convey more information.

The content should be as short, catchy and vivid as possible in order for people to notice it. For example, Thessaloniki Pride from Greece developed 6 shorts Reels under their UberTalks concept to promote the EuroPride. They were short and fun, they engaged people, and they got the message across. Similarly, the Transgender Equality Network from Ireland signalled the importance of having “a catch” in the content, and preferably having the main message in one powerful and strong sentence. They also paid attention to the inclusion of vivid colours. Alternatively, ICEI from Italy used the POV (Point of View) narrative style for their graphic campaign, which featured typical language of Gen Z. Finally. It is also recommended to include a continuous positive messaging, as was described in detail in Section 2. This strategy was followed by ICEI, who shared non-divisive (pop) messages filled with optimism and hope. They considered this one of the success factors for the positive engagement with their target audience.

2.1.3 Campaign language

Using a specific language can shape the effectiveness and public perception of campaigns based on CANs. Using the analysis function in the WhoDis tool with data from APICE, Justice for Prosperity analysed the spread, dissemination and reaction to specific keywords and hashtags of APICEs campaigns. According to the analysis, "Discorsi d’odio" (hate speech) often appeared in a negative context and was frequently used by social media accounts promoting a hate agenda. Whereas the hashtag "#noallodio" (no hate) was linked to positive, anti-hate campaign messages, which demonstrated public support for hate speech countermeasures. Based on this, it is important to reflect on how the language used in a campaign can shape the public perception of it. Furthermore, posts which target high-engagement terms tend to have more visibility as these are favoured by the algorithms.

2.1.4 Campaign timing

Ideally, a campaign will take four to six weeks to plan. This is enough time to work through the creative process and ensure that all team members acquire the necessary skills.

The timing of a campaign depends on the intention it has. If the campaign has a reactive intention, then when something 'hot' or controversial occurs, the campaign should be started as soon as possible. If the campaign has a more awareness raising intention, then perhaps campaigning continuously throughout the year would be more appropriate. There is added value to linking parts of the campaign to special events that are happening locally or nationally. If a campaign intends to engage and seeks fruitful conversations, then there needs to be time and space for constructive dialogue. This means that it is not possible to deconstruct a topic when the topic is 'hot' or very current. Generally, before elections, the awareness of hate speech is higher so online campaigning should be intensified. This is something that Jugendstiftung Baden-Wurttemberg noted. Concerning the worst periods of the year to publish, all the CSOs agree that summer is not ideal as people tend to be away. Also, during the Christmas period, many brand promotions are competing with their content on the same algorithms.

Based on a digital marketing expert from the Transgender Equality Network, the period that generates the most engagement with content is during the morning commute. This is followed by the period of lunch time. Another peek time is at 17:00-18:30, again commuter time. Finally, there is a period after dinner that also produces engagement.

2.1.5 Platform specific considerations

Depending on the level of engagement that the campaign is striving for, an organisation can concentrate their efforts on one platform or another. If the aim is to have high visualisations and low awareness, Instagram or X might be the most interesting to explore. Whereas, if the aim is for high awareness, platforms including Facebook and LinkedIn, or even X might be more useful. This is based on the CSOs' experiences.

It is important to understand different audience engagement with content on the different platforms used. There are positive or negative outcomes and potential risks and criticisms with each platform. For example, despite the high audience interactions on Facebook, the platform produces polarised opinions. On the other hand, on X there appears to be more in-depth discussions and debates, especially with real-time political issues. Therefore, if the intentions are more reactionary, as explained above, then X might be an interesting platform to consider. Alternatively, LinkedIn offers a more professional networking platform, thus if a campaign has an engaging intention and seeks fruitful conversations, LinkedIn might be the best option as audience reactions will rarely escalate into hateful critiques. It additionally appears to be a useful platform to obtain high clicks on links. All these are aspects to consider before deciding on which platform to deploy the campaign content on. Rutgers from the Netherlands advises creating specific key messages for each medium, as well as for each target audience if there is more than one group.

Furthermore, whenever possible, CSOs should engage with Analytical Tools to understand the spread of hate speech around the topic that they are addressing and the platforms where it prominently appears. Knowing this, CSOs can tailor their content according to the platform used and be able to see which platforms are most relevant for the type of awareness raising they seek to achieve. It is also useful to see when certain peaks of hate or the use of certain hashtags emerge, and whether they are connected to any specific key dates, for example, international days. This analysis will serve as a strategic guide for crafting impactful messages that resonate within diverse online communities.

Insights from Transgender Equality Network's digital expert on how the algorithm works

According to a digital expert and board member of the Transgender Equality Network, algorithms select and choose the content to show to a specific user. This is because in general there is too much content and too

few people to see it. When publishing a new post, the platforms algorithm chooses a fraction of the followers to share the post with (each platform has different benchmarks for what an average post gets in terms of engagement). For example, when an organisation's content gets 10 likes out of 100 users, this becomes the benchmark. If a post only gets 5 likes, it is below average, and the platform will not diffuse it. However, if the post gets 30 likes, it is above average, and the algorithm will show it to more people. This is how a post can go viral. If the post is not successful and the followers do not like it, the algorithm platform will 'bury' the post.

2.1.6 Target Audience

When designing counter-narratives, it is important to distinguish between age groups. It is especially important to separate young people and adolescents from adults because they have different ideas about what is compelling. In order to convince and persuade a target audience to adopt a desired behaviour, the organisation must appear as a credible messenger. The Study results are not generalisable as they were based on Jugendstiftung Baden-Wurttemberg's survey of the reporters of hate speech on their portal in Germany. It appears that there are lower levels of trust in social media platforms and the internet in general than the traditional media, newspapers, and the legal and police systems. These findings directly contradict those presented in Section 2, so further research should be done.

A way to involve the target audience is by co-designing the campaign concept and message through a bottom-up approach. This approach was adopted by ICEI who involved young people in their DiversaMente project. Participatory approaches can represent a key strength for developing and fine-tuning the right campaign messages, and for getting relevant insights during the entire process of campaign development. They can also ensure that the emotions of the campaign, specific language, vocabulary, and platforms are relevant to the target audience. This increases the probability of higher visualisation, and it could even contribute to reducing backlash. For example, intergenerational teams will make sure that the humour used is not 'cringy.' Young people co-designing the campaign with adults, can warn them that some joke will probably not be funny to young people and so will not have the desired effect.

2.1.7 Risk mitigation

The organisations should be aware of those opposing their counter-narratives and target their campaign towards the most homogenous audience possible in order to avoid backfiring. Performing a stakeholder mapping exercise, identifying the allies and enemies to a specific cause can help in tailoring the communication activities. Additionally, researching examples of how campaigns have backfired can help to mitigate those risks for their own campaigns.

Investing in crisis preparedness and response is key. CSOs should have a crisis management plan, with enough financial and human resources to manage any such responses. Investing in partnerships can be effective for gaining access to resources that an individual organisation may not have. CSOs also recommend maintaining close cooperation with the police to stay further protected whenever incidents emerge.

2.1.8 Meaningfully evaluating CAN effectiveness

To evaluate the effectiveness of the CANs in combating hate speech, it is imperative to measure the campaign's impact. An impact could be if the campaign has stimulated or achieved behavioural changes amongst the population. This is complicated to estimate unless there are examples of people mobilising or

changing opinions based on the campaign. Another idea to assess is if these changes are due to the campaign being impactful or other external factors.

Behavioural change

Behavioural changes can only be seen when the population mobilises. Yet, this is difficult to measure online due to the general lack of interaction with the target audience. To measure behavioural change, impressions or likes are not enough.

It is easier to measure behavioural changes in the offline space, for example, ICEI conducted focus groups and workshops with young people. These were very relevant and important, and helped them to determine whether the young people had changed their ways of thinking and interacting with hate speech. Additionally, by directly interacting with the young people, the CSO can understand how the change has happened and whether it was due to the content they were exposed to through the campaign, or due to other external factors, for example, a political context. Without interacting and engaging with the target audience, it is extremely complicated to see whether there has been an attitudinal or behavioural change.

To effectively measure behavioural changes in the online environment, there should be a direct call to action. If the audience follows this, it is evident that there has been change. For example, if a post invites people to attend a conference and it receives 500 likes but only a few people come, this is not impactful. Clicking like is not an impact indicator. What is needed is for the audience to engage with the post by responding to the call to action and attending the conference. The same applies for cases like reporting hate through online portals. An example here of behavioural change is when the number of reports of hate speech consistently grow, especially in comparison to previous campaigns.

Combining quantitative and qualitative indicators

To measure a campaigns impact, there are different indicators and platforms that can be used. This was done in Justice for Prosperity's data collection spreadsheet and end of Study evaluation report. By monitoring quantitative platform-specific data, it is possible to compare between audience impressions (number of views) vs audience interaction and engagement (likes, shares, reposts, commenting, etc.) with the content. Indicators such as comments, clicks on links, reposts/re-shares, and likes are appropriate estimators on whether a CAN has been effective because they imply that the audience has engaged with the content. Quantitative indicators monitor audience engagement across different platforms carefully and compare the results with those of previous years. This enables the identification of when and on which platforms the campaign was more successful and allows the organization to see variations across platforms as well. Thanks to such close monitoring, Jugendstiftung Baden-Wurttemberg realised that, despite their online campaign being mainly deployed on Instagram, the majority of users who heard about them through social media did so through Facebook and X. As an organisation they do not use Facebook, and they only used X via their partners' accounts. They are now considering using Facebook and X for future online campaigns.

For qualitative reflection and evaluation of a campaign, tools such as open question surveys, individual interviews, and focus groups can be used. These multifaceted evaluation formats allow for the assessment of an individual or groups' possible behavioural changes and/or beliefs.

Although it is ideal to be in close contact with the target audience throughout a campaign and during an evaluation, it is not always easy or possible. For these cases, the sentiment analysis method is especially effective. The method consists of collecting data on the internet through an automation mechanism.

Subsequently, Natural Language Processing can track the increase and decrease of hate speech online, both before and after a campaign. Tools such as the WhoDis Tool can conduct such analysis.

Audience monitoring

By monitoring their audience closely, CSOs can study through which means and platforms they reach the most people. For example, APICE is very well known among teachers and young educators for their excellent quality tools, but they are not very popular on social media. Knowing the organisation's assets and strengths is a key element to make a counter-narrative successful and impactful. Additionally, knowing which platform the audience uses to interact with the CSO helps them to adapt their campaigns better to the online and/or offline environments.

Additionally, CSOs should also analyse the differing levels of engagement of their audience. With this knowledge, they can target their campaigns accordingly, and potentially create different messaging for each level of engagement. Such an analysis was done by Jugendstiftung Baden-Wurttemberg, who asked their users how many times they had used their Reporting Portal. This allowed them to know how often their tool was used by the users, and the percentage of the audience with high, medium, and low engagement. This is only possible with a direct call to action that will measure the desired concrete behaviours.

Justice for Prosperity's Spreadsheet (Quantitative) Monitoring Tool

Justice for Prosperity designed a monitoring tool based on 13 Key Performance Indicators (KPIs). This tool is composed of seven chapters and over 130 data inputs. As part of this Study, the CSOs were asked to share their experience using the tool.

Figure 36 shows the potential of the Spreadsheet which was appreciated by the CSOs. The organisations agreed that the tool could be used in the future, despite the fact that it consisted of many data collection points, causing it to not be very straight-forward

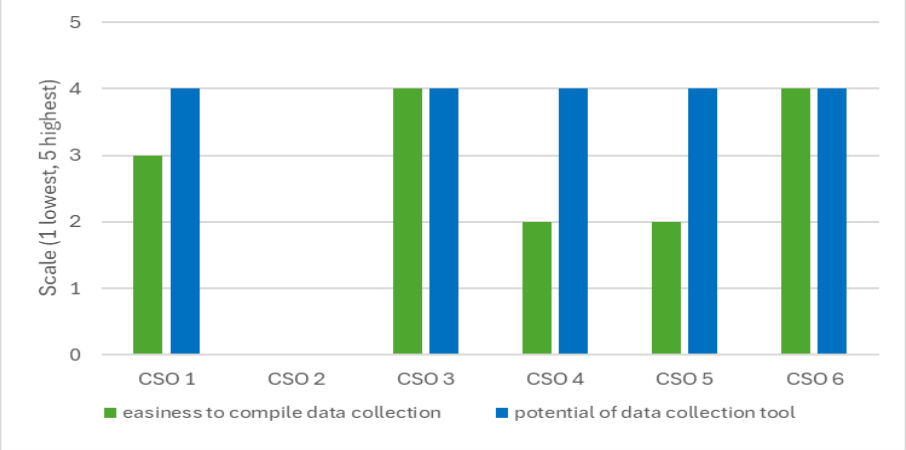


Figure 36: CSOs' perception of easiness to compile data collection, and potential of the tool.

APICE shared that despite the explanation on how to use the tool being "very clear and straightforward," the Spreadsheet itself initially presented an obstacle since it introduced the group not only to a new task, but to a new perspective on the creation of the content. Yet, this new perspective stimulated critical thinking, became a source for 'competence development,' and led to improved outputs. Additionally, the Transgender Equality Network also complained that compiling the data took longer than expected and was burdensome.

For the ICEI, the complexity of compiling the detailed data collection increased as they partnered with the media. The problem here was that these partners only shared the most basic data.

For their next campaigns, the CSOs have all expressed willingness to continue to use such a data collection tool. They appreciated that the tool made them actively think about parts of their CANs that they had previously 'mechanically overlooked,' and that it stimulated conversations about what else could be done. The CSOs proposed to scale the tool down to the specific indicators that they are able to produce and that are relevant according to their type of campaign. Additionally, adding a drop-down menu could improve the tool's accessibility.

2.2 Fostering behavioural change

Based on the experiences of the CSOs participating in this Study, offline campaigning and deployments were more successful in fostering behavioural changes than online campaigning. Online campaigns tend to be a mechanism to spread awareness and increase visualisations of issues and narratives. However, the online campaigns failed to foster the same level of behavioural changes as offline campaigns. They might therefore be seen as more of a complement to the offline campaigns, where the core efforts must be concentrated.

For example, if Jugendstiftung Baden-Wurttemberg would have only spread awareness about their online reporting portal through their social media channels, considering their current levels of audience impressions and engagement, their portal would not have been very popular. Jugendstiftung Baden-Wurttemberg was so successful due to their key collaborations with law enforcement authorities that put posters about their reporting portal in police stations in Bavaria. Furthermore, they are mentioned in many press releases with state cooperation authorities. All this offline deployment gives visibility and makes the CSO and their portal known, ultimately bringing success. Their social media campaign is therefore a compliment and not a central part to their modus operandi. A staff member from this CSO said bluntly that "it is clear that the campaign work is [an] 'add-on' to the work of the reporting centre and has hardly had any effect." It would of course be beneficial if they chose to invest more in social media, but they should not make it the centre of their efforts and funding as it has proven to be substantially less successful than all their very impactful offline work.

Similarly, APICE has not been very successful with their awareness-raising campaigns on social media. However, the organisation could be considerably more impactful if they changed their modus operandi to more educational offline or hybrid workshops, presentations, or initiatives. CSOs do not necessarily need to have their entire modi operandi revolving around online campaigning; they can be equally or even more impactful if they deploy their efforts in person-to-person activities for the population. These activities, if attended by the population, can result in behavioural changes and provide indicators of success. This CSO could use its social media campaign to give further visibility to their created tools and offline deployments.

2.3 Strategic partnerships for your CANs

Investing in social media and digital expertise for a CAN campaign significantly increases the campaign's visibility, as organisations will gain insight into how to better use social media algorithms and apply persuasive communication models. Investing in digital expertise can take different shapes and forms. One approach is to train staff in social media skills and build social media infrastructure within the CSO. This approach was taken by the Transgender Equality Network who used paid promotions to form a larger audience and made a modest investment in advertising. Another way is to work with a communications

agency to co-create the campaign, an approach taken by both the Transgender Equality Network and the ICEI. If organisations have the funds for it, another avenue to explore would be to work with influencers, especially on the development and promotion of the CSOs' campaign. Based on the ICEI's experience, working with influencers is beneficial, especially because they spend much of their time working with young people. These influencers are well known which makes them experts on audience engagement. Influencers can also support organisations in taking the right social media approach. Furthermore, their appearance directly in the campaign can contribute to increased audience engagement and help it potentially to go viral. The influencers or other relevant third-party media can help reach the target audience by disseminating the campaign in their own networks and to maximise coverage. However, this approach has several challenges for the CSOs, including the need for budgets to be allocated in order to pay the influencers and other media services, and the complexity of collecting detailed quantitative performance data on the part of the campaign being run on third-party communication channels.

Many CSOs participating in this Study had staff, volunteers, communication agencies, and/or influencers working for them pro-bono. CSOs, especially those with less financial resources, should look for such pro-bono opportunities as a way to invest in digital expertise.

Furthermore, based on the findings from the Transgender Equality Network, partnering or having closer links with public authorities and governmental bodies results in more public acceptance of the organisation and its mission. This partnership could lead to a higher level of promotion when the cooperating partners speak about the organisation and campaign in press releases. Likewise, Rutgers argues that being featured in the mainstream media can also be positive, as it gives the campaign public visibility and allows the organisation to choose how their work and campaign is framed in the media.

2.4 Receiving criticism - Human Rights narratives can provoke replies, questions, challenges, and even more hate speech.

Even if it might be perceived as common, not all the CSOs participating in this Study faced criticisms during their campaigns. For example, the Transgender Equality Network did not face any backlash, and they argued that this was because the users who saw their campaign were also their allies. It may be that criticism arises as the CSO continues to grow and their content appears on a broader range of users' feeds, including that of opponents.

Receiving criticism may be influenced by the organisation's popularity and how long it has been operating. For instance, Jugendstiftung Baden-Wurttemberg shared that when they started their organisation, they received quite harsh opposition on X, but when they partnered with the police, the personal threats and emails significantly reduced. However, since they were granted the Trusted Flagger Award in October 2024, hostilities against them have grown exponentially, mainly through right-wing extremist disinformation accusations. This indicates that when a CSO receives a sudden high level of exposure, criticisms against the organisation and its staff increase.

Receiving criticism may also be influenced by the type of narrative the CAN is trying to promote. Thessaloniki Pride claimed that if their narrative was to pass legislation instead of solely defending a human rights principle, the amount of hate speech received would grow.

To reduce the amount of criticism on the campaign, Rutgers recommends having a non-controversial theme, so it is less prone to opposition. As part of their stakeholder support strategy, they created support for the campaign audience by developing a website targeted directly at parents, their most important stakeholder. Additionally, they sent newsletters to schools and partner organisations.

2.4.1 What are the best strategies to respond to replies containing hate speech?

It is crucial to recognise when to engage with criticism and when to refrain. According to the Transgender Equality Network and the ICEI, engaging works best when the comments are open to discussion and reflect misunderstanding, rather than outright hostility. When comments are openly hostile, it is better not to engage with such blatant hate speech because it is intended to provoke and/or incite further negativity. Thessaloniki Pride and APICE additionally add another distinction between young people who are still likely to change their minds and radicalised adults.

Once organisations decide to engage with criticism, there are different approaches. Jugendstiftung Baden-Wurttemberg employs a clear description of their activities when professionally answering news and media inquiries, and maintains a 'neutral informative tone,' attempting to calm down instead of 'feeding' the troll. They believe that when there is a wide range of accusations, organisations must react immediately with a lot of energy and in a sensible manner. The ICEI engages in conversations that highlight common ground and attempts to transform hostile exchanges into opportunities for education and dialogue. Similarly, APICE tries to interact with the 'troll' and counter their statements. The Transgender Equality Network and Thessaloniki Pride adopt somewhat more radical approaches, the former with a 'policing and comment-eliminating' approach, and the latter by taking legal action whenever the hate speech includes threats and violence. The Transgender Equality Network is only involved in Facebook and Instagram, as they have control over the platform and can delete hateful comments, whereas they are not active on X as they cannot do this there. Finally, Rutgers has a Crisis Management Team which acts depending on the situation. Sometimes this Team does not respond, so when things escalate, Rutgers often involves the authorities. This Team is now trying to anticipate pushback beforehand by including the social media approach in advance so they can adjust their theme and approach.

2.5 The Media

2.5.1 The Media's Interests

Based on conversations with Jugendstiftung Baden-Wurttemberg, the media is often interested in specific results from an action (i.e. the end of the process, which in their case is the criminalisation of hate speech offenders). When someone goes to jail or has to pay a fine for the hate speech they spread, this attracts interest and is likely to get published in the media. Despite Jugendstiftung Baden-Wurttemberg contributing to the reporting of criminalised hate speech through their reporting portal, once they pass a report to the police, they do not know what the end result of this report will be. This creates a gap between what the media is interested in and the information that the CSO can provide to them.

2.5.2 Engaging with The Media

As part of this study, Rutgers shared some of their decade long expertise in engaging with the media and dealing with opposition.

Rutgers positions themselves as a centre of expertise in the domain of sexual and reproductive health and rights. Part of their role is sharing knowledge and informing public debate about these issues through the media. Additionally, by engaging with the media, they ensure that correct information reaches their target audience. By expanding their media network, they can also influence policies and political discussions on sexual and reproductive health and rights issues.

They believe that it is good to proactively approach the media, especially to share results of research or before launching a large public campaign. However, there are also benefits of reactive press information, as it allows them to see which media channels are interested in their organisation. To minimise any risks whenever the media reaches out to them, they always check the context in which the article featuring their organisation will be written, and the actual reason behind the media wanting to engage with their organisation.

Concerning which media to engage with, they argue that both social media and traditional media (newspapers, radio, TV) are important. This CSO additionally values a diversity of media, national and regional media, and media that reaches different target groups. This is different from the recommendations provided in Sections 2 and 3, which propose reaching out to the most homogenous target audience possible. Rutgers is selective and chooses not to work with media who does not apply the principle of 'Audi Alteram Partem' (listening to the other side). This principle is based on the notion that no person should be judged without a fair hearing in which each party is given the opportunity to respond to the evidence against them. This is along the same meaning as the two-sided message that was explored in detail in Section 2.

When the CSO engages with the media, they make statements about facts and topics of their expertise. Additionally, they never make assumptions or statements about areas outside of their expertise, and actively reach out to journalists to make sure that they have the facts right. Furthermore, before anything is published, they agree in advance to see the written article, checking for factual inaccuracies. For TV recordings, they ask if it is possible to read the script and/or actively contribute to it, and for TV interviews or radio appearances, they have a preliminary interview with the editor and ask them for the questions in advance.

The organisation considers timing important. They give big interviews well before their campaigns, ensuring that they have enough time to deal with any pushback before the official launch. By doing this, when their campaign starts, they are able to focus more on the content and less on the media. However, they recognise that some topics can generate resistance no matter what.

To avoid disinformation being spread about their 2024 campaign, Rutgers revealed the disinformation that was used against their previous campaign. They explained how it was incorrect and provided the accurate information. This enabled people to realise that what was said about their campaign was not correct.

They claim that the more media that the organisation has on their side, the more positive public perception can be obtained.

3. Key Take-Aways and Recommendations

This summary provides the main recommendations from Section 3, based on the quantitative and qualitative results, and lessons learnt from the experiences of the six participating CSOs. There are specific online campaigning recommendations, as well as recommendations that apply to both online and offline campaigns. To prove the scientific validity or causality of the claims which are be made below, further research involving more CSOs and with a prolonged period of time is needed. The authors of this study will therefore exclusively say what the CSOs experiences have indicated, but will not establish a scientific correlativity.

3.1 Key Takeaway

Just because hate speech is mainly spread online does not mean that the most effective way of countering it is online. Unless the CSO is a digital marketing or social media expert in online campaigning, in person campaigning efforts can be more impactful for fostering behavioural changes.

This key takeaway is based on the principles mentioned in Chapter 5 about behaviour change, and how in order to foster and measure behaviour change, an organisation must either 1) physically interact and engage with the target audience, or 2) make direct calls to action in online campaigns. In the first instance, physical interaction at different stages allows the CSO to see and hear how opinions and behaviours have or have not changed. In the second instance, making a direct call to action means being able to measure whether people respond or not to the call.

True human interactions and appealing to emotions is what seems to matter most. CSOs must make calls to action to humanise individuals and encourage interaction with one another. Hate speech is present online, but some campaigning strategies can be focused on the in-person (offline) activities. For example, it tends to be more impactful to promote human rights through in-person activities or screening movies rather than through an online post. In-person activities and human interactions include discussions which can be meaningful and sustained, providing opportunities for the opinions of the attendees to be challenged.

The ICEI is doing this with young people through focus groups in which they talk about certain stereotypes and discrimination and share emotions and opinions about how it makes them feel. Additionally, they use slang, memes, and references that resonate with the experiences of their target audience - young people - which helps them connect better. Similarly, Thessaloniki Pride were particularly proud of their Pride march, where they doubled the expected participants and stimulated emotions even in the elderly, who danced on their balconies as the parade passed by their houses. This mobilisation of society is not only an impactful, but also produces greater self-satisfaction for the CSOs than posts or likes online.

3.2 Specific Online Campaigning Key Take-away

Takeaway 1:
Online campaigning might be more impactful if done with a digital marketing/social media expert or influencer. Investing in social media and digital expertise for a campaign can significantly increase its visibility. There are also benefits to be had from theories and practical tools on these topics.

Takeaway 2:
When designing online campaigns, it is important to understand platform specific differences, to adopt a bottom-up approach, and to monitor the peaks of hate speech on the different platforms.

Takeaway 3:
The campaign must include short, catchy, vivid and positive content for people to notice it and engage with it more.

Takeaway 4:
Campaign effectiveness can be monitored by measuring behavioural changes, platform-specific engagement, and the audience's differing levels of engagement with the content.

In order to monitor behavioural changes, adding a call to action in the messages for the target audience is crucial. This can show whether and how often users are engaging with the call to action and/or adopting the desired behaviours.

Additionally, keeping records of a previous year's performance is recommended. This will provide yearly comparisons so the CSOs can see whether they are gaining or losing momentum and success.

It is also interesting to research how the target audience find out about the CSOs. From this, it can be seen whether it was through the social media channels. If it was through social media, then these are the spaces where the content is gaining visibility.

Takeaway 5:
The campaign is only being viewed by allies is not necessarily a negative thing. The Transgender Equality Network's content is predominantly viewed by their allies. Their audience engagement tripled in 2024, and they are not receiving any criticism. Do not be afraid to remain within your allies' bubble, it also comes with its positive side!

3.3 Both Online and Offline Campaigning Take-Aways

Takeaway 1: A mix of expertise and experience is key.
Expertise is crucial when it comes to developing campaigns to combat hate speech on social media. The higher the quality of the campaign content and the higher the social media planning and design are, the more likely the campaign is to be successful. Experience means being able to prepare to mitigate risks and have resilience when a narrative backfires. It also helps to create campaigns which take into account previous monitoring, evaluation cycles, and lessons learnt.

Takeaway 2: Specific language can shape the effectiveness and public perception of CAN campaigns.
It is important to reflect on the way in which language will be used in the campaign. The language can shape public perception of the campaign and the organisation.

Takeaway 3: Educate, do not counter.
The most successful CSOs were the ones that did not counter hate speech, but rather educated through training and raising awareness on alternative and positive narratives.

Takeaway 4: Partnerships can considerably lead to campaign success.
It is recommended that CSOs form partnerships with ally organisations, individuals, and/or governmental bodies. This is especially important if the target audience trusts these bodies, as this could potentially increase the chances of the campaign reaching higher engagement and receiving more public acceptance.

Takeaway 5: Reflect about possible risks and invest in crisis preparedness and response.
CSOs should develop crisis management strategies and create a crisis management team where possible. It is important to invest in crisis preparedness and to take as many precautions as possible in order to be ready for when it is needed.

Takeaway 6: Just because things have always been done a certain way does not mean they have to always be done that way, especially when they do not produce results.

APICE staff classified the situation as 'depressing' when it came to their high efforts and low results. The activists did not want to give up their way of doing things and kept with the same approach. Yet their efforts were not being reflected in social media audience engagement. However, they did manage to create long-term positive effects through a different type of intervention: educational work. This has become their strength and a direction where they may be much more impactful and obtain more self-satisfaction for their members. It is never too late to change directions. It is important to think outside the box and to continuously experiment with new approaches to see what works and what does not.

3.4 Future research recommendations based on the findings from this Study

Further studies should continue to be supported so that stakeholders combating hate speech can continue to grow their expertise and pool of knowledge about how to most effectively use counter-narratives and how campaigns are most effectively developed. In doing this, they can achieve increased audience engagement and foster behavioural change.

Further research needs to look specifically at the risks and potentials of using CANs. As mentioned in Section 2, the studies of CANs often focus on how to combat radicalisation and prevent violent extremism. Whilst lessons can be transferred from one field to another, violent extremism and hate speech are different in character, and therefore there is a need for dedicated studies that focus on hate speech exclusively. Currently existing literature is lacking such studies, with the exception of this Study.

Furthermore, future research and campaign development need to pay attention to platform-specific preferences and academia. CSOs need to generate more knowledge on the types of audiences and audience interactions amongst the different platforms. Targeting high-engagement terms and understanding platform-specific preferences are crucial for NGOs to effectively shape CANs to combat hate speech. This will allow CSOs to do tailored platform campaigns and choose the platforms where they wish to deploy their content depending on the type of awareness raising and audience interaction they want. More research is needed to look at the potential of combating online hate speech through offline deployment and actions. This Study has shown that offline actions have more impact than online actions. Appealing to emotions is important when physically engaging with the target audience. Making direct calls to action can support the measuring of any behavioural changes. Further research needs to study the hypothesis that 'hate speech majorly emerges and is spread online but can most effectively be countered offline.'

Even if the study results based on Jugendstiftung Baden-Wurttemberg's survey of the reporters of hate speech on their portal in Germany cannot be generalised to other CSOs, it appears that there are lower levels of trust in social media platforms and the internet than the traditional media, newspapers, and the legal and

police systems. These findings directly contradict those presented in Section 2, so further research is important here.

Future research should engage in a longer-term project and should also study more than six CSOs to have a bigger sample and be able to develop more reliable and scientifically valid conclusions.

Finally, future research should also explore the role of content-moderation, AI, and social media algorithms, as the CSOs expressed major interest in understanding the variations better.

Concluding Remarks

This Study has spent eight months observing six different CSO across five EU countries. This has enabled Justice for Prosperity to make evidence-based recommendations on how to create effective CANs and measure the effectiveness and impact of such campaigns.

CANs appear to be the most effective when they directly engage with people, preferably physically, but if not, then with a strong emotional appeal, direct communication, and calls to action with the audience. CAN campaigns must be planned and take into consideration particular platform-preferences, including the type of audience and engagement that there tends to be on that specific platform. They also must be willing to adjust the messaging according to the platform that the content will be deployed on. Using narratives which use persuasive communication theories (namely Transportation, Identification or Parasocial) and are emotionally compelling will make a CAN campaign more persuasive and reduce reactance. However, specific emotions such as humour, fear and regret must be used with caution, as discussed in Section 2. Nonetheless, as shown with the ICEI's experience, using emotions can also be extremely helpful. The use of humour and irony in the ICEI's messaging allowed them to engage the audience (young people) in an entertaining and thought-provoking way, fostering critical reflection without being overly didactic. This combination of elements, if used properly, can inspire lasting change.

Additionally, it is often effective to incite emotion and encourage the audience to identify with a character in the narrative, creating a relationship of trust between the audience and the character, and therefore making the narrative easier to understand. Sometimes, to create greater proximity with the audience or increase the level of trust with the target audience, it is useful to partner with a third party which the target audience trusts (for example, an influencer, the government, law enforcement authorities, etc.). Additionally, developing campaigns with social media experts can also be particularly useful, as they know how to engage audiences and influence them.

Furthermore, it is advisable to keep a generally positive tone in messaging. By emphasising positivity in narratives, one creates an environment that encourages people to participate in constructive conversations rather than resorting to negativity.

It is recommended that campaigns are planned over a period of 4-6 weeks. This should provide adequate time for the creative process and the skilling up of staff to occur. Furthermore, online campaign posts or Reels should be deployed in moments of highest audience engagement, including in the morning commute, followed by 11:00 lunch traffic. The sweet spot is between 17:00 and 18:30, the period of commuter traffic. High engagement is reached again in the evening from 19:00 to 20:00.

When it comes to receiving criticisms, it is advisable for organisations to have a crisis management team who can decide whether to engage with the criticism or not. If they decide to engage with the criticism, different approaches can be taken. The following are some ideas performed by the CSOs participating in this Study: present a clear description of the activities, engage in conversations that highlight common ground, try to interact with the Troll Account and counter their statements, adopt a 'policing and comment-eliminating' approach, or, if the criticism escalates, take legal action.

From this study, it has become clear that there is not one perfect methodology to measure whether CANs are impactful. This methodology must be open, flexible, and able to adapt to the different CSOs' modi operandi. Yet, Justice for Prosperity's (holistic) methodology, has proven successful in measuring effectiveness, even if there tends to have been a lack of (online) success amongst the CSOs. Furthermore, the methodology was also effective by combining quantitative and qualitative elements and gaining extremely detailed insights into platform-specific differences regarding audience visualisations and audience engagement. Additionally, through this methodology, behavioural changes were measured for some of the CSOs, demonstrating that the tools employed for this Study have potential. Finally, this methodology also fostered critical thinking and self-reflection amongst the participating CSOs, who at the end of the campaign thought deeply about lessons learnt and avenues to explore for following campaigns.

Justice for Prosperity therefore managed to get each participating CSO to monitor and evaluate all their campaigning processes, from designing, to deploying, to monitoring the success of the campaign, making it an extensive and holistic campaign impact evaluation process. This has proven successful in creating a more complete and systematic monitoring tool, which the CSOs can now incorporate into their institutions. For some of the CSOs including the Transgender Equality Network, participating in this Study not only contributed to improving their monitoring and evaluation cycles, but also to creating institutional knowledge on social media and digital markets. Participating in this Study has therefore been a considerable transformative process for all the CSOs involved.

However, this study must not finish here. The spread and rise of hate speech is still increasing exponentially and is creating major harm to those it targets. Research must still be done on the issue of hate speech specifically, to find even more innovative and effective ways to impactfully combat it and make the world a better and safer place for all. Further support from EU and international institutions is needed to continue to fund projects such as these. Financing is needed for more research so that organisations can continue to discover more successful ways of addressing hate speech, and for the building of capacities to support the fight for human rights.

This Study was a pilot study taking place within a European Union and Council of Europe joint project. It is hoped that many CSOs and Human Rights defenders may benefit from it, and that together we make the world a safer place for all. And remember, even though hate may be online, it is necessary to counter it offline as well!

APPENDIX

Key Performance Indicators and Chapters measured with Monitoring Tool

Table 16. The 13 KPIs to be measured in each CSOs campaigns:

KPI 1: *AUDIENCE ENGAGEMENT*	This KPI focuses on measuring audience interaction with the counter-narrative content. It includes the number of likes, shares, and comments on social media platforms. CSOs must also distinguish if people react with emojis or respond with messages. Responding with written messages may suggest higher engagement than quick-emoji reactions. Higher engagement indicates that the content resonates with the audience and has the potential to influence perceptions.
KPI 2: *(NON-)HOMOGENEITY OF AUDIENCE*	This KPI emphasises the demographic diversity of the engaged audience, e.g. age, gender, and location. Diverse/distinct audience categories/groups/profiles ensures that the counter-narrative resonates with various population segments, promoting inclusivity, sustainability, and educating people across generations.
KPI 3: *AUDIENCE REACH*	Audience reach assesses the extent to which the counter-narrative messages are disseminated. This KPI encompasses impressions, shares, and using relevant keywords or hashtags. A broader reach indicates a wider impact and increased visibility of the alternative narrative
KPI 4: *WEBSITE AND SOCIAL MEDIA ANALYTICS*	Digital analytics focuses on metrics such as website traffic related to counter-narrative content and click-through rates on links, leading to more information. These metrics provide insights into the online engagement and interest generated by the alternative narrative.
KPI 5: *PARTNERSHIP EFFECTIVENESS*	This KPI evaluates the effectiveness of collaborations with organisations or influencers. Successful partnerships amplify the reach and impact of counter-narratives, leveraging collective efforts to combat hate speech more effectively.
KPI 6: *MEDIA EXPOSURE*	Media exposure assesses the extent to which counter-narrative initiatives, preferably focusing on the CSOs' CANs, are covered in the media. Positive and widespread media coverage shapes public discourse, and challenges hate speech.
KPI 7: *SENTIMENT*	Sentiment analysis evaluates the emotional tone of audience reactions and comments to the counter-narrative content. A higher ratio of positive sentiment indicates that the content is

	successfully fostering a constructive and supportive online environment.
KPI 8: *BEHAVIOUR CHANGE*	Behaviour change is a critical indicator of effectiveness. Surveys and testimonials assess whether exposure to the counter-narratives leads to positive shifts in attitudes and behaviours. Successful counter-narratives should contribute to tangible and positive changes in individuals' perspectives.
KPI 9: *EFFECTIVENESS PERCEPTION*	Surveys measuring the perceived effectiveness of (the CSOs') counter-narratives offer insights into how the audience perceives the impact of these initiatives. Positive perceptions contribute to building credibility and trust in their counter-narratives lead to the belief that CAN intervention is effective. Here, the perceptions of victims can be included to gain insight into their experience with CAN campaigns and hate speech prevalence.
KPI 10: *INCREASE IN AWARENESS AND UNDERSTANDING*	Educational impact measures the increase in awareness and understanding resulting from counter-narrative initiatives. Pre- and post-campaign knowledge assessments can gauge the effectiveness of educational efforts in challenging stereotypes and misconceptions.
KPI 11: *INCIDENT REPORTING*	Incident reporting involves tracking the number and nature of reported hate speech incidents. This KPI helps assess counter-narratives effectiveness by analysing the amount of people registered to have reported a complaint (on a social media platform) and thus are reticent towards this content, preventing further escalation.
KPI 12: *TIMELY RESPONSE TO HATE SPEECH INCIDENTS*	This KPI measures the responsiveness of the CSO to reported hate speech incidents. A swift and effective response indicates a rapid and proactive approach to mitigating the impact of hate speech, demonstrating the commitment to maintaining a safe online environment.
KPI 13: *RESPONSE TO CAN CRITICISM*	CAN campaigns may receive negative responses. This KPI measures the responsiveness of the CSO to criticism towards their campaign. Post-campaign criticism responses can change the effectiveness of the CAN, as it will influence public perception of the CSO and how it performs when responding to criticism. Successful responses may include responding proactively, with the aim to mitigate the criticism.

Table 17. Chapters of the Monitoring Tool

Chapter 1: *DEVELOP AND DESIGN*	The development and design of a CAN provides data points on content ideation and creation. It therefore answers what the goal of the CAN is, who its audience is, how it was set up, how long it takes (or took), and which steps are involved. For campaigns targeted at prevention or education, reaching more people, hence more people engaging with the content, could be used as a measure of impact. The nature and duration of engagement is extremely important to register for a correct evaluation of whether engagement is sustained or not.
Chapter 2: *ONLINE DEPLOYMENT*	Online development gathers information about the social media platforms users are active on, and metrics on views, impressions (and impression frequency), reach, timing, posting frequency consistency, clicks on links, etc.
Chapter 3: *AUDIENCE ENGAGEMENT*	This involves using data points to monitor engagement. To do so, the sentiment of comments as well as incidents, likes, saved content, reactions, shares, Click-Through Rate (CTR), web traffic, bounce rate, and time spent on the site are tracked. These metrics are essential as they show whether the message and tone of a campaign was compelling and resonated with the audience. If a campaign is disseminated in video format, viewer retention and drop-off rates should be specified.
Chapter 4: *OFFLINE ACTIVITIES*	Offline activities measures data on all types of offline activities such as marches, panel discussions, conferences, workshops, and more.
Chapter 5: *BEHAVIOURAL CHANGE*	Behavioural change collects data from survey responses, social listening, and testimonials, providing insight into the effect of a CAN Did the content change its audiences' opinions, attitudes, or behaviour in any observable way?
Chapter 6: *HATE SPEECH INCIDENTS*	Hate speech incidents is a key indicator of how effective CANs are in responding to hate speech incidents. It shows how the incident was reported, the response time, the content of the response, and the technique used in responding.
Chapter 7: *CAMPAIGN CRITICISMS*	Should a campaign receive criticism, this tool collects data about the demographics of who criticises campaigns and how the campaign responds. It may even provide precautions for an CSO to take to prevent criticism and guidelines on how to respond to criticism. It also allows space for the campaign to ensure that its message is supported.