SELMA

HACKING HATE

Hacking Online Hate:

Building an Evidence Base for Educators

Funded by the Rights, Equality and Citizenship Programme (2014-2020) of the European Union

European Schoolnet

SWGfL

LMK

CfDP

Please cite this report as:
SELMA (2019) Hacking Online Hate: Building an Evidence Base for Educators. www.hackinghate.eu.

This research report is published under the SELMA (Social and Emotional Learning for Mutual Awareness) project by European Schoolnet. It is the result of a collaborative effort from European Schoolnet, For Adolescent Health (FAH), South West Grid for Learning (SWGfL), The Diana Award, Media Authority for Rhineland-Palatinate (LMK) and Centre for Digital Youth Care (CfDP). More information about the SELMA project and partners is available at www.hackinghate.eu.

The publication was funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020). The contents of the publication are the sole responsibility of the authors and can in no way be taken to reflect the views of the European Commission.

Cover image and icons: Freepik, George Studio/Shutterstock.com

Any questions can be addressed at info@hackinghate.eu.

SELMA
HACKING HATE

TABLE OF CONTENTS

GENERAL PREFACE
THE SELMA PROJECT AND ITS RESEARCH PROGRAMME

Online hate speech is a growing problem. People often experience the internet to be a hostile space. Hateful messages are increasingly common on social media. To complement existing initiatives to regulate, monitor or report online hate speech, a more pro-active approach is needed.

SELMA (Social and Emotional Learning for Mutual Awareness) is a two-year project co-funded by the European Commission[1] which aims to tackle the problem of online hate speech by promoting mutual awareness, tolerance, and respect.

The overall vision of the SELMA project is captured by its slogan: Hacking Hate. It builds upon social and emotional learning (SEL), media literacy and citizenship education approaches to empower young people to become agents of change. It helps them to better understand the phenomenon of online hate. It provides them with tools and strategies to act and make a difference.

1 For more information about the SELMA project, see http://hackinghate.eu. The SELMA project is funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020). The contents of this publication are the sole responsibility of its authors and can in no way be taken to reflect the views of the European Commission.

"Hacking is any amateur innovation on an existing system, and it is a deeply democratic activity. It's about critical thinking. It's about questioning existing ways of doing things. It's the idea that if you see a problem, you work to fix it, and not just complain about it."

Catherine Bracy, TechEquity Collaborative

In more concrete terms, SELMA targets young people (age 11-16), primarily in schools, but also in the out-of-school communities that impact on their wellbeing. It engages them – together with their peers, teachers, parents and other professionals and carers – in a multifaceted learning journey. It fosters a wider dialogue with education stakeholders (including Ministries of Education), civil society organisations and industry.

In order to take an evidence-based approach to prevent and remediate online hate speech, it is fundamental to first understand what online hate speech is and how it works. Therefore, the SELMA Toolkit is built on a comprehensive research programme which comprises three interrelated components enriching each other: a literature review, a series of qualitative focus groups and an online quantitative survey. The current report synthesises our main research findings. It is created by professionals with a variety of academic and organisation backgrounds, in an effort to achieve a holistic understanding of the online hate speech phenomenon. It provides a theoretical and empirical backbone for the education and awareness-raising activities carried out as part of the SELMA project.

PART I:
LITERATURE REVIEW

1. Introduction

1.1. Overall objectives and structure of the literature review

Hate speech is not a new phenomenon. It is as old as the formation of human societies and the organisation of people into groups. Historically, it did not always affect the same people, nor was it always expressed in the same way. Yet, concerns about its impact on individuals and society have largely been the same: hate speech may affect individuals' wellbeing; it may weaken, disintegrate or even destruct social cohesion; it might lead to violence between individuals, groups and communities, sometimes even threatening peace. In a digital world, these kind of concerns have only increased, because of the rapid pace at which online hate speech messages can be uttered and spread.

Over the years, a large number of academic and policy publications have emerged which define hate speech in relation to these kind of societal concerns. Reference is often made to the historical or socio-economical context in which hate speech occurs, with possible solutions being explored. Due to its applied and multidisciplinary nature, an eclectic mix of concepts and ideas are being used, interchangeably or not, to describe its nature and dimensions. Is hate speech similar to dangerous speech, or inflammatory speech or even bullying? And how does this all translate to the online environment in which children and young people are nowadays growing up? If researchers and policy makers struggle to agree and understand what hate speech is, how can we possibly expect teenagers to come to terms with how online hate speech may affect their everyday lives?

In this literature review, we recognise the existing diversity of perspectives, but equally try to end up with a comprehensive online hate speech definition. From an education point of view, children and young people need a meaningful starting point from which they can explore and reflect upon their own views and experiences. They need to be equipped for meaningful dialogue and peer-to-peer discussion, while building towards mutual tolerance and respect. In trying to capture the full complexity of the online hate speech phenomenon, we will therefore move beyond academic discussions about what online hate speech

is, diving into its prevalence, causes and consequences, while giving a more anecdotal account of how this all translates across SELMA partner countries.

At the end of this report, we will explain in more detail how this literature review – together with the subsequent empirical sections – has informed the pedagogical model underpinning the SELMA Education Toolkit.

1.2. Research methodology

As hate speech clearly predates the internet, our desktop research started from the available literature concerning "offline" hate speech. Subsequently we looked into the connection of offline to online dynamics, with a particular focus on how the nature of hate speech is changing in a digital society.

For this purpose, a comprehensive search on PubMed and Google Scholar was performed to identify relevant papers on (online) hate speech. We complemented this corpus with relevant Recommendations, Declarations, Reports, Guidelines and Factsheets from international bodies and organisations such as the European Commission, the European Union Agency for Fundamental Rights, the European Commission against Racism and Intolerance, the Council of Europe and the United Nations. For the needs of the study we focused on the phenomenology of online hate speech, its causes and consequences, as well as relevant programmes already implemented, particularly in the Europe Union. We also integrated documents from a landscape review of relevant policies and practices in the following SELMA partner countries: Germany, Greece, Denmark and the United Kingdom.

We used various search strings, with simple key words such as: *"hate speech", "dangerous speech", "inflammatory speech", "hateful speech", "hate rhetoric", "illegal hate speech", "harmful speech", "offensive speech", "dehumanizing speech", "sexist hate speech", "anti-refugee hate speech", "hate speech online", "cyberhate", "cyber harassment", "expression of hate online", "hate crime", "cybercrime", "hate crime online", "combating hate speech online"* but also more elaborate combinations, such as *"freedom of speech AND hate speech online", "freedom of speech AND internet", "LGBT AND hate speech online", "(Jews OR anti-semitism) AND hate speech online", "Muslims AND hate speech online",*

"Islamophobia AND cyberspace", "racist speech AND (internet OR online)", "(Racism OR Xenophobia) AND cyberspace", "cyberhate AND social media", "hate speech online AND counter-narrative campaigns".

We completed our desktop research using a snowball procedure, drawing upon the references in the articles we found in our primary searches.

2. What is hate speech?

2.1. Introduction

To understand and combat online hate speech, elaborating a definition on what is hate speech is fundamental. Discussions on hate speech cannot be understood without exploring the fine and delicate line which differentiates hate speech from free speech. Meanwhile, the way in which hate speech is defined largely depends on contextual factors. As such, it is bound to evolve over time and place, with broader societal developments playing in the background.

2.2. Drawing the line: Hate speech versus freedom of expression

In ancient Greece, the freedom of expression was a fundamental right of free citizens. *Isegoria – the fact that every citizen has an equal right to express his opinion – was a quintessential notion in Athenian Democracy (Finley, 1996). The freedom to question authorities, under some restrictions, became a recurring theme in the writings on politics, arts and education emerging from this seminal tradition. In the more recent European history, John Milton defended – in a similar vein – the freedom of speech in his Areopagitica pamphlet: "Give me the liberty to know, to utter, and to argue freely according to conscience, above all liberties." Voltaire is famously quoted along the same lines: "I disapprove of what you say, but I will defend to the death your right to say it" (Hall, 1906). Ever since, oppressed groups have fought for and drawn upon their freedom of speech, as illustrated in the 20th century time and time again, for instance by the civil rights movement, the women's movement, and the gay liberation movement, who all struggled in their own way for social justice (Cowan, Resendez, Marshall, & Quist, 2002).*

The Freedom of Expression and of Opinion became a fundamental human right in 1948 as it was integrated into Article 19 of the Declaration on Human Rights of the United Nations, which declares the right of every person to seek, receive and impart information and ideas of all kinds freely, regardless of frontiers, *"either orally, in writing or in print, in the form of art, or through any other media of his choice" (UN General Assembly, 1948). Importantly however, Article 19 is directly followed by Article 20 that expressly limits freedom of expression in cases of "advocacy*

of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence." Likewise, in the European Convention on Human Rights it is stated that everyone has the right to freedom of expression. Yet, here again, this right carries its duties and responsibilities and "may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary" (Council of Europe, 1950).

In other words, the freedom of expression constitutes one of the essential foundations of democratic societies, one of the basic conditions for its progress and for the development of every individual. It is applicable not only to information or ideas that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the state or any sector of the population (McGonagle, 2013). This does not equate the freedom of expression with a right to offend. While freedom of expression is everyone's right, it comes with certain restrictions. The kinds of hatred which constitute incitement to discrimination, hostility or violence in particularly tend to be prohibited by law (Gagliardone, Gal, Alves, & Martinez, 2015).

Approached in this way, hate speech inevitably brings along questions about how absolute freedom of speech is, what kind of democracy citizens need, and how to strike a balance between the right to express oneself and the right to not be discriminated against (Cammaerts, 2009). Hate speech can thus be understood as a specific type of expression, which might undermine safety, health, morals or reputation, and sometimes might even turn into a violent act undermining the human rights of others. This is illustrated in *Figure 1, which was drawn from the Council of Europe publication Bookmarks: A manual for combating hate speech online through human rights education (Keen & Georgescu, 2016). In accordance with international human rights law, hate speech in many forms cannot be restricted – this is the subgroup of "mild" hate and intolerance. Yet, in its more extreme forms, hate is likely to provoke individual or societal harm, making it illegal. Because indeed, "there needs to be a balance between allowing people to express their inner thoughts, and ensuring that this does not undermine the rights of others, or cause greater damage to society" (Keen & Georgescu, 2016, p. 162).*

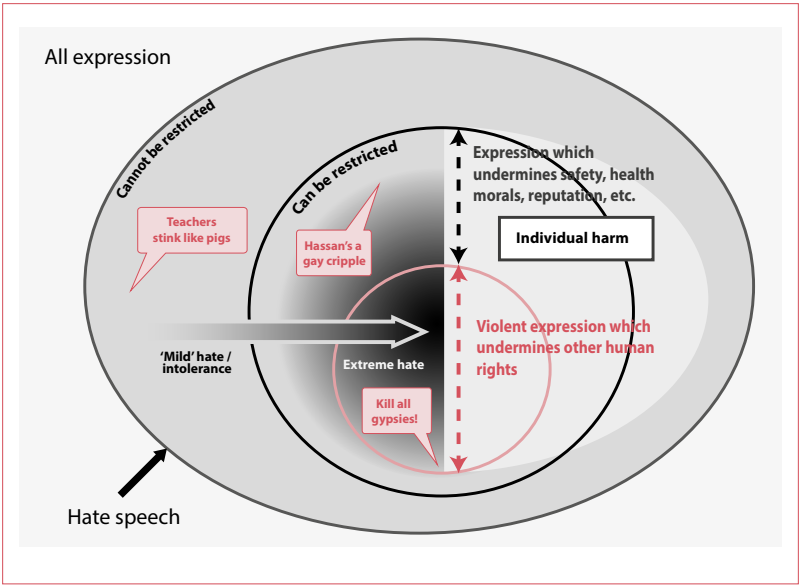

***Figure 1**. The limits to freedom of expression (Keen & Georgescu, 2016)*

To make this more specific, the Rapporteur of Minority issues of the United Nations (UNHRC, 2015) differentiates three types of hate speech expression: (a) expression constituting an offence under international law that can be prosecuted criminally (the red circle presented in the No Hate Speech schema above); (b) expression not criminally punishable but that may justify a restriction and a civil suit (the smaller black circle of the above schema); and (c) expression that does not give rise to criminal or civil sanctions but still raises concerns in terms of tolerance, civility and respect for others (the bigger black circle of the schema). By contrast, forms of expression such as satire or objectively based news reporting and analysis – even if it offends, hurts or distresses – are typically excluded from any hate speech definition (see for example Recommendation No 15 from the ECRI (2016)).

2.3. An eclectic mix of hate speech definitions...

...in the research literature

Against this background, one can identify a large variety of hate speech definitions in the academic literature. No universally accepted definition seems to exist. Rather, hate speech presents itself as a broad umbrella term, which covers many different types of hateful and harmful expressions, typically targeted at groups or classes of persons with certain characteristics (Brown, 2017). What is clear from these definitions is that "hate" is more extreme than mere dislike, and tends to be discriminatory or abusive in nature. Meanwhile, "speech" can refer to the expression of opinions, ideas or emotions, not only verbally but also through other forms like images, video or sound.

For instance, for Leets and Giles (1999), hate speech is a subcategory of harmful speech, the latter being defined as *"utterances that are intended to cause damage, and/or irrespective of intent, that their receivers perceive to result in damage"* (Leets & Giles, 1999, p. 95). Smolla (1990, p. 195) states that *"hate speech is the generic term that has come to embrace the use of speech attacks based on race, ethnicity, religion, and sexual orientation or preference"*. Cowan and Hodge (1996) describe hate speech in relation to racist hate speech, sexual harassment and anti-gay speech. Sedler (1992) holds that hate speech is used to denigrate persons on the basis of their race or ethnic origin, religion, gender, age, physical condition, disability or sexual orientation. While the incitement to hatred, violence or segregation is a recurrent element across hate speech definitions, the range of individual or group characteristics which may be targeted appears sheer endless, including real or perceived race, ethnicity, skin colour, language, nationality, religious beliefs or lack thereof, gender, gender identity, sexual orientation, political beliefs, social status, age, mental health, disability, disease, being Roma or refugee (Berecz & Devinat, 2016abc, 2017abc). According to Duffy (2003), hate speech is commonly used by a hate group, which shows unreasonable violence, verbal or sometimes physical, against other groups defined by their race, ethnicity, religion, sexual orientation, gender or other characteristics.

Hate speech is often discussed in close association (or interchangeably) with other concepts, which may point to nuances in perspective. For instance, some align hate speech to racist speech, describing how this kind of speech delivers the message of racial inferiority, is directed against a member of a historically

oppressed group, and is persecutory, hateful and degrading. In this sense, it addresses individuals or groups of people with the goal to spread violence, racist and discriminative attitudes in societies, and to give justification to emotions, words or actions of this kind (Matsuda, Lawrence, Delgado, & Crenshaw, 1993). Others use the term inflammatory or dangerous speech as a subcategory of hate speech arguing it grows the chances of catalysing violence by one group against another. The amount of violence possibly resulting from inflammatory speech may depend on various factors, such as the capacity of the speaker to exercise influence among others, the nature of the language used, the context in which it is uttered, the medium used, and the kind of the audience reached (Yanagizawa-Drott, 2014; Benesch, 2013).

...in European law and policies

In Europe, the Council of Europe has traditionally played a leading role in shaping hate speech law and policy making. The Committee of Ministers, at their October 1997 meeting, first recommended that the term hate speech refers to *"all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin" (Council of Europe, 1997, p. 107).*

In 2002, the European Commission against Racism and Intolerance (ECRI) was founded; a human rights monitoring body which specialises in questions relating to the fight against racism, discrimination on grounds of race, ethnic or national origin, colour, citizenship, religion or language (racial discrimination), xenophobia, anti-semitism and intolerance. ECRI performs regular country monitoring across all 47 Council of Europe Member States, while maintaining special relations with independent authorities responsible for combating racism, racial discrimination, xenophobia, Anti-semitism and intolerance at national level and with relevant intergovernmental bodies, such as the European Union and the United Nations. Hate speech is a specific point of attention in the country monitor reports, which specific policy guidance also being issued on how to combat it at national level (ECRI, 2016).

When hate speech is defined in a regulatory or legislative context, it may imply that certain characteristics become legally protected. This list of protected

characteristics has grown over the years. For instance, building on the initial focus on race and ethnicity, in 2007, the Parliamentary Assembly of the Council of Europe reaffirmed that *"national law should [also] penalise statements that call for a person or a group of persons to be subjected to hatred, discrimination or violence on grounds of their religion."* A recommendation from the Council of Europe Committee of Ministers (2010) further included all forms of expression *"which may be reasonably understood as likely to produce the effect of inciting, spreading or promoting hatred or other forms of discrimination against lesbian, gay, bisexual and transgender persons".*

The 2016 ECRI General Policy Recommendation on combating hate speech further widened the Council of Europe net, broadly defining hate speech as *"the advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatisation or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of 'race', colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status"* (p. 3). Here again, the need for a balanced approach is consistently addressed, reaffirming the fundamental importance, or a democratic and pluralistic society, to find a balance between the freedom of expression and opinion with tolerance and respect for the equal dignity of all human beings.

Within the European Union, illegal hate speech is defined by the 2008 Framework Decision on Racism and Xenophobia, which requires EU Member States to take the measures necessary to ensure that the intentional public incitement *"to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin"* is made "punishable" (p. 56). As we will see later on, this still forms the legal basis for ongoing multi-stakeholder efforts to combat hate speech in a digital environment.

2.4. The key features of hate speech

As we have seen, at a basic level, hate speech is any expression of discriminatory hate towards people. Beyond this simple understanding, its meaning becomes rapidly contested. Still, it is possible to distil a core set of key features which can

help individuals – including children and young people – to identify or recognise hate speech.

For the purpose of the SELMA project, we will primarily focus on the "protected characteristics" of the groups against which hate speech is typically addressed. To illustrate, according to the Equality Act of 2010 of the Equality and Human Rights Commission of Great Britain, there are nine characteristics which trigger the prohibition against discrimination, namely: age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex and sexual orientation.

Throughout the SELMA Toolkit, we will consider these characteristics as being protected within a specific historical, social and political context. For instance, historically, we can easily distinguish a difference between the United States and European countries concerning the target groups of hate speech. The American literature has often framed hate speech in relation to people of colour, with anti-black or anti-Indian theories of white supremacy being relatively widespread. In comparison, the European literature is more often inclined to focus on people with immigrant origin (Daniels, 2008; Timofeeva, 2003).

Apart from the protected characteristics, a number of other elements stand out as relevant hate speech features:

- The content and tone of the expression.
- The intention to cause harm and whether the expression is considered to have a potential or actual harmful consequence.
- The extent to which the expression of hate is publicly disseminated.

In the next chapter, we will come back to the question on how to identify or recognise hate speech in an online environment. Rather than being prescriptive, we will integrate the most relevant features and dimensions as relevant points for pedagogical reflection and debate. This will help to foster a meaningful discussion about what hate speech is and how it relates to children and young people's everyday online media experience.

3. The online hate speech phenomenon

3.1. Introduction

In a democratic society, individuals and groups should be free to meet in public spaces, expressing their critical views and constructively exchanging informed opinions. Nowadays, these public spheres are largely moving online, with commercial platforms facilitating the mass sharing of information, views and opinions, raising both opportunities and challenges in terms of active citizenship (Williams & Burnap, 2016). Hate speech has gone digital in a very similar fashion. As we will see, this has greatly amplified its prevalence and visibility. It also made the phenomenon more difficult to grasp and combat.

In this chapter, we will first present the transition from offline to online hate speech from an academic, policy and wider stakeholder perspective. We will again compare a number of relevant definitions, before exploring what is known about the specific nature of online hate, its causes and its consequences. We will also reflect on how this is relevant for the SELMA project's key target audience: children and young people.

3.2. Defining online hate speech…

… in the research literature

As with offline hate speech, a variety of online hate speech definitions exist. From a content point of view, these definitions typically relate to the key features we have identified in the previous chapter. Hate speech still refers to hateful public messages directed at individuals or groups with certain characteristics, with varying levels of emphasis given to the particular tone of expression, the intention to cause harm, and so forth. However, what is clearly different is the online means of expression, the digital medium used for hate speech to reach its victims, possible supporters and the general public. This brings along some specific characteristics which make online hate speech substantially different in nature.

In 1984 already, Kiesler, Sigel and McGuire described the phenomenon of uncivilised behaviour and angry messages in anonymous computer-mediated

communication, which they called flaming. This could be considered an early attempt to describe hate speech online. Nowadays, online hate speech is spread through websites, social media, online games and, in general, through online systems. In an online world, people aiming to express hate speech can more easily seek and contact potential audience members, increase their influence in the digital and physical world (Timofeeva, 2003; Duffy, 2003). Hawdon, Oksanen and Räsänen (2015, p. 30) define online hate speech as a form of cyberviolence, which uses information communication technology to *"advocate violence against, separation from, defamation of, deception about or hostility towards others." Others refer to "any use of electronic communications technology to spread anti-semitic, racist, bigoted, extremist or terrorist messages or information. These electronic communications technologies include the internet (i.e., websites, social networking sites, 'Web 2.0' user-generated content, dating sites, blogs, online games, instant messages, and e-mail) as well as other computer - and cell phone-based information technologies" (Anti-Defamation League, 2010, p. 4).*

Hate speech online, because of its dissemination through computers, digital devices and the internet, may appear in the form of text, music, online radio broadcasts, or visual images (Keipi, Näsi, Oksanen, & Räsänen, 2017; Thiesmeyer, 1999). This material is typically used as an expression of harmful or threatening content against individuals or collectives of people (Oksanen et al., 2018) and may induce users, directly or indirectly, to act against these target groups (Thiesmeyer, 1999).

... in European law and policies

In 2003, the Council of Europe published an additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems. It refers to *"any written material, any image or any other representation of ideas or theories, which advocates, promotes or incites hatred, discrimination or violence, against any individual or group of individuals, based on race, colour, descent or national or ethnic origin, as well as religion if used as a pretext for any of these factors" (p. 2).*

Similarly, in the 2016 ECRI Recommendation previously mentioned, hate speech is explicitly understood *"to cover speech and publications in any form, including through the use of electronic media, as well as their dissemination and storage. Hate speech can take the form of written or spoken words, or other forms such as*

pictures, signs, symbols, paintings, music, plays or videos. It also embraces the use of particular conduct, such as gestures, to communicate an idea, message or opinion" (ECRI, 2016, p. 17). Further on, it is noted that the use of hate speech is "a particular feature of some electronic forms of communication, with web pages, forums and social networks forums having that as a primary purpose and some using such speech even when they are hosted by local government bodies" (ECRI, 2016, p. 20).

Meanwhile, in the European Union, while the 2008 Council Framework Decision does not explicitly address the internet, the European Commission and EU Member States have built on this legal basis to ensure that national laws to prevent the spread of illegal hate speech in the offline as well as the online environment are fully enforced. This has largely taken the form of a multi-stakeholder approach, where both IT companies and civil society organisations become involved in a joined-up effort. Most notably, the European Commission agreed in 2016 with IT companies Facebook, Microsoft, Twitter and YouTube on a Code of Conduct on countering illegal hate speech online2 to help users to notify illegal hate speech on social media platforms, while improving the support to civil society as well as the coordination with national authorities. Meanwhile, important platforms such as Instagram, Google+, Snapchat and Dailymotion announced their intention to join this Code of Conduct. In 2018, the European Commission also adopted a Recommendation on measures to effectively tackle illegal content online along the same lines.

... from the perspective of industry

As online hate speech is in many cases associated with social media platforms (Gerstenfeld, Grant, & Chiang, 2003), social network providers have put forward their own definitions for online hate speech. These typically form the basis for the processes IT companies put in place to review notifications regarding illegal or possibly harmful hate speech on their services, so they can remove or disable access to such content, in correspondence to the Terms of Service or Community Guidelines they have in place to prohibit the promotion of or incitement to violence and hateful conduct.

2 https://ec.europa.eu/info/files/code-conduct-countering-illegal-hate-speech-online_en.

Google for example, in its User and Conduct Policy[3] says *"Our products are platforms for free expression, but we don't permit hate speech. Hate speech is content that promotes or condones violence against, or has the primary purpose of inciting hatred against, an individual or group on the basis of their race or ethnic origin, religion, disability, age, nationality, veteran status, sexual orientation, gender, gender identity, or any other characteristic that is associated with systemic discrimination or marginalisation." In its Hate Speech Policy4, YouTube additionally points to the "fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticise a nation-state, but if the primary purpose of the content is to incite hatred against a group of people solely based on their ethnicity, or if the content promotes violence based on any of these core attributes, like religion, it violates our policy."*

Likewise, Facebook in its Community Standards[5] defines (and indicates it does not allow) hate speech as *"a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation."* By contrast, humour and social commentary related to these topics is allowed. Also, people can share *"content containing someone else's hate speech for the purpose of raising awareness or educating others. Similarly, in some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. When this is the case, we allow the content, but we expect people to clearly indicate their intent, which helps us better understand why they shared it."*

Finally, Twitter in its Hateful Conduct Policy[6] explains that users *"may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories."*

3 https://www.google.com/intl/en_uk/+/policy/content.html, consulted in January 2019.
4 https://support.google.com/youtube/answer/2801939?hl=en, consulted in January 2019.
5 https://www.facebook.com/communitystandards/hate_speech, consulted in January 2019.
6 https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, consulted in January 2019.

3.3. The nature of online hate

a) From offline to online dissemination channels

Through history, we have seen how traditional mass media – from books or newspapers to radio, film and television – can be abused as an instrument for stigmatisation, discrimination, exclusion, and incitement to hatred and violence. In its most extreme form, mass media propaganda contributed to the widespread marginalisation of groups and even mass murder and genocide.

The most well-known example in this regards is that of Nazi Germany, where Joseph Goebbels, Reich Minister of Propaganda recognised radio and film as the eighth great power; the most modern and scientific means to influence the masses (Timmermann, 2008). More recently, in 1994, Radio Rwanda and Radio Télévision des Mille Collines played a significant role in the Rwandan Genocide, broadcasting anti-Tutsi propaganda, while inciting to hatred and violence. The newspaper Kangura further helped to frame Tutsis as "cockroaches" to be extinguished, with hundreds of thousands of people being killed as a result (UNHRC, 2015; Yanagizawa-Drott, 2014).

Similar dynamics continue to be at play – in less dramatic but still pervasive forms – in books, magazines or newspapers, on the radio or on television. At the same time, the media landscape is undergoing a digital transformation, with the internet and social media having a far-reaching impact on how mass media reach and influence their audiences. Of great concern is that, historically, hate groups have been technologically sophisticated and often know earlier and better than the authorities or the wider public how to use technology in their favour (Daniels, 2008; Duffy, 2003). This can be illustrated with a number of examples of how hate groups over time have embraced new media to amplify their key messages.

i. Websites promoting extremist groups

Online hatred has been spreading rapidly on websites, blogs and forums. Hate groups have also been early adopters of (chain) emails, mailing lists, electronic discussion groups and bulletin boards (Berecz & Devinat, 2016abc; Perry & Olsson, 2009; Thiesmeyer, 1999).

Early in the World Wide Web history, in 1995, Don Black, a grand wizard of the Ku Klux Klan (KKK), famously created one of the first hate websites, Stormfront.

org (Levin, 2002). Many other hate groups followed suit, often with a right-wing flavour, with hundreds of neo-Nazi, racist, skinhead, Christian identity but also some black separatist hate sites emerging. By 2000, an estimate showed more than 5,000 websites advocating hatred, with numbers further increasing dramatically over the years (Perry & Olsson, 2009). Meanwhile, Stormfront.org took the genre to a new level, with its audience rapidly expanding. By 2009, the site had over 159,000 members and by 2015 the number of registered users was approximately 300,000 (Bowman-Grieve, 2009; Potok, 2016).

Many hate speech websites have found a "home" on US ground, because of its firm stance prioritising freedom of speech against the rhetoric of hatred.

ii. Cloaked websites

In its most infantile form, hate websites overtly display hate speech messages, clear and easy to detect, using direct words with the intention to hurt. Yet, over time, more subtle means of manipulation started to appear alongside. Search engines have been fooled by their own algorithms, with bigoted websites being ranked among leading results for a given topic, thereby achieving the kind of worldwide readership once reserved for mainstream media messages. Extremist groups have taken advantage of this characteristic of the internet to create "cloaked websites" – pages that masquerade as impartial or factual information about social, historical, or political topics but are actually founts of hate-filled propaganda. Often, the goal is to lure students and other curious would-be researchers into reading and believing content that would almost never find its way onto the shelves of a reputable school, college, or public library (Foxman & Wolf, 2013).

Cloaked websites disguise their purpose, using statements which may seem objective and neutral at first sight, while providing links to other sources of information where messages of discrimination, hatred and violence appear in abundance (Daniels, 2008; MgGonagle, 2013; White & Crandall, 2017). Perhaps the most infamous example was www.martinlutherking.org[7] a website ran by Stormfront which has long figured prominently among the top result for searches of "Martin Luther King" on Google. The site purports to be *"A valuable resource for teachers and students alike"* where you can read *"the truth"* about King —

7 At time of writing this report, the website was down, but pages can still be retrieved at https://web.archive.org.

communist, wife-beater, plagiarist and sexual deviant. There are even flyers to the same effect that children can download, print and bring to school. One of them reads: “GET THE FACTS! Martin Luther King Jr. Was A Fraud!”.

iii. Social media pages

More recently, social media platforms became increasingly fertile ground for groups to spread hateful messages (Gerstenfeld et al., 2003). Facebook, YouTube, Twitter, Tumblr, Pinterest and other social media sites are used by hundreds of millions of people around the world. They are free and simple to use. They can help rabid haters to reach a large audience, spewing lies and vitriol against those they deem less worthy – and encourage others to do the same (Berecz & Devinat, 2016abc, 2017abc; Keipi et al., 2017).

Conducting an online mapping of the populist right and right-wing extremism in thirteen European countries, Rogers (2013) found how right extremism has moved indeed from websites to blogs and social media. It is not only the youth who have left the web behind and now use social media, but also the new right. To give one example of how this can look in practice: on the Fourth of July, 2010, while Americans were celebrating their nation’s birthday, a new event was announced on Facebook: “Kill a Jew Day”. A Nazi Swastika adorned the official event page on Facebook, with the host for the event writing “You know the drill guys,” urging followers to engage in violence “anywhere you see a Jew”. The posting prompted a whole range of anti-semitic rants in support of targeting Jewish people (Foxman & Wolf, 2013). Many similar examples exist, where social media companies like Facebook, Twitter, YouTube or Tumblr struggle with hate speech. People use both public and private spaces on these platforms to promote intolerance and fear, often crossing the line with offensive and threatening language, with racist mockery and attacks, with “jokes” about sexual harassment and rape, denying the Holocaust, or identifying Islam with terrorism.

As with cloaked websites, hate groups have explored more subtle avenues to attract new audiences. Berecz and Devinat (2016c) give the example of the activist group “Matefaschisten” (mate fascists, “mate” is a kind of tea from South America) which streamed a live video via Facebook that showed them baking a mate cake. “Referring to a very popular beverage among young people in Germany was a strategy to address a young target group with right-wing propaganda coming along in an unsuspicious format. During the baking, the

activists answered questions from the audience. Their statements alternated between fooling around and neo-Nazi propaganda" (p. 4). In this way, extreme groups exploit savvy social media tactics to hook in naïve web users. They use a mix of humour, emotions and disinformation to attract followers – for instance, appropriating the popular internet meme Pepe the Frog, or taking images out of context to sell bogus narratives. Once in people's social media feeds, they end up sharing messages of a more troubling nature. In parallel, online trolling attacks, with erroneous or antagonistic messages being posted to elicit hostile or corrective response, are sometimes being launched and coordinated from behind the scenes of closed community groups or just on message boards like 4chan's /b/, to throw possible adversaries of balance, intimidating them for the purpose of policing behaviour (Eckstrand, 2018).

Even terrorist organisations such as ISIS started using social media to recruit significant numbers of jihadists, while informing their audiences of their actions or contacting potential donors (Conway & McInerney, 2008).

iv. Games

Hate speech has also found its way into online gaming, with many gamers exchanging hate speech over their headsets as they fight each other in the virtual battlefield. *"The violent nature of the games themselves, combined with the anonymity prevalent in online gaming sites, encourages players to indulge freely in fantasy behaviours that would be unacceptable in real life. These behaviours can include the use of hate speech – such as racist, ethnic, anti-semitic, misogynistic, and homophobic slurs – against opponents" (Foxman & Wolf, 2013, p. 20).*

b) The online stratification of hate

Alongside the content and tone of the expression, its intention to incite hatred, and the fact that it is targeting an individual or group based on its protected characteristics, a number of specific online hate speech features can be identified.

Once it is out, online hate speech is not likely to go away. What is published often stays there for days. Even when taken down, messages can easily be repeated, while inspiring similar messages to appear in different forms and places. If a user account is blocked or suspended, another can be created. If a website is shut down, its owner can look for an alternative web-hosting service, with less stringent terms and conditions, or reallocate to a country with higher hate speech

thresholds. This also illustrates how online hate groups are no longer restricted by geographical boundaries. This makes it all the more difficult to combat it through legal mechanisms (Burris, Smith, & Strahm, 2000; Duffy, 2003; Gagliardone et al., 2015; McGonagle, 2013).

Against this background, Keipi et al. (2017) describe what they call "the stratification of online hate": the multiplication and visibility of hate speech has increased over time; the growing popularity of social media has been quintessential in this regard. At the core, traditional hate groups continue to be active both offline and online, maintaining websites to inform their audiences of their basic ideology, granting them access to comment sections on news or related events, and providing information regarding potential social gatherings, protests and so forth. These websites often have only a handful of visitors and a relatively small following due to the necessity for audiences to search specifically for a particular site. While this layer is well structured, its impact potential is low. On top of this first layer, a second one emerged, which tends to be less structured and more active, with larger visibility and a higher impact potential. These are the vibrant groups of like-minded users tuning into internet fora, online radio programmes, podcasts, user-generated blogs, sharing or chat services. Finally, social media hate constitutes the third and newest layer. This is by far the most visible and the least structured hate sphere. On social media, a multitude of users are granted high levels of visibility to express hateful ideas and opinions, free from restrictions of an enforcing community as can be the case with the other layers (Foxman & Wolf, 2013; Keipi et al., 2017).

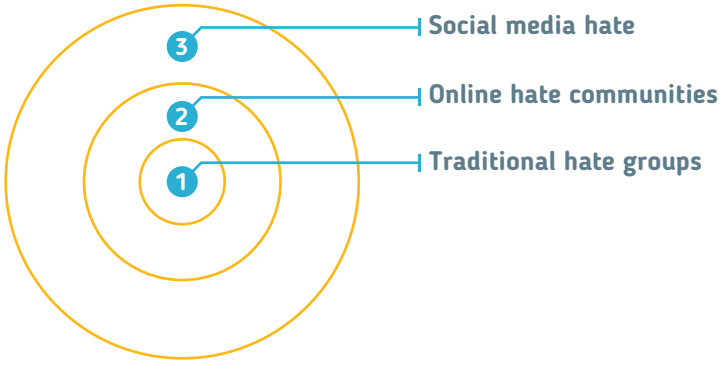

***Figure 2**. The stratification of online hate (Keipi et al., 2017)*

Because of the increasing popularity of social media, the visibility of hate has substantially increased. Meanwhile, online discussion and debate on topical issues has the tendency to rapidly spiral into heated disagreement and polarisation. Justine Sacco's infamous tweet *"Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!"* provides a perfect case of how statements on social media can result in stronger negative messages than was perhaps originally intended. *"This quickly stirred social media into a frenzy. This misstep eventually resulted in her losing her job while drowning in a sea of tens of thousands of hate-filled messages from strangers around the world. Justine Sacco claimed her tweet was intended as an ironic statement of the existing class differences between the Western world and developing nations rather than as a racist remark. Regardless of initial intent, the pace and scale of escalation in this example represent a new phenomenon of relatively small actions taking hold online and creating significant consequences. Context rarely counts for much when things get out of control in social media, whether in a positive or a negative sense" (Keipi et al., 2017, p. 112).*

3.4. Risks, causes and consequences

While the basic principles on which hate speech operates are still the same, the shift from offline to online environments has an impact on how it spreads and the consequences it can have. Individuals do not become someone entirely different when going online. Yet, the sources of available information and the kind of interaction people can have do significantly differ. Within this context, a number of theoretical frameworks explain why young people in particular may be more likely to experience or engage with online hate speech.

a) Young people and online risks

Risk of varying degrees is a normal part of everyday life. Individual choices can affect the outcome of one's routine in any number of positive or negative ways. This is particularly true for young people going online. Young people tend to be early adopters and active users of new technologies, as illustrated by the rapid take up of email, chatroom, texting, instant messaging, blogging and social networking. At the heart of these developments is their desire to explore and develop their identity; to connect with peers anywhere and anytime; to stay in touch, express themselves and share experiences, while having fun together. In this way, youth mix and explore multiple forms of communication, with online technology primarily

used to sustain or extend existing interests or friendships already established offline (Livingstone & Brake, 2010). Yet, while doing as such, young people also encounter online materials accidentally. They become lost in the interconnections of the web. Because they are, in many ways, less mature and less experienced when it comes to critically reflecting on the type of content, users and behaviours they encounter, the risk can more easily turn into harm (Keipi et al., 2017).

From an empirical point of view, ample evidence exists about the widespread availability of hate speech material, even if it can be very hard to track it down as it moves and is replicated across websites and platforms. The most obvious starting point for grasping the scale of the problem is to look at data recorded as part of user notifications, alert or complaint schemes. More innovative approaches also draw upon machine learning techniques, scraping for hate speech messages on social media, taking into account sentence structure and other contextual factors (Gagliardone et al., 2015; Silva, Mondal, Correa, Benevenuto, & Weber, 2016)

Based on these kind of data, Gagliardone et al. (2015) claim that the majority of online hate speech cases worldwide target individuals because of their ethnicity and nationality. A study about hate speech in Twitter and Whisper showed that people were targeted mostly because of their race and behaviour and less because of physical factors, sexual orientation, class, ethnicity, gender, disability and religion (Silva et al., 2016). Marwick and Miller (2014) suggest that apart from people of colour, women and sexual minorities are most likely to be the victims of hateful speech online. Within Europe, the Muslim community is often identified as a key target group of online hate speech, with Islamophobia on the rise more generally (Jubany & Roiha, 2015). Results from a first monitoring exercise carried out under the Code of Conduct on countering illegal hate speech online identified the following grounds for reporting hatred: race, colour, national origin, ethnic origin, descent, religion, anti-Muslim hatred, anti-semitism, sexual orientation or gender-related hatred. Here also, a large number of cases corresponded to some form of anti-migrant speech identified on the grounds of anti-Muslim hatred, ethnic origin or race (European Commission, 2016). Ongoing ECRI (2018) monitoring work points to Jewish, Muslim and Roma communities being particularly affected by hate speech. The International Network Against Cyber Hate (INACH) analysed data from Austria, Spain, Belgium, France, Germany and the Netherlands showing that anti-semitism, racism, anti-Muslim hate and anti-refugee hate were highest on the list (Berecz & Devinat 2016abc, 2017abc).

Relevant as these findings might be, the prevalence of online hate speech in absolute terms tells little about the risk of children and young people being exposed to it. For this, survey research is much more informative. In a large cross-national comparison across EU countries, Livingstone and Haddon (2009) identified "seeing violent or hateful content" as the third most common online risk, experienced by approximately one third of teenagers, ranking it after "giving out personal information" and "seeing pornography online", but before "being bullied", "receiving unwanted sexual comments" or "meeting an online contact offline". Keipi et al. (2017) present comparative cross-sectional survey data from teenagers and young adults aged between 15 and 30 in Finland, Germany, the UK and the US. Respondents were asked, "In the past three months, have you seen hateful or degrading writings or speech online, which inappropriately attacked certain groups of people or individuals?" The average rate of exposure in the four countries was approximately 42 percent. Compared to these figures, very few young people (admit to) (be member of a group that) produce(s) online material that other people would interpret as hateful or degrading.

In sum, regardless of the many positive affordances of digital technology, hate speech has become relatively commonplace online, with young people being exposed it alongside other types of offensive and possibly harmful content, conduct or contact. The nature of social media further exacerbates matters, as they play a significant role bringing together like-minded people in the context of negative or risky behaviour.

b) Identify formation and group dynamics

As implied, young people are eager to explore their personal identity online, consciously creating an idealised self-image, looking for mutual recognition in their peer networks. At its core, this is not a new phenomenon – teenagers have long decorated their bedroom walls with images to express their identity, while keeping a diary or photo album, sending notes and chatting to friends. Yet, the affordances of social media in particular have changed *"the quantity and, arguably, the quality of communication: these include the ease, speed and convenience of widespread access and distribution of content, connectivity throughout a near-global network, the persistence and searchability of content over time, the facility to replicate, remix and manipulate content, and settings for managing conditions of privacy, anonymity and exchange" (Livingstone & Brake, 2010, p. 77).*

The concept of self-presentation helps to explain how individuals explore and tailor the self to please the group they want to belong to. This also happens in the offline world; yet, the scope and ease of access to others is dramatically different, as well as the degree to which the self can be freely tailored and polished, with the accuracy of displayed identity characteristics often not being verifiable due to the lack of physical presence. In this sense, young people's social identity is what connects young individuals to the social spheres they explore in an online environment. As they navigate the social landscape online, they continuously explore more or less favourable identity characteristics, in interaction with known and unknown partners who function as points of comparison (Keipi et al., 2017).

At the same time, a process of self- and group-categorisation is likely to take place. While categorisation is a fundamental human cognitive process to recognise and understand reality, it can also lead to the formation of stereotypes and, by extension, prejudice. As identity groups become more defined and distinct from one another, how one views oneself and others can shift and develop to a point otherwise unlikely. That is, a process of categorisation is likely to occur where distinctions between "we" and "others" develop. As such, complex individuals risk being simplified into stereotypical caricatures, perceived as members of a specific group, based on characteristics such as ethnicity, skin colour or gender. This paves the way for polarisation and prejudice, where the in-group ends up being more favourably considered, standing above the other(s) who are being denigrated (Assimakopoulos, Baider, & Millar, 2017).

In its more dramatic forms, group identification processes may move the individual to depersonalisation, acting purely in accordance with group characteristics and norms. Nowhere is this more apparent than in the message boards and comment sections of the internet where heated arguments routinely occur among strangers. *"The online setting truly offers a fresh glimpse into the dynamics of social identity processes: the comment section of a benign YouTube video can turn into a battlefield of philosophy, religion, politics or anything else for that matter. Stereotyping and depersonalisation are centrally important to the dynamics of online hate, where categorisation and group norms work to damage a targeted group or individual. As individuals are grouped into negative categories, targets are no longer seen as unique individuals but rather as representatives of a hated group whose concept is driven by an oversimplified prototype. Skin colour, sexual orientation and religious*

belief are all examples of characteristics used to trigger hate despite the multitude of other characteristics held by the victims in question" (Keipi et al, 2017, p. 27).

Finally, anonymity – which can range from less anonymous to more anonymous, that is, from visual anonymity to pseudonymity, then full anonymity (Keipi, 2015) – may further enhance stereotyping and polarisation between groups by limiting the visibility of the complex "other" while also potentially lessening the accountability of the users in question (Keipi et al., 2017).

In sum, through processes of self-presentation, self-categorisation, depersonalisation and anonymity, the online world provides a fertile ground for hate speech and polarisation to thrive. As these processes take place, harmful content or action can end up being justified by group norms rather than individual agency.

c) From causes to consequences

As we have implied, nowadays, the risk for young people to somehow encounter hate speech online is substantial. While risk does not inevitably constitute harm (see Livingstone & Haddon (2009)), exposure to online hate does increase likelihood of individual or societal damage, with a distinction to be made in terms of the various forms harm can take.

In general terms, online hate speech may cause direct and indirect effects on individuals' psychological wellbeing, short and long term, with the amount of damage significantly bigger in case of victimisation in comparison with mere witnessing (Oksanen et al., 2018). Once could argue that the consequences of hate speech are similar in form (but sometimes not in intensity) to the effects experienced by recipients of traumatic experiences. That said, responses will be mediated by past experiences, psychological and physical strength, the available sources of support, and so forth (Leets, 2002). More specifically, victims of online hate speech may show low self-esteem; they may feel lonely or isolated, or suffer from sleeping disorders, increased anxiety and feelings of fear and insecurity; their human dignity might be violated, no longer seeing themselves as good and appropriate, in accordance with socio-cultural norms (Keipi et al, 2017; Leets & Giles, 1999; White & Crandall, 2017).

Apart from its impact on individual wellbeing, one should also factor in how online hate speech has a wider societal cost. It may lead to the normalisation of

discrimination, intolerance and hateful attitudes and behaviour. In this regards, Jubany and Roiha (2015) describe a number of narratives from young people they interviewed, reflecting a widespread laissez-faire attitude, of being indifferent, seeing hateful comments as jokes, minimising the impact, or linking hateful content to freedom of speech and everyone's right to express their opinions. In addition, online hate can create a wider climate of fear and polarisation, with social cohesion being threatened by hostility (Oksanen et al., 2018).

Finally, in its most aggressive form, hate speech may encourage – and therefore lead to – abusive, harassing or insulting conduct, including physical violence (Citron & Norton, 2011). When hate speech takes the form of conduct that is in itself a criminal offence, it may be referred to as hate crime (ECRI, 2016). In the latter sense, one can envisage a ladder of harm, with many gradations, starting from acts of bias and discrimination, moving up towards bias motivated violence such as murder, rape, assault or terrorism (Anti-Defamation League, 2018).

3.5. Towards a contextual understanding

In this chapter, we have developed our understanding of what hate speech is further, firmly placing it within a digital environment. For the purpose of the SELMA Toolkit, we will integrate the various elements touched upon, broadly defining online hate speech as:

- Every form of expression, including text messages, images, music, videos, games, or other symbols and signs;
- Disseminated by any possible form of digital media, including websites, forums, blogs, email, social media platforms, or other online communication channels;
- Targeting an individual or a group of people based on protected characteristics, such as race or ethnic origin, gender or gender identity, sexual orientation, religious affiliation, or disability;
- With the intention of inciting, spreading or promoting hatred or other forms of discrimination, or when the message can be reasonably understood as likely to produce that effect.

Or, in simplified form:

- Any online content targeting someone based on protected characteristics with the intent or likely effect of inciting, spreading or promoting hatred or other forms of discrimination.

While this definition provides guidance on how to identify and recognise online hate speech, it only constitutes a first step towards a more nuanced contextual understanding of how children and young people may relate and respond to it. To that effect, we have analysed in further depth a number of key online hate speech features and dimensions, while providing insight into its causes and consequences. As we have seen, online media play a crucial role in the stratification of hate. Meanwhile, the risk for harm largely depends on identity formation processes and online group dynamics.

These lines of thought will provide a springboard for the SELMA Toolkit. They will help to move beyond rigid discussions of what hate speech is, aiming for a more dynamic approach, enabling children and young people – and their educators and carers – to understand and respond to online hate speech in context.

4. National perspectives: Denmark, Germany, Greece and the United Kingdom

As part of the SELMA literature review, partners carried out a landscape review on how online hate speech is defined and which relevant policies and legislation are in place in each of their respective countries: Denmark, Germany, Greece and the United Kingdom. While partners were asked to develop their contribution around the same set of questions, they were expected to provide answers in line with their interpretation of how the subject is viewed and addressed in the countries in which they are based, and the role their organisation could possibly play within this context. This resulted in pieces which vary in scope and tone, but all give more concrete insight on how online hate speech is currently approached across a subset of EU Member States.

4.1. The Danish perspective (by CfDP)

Prologue about the Danish context

As a prelude to the following pages, it should be noted that in Denmark – outside of some academic circles – hate speech, online or offline, is not a very clearly defined, understood and used term in the public debate. This does not imply that people living in Denmark are not subject to the problems and consequences associated with hate speech as a phenomenon. However, it does mean that giving a precise overview is challenging, as the phenomenon is only rarely investigated and discussed on its own, but rather as part of (or in relation to) wider societal issues.

This is also a likely reason why, when we at Centre for Digital Youth Care (CfDP) ask young people if they have ever experienced being the target of online hate speech, they are very often uncertain of exactly what that question means. If they give examples, it is often comments which, although unkind, would not normally be considered hate speech (individual bullying online, rooted in school yard disagreements and the like).

The phenomenology of online hate speech

With all that said, a study on Hate speech in the public online debate published by the Danish Institute for Human Rights in 2017 takes the ERCI definition as a starting point, and expands it:

> ***"Publically voiced stigmatising, derogatory, offensive, harassing and threatening statements that are directed at an individual or a group based on the individual's or group's gender, ethnicity, religion, disabilities, sexual orientation, age, political beliefs or social status."***
>
> *(Zuleta & Rasmus, 2017, p. 22)*

This definition is somewhat broader than both the relevant legislation of the Danish criminal code, and the ECRI definition. Still, being one of the more significant (and recent) empirical studies on the topic in Denmark, it is our most relevant and up-to-date point of reference.

The evolution of the phenomenon through time

The Danish Institute for Human Rights study largely draws upon an analysis of 2,996 comments taken from the Facebook profiles of the two largest news channels (DR and TV2). According to this analysis, about 1 in 7 (15 per cent) of the comments which were allowed to remain on the profiles (after the official moderation had taken place) could be characterised as a hateful utterance, in accordance with the above definition. Considering that Facebook is by far the most widely used social media in Denmark, these numbers do give an indication of the current scope of the phenomenon.

Furthermore, the study contains data from the years 2013-2015 regarding the number of reported violations of paragraph 266b (commonly known as "the racism paragraph", see elaboration below) which indicate that there has been a significant rise in 2015 (45 reports, compared to 30 in 2014).

It should be noted that most of these cases were dropped before reaching the court. Therefore, it is uncertain if the numbers translate into an increase in the number of violations of the law. Still, the numbers seem to suggest an increase in the number of utterances that are (at least initially) perceived as potentially being unlawful. That said, as significant as a 50 per cent increase in the number

of reported violations may seem, three years of data (which was all we were able to find within our time frame) is a limited period over which to observe meaningful trends.

As a side note to this, the report points out that there is a clear indication that news reports containing hateful utterances (e.g. because of quotes) tend to generate significantly more hateful comments than news reports which do not.

The different target groups of online hate speech

According to the same study, in Denmark, news reports on social media are most likely to generate (some) hateful comments, when they cover one (or a combination) of the following subjects (listed in a random order):

- Religion (particularly in relation to Islam).
- Refugees, migrants and asylum seekers (particularly in relation to people from non-western parts of Europe and people from the Middle East).
- Gender equality (women are often targeted, on account of their gender).
- Violence.
- Crime.
- The judiciary system.
- Foreign affairs (including EU politics).

In addition, most of the hateful comments (but by no means all) are directed at groups of people, rather than at individual persons. That being said, quite a few individual politicians are often targets of hateful utterances also, and the same is often the case for other people who are open about their political affiliations.

Finally – although not limited to specific groups or individuals – politicians and regular citizens alike claim that the harsh tone of certain online debates sometimes keep them from participating. To illustrate, the study refers to results from a recent survey wherein one in four of the local politicians who responded indicated that the harsh tone of the public debate had given them cause to consider whether or not running for office was worth it.

The national legislation concerning online hate speech

As alluded to earlier, paragraph 266b of the Danish Criminal Code (the so-called "racism paragraph") is the chief piece of national Danish legislation commonly used when dealing with online hate speech legally. According to paragraph 266b:

> ***"A person who, either publicly, or with the purpose of dissemination to a large group of people, makes a statement or another kind of message, wherein a group of people are threatened, ridiculed or degraded, on the basis of their race, skin colour, national or ethnic origin, religious belief, or sexual preference, is liable to being punished by being fined, or by being imprisoned for up to two years."***
> *(danskelove.dk/straffeloven/266b, 2018)*

In relation to this, the Danish Institute for Human Rights notes that, as it stands, the paragraph does not in any way cover hate speech with regards to, for example, gender or disability.

Meanwhile, the study also mentions a number of other relevant paragraphs from the Danish Criminal Code (paraphrased and abbreviated to the best of the authors ability) (danskelove.dk, 2018):

- 266: Making it unlawful to force others to specific actions or inactions, through threat of harm, forceful loss of liberty, slander, disclosure of private details, or submission of reports of unlawful activity not reasonably related to the matter which the threat concerns.
- 267-274: Dealing with various unlawful acts of public slander.
- 81, chapter 6: Defining it as an aggravating circumstance if an unlawful act was motivated by the victim's ethnicity, religion or sexual preference.
- 81, chapter 7: Defining it as an aggravating circumstance if an unlawful act was motivated by a victim's lawful contributions to the public debate.

The report points out, however, that these days the courts have a clear tendency to put a high priority on freedom of speech, meaning that in practice only very clear and egregious violations of the law will result in convictions.

The national programmes of prevention or intervention related to online hate speech

Referring back to our prologue, it should be noted again that – outside of certain academic circles –hate speech is rarely discussed as a stand-alone issue, but rather within the context of wider societal issues. For this reason, in Denmark, the number of national programmes aimed at prevention or intervention in relation to hate speech online is rather limited.

One example we were able to find is the (now completed) campaign/initiative "Hack Hadet" (Hack the Hate – www.hackhadet.dk), in which young people from around the country were invited to participate in hackathons, aimed at developing measures and campaigns for countering and "combating" hateful comments online. The campaign was initiated by the Danish Center for Prevention of Extremism. This, of course, is not a knock against the campaign in any way, but it is, in a way, rather telling, when it comes to how online hate speech is (or isn't) usually debated on its own, outside of specific contexts.

Perhaps the most prevalent of these specific contexts is that of religious and political extremism, which is also one of the main focuses of the authorities, when it comes to preventive measures and intervention. While certainly relevant and related to the topic of online hate, it is mostly treated as a security issue in relation to perceived potential terror threats.

Some other examples exist, where the prevention of hate speech has become part of education of awareness programmes promoting "good digital etiquette, online citizenship and 20th century skills." Here, young people learn how to behave and to be careful when engaging with, and making use of, the online world, with online hate speech becoming one small part of the many aspects to be covered.

4.2. The German perspective (by LMK)

Online hate speech is a well-known phenomenon in Germany. A recent survey study conducted among a representative sample of German internet users demonstrated that 78 per cent of the participants have already witnessed hate speech online. Young people between 14 and 24 years old are, on average, more often exposed to online hate speech than the other age groups (LandesanstaltfürMedien NRW,

2018a). Despite the pervasiveness of the phenomenon in Germany, to date, there is no overarching definition of the term hate speech (German: Hassrede). The term is rather considered as a political term instead.

In an attempt to specify the term hate speech, the non-governmental organisation No-Hate-Speech-Movement Germany formulated the following definition: *"As hate speech, we define verbal actions against individual persons and/or groups, which aim to devalue or threaten them because they belong to a disadvantaged group in society. This person or group does not necessarily have to be a minority, and minorities are not necessarily disadvantaged. For us, examples of hate speech are: sexism, racism (against Muslims), anti-semitism, antiziganism, neo-Nazism, classism (i.e., discrimination of people from "lower" social strata), discrimination of disabled people, homo- and transphobia" (own translation, No-Hate-Speech-Movement Germany, 2018). These hate speech examples are similar to those provided by the German Federal Agency for Civic Education (FACE), except that FACE also considers lookism, that is, discrimination based on looks, as a form of hate speech, too (FACE, 2018).*

Currently, hate speech is not a legal term in Germany. The German justice system only differentiates between admissible and inadmissible freedom of expression. Freedom of expression is a fundamental right in Germany, however, it is not granted without any restrictions. In fact, actions that threaten human dignity, such as slander, offence or incitement of people, are not covered by the fundamental right to freedom of expression (Arbeitsgemeinschaft Kinder- und JugendschutzLandesstelle NRW e.V. (AJS), 2016). In Germany, the most relevant statutory offences listed in the penal code (Strafgesetzbuch (StGB), 1998) and related to hate speech are:

- **Incitement of people** (Volksverhetzung; § 130 StGB): Is considered to be an offence when someone disturbs public peace by inciting hatred and violence against individual people or whole groups because of their origin, ethnic or religious affiliation. Incitement is punishable by a fine or imprisonment of three months to five years.
- **Offence** (Beleidigung; § 185 StGB): Is an attack on the honour of a person by disregard. If evident, people convicted of offence can expect fines or jail sentences of up to one year.

- **Defamation** (ÜbleNachrede; § 186StGB): Relates to spreading a false claim about someone. The perpetrator spreads the false claim in the belief that it is true. Defamation may be punished by a fine or imprisonment of up to one year.
- **Slander** (Verleumdung; § 187 StGB): Means that lies are deliberately disseminated. The perpetrator knows exactly that his/her assertion is not true, therefore, slander is classified as particularly sneaky. It is punishable by imprisonment of up to five years or by fines.
- **Coercion** (Nötigung; § 240 StGB): Relates to someone pronouncing or writing death threats or assault threats to force someone to do something that (s)he does not want to do. Even the attempt of coercion is punishable. It is punishable by imprisonment of up to three years (in severe cases up to five years) or fines.
- **Threat** (Bedrohung; § 241 StGB): Threatening somebody is punishable by imprisonment of up to one year or fines.
- **Public call for crimes** (ÖffentlicheAufforderungzuStraftaten; § 111 StGB): Public calls for committing crimes are punishable by imprisonment of up to five years or fines.

In addition, the German government aims to combat online hate speech by means of the recently introduced Network Enforcement Act (NEA) (German: Netzwerkdurchsetzungsgesetz “NetzDG”). In short, the NEA, which entered into force on 1 January 2018, obliges operators of profit-oriented social networks to delete «obviously criminal content» within 24 hours after receipt of a complaint. Failure to comply with this requirement may impose fines of up to five million euros on companies (Federal Ministry of Justice and Consumer Protection, 2017). The NEA is a highly controversial topic in Germany – on both sides of the political spectrum. While the far-right political party Alternative for Germany (Alternative für Deutschland, AfD) considers the act as a “Stasi method”(referring to the censorship in the former communist East Germany), left-wing critics accuse the German state of outsourcing work to private companies that should be carried out by judicial bodies (Oltermann, 2018).

In response to these points of critique, the State Media Authority of North Rhine-Westphalia (NRW), the Central and Contact Point Cybercrime NRW, the State Office of Criminal Investigation NRW, the Police Headquarter Cologne, and several regional media companies have initiated the working group “Prosecuting

instead of deleting – Law enforcement on the internet" (Verfolgenstattnurlöschen – Rechtsdurchsetzungim Internet). The initiative, which was launched in 2017, aims to ensure effective law enforcement on the internet by coordinating policy, oversight, law enforcement agencies, and media outlets in their efforts to prosecute online hate speech (LandesanstaltfürMedien NRW, 2017). During the first 70 days after the initiative's operative work began, more than 130 hate speech postings have been reported to the Central and Contact Point Cybercrime NRW. If successful, this approach might become a model for the rest of Germany (LandesanstaltfürMedien NRW, 2018b).

4.3. The Greek perspective (by FAH)

Greece is among those countries where respect for human rights is a cornerstone of democracy. Any means of expression that offends, insults or slanders may be subject to restrictions by law.

Since the 1950s, provisions have been made in the Penal Code, which framed restrictions on freedom of expression when it caused disruption of the citizens' peace (articles 191-192) or constituted a threat (article 333), vulgarisation (article 361) or defamation (article 362). The European Convention on Human Rights (ECHR) became law of the Greek State two years after the Penal Code (with Law 2329/1953) and included, together with the definition of freedom of expression, a list of possible restrictions on it.

In 1970, Greece ratified the International Convention for the Elimination of All Forms of Racial Discrimination (ICERD) by Law 494/1970, which required the contracting parties to prosecute discrimination, including racist speech. Greece validated the ICERD without any reservation, unlike other countries which expressed reservations about the articles criminalising racist speech, such as the United States.

In accordance with these ICERD articles, Greece introduced the anti-racist Law 927/79, according to which behaviour targeting individuals or groups based on their racial or ethnic origin is a crime. The law put forward the concept of public expression of racist discourse, which can be performed orally, by written texts and illustrations or by any other means. This formulation can easily include the internet. In addition, the law covers all instances where a perpetrator incites,

provokes, exacerbates or exhorts acts or actions that may cause discrimination, hatred or violence, harm or damage to property to persons or groups of persons that have been targeted (Article 1 paragraph 1, 2 of Law 927/1979). Later, with Law 1419/1984, targeting religion was added to the offences.

Under Law 2462/1997, Greece ratified the International Covenant on Civil and Political Rights with which *"Any invocation of national, racial or religious hatred, which is the cause of discrimination, hostility or violence, is prohibited by law".*

Greece then acquired a new institutional framework for the criminal treatment of racism and xenophobia, with a new anti-racist Law 4285/2014, now making explicit reference to the use of the internet as well. This law extends the basis of criminal discrimination from racism and xenophobia to include gender, sexual orientation, gender identity and disability. Broadening the legislation was not without its challenges. The process annoyed groups of citizens targeting homosexuals and putting pressure on the non-voting of the bill, citing freedom of speech.

Law 4411/2016 ratified the Council of Europe Convention on Cybercrime and the Additional Protocol on the criminalisation of acts of racist and xenophobic nature committed through computer systems.

Greece, due to its geographic location, which is a crossroad between East and West, has experienced the arrival of large waves of immigrants in recent years from countries mainly in war zones. As a result, various political figures and parties have exploited fear and hatred for foreign identities; a rhetoric of hatred flourished inside and outside the parliament. Meanwhile, debates on the reconciliation agreement of homosexual couples, as well as a bill on gender identity change, illustrated the prevalence of homophobic views among several political figures and religious ministers.

In January-December 2017, the Incident Reporting Network of Racist Violence recorded 102 incidents of racist violence with more than 120 victims. In 34 cases, immigrants or refugees were targeted due to ethnic origin, religion, colour and/or gender identity. In seven cases, human rights defenders and employees of organisations or agencies providing services to refugees were targeted. In 47 cases, LGBT people were targeted. In 11 incidents sacred or symbolic sites and the

Jewish community were targeted, and in two incidents native people were targeted due to religion. In one incident, a Roma man was targeted due to his ethnic origin.

However, these figures differ greatly from official police figures. The Ministry of Justice, Transparency and Human Rights, alarmed by this mismatch, and taking into account the very few prosecutions, asked for the assistance of prosecuting authorities to carry out swift and in-depth investigations against the perpetrators.

Legal measures are very important, but not enough. Education is a long-term solution, as it can prevent and denounce hate speech, while promoting solidarity with its victims. In light of this, the Greek Ministry of Education, Research and Religious Affairs has led the national implementation of the No Hate Speech Movement campaign, providing the necessary tools and materials for stakeholders who wish to undertake action. Under the auspices of the Ministry of Education, a diagnostic tool was also created, which investigates the penetration of the rhetoric of hatred among students, measuring exposure, perception and response based on both quantitative and qualitative criteria (Bakuros, V., 2015).

4.4. The UK perspective (by The Diana Award and SWGfL)

The right to free expression

In a recent interview, Conservative Party politician and Chair of the Digital, Culture, Media and Sport Committee, Damian Collins, alluded to the United Kingdom's (hereafter "UK") strong liberal tradition when it comes to free expression, stating: *"In terms of [the UK's] philosophy and our attitude towards free speech and the net, we're probably philosophically closer to the Americans than the French and Germans" (Kanter, 2018).*

Yet, despite the UK's strong liberal philosophical tradition, it has also been noted that "English law has traditionally taken little or no notice of freedom of speech" (Barendt, 2009, p.851). In practice, the risk of speech being stifled was offset somewhat by the fact that the UK had no system of administrative press censorship since the late 17th century. Also, the courts sometimes did draw on the free speech principle in their interpretation of statutory and common law. Nevertheless, it was only in 1998, with the introduction of the Human Rights Act, that freedom of expression gained formal legal protection in the UK. The Human

Rights Act (1998) enshrines the right to free expression under Article 10 of the European Convention on Human Rights, making clear that this comes with "duties and responsibilities" as well as "formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary" (The Human Rights Act, 1998, Art 10:2).

The hate crime phenomenon

Online hate speech in the UK, as well as efforts to address it, must be understood within the broader context of hate crime. As stated on its official website, The Crown Prosecution Service (CPS) defines hate crime as *"a range of criminal behaviour where the perpetrator is motivated by hostility or demonstrates hostility towards the victim's disability, race, religion, sexual orientation or transgender identity."*

While hate crime is not a new phenomenon in the UK, there have been concerns about rising levels in recent years.

i. England and Wales

Figure 3 shows the number of recorded racially or religiously aggravated offences in England and Wales since 2013.

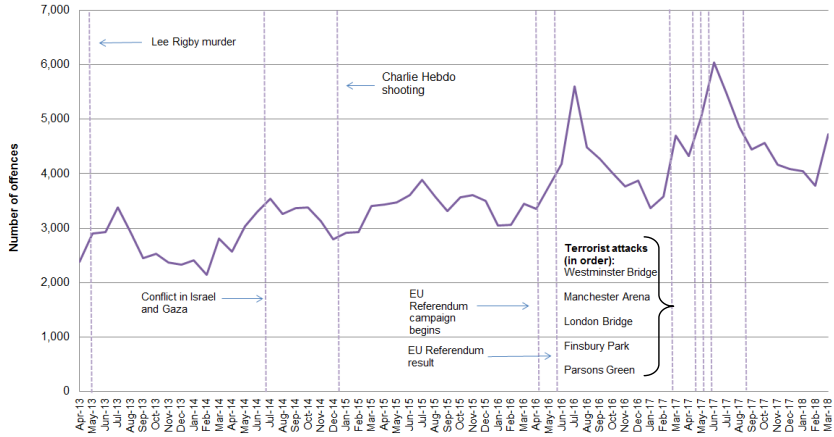

***Figure 3**. Number of racially or religiously aggravated offences recorded by the police by month, April 2013 to March 2018 (Home Office, 2018)*

As the data suggests, there seems to be a link between the offences and socio-political events that occurred over the same period, including, but not limited to, the EU referendum in 2016. Clear spikes in recorded crimes can be seen at the time of, or immediately after the event, with a subsequent reduction thereafter. Generally speaking, over time these crime statistics record an increase in the number of reported offences. However, it should be noted that these figures are not solely online hate crime, but any crime where race or religion is an aggravating factor.

ii. Northern Ireland

The high number of hate crimes witnessed in Northern Ireland have led some to describe it as "the race hate capital of Europe". The most recent police figures show that race is the primary factor behind all recorded hate crimes, followed by sectarianism and homophobia (Police Service of Northern Ireland, 2018). The "stubbornly high" (Criminal Justice Inspection Northern Ireland, 2017, p.5) levels of hate crime in Northern Ireland have been partly attributed to the nation's history of ethno-nationalist tensions, dating back to the three decades known as "The Troubles" that began in the 1960s. Some have argued that the "peace" that followed the 1994 Good Friday Agreement has not adequately addressed those tensions (McVeigh & Rolston, 2007), which continue to be felt in Northern Ireland. These tensions have been amplified by contemporary concerns such as the wave of immigration from EU countries and fears over resources and employment, creating an environment in which hate crime is more likely (Doebler, McAreavey, Shortall, & Shuttleworth, 2016; McVeigh & Rolston, 2007; Montague & Shirlow, 2014).

iii. Scotland

As in Northern Ireland, recent official figures show that racist hate crime is the most common type of recorded hate crime in Scotland, although it has fallen over the years and currently is the lowest it has been since the police began collecting data in 2003 (Crown Office and Procurator Fiscal Service, 2018). However, Scotland has witnessed an increase in recorded hate crime on the basis of sexual orientation and disability since legislation covering the two protected characteristics was adopted in 2009 (Crown Office and Procurator Fiscal Service, 2018). In 2017, Bridger, Bachmann and Gooch reported that one in five LGBT people in Scotland had experienced either a hate crime or a hate incident over the previous year. Concern over the role of social media in exacerbating hate

crime has led Scottish police to launch a "Be Greater than a Hater" campaign that specifically targets children and young people (BBC News, 2018a).

Hate speech legislation in the UK

Turning to hate speech, just as there is no single hate crime legislation in either England, Wales, Scotland or Northern Ireland, there is no single hate speech legislation in any of the nations. Moreover, cases of hate speech frequently get lumped together with hate crime cases making the concepts difficult to disentangle.

The "stirring up" offences[8], which legislate around public conduct (including speech or published material) intended or likely to incite hatred, would most clearly map onto the hate speech concept (Walters, Brown & Wiedlitzka, 2016). However, it would be misleading to only focus on this legislation. This is due to the fact that wider criminal legislation in each of the UK nations specifies hatred on the basis of a victim's protected characteristic as an aggravating factor when it comes to prosecuting and sentencing. Since the use of hate speech is a clear way for offenders to demonstrate or reveal that they were motivated by hatred when engaging in criminal conduct, cases of hate speech are in practice frequently prosecuted as crimes with hate as an aggravating factor, that is, hate crimes.

As Lord Bracadale (2018) noted in his independent review of Scottish hate crime legislation, "almost every case which could be prosecuted as a stirring up offence could also be prosecuted using a baseline offence and an aggravation." For example, somebody who harassed someone while also using hate speech against them might be prosecuted under the Protection from Harassment Act (1997) with racial hatred as an aggravating factor (Bakalis, 2018).

Below, we summarise the key pieces of legislation that can be used to prosecute hate speech in the UK and briefly summarise some of the commentary around their effectiveness.

8 England, Wales and Scotland: The Public Order Act 1986, Part III, Articles 18-23; Northern Ireland: Public Order (Northern Ireland) Order 1987, Part III.

England and Wales

LEGISLATION	ARTICLE/SECTION	SUMMARY
Public Order Act (1986)	Part III, s 18-23: Acts intended or likely to stir up racial hatred	Makes it an offence to use "threatening, abusive or insulting words or behaviour, or displays [of] any written material which is threatening, abusive or insulting"9 in a public place if "he intends thereby to stir up racial hatred, or having regard to all the circumstances racial hatred is likely to be stirred up thereby". The scope of the law was extended to also cover religion10 and sexual orientation11 however in the case of religion and sexual orientation, the threshold is set higher as material must be threatening (not simply abusive or insulting) and there must be evidence of intention to stir up hatred (likelihood is not sufficient).
Malicious Communications Act (1998)	S 1: Offence of sending letters etc. with intent to cause distress or anxiety	Covers the sending of a "letter, electronic communication or article of any description" containing "a message which is indecent or grossly offensive; a threat…" or "any article or electronic communication which is, in whole or part, of an indecent or grossly offensive nature… if his purpose" is to "cause distress or anxiety to the recipient or to any other person to whom he intends that it or its contents or nature should be communicated." This legislation is subject to the provisions of the Criminal Justice Act 2003.

9 The Public Order Act 1986, Part III, Articles 18-23.

10 Public Order Act 1986 - Part 3A Hatred against persons on religious grounds or grounds of sexual orientation. (Inserted by The Racial and Religious Hatred Act 2006.)

11 Public Order Act 1986 - Part 3A Hatred against persons on religious grounds or grounds of sexual orientation. (Inserted by The Criminal Justice and Immigration Act 2008.)

LEGISLATION	ARTICLE/SECTION	SUMMARY
Communications Act (2003)	S 127 (1): Improper use of public electronic communications network	Makes it an offence to use "a public electronic communications network" to send "a message or other matter that is grossly offensive or of an indecent, obscene or menacing character; or… causes any such message or matter to be so sent." This legislation is subject to the provisions of the Criminal Justice Act 2003.
Crime and Disorder Act (1998)	Part II: Racially or religiously aggravated offences: England and Wales	If just before, during or right after committing an assault12, criminal damage13, public order offence14, or harassment15 "the offender demonstrates towards the victim… hostility" based on the victim's race or religion, or there is evidence that the offender was motivated by such hostility then this will lead to higher penalties than would normally be applied for the baseline offence.
Criminal Justice Act (2003)	Part XII, s 145: Increase in sentences for racial or religious aggravation Part XII, s 146: Increase in sentences for aggravation related to disability, sexual orientation or transgender identity	s 145: Enhanced sentencing will be applied to offenders that demonstrate or were motivated by hostility against a victim on the grounds of race or religion for any offence other than the four listed in the Crime and Disorder Act (1998). s 146: Enhanced sentencing will be applied to offenders than demonstrate or were motivated by hostility against a victim on the grounds of sexual orientation, physical or mental disability, or transgender status.

12 Crime and Disorder Act (1998), Part III, s 29.
13 Crime and Disorder Act (1998), Part III, s 30.
14 Crime and Disorder Act (1998), Part III, s 31.
15 Crime and Disorder Act (1998), Part III, s 32.

Scotland

LEGISLATION	ARTICLE/SECTION	SUMMARY
Public Order Act (1986)	Part III, s 18-23: Acts intended or likely to stir up racial hatred	Makes it an offence to use "threatening, abusive or insulting words or behaviour, or displays [of] any written material which is threatening, abusive or insulting"16 in a public place if "he intends thereby to stir up racial hatred, or having regard to all the circumstances racial hatred is likely to be stirred up thereby".
Crime and Disorder Act (1998)	S 96: Offences racially aggravated	If, just before, during or right after committing an offence the offender showed "malice and ill-will based on the victim's membership (or presumed membership)" of a protected group or if there is evidence that such malice or ill-will was the motivating factor behind the offence, then this is considered an aggravating factor.
Criminal Justice (Scotland) Act 2003	S 74: Offences aggravated by religious prejudice	
Offences (Aggravation by Prejudice) (Scotland) Act 2009	S 1: Prejudice relating to disability S 2: Prejudice relating to sexual orientation or transgender identity	By means of the cited legislation and sections, the protected categories are race, religion, disability, sexual orientation and transgender status.
Criminal Law (Consolidation) (Scotland) Act 1995	S 50A: Racially-aggravated harassment and conduct	Harassment by an offender who demonstrates ill-will or malice towards their victim or is motivated by ill-will or malice towards their victim on the basis of race, as per the terms set out in Crime and Disorder Act (1998), is an aggravating factor.

16 The Public Order Act 1986, Part III, Articles 18-23.

LEGISLATION	ARTICLE/SECTION	SUMMARY
Criminal Justice and Licensing (Scotland) Act 2010	S 25, 28 and 39	Section 25 amends Crime and Disorder Act 1998 and Criminal Justice (Scotland) Act 2003 regarding offences aggravated by racial or religious prejudice. Section 38 introduces new offence to combat threatening or abusive behaviour. Section 39 introduced specific criminal offence of stalking.
Communications Act (2003)	S 127 (1): Improper use of public electronic communications network	"A person is guilty of an offence if he sends by means of a public electronic communications network a message or other matter that is grossly offensive or of an indecent, obscene or menacing character; or… causes any such message or matter to be so sent." This legislation does not address hate directly, but is subject to statutory aggravations.

Northern Ireland

LEGISLATION	ARTICLE/SECTION	SUMMARY
Public Order (Northern Ireland) Order 1987	Part III: Stirring up hatred or arousing fear	"Stirring up hatred" or "arousing fear" against a group of people due to their race, religion, sexual orientation or disability is an offence. Sexual orientation and disability were inserted by The Criminal Justice (No. 2) (Northern Ireland) Order 2004.
The Criminal Justice (Northern Ireland) (No.2) Order 2004	S 2: Increase in sentence for offences aggravated by hostility S 3: Inciting hatred or arousing fear on grounds of sexual orientation or disability	If just before, during or right after committing an offence "the offender demonstrates towards the victim… hostility" based on the victim's race, religion, sexual orientation, or disability or there is evidence that the offender was motivated by such hostility then this will lead to higher penalties than would normally be applied for the baseline offence.

LEGISLATION	ARTICLE/SECTION	SUMMARY
The Malicious Communications (Northern Ireland) Order 1988	A 3: Offence of sending letters etc. with intent to cause distress or anxiety	Makes it an offence to send "a letter or other article" containing either "a message which is indecent or grossly offensive" or "a threat". Does not directly address hate but can be aggravated by hostility under The Criminal Justice (Northern Ireland) (No.2) Order 2004.

Challenges and developments in existing legislation

Criticism of the UK's hate crime and hate speech legislation has centred mainly around its fragmented nature, with calls to consolidate it into a single legislative framework (Bakalis, 2018; Walters et al., 2017; Lord Bracadale, 2018). Further calls have been made for there to be an enhanced understanding of how the different pieces of the legislation should be used (McVeigh, 2018). Concerns have been raised about the fact that the lack of consolidated legislation means some hate crimes are prosecuted more or less harshly depending on the protected characteristic of the victim, effectively creating a "hierarchy of hate" (Isaac, 2016). For example, the five strands of documented hate crime across the UK are currently race, religion, sexual orientation, gender identity and disability. But the "stirring up" offences in the Public Order Act (1986) only cover race, religion and sexual orientation in England and Wales, and only race in Scotland. This unevenness has sparked debate and in a 2018 independent review of Scottish Hate Crime legislation (Lord Bracadale, 2018) calls were made for "stirring up" offences to cover the same characteristics as the aggravation laws. Moreover, the review went even further by recommending extending both "stirring up" laws and statutory aggravation laws to also cover age and gender.

With regard to England and Wales, a 2014 Law Commission called for a full review of the aggravation offences and in the absence of a review, proposed adding gender identity, sexual orientation and disability under the terms of the Crime and Disorder Act 1998 (currently the Act only covers race and religion). Hate crimes falling under the Crime and Disorder Act 1998 carry greater penalties than those under the Criminal Justice Act 2003, since the hate element is considered from the start of the case and not only at the time of sentencing (Bakalis, 2018). The government "continues to carefully consider" the Law Commission's proposals (Home Office, 2016, para 111). Controversially, however, the 2014

Law Commission advised against extending the "stirring up" offences to cover additional characteristics beyond race, religion and sexual orientation.

Concerns that the gap between existing aggravation legislation and "stirring up" legislation effectively results in a hierarchy of hate are arguably reinforced further by the fact that even among the characteristics included within the "stirring up" offences, the thresholds for prosecution are higher for some characteristics than others. For example, in England and Wales, the threshold for "stirring up" offences on the grounds of religion and sexual orientation is higher than it is for race, insofar as speech targeting someone on the basis of their religion or sexual orientation must be threatening (not simply abusive or insulting) and there must be proof of intent[17] (Barendt, 2009; College of Policing, 2014; CPS, 2010). However, it should be emphasised that this is very much an evolving landscape. For example, in 2018, as this report was being compiled, a review had been launched into whether misogynistic behaviour should be considered a hate crime – potentially extending the protected characteristics to include gender (BBC News, 2018b).

There has been substantial media attention placed on the number of police referrals and prosecutions for hate crime (BBC News, 2016; McDonald, 2017). When the Crown Prosecution Service (CPS) reported that police referrals in England and Wales had fallen in the two years between 2014 and 2016 compared to previous years, and that the number of prosecutions were down from 15,542 in 2015/16 to 14,480 in 2016/17 (CPS, 2017), this attracted widespread criticism. It was seen as particularly worrying considering the rise in reported hate crime that occurred in the period following the EU referendum (Baynes, 2017; Dearden, 2017; Morley, 2017). Reflecting on these figures, the CPS' report (2017) noted that over the last ten years, conviction rates for hate crimes had increased, as had guilty pleas. Moreover, focusing solely on the overall fall in hate crime prosecution rates over 2015/2016 and 2016/2017 risked masking success stories such as the increase in prosecution of certain specific hate crime strands, notably disability hate crime, where prosecutions increased by over 7 per cent in this period. Nevertheless, the CPS recognised, and laid out plans for, addressing low referral rates and improving the prosecution process.

Criticism of the prosecution process has included those arguing that hate speech laws are poorly understood by authorities. For instance, following interviews with

17 Public Order Act 1986 Part 3, 3A.

the Police Service of Northern Ireland (PSNI) as part of its research, the Northern Ireland Human Rights Commission (2013, p.45) reported that the criminal justice agencies showed "minimal knowledge" about hate speech laws which called into question their ability to properly implement them. In England and Wales, attention has been drawn to the challenges around prosecuting cases of disability hate crime due to disagreements about distinctions between targeting a person with a disability based on their vulnerability versus hostility (Walters et al., 2017). Others have lamented the tendency of victimised minorities to under-report hate incidents (Hall, 2017; Hume, 2017) and most pertinently for the SELMA project, argued that legislation is not fit for purpose for the digital age, resulting in an under-prosecution of online hate crime (Bakalis, 2018; Stray, 2017).

Indeed, significant discussion continues about the unique challenges involved in effectively tackling hate crime that occurs online. In England and Wales, the Government's Action Plan to tackle hate crime (Home Office, 2016) includes, among other things, a commitment to improve the recording of online crime more generally by including a flag next to any recorded crime that included an online component, more police guidance for dealing with online hate crime, a committee to address Violence Against Women and Girls (VAWG) online, and support for children and young people dealing with online hate. In 2015, a Minister for Internet Safety and Security was appointed, with tackling online hate included within their portfolio. In January 2018, a National Online Hate Crime Hub was launched, staffed by a team of specialists tasked with assessing and managing cases of online hate and referring to the appropriate authorities for handling. The expectation is that the Online Hate Crime Hub will increase the number of prosecutions for online hate crime (Home Office Press Release, 2017).

At the start of 2018, the Department for Digital, Culture, Media and Sport introduced the Digital Charter with the stated goals of making the UK "both the safest place to be online and the best place to start and grow a digital business". Listed among its key focus areas was the issue of "online harms", and specifically "protecting people from harmful content and behaviour". Growing out of this focus was the government's Internet Safety Strategy Green Paper (2018) which sets out a number of planned measures to tackle harmful content online – some of which will be elaborated on and formalised in an upcoming White Paper being developed by the Department for Digital, Culture, Media and Sport and the Home Office. Notably, the Internet Safety Strategy puts forward a draft code of conduct for

social media companies. The code emphasises a number of core principles that social media companies should follow. These include having clear and accessible terms and conditions, clear reporting procedures and signposting, as well as providing feedback to those who report content (following the "comply or explain" principle) so that users know what has happened to their report. The Strategy also establishes a first framework for an annual transparency report that social media companies will be asked to submit to the government. The framework will require companies to provide UK-specific data including information such as the number of reports they have received and how these were handled. The draft framework was rolled out for testing and feedback from social media companies in 2018, and it is stated in the Strategy that this will be formalised in 2019. Finally, the Strategy also highlights the possibility of introducing a social media levy – a tax on social media companies which would fund digital literacy programmes, however the exact workings of this have not yet been agreed.

2018 also saw the CPS release Guidelines on prosecuting cases involving communications sent via social media, including "stirring up" offences. There is an ongoing Sentencing Council consultation on increasing sentences for those committing stirring up offences if they have significant influence, which may include having a large social media following or using multiple social media platforms to spread their hate messages (Sentencing Council, 2018; Dearden, 2018). Furthermore, a piece of work was recently launched to review existing legislation around "offensive online communications" (Law Commission, 2018).

National programmes of prevention or intervention related to (online) hate speech

PROGRAMME NAME	ORGANISATION LEAD	LINK
Be Internet Citizens	ISD + Google + YouTube Creators for Change	https://www.isdglobal.org/programmes/education/internet-citizens-2/
Online hate speech guides for young people, practitioners and general public	Cardiff University Social Data Science Lab + Welsh Government + ESRC	http://socialdatalab.net/guides

PROGRAMME NAME	ORGANISATION LEAD	LINK
No Love for Hate, Stop Extremist Content Spreading Online	Educate for Hate	https://educateagainsthate.com/teachers/?filter=classroom-resources
Workshops for young people tackling racism and various other forms of discrimination	EqualiTeach	http://www.equaliteach.co.uk/our-work/#Workshops
Positive Messengers	Languages Company (in the UK)	https://www.languagescompany.com/projects/positive-messengers/
Education and consultancy	Stop Hate UK	https://www.stophateuk.org/
Myths of Immigration school resources	Education Institute of Scotland	http://www.sec-ed.co.uk/news/resources-released-in-bid-to-challenge-asylum-seeker-and-immigration-myths/
School classes and workshops about causes and consequences of racism	Show Racism the Red Card	http://www.theredcard.org/education/ http://www.theredcard.org/noplaceforhate
Free training on tackling hate speech in youth work	YouthLink Scotland	https://www.youthlinkscotland.org/news/march-2018/outside-in-free-training-on-tackling-hate-speech-in-youth-work/
Education resources for tackling hate crime	True Vision (police-funded website where people can learn about and report hate crime)	http://www.report-it.org.uk/education_support
Switch off Prejudice workshop	Anne Frank Trust UK	https://annefrank.org.uk/switch-off-prejudice/

PART II:
EMPIRICAL FINDINGS

5. Qualitative research: focus groups with teenagers

5.1. Introduction

As explained in the literature review, despite ongoing research and policy making efforts across Europe, there does not exist a universally accepted online hate speech definition. The national landscape review from SELMA partners further reinforced this picture; even at national level, it proves difficult to agree on what online hate speech is, let alone on what should be done to regulate or remediate it.

The purpose of PART I of this research report was to draw upon the available literature to move beyond rigid notions of what online hate speech is, trying to grasp the phenomenon in its full contextual complexity. Complementary to this, a piece of qualitative research was carried out as part of the SELMA project, to further tailor and translate the SELMA Toolkit to the perspective and needs of children and young people. In line with this, PART II starts by presenting key findings emerging from a series of qualitative focus groups carried out with 11-16 year olds across SELMA partner countries. These focus groups were primarily designed to explore and analyse the perspectives and experiences of children and young people with regard to online hate speech, while also asking their views and expectations on possible sources of help and coping strategies. In the subsequent chapter, we will bring in further findings from a quantitative survey targeting young people from 13 to 18 years old.

5.2. Methodology

a) Research questions

The main objective of the focus group interviews was to investigate adolescents' views on and understanding of:

- Offline and online hate speech and how these may differ from one another.
- Causes and consequences of online hate speech.
- Sources of help and their effectiveness.
- Strategies on prevention and intervention.

- Anything else of concern that the adolescents wanted to discuss.

In line with this, a number of primary and secondary follow-up questions were shaped and agreed upon in advance.

	PRIMARY QUESTIONS	SECONDARY QUESTIONS
1	How do you understand the notion "freedom of speech"?	Should there be any kind of restrictions as to what extent one expresses him or herself? What if somebody feels insulted?
2	What comes to your mind when you hear the phrase "hate speech"?	Could you tell me some synonyms of this phrase? Could you provide me with a relevant story or describe a message or image referring to hate speech?
3	Do you believe that hate speech can be found in the real world only, or also in an online environment?	Do you think that people who create or spread hate speech online do this in the real world as well?
4	Have you come across hate speech on the internet while surfing? Give some examples.	If yes, where?
5	Can you distinguish different types/ categories of online hate speech?	Which of the people's characteristics does it refer to?
6	Why do you think some people harass others verbally or through images, online?	Do you think that people who use hate speech online have special characteristics or could it be anyone? Why do you think others tolerate it?
7	Why do some people become victims?	Do you think people being the victim of online hate speech (should) receive any kind of support?
8	What kind of emotions do victims feel?	What kind of emotions do people who create and spread online hate speech against others feel?

	PRIMARY QUESTIONS	SECONDARY QUESTIONS
9	What are the consequences of online hate speech on society and young people themselves?	
10	What do you think would prevent and solve this phenomenon?	

b) Data collection

SELMA partners used this set of questions to carry out semi-structured focus groups. A distinction was made between two age categories to be covered, in line with the SELMA Toolkit core target groups: teens 11-13 years old and teens 14-16 years old. Overall, fifteen focus groups were conducted – three in Denmark, and four in Germany, Greece and the UK – with 159 young participants involved in total.

In terms of group characteristics, apart from age, a broad mix of socio-demographic features was aimed for. Participants came from different economic status, education level, age and geographical area of residence. This was a useful strategy to facilitate group dynamics with a variety of views and perspectives being put forward. An effort was also made to have groups where boys and girls were equally represented.

A snowball procedure was followed to recruit participants, drawing upon each partner organisation's professional networks, with teenagers typically being invited through a teacher or school. While most focus groups took place inside the school, some out-of-school locations where teenagers felt free to express themselves were included also. The duration of each focus group ranged from 60 to 90 minutes, with brief breaks if required.

All adolescents participating in the focus groups, as well as their parents or carers, had to sign a consent letter, where information about the SELMA project was presented, including the aim of this particular study, how the information collected would be analysed and managed. The same protocol was followed across all partner countries, ensuring compliance with a broader set of ethical, child protection and data protection policies and principles. The respective points of contact were asked to confirm all participants possessed adequate

communication skills and language competence. It was also stressed that participants had the right to withdraw from the group at any time.

For the actual focus group interviews, a facilitator or moderator each time ensured a similar flow of discussion with key areas of interest begin covered. Focus group discussions were audio recorded, with additional notes taken by a research assistant. Furthermore, all participants had to fill in a basic demographic survey including information on socio-economic background, gender, age, and ethnicity, as well as the average time spent on the internet per day, most visited websites and social media, among others.

c) Thematic analysis

Based on translated transcriptions of the audio recordings, and a first line analysis of verbatim quotes, key words, observations and reactions by SELMA partners, a thematic analysis was carried out, following a number of analytical steps, such as: coding; dividing the codes into categories; searching for connections across the categories; formulating lower and higher level categories; creating broader themes which encompass the various categories for each focus group; combining the themes that were derived from each country's focus group; creating a thematic pap.

5.3. Results

A comprehensive set of themes and subthemes emerged from the thematic analysis, covering a rich variety of adolescent ideas and perspectives in relation to online hate speech, as synthesised below in the form of six master themes:

1. Freedom of speech
2. Offline and online hate speech
3. Hate speech versus bullying
4. Victims, perpetrators and bystanders
5. Perceptions on ways to address online hate speech
6. The social impact of hate speech

a) Freedom of speech

In order to introduce focus group participants to the topic of hate speech, they were asked to present their own understanding of what freedom of speech is and whether there should be any restrictions to it.

As we have seen, freedom of expression constitutes one of the essential foundations of democratic societies. Yet, this fundamental right has certain boundaries, for instance in cases of hatred which constitute incitement to discrimination, hostility or violence. In that sense, the underlying aim was to detect in how far adolescents understand freedom of speech and whether they know the difference between, for instance, objectively-based news reporting and analysis – even if it offends, hurts or distresses – and speech which incites hatred or discrimination on the basis of protected characteristics.

Participants from all age groups tended to perceive freedom of speech as the right to "express your opinion", "say or think what somebody likes" without the fear of being criticised or made to feel wrong. They valued the right to express what they want without being forced to say something they do not want to say. Moreover, they related freedom of speech to somebody's right to voice his or her opinion about religion and politics. According to many of the young people, the concept of freedom of speech is not only limited to the act of speech itself but also related to the freedom of expressing emotions, beliefs, thoughts and actions in general:

"You can say whatever you want. You're free to say what you want."
(Group 1, UK)

"I'm allowed to state my opinion without... permission."
(Group 1, Denmark)

Some of the Greek children from the younger age groups (11-13 years old) made a connection between freedom of speech and other human rights, such as the right to learn, to play, to friendship, as well as medical care. One child from the UK equated the absence of freedom to speech to the feeling of being "imprisoned":

"Learning. Play. The need for friends."
(Group 3, Greece)

"It's like when you express yourself and instead of being uh, like, eh, say for example being behind bars and there's no air..."
(Group 2, UK)

While most children recognised the right to express one's opinion to be very important, a lot of attention was also given to the fact that this should happen in an acceptable and civic manner, without offending or insulting. Both younger and older participants emphasised the need for certain restrictions. For example, posing clear rules which prohibit degrading comments based on race, health, disability, sexual preference, religion, nationality, gender or appearance:

"There need to be restrictions otherwise you can say anything, use bad insults or racist comments and say 'I'm free to say what I think because of the freedom of speech'; that's not okay."
(Group 4, Germany)

"Kind of the same thing, except I think there're some special rules when it comes to religion or something. That you can't say that Muhammad lies or something like that..."
(Group 3, Denmark)

"And, like being racist with people that come over from countries that have wars, and came over to us for like, to comfort them and welcome them here. And then, some people are just, get, like, hating on them because they've came over here for, like, safety."
(Group 2, UK)

While these kind of restrictions come close to the freedom of speech and hate speech principles put forward by international bodies such as the Council of Europe or European Commission, it is important to add that participants typically articulated them in a vague, general and sometime inconsistent manner. Still, most participants did demonstrate a basic appreciation of the fact that the need

for speech to be free does not prevent it from being restricted or questioned under specific circumstances.

b) Offline and online hate speech

A systematic finding emerging from all partner countries' focus groups is that while most participants may have come across the phenomenon of hate speech in one way or another, they have difficulty in defining it accurately, and have a limited understanding of its nature, causes and consequences.

Those who can elaborate on a more specific definition express it as *"Saying bad and degrading things about others"* or *"Hating, or speaking offensively to other people online and on social media"*. Further reflecting on the concept, they recognise that hate speech can be directed both at individuals and groups of people with specific characteristics. Interestingly, participants from both the UK and Greece also talked about the existence of hate comments between *"rival football teams"*, while referring to *"competitive activities at school"* as well.

While most of the concepts discussed in the focus groups were somewhat akin to the definitions we derived from the literature review, they were typically presented in equivocal form, as illustrated with some examples below:

"When I am watching a YouTube video, for example, below the video there are comments saying like 'You are so stupid!' or 'You bit**'."

(Group 1, Germany)

"I haven't heard the term. But I guess it means to leave comments to hurt someone you don't even know."

(Group 2, Germany)

"Hatred – talking trash to others online – for instance a group for all of those who hate K... Something like that. If you're in a public group."

(Group 2, Denmark)

"Verbal violence... Sister. Animal, Nigger...Fascists...Fat. Ugly... Gipsy, Origin...Disabled."
(Group 2, Greece)

"Like if something bad happens and someone writes a post about it on Facebook, and everyone shares it and then it's all over Facebook."
(Group, 1 UK)

Young participants in particular often took a very broad view of what hate speech might be, diluting its relation to protected characteristics and possible harmful consequences. While they seemed to understand the intention to hurt somebody's feelings is part of it, they were not aware that hate speech typically goes beyond feelings of dislike and might lead to abusive, harassing or insulting conduct, including physical violence.

When subsequently asked to compare offline with online hate, respondents did point to some relevant distinctions. For example, many recognised that anonymity can foster online hate, because online haters can hide behind their screen and therefore can avoid any repercussions. In addition, focus group participants were aware of how online media make it possible for online haters to reach a lot of people, while making haters feel more comfortable, because they don't have to face the other in person:

"Online there are no consequences – it's also easier for people when they sit behind a screen and aren't able to see the people they put down."
(Group 1, Denmark)

"I think those who do it online, don't do it as often in real life. Because maybe it's harder looking the person in the eye."
(Group 3, Denmark)

"They do that on the internet because they can do it anonymously [...] in the real world, everybody would know who is using hate speech."
(Group 1, Germany)

"They feel safe at home behind their computers and think they can say whatever they want; they wouldn't dare to say it out loud in the street as they'd be scared people might react."
(Group 2, Germany)

When asked where on the internet they are most likely to come across hate speech, social media is most frequently mentioned, in particular Facebook and Instagram. Twitter was mentioned in fewer cases, because Twitter is more widely used by adults than youth. Participants from all countries considered YouTube as a place where you can find a lot of hate also, for instance in the comment sections under popular videos or songs. Other places mentioned were WhatsApp, Tinder and the Snapchat game "Show and Cover". Participants from the UK also referred to online gaming sites. The general assumption is that online hate speech is "actually everywhere".

c) Hate speech versus bullying

One issue which became apparent across countries was the difficulty to distinguish hate speech from bullying, particularly in an online environment. To be sure, even in the academic literature, the similarities between cyberbullying and online hate speech are often emphasised. Still, a number of key differences exist, in terms of the reason behind the attack, the target, the means of expression, how it is disseminated, and measures to combat the issue.

Some of the participants were able to hint at some of these nuances. For example, a participant from Denmark stated that *"hate speech could be bullying, but bullying is not necessarily hate speech"*. One group from the UK (11-13 years old) understood "hate speech as a form of bullying". Several also agreed that, compared to hate speech, bullying is more personal and it usually happens among people who know each other in real life. In some cases, they implied that bullying typically targets individuals, while hate speech is more directly linked to groups with certain characteristics.

Still, by and large, participants tended to find it difficult to keep both terms clearly separate. They also confused it with other cyber phenomena such as "revenge porn", "cat fishing", "stalking", "gossip" and "chain letters".

d) Victims, perpetrators and bystanders

i. i. Depiction of haters

When focus group participants were asked to identify some general characteristics of "haters" and their possible motives, diverging and sometimes contradicting profiles were depicted.

Some described haters as *"shy and quiet in real life", "lonely souls" who "are feeling bad about themselves", "they are posting hate comments as a way to feel more powerful."* Others see them as *"popular people who have doubts about themselves"* with *"a superiority complex"* who *"need to feel strong"*.

Another depiction was that of individuals with negative experiences or in a vulnerable situation, with reference to their sadness, fear, loneliness, anger, jealousy, or low self-esteem. These type of haters feel excluded. For them, hate comments are "a way to become part of the society", "they want attention and a bit of fame", or to "project their inner pain onto others".

The young people from Germany, Denmark and the UK agreed that haters could be anyone: *"Anybody could do it via social platforms", "it could be anyone of us, they don't have to have special characteristics."* A younger participant argued: *"you would think that only a certain group would be mean to people online or anything like that, but really everyone does it"*. In some cases, it is not even intentional: *"They just do it in the heat of the moment," because they "want to express something"*.

While psychological motives and negative feeling of the haters were often mentioned, some rather suggested that haters just want to make fun of the victim "because it is their hobby" or "because they find it cool or funny".

ii. Depiction of hate speech victims

As for hate speech victims, they are primarily characterised as people who are "different", with reference to physical or group characteristics:

> ***"Because they are different than others."***
> *(Group 1, Germany)*

"Most of the time it's just people who are different, people who do different things, people who look different... that's most of the time who would have hate speech put towards them online."
(Group 2, UK)

"Because you are fat."
(Group 1, Greece)

"Your sexual orientation, like if you're gay or straight, or... other things like that."
(Group 2, UK)

"When you are good at something, for example you paint well, and others are jealous. This spoils your plans."
(Group 3, Greece)

In line with some of the depictions of haters we have previously seen, victims are also described as being "shy, vulnerable, psychologically weak, and sensitive; people who are easy to upset as they take things seriously." From a very different point of view, participants also thought about popular music artists or famous people, such as Selena Gomez and Bibi. Others just say "it could happen to anyone."

In the follow-up questions, participants were also asked to share their views about possible consequences:

"They feel left out, like they've got no friend."
(Group 1, UK)

"Because of that online hate... they'd just be scared to like even express themselves."
(Group 2, UK)

"They could feel lonely...Maybe they are mad at the others... Become haters/fight back."
(Group 3, Germany)

"You feel kind of lost, the whole world's against, feel like you're being withdrawn...get stuck and stagnate in their social development...develop social anxiety."
(Group 3, Denmark).

"You're saying to someone hate and that could later lead to like depression and like suicidal thoughts."
(Group 3, UK)

Thus, feelings of depression, isolation, paranoia, social anxiety, self-doubt, disappointment, loneliness and lack of confidence were identified in terms of psychological wellbeing. Poor school performance was also mentioned. As illustrated, some respondents pointed to the possibility of behavioural harm, with victims possibly committing suicide or cutting themselves. This is all very much in line with the consequences we have identified in the literature review.

When discussing the support that hate speech victims might need, in most focus groups, respondents were able to list a variety of possible sources for support, including parents, teachers, peers, and even internet providers. At the same time, participants acknowledged that victims are likely to hesitate to ask for help since "they are afraid that the problem will be dismissed" or "they would not be understood". They also "do not want to make the problem official" or they just "feel embarrassed". Therefore, they want "to deal with the problem by themselves".

iii. Depiction of bystanders

Another theme emerging through the focus group discussions regarded the role of a third group of people that are not victims or perpetrators, but commonly referred to as "bystanders". They are the ones witnessing hate speech situations. They may ignore or choose not to get involved which could reinforce the feeling of helplessness and isolation of the victims or empower the perpetrator. Many campaigns have therefore aimed to encourage bystanders to take a stance, for instance through counter speech or by supporting the victims in other ways.

Participants provided a variety of reasons for why this group of bystanders might not get involved in hate speech situations, with the most popular being that bystanders do not "feel it is their responsibility". Many participants suggested that

in case of interfering or taking action "maybe they'd be a target too, and it's… a bit scary almost to get yourself into that sort of thing…".

Moreover, according to some of the young people, bystanders' actions are largely ineffective:

> ***"It will always be like this. There will always be someone posting hateful comments… There will always be fights and wars, because this is how humans are…You can't do anything."***
> *(Group 1, Germany)*

To a certain extent, this speaks to the increasing normalisation of hate speech we have discussed in the review of the literature, with victims and bystanders having a "laissez faire" attitude, not getting involved or taking any action. In this regards, one group from Greece made an interesting point about bystanders not taking action unless they felt somehow connected to the victim. This alludes to the role empathy can play in affecting change, as we will argue in further detail towards the end of this chapter.

e) Perceptions on ways to address online hate speech

As part of the focus groups, participants had some time to brainstorm on possible prevention and intervention initiatives. Three different types of response to online hate speech were broadly identified.

The most frequent solution mentioned was legislation. This should make it possible to regulate hate speech on websites and social media platforms, while giving the opportunity for users to report violations to law enforcement.

Meanwhile, according to the young people, the different kinds of internet and social media service providers have a role to play as well, putting in place clear rules and guidelines, adequate moderation and reporting mechanisms, and rethinking some of their platform features:

"I think it should be people (not bots) who look at the comments and make decisions, for instance that you're not allowed to write comments for some time or even that you aren't allowed to use social media services at all. I think that would be better than charging money, otherwise they could say 'I can afford it'..."
(Group 1, Germany).

"In games, the way to prevent cyberbullying right, is when you write something and then you send it, the other person can report it...if they think it's not... appropriate... and the person can get banned from using the app."
(Group 2, UK).

"There should be a more obvious, easier way to report something. So maybe something on each page like at the top like on the corner of the screen saying like 'Report this message' or something."
(Group 2, UK).

"Establishing terms concerning privacy policy."
(Group 4, Germany).

That said, some point out that online moderation and reporting tools for online services might not be the magic solution, as the issue will still prevail in the offline environment: "Close Facebook. Close Snapchat. Close Instagram. All of it... It would still happen in real life. It doesn't help just shutting it down".

Another strategy proposed by the young people was to establish formal and informal support spaces for victims. This could take the form of online counselling services, individual therapy, group spaces and teen focus groups. One Danish group explained how the problem of hate needs to be addressed at its roots, for instance by providing support to haters as well, trying to better understand and address their motivations. Participants also acknowledged the relevance of strengthening human relationships among the community and making sure that people take action when hate speech incidents happen.

More broadly, participants talked about the importance of awareness raising and education programmes, ideally from early childhood onwards, while pointing to the role of parents also:

"I think there should be awareness raising about freedom of speech and hate speech."
(Group 2, Germany)

"Introducing some kind of conference at every school, where people talk about it."
(Group 3, Germany)

"Or you could try to strengthen the community. Maybe you could create events helping everybody to feel safer with each other. Maybe people would find it more difficult to bully and not have a reason for it."
(Group 1, Denmark)

"Meetings with parents."
(Group 2, Denmark)

"I think, if people learn from home how to talk to people and how to act, from early on, I don't think it'd be as much of a problem."
(Group 3, Denmark)

In some groups, there was discussion about how haters could be encouraged to apologise to the victims. One teen proposed the creation of forums where haters and victims are brought together, possibly with their educators, parents or carers, while looking for reconciliation.

f) Social impact of hate speech

Finally, respondents across all focus groups discussed the impact that online hate speech can have beyond the individual victims on a broader societal level.

Participants primarily referred to the normalisation of hate speech and people becoming desensitised to it. This was perceived to be a risk for young children in

particular, because they are more prone to routinely imitate roles and behaviours from others, as they learn from observation in their own context. Or, to put it in their own words:

"The more it happens, then people take it as a habit, as a given."
(Group 2, Greece)

"Maybe one day you get the feeling hate is normal and swearwords become normal as they're used all of the time... I'm not sure if young children know what it means but they use really bad expressions, they learn it on the internet with all the hate they see."
(Group 2, Germany)

"Hate comments have become normal and people think as individuals they can't change anything about it."
(Group 4, Germany)

Meanwhile, in line with our findings in the literature review, some participants pointed to social and political consequences, where hate can divide people, since they will have to choose sides in a polarised climate. In this regard, lack of respect and increased prejudice were also mentioned:

"Because of hate speech, freedom of expression is abolished. A society that does not express opinions and ideas will not progress."
(Group 2, Greece)

"It could lead to... a big group of people... leaving that certain crowd and then that separates... those people from other people and then... the young people grow up to believe that same thing and it just sort of like... separates two peoples."
(Group 1, UK)

"Real-world fights...Gang violence."
(Group 1, UK)

5.4. Discussion

The main purpose of our qualitative research was to complement key insights from the literature with a more descriptive account of the perspectives and experiences of children and young people with regard to online hate speech.

Our analysis of six master themes paints a picture of savvy young people who are somewhat familiar with the topic of hate speech and are able to relate it to their personal online media experiences. Yet, while a basic notion of concepts such as freedom of speech and incitement to hatred clearly exists, most participants failed to systematically articulate what online hate speech is, and how it transpires in the current social media landscape.

This is in no way meant to underestimate the pre-existing knowledge and skills these young people have. Because, indeed, respondents were able to discuss the various themes and questions in a very rich manner, often providing concrete and intriguing examples. Meanwhile, the emerging description of victims, perpetrators and bystanders was quite compelling, with many of the views and ideas which were expressed resonating with what has been said in the academic literature. Most respondents also seemed to grasp some of the possible consequences online hate speech may have, both at individual and societal level.

Nevertheless, as the young people argue themselves, apart from regulation and adequate monitoring and reporting procedures, awareness raising and education seem essential. Ideally, this would start from an early age onwards. These programmes should help learners to systemise their thinking about the nature of online hate, its causes and its consequences. They should also address the normalisation of hate, and the indifference and polarisation it may cause. For this, a wholly different approach is needed, one which helps young people to understand and regulate their own thoughts, emotions and behaviours, while empathising with others, including those who are perceived to be different.

6. Quantitative research: an online teen and teacher survey

6.1. Introduction

In the literature review, we provided an overview of what is empirically knows about the risk of exposure to online hate, and the harm if might cause. Unfortunately, there is only a limited amount of empirical data available on this topic, particularly if one wants to focus on the perspective of children or young people. Therefore, a quantitative online survey was developed to better understand two key SELMA target audiences: teens and teachers. Results from this survey further complement the findings which already emerged from our qualitative focus group.

6.2. Methodology

As we have explained, to better understand the nature of online hate speech and how this affects children and young people's everyday online media experience, a contextual approach is needed. Therefore, the survey questionnaire mainly focused on items which provide further insight into teens' and teachers' digital media use patterns, the knowledge and experience they have in regards online hate speech, and the extent to which this topic being dealt with in an educational context.

The online survey was launched in August 2018 with respondents being able to respond until the end of October 2018. No sampling methods were used. Rather, each SELMA partner drew upon a variety of communication strategies to promote the survey in their respective countries. While this inevitably raises a number of issues in terms of representativeness, partners did succeed to reach (after data cleaning) a large number of respondents, more specifically a total of 776 teens and 333 teachers across (in order of response numbers) Greece, Germany, Denmark, the UK and other EU countries. Importantly, the demographic data captured as part of the survey indicated that a good mix of sociodemographic backgrounds was covered.

6.3. Teen survey results

In the review of the literature, we explained how children and young people increasingly use digital media to explore and develop their identity, to connect with peers anytime on their mobile phone, and to stay in touch, express themselves and share experiences, while having fun together.

This seems overwhelmingly true indeed for the teenagers we surveyed. If asked how much time they spend online, 77 per cent or our teen respondents indicate they are using websites or apps on a mobile phone, tablet or computer "several times each day" or "almost all of the time". As illustrated in Table 1, they primarily do this to chat with people or because they are watching videos.

ONLINE ACTIVITIES	n (%)
Do schoolwork	294 (37.9)
Chat with people	614 (79.1)
Read the news	316 (40.7)
Use blogs and/or forums	134 (17.3)
Create and upload my own videos or pictures	193 (24.9)
Play games	365 (47.0)
Watch videos	620 (79.9)
Search for information	392 (50.5)
Sell or buy things	154 (19.8)
I don't know/I prefer to not answer	11 (1.4)
Other	63 (8.1)
Total	**776 (100.0)**

***Table 1**. When you go online, what activities do you do? (more than one option)*

Most of this time is spent on Instagram and YouTube and, to a somewhat lesser extent, on Snapchat and Facebook. For the teens we surveyed, Twitter again seems less often the platform to go to connect with others, as indicated in Table 2, and as already implied based on our qualitative focus group data.

SOCIAL MEDIA WEBSITES OR APPS	n (%)
Facebook	327 (42.1)
Twitter	97 (12.5)
Instagram	634 (81.7)
WhatsApp	261 (33.6)
Snapchat	399 (51.4)
YouTube	594 (76.5)
None, I do not use any of these websites or apps	13 (1.7)
I don't know/I prefer to not answer	9 (1.2)
Other	122 (15.7)
Total	**776 (100.0)**

***Table 2**. These days, do you use websites or apps such as Facebook, Instagram, Snapchat, WhatsApp… to communicate or share information/pictures/videos online with others? (more than one option)*

On the specific topic of online hate speech, we asked respondents if they had ever heard of this concept, to which only 39 per cent of respondents answered yes. While this percentage might seem low, it is interesting to note that among those respondents, several were able to provide some of the key characteristics we have identified in our own review of the literature, when asked to briefly explain "in their own words what [they] think hate speech is", as illustrated in the word cloud in Figure 4.

Figure 4. In their own words... what do teens think hate speech is?

At this point in the survey, we provided all respondents with further background information, stating that "when people talk about hate speech, they often refer to the fact that people sometimes say hurtful or nasty things when talking about groups of people or communities with a certain culture, religion, colour of skin, sexual orientation, or other group characteristics. This may happen in the offline world; at school, in your neighbourhood or even within your family. It may also happen online; on websites, apps, blogs, social media or services you use, or on your email or instant messaging service." In this way, we were able to ask all survey respondents further questions about their experiences of online hate speech, based on a shared understanding of what it is.

Once defined in this manner, respondents to our teen survey reported about their hate speech experiences as follows:

- In the past three months, 57 per cent of teens had encountered hate speech online, once or several times; this is slightly more than the 51 per cent of teens who had encountered hate speech situations in person (i.e. offline/face to face).

- When asked if, in the past three months, they had been treated in a hurtful or nasty way online, 18 per cent of respondents said yes, which is substantially less than the 30 per cent of respondents who indicated that they have been treated this way in person.
- When asked to think about the reasons why someone was being hurtful or nasty to others in an online environment, respondents indicated this was because of "Physical appearance, for example your weight, height, etc." (58 per cent), "Sexual orientation, for example being gay, bi-sexual, etc." (56 per cent), "Ethnicity or nationality, for example being Roma, Refugee, Immigrant, etc." (44 per cent) and "Sex/gender, for being a woman, transgender, etc." (43 per cent) (with multiple answers per respondent being possible). Interestingly, in comparison, the instances of offline hate speech our respondents encountered were mostly confined to physical appearance (60 per cent), with far fewer respondents making a link with sexual orientation (41 per cent), ethnicity (33 per cent) or sex/gender (25 per cent). This may point to an important difference between the nature of offline and online hate young people witness.
- If respondents had encountered hate speech online, it most often happened on a social media website or app (such as Facebook, Twitter, YouTube, Instagram, etc.) (74 per cent), in the comment sections (for instance next to online videos, news articles, etc.) (43 per cent), and (to a lesser extent) in a chatroom (17 per cent), by instant messaging (17 per cent) or while playing video games (16 per cent). When asked specifically about social media websites or apps, YouTube (56 per cent), Instagram (54 per cent) and Facebook (40 per cent) were mentioned most often.
- A majority of respondents (62 per cent) said that when they encountered hate speech online they accidently came across it. 46 per cent of respondents said it was posted/shared by a person they don't know. 20 per cent indicated it was posted/shared by a friend/person they know, which is lower, but still a substantial proportion. Only 8 per cent of respondents said they were actively looking for it.
- If asked how they responded, 30 per cent reported it to the website or app, for example Facebook, YouTube, Instagram, WhatsApp, etc. Yet, many respondents also indicated they rather decided to ignore it, because they didn't care (29 per cent) or because they didn't know what to do (21 per cent). Interestingly, while telling a friend of their age (24 per cent) or supporting the victim by saying something positive (22 per cent) seemed a

viable option for a substantial amount of teenagers, very few respondents indicated they would tell a parent/guardian/teacher (6 per cent). More positively, only 5 per cent said they responded in the same way, with a similar hate speech message.

In Figure 5, we illustrate how online hate speech knowledge and experience pan out as teenagers become older. On average, older students are more likely to have heard about the (online) hate speech phenomenon and they have encountered if more often. By contrast, the percentage of students who reported experience of teachers (or any other educator) addressing the topic in an education setting is low across all age groups.

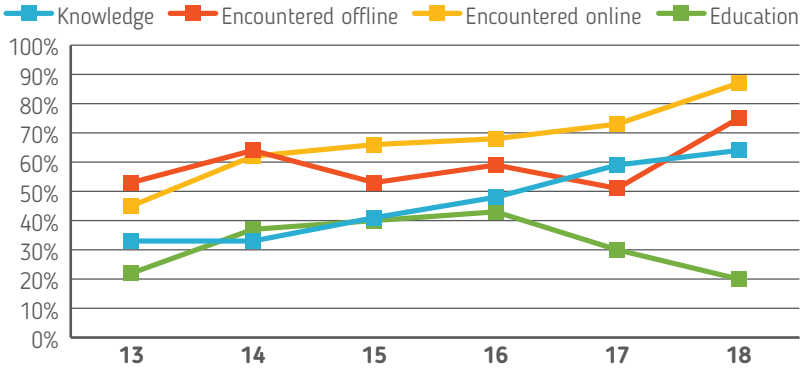

***Figure 5**. Breakdown per age of the percentage of respondents who have knowledge about what hate speech is, have encountered it offline or online, and who had any of their teachers (or other educators) ever talk to them about it.*

As illustrated in Figure 6, we also asked respondents to indicate to what extent they agreed or disagreed with a number of statements. Most respondents rejected the idea that "it is OK to send hateful or degrading messages against someone online if they start to attack you, your friends or family first", that it is "OK to send hateful or degrading messages against someone online because hate speech is fun", or that "hate speech online is just words". Meanwhile, a majority of respondents feels "hate speech online reflects the tensions within a society".

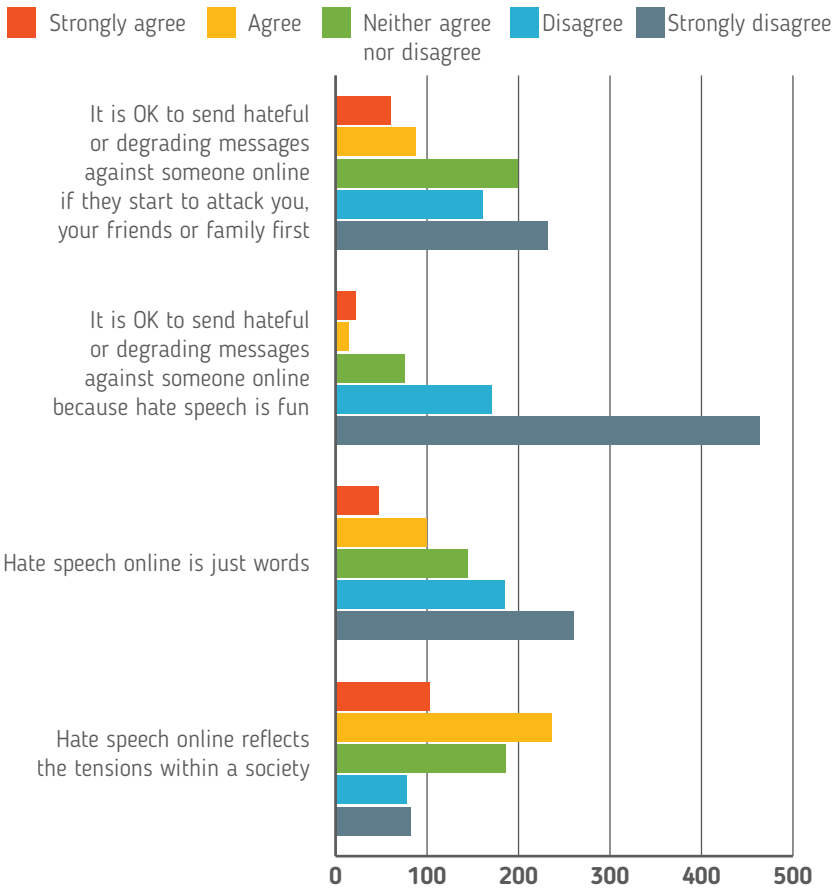

Figure 6. *Thinking of hateful or nasty messages and comments on the internet, please tell us how much you agree or disagree with each of these phrases.*

Finally, we asked respondents some additional questions about online hate speech education (see also Figure 5). In total, only 33 per cent of respondents indicated that any of their teachers (or any other educator) has ever talked to them about it. From those respondents, 77 per cent said it was interesting, while 73 per cent found it useful. That said, only 30 per cent said it changed their opinions and only 27 per cent indicated it changed how they behaved.

6.4. Teacher survey results

While the teens in our survey mostly use the internet for communication and entertainment purpose, teachers who responded are clearly more driven by information seeking purposes. When they go online, they do this to search for information (91 per cent), to read the news (85 per cent), to do schoolwork (81 per cent), and only to a lesser extent to chat with people (54 per cent) or to watch videos (47 per cent). A similar gap becomes apparent when comparing their preferred social media websites or apps. For teachers, Facebook (81 per cent) is much more popular compared with teenagers. By contrast, Instagram (33 per cent) and YouTube (48 per cent) are used by a much lower proportion of teachers, and Snapchat (11 per cent) is hardly used at all.

When asked if they have ever heard about hate speech, 75 per cent of respondents indicate they have. In this sense, the topic is much better known by our teacher respondents compared to the teenagers, and this was also clear from the relevant concepts many teachers were able to provide when asked to explain in their own words, as illustrated in Figure 7.

Figure 7. *In their owns words... what do teachers think hate speech is?*

Teachers were then asked to indicate if any of their students has been involved in hate speech incidents or situations, both in person and online. For instance, a student might have been the target of hate speech or might have been the one expressing or spreading these kind of messages. Table 3 provides an overview of figures for both offline and online incidents. At first sight, these figures suggest that hate speech incidents more often takes place offline. However, in our view, the data rather reveal that teachers have less knowledge of the hate speech experiences their students have in an online environment. Because, indeed, as we have previously seen, pupils are not likely at all to report these kind of online incidents to their teachers.

STUDENTS INVOLVED IN HATE SPEECH	IN PERSON (%)	ONLINE (%)
Yes	38.1	24.9
No	44.1	33.9
I don't know/I prefer to not answer	17.7	41.1

Table 3. *The percentage of students involved in hate speech incidents, as reported by their teachers.*

When asked how they responded the last time they became aware of an online hate speech situation involving one or several of their students, most teachers indicate that they supported the victim by saying something positive (61 per cent). Many teachers also discussed it with another colleague, the school principal, or another person whose job it is to help children. Far fewer respondents thought about reporting the incident the website or app, as illustrated in Table 4 below.

RESPONSE TO ONLINE HATE SPEECH	n (%)
I ignored it, because I didn't care	2 (2.4)
I ignored it, because I didn't know what to do	3 (3.6)
I discussed it with another colleague	40 (48.2)
I discussed it with the school principal	34 (41.0)
I discussed it with another person whose job it is to help children	34 (41.0)
I reported it to the website or app, for example Facebook, YouTube, Instagram, WhatsApp, etc.	17 (20.5)

RESPONSE TO ONLINE HATE SPEECH	n (%)
I reported it to the police	5 (6.0)
I supported the victim by saying something positive	51 (61.4)
I don't know/I prefer to not answer	1 (1.2)
Other	23 (26.5)
Total	**83 (100.0)**

***Table 4**. The last time you became aware of an online hate speech situation involving one or several of your students, what did you do? (more than one option)*

To develop a more in-depth understanding of teachers' expertise and confidence in dealing with hate speech issues or situations in relation to their students, a number of education-oriented questions were included. First of all, we asked teachers to evaluate a number of statements, as illustrated in Figure 8. Teachers are, on average, quite confident to discuss hate speech with both students and parents. However, they have less knowledge about how to report a hate speech issue on a website or app, and they feel less confident developing educational or awareness-raising programmes or activities in their school.

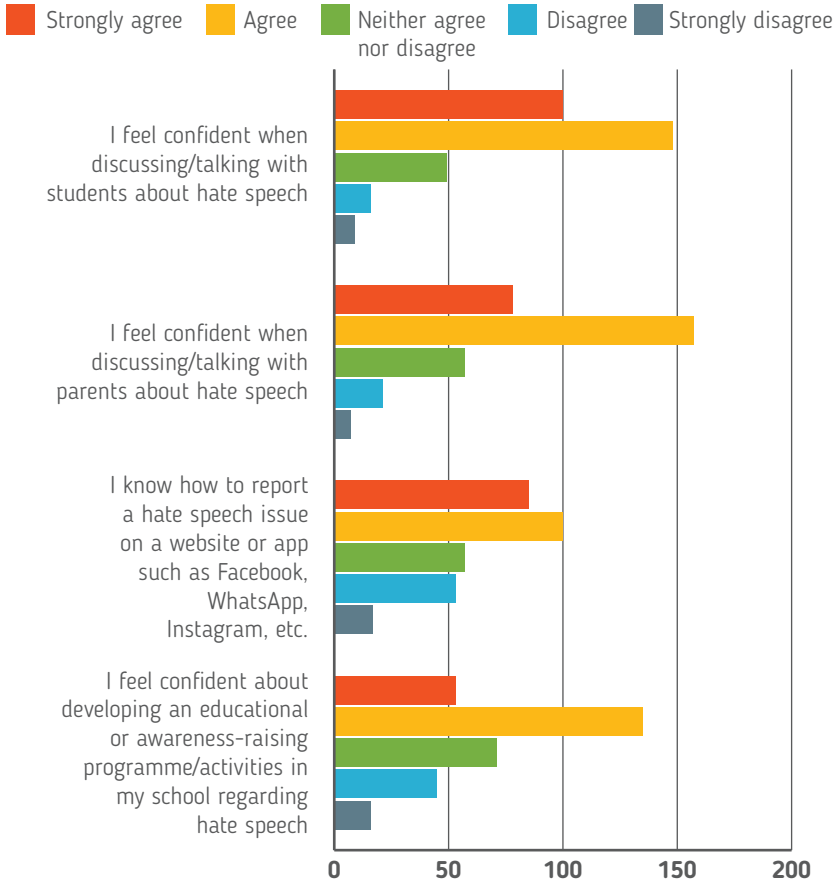

***Figure 8**. For each of the following statements, please tell us how much you agree with them.*

In terms of possible education strategies, we asked in how far teachers were aware of or familiar with social and emotional learning (SEL) or media literacy

(ML) approaches.[18] Only one in three teachers had ever heard about SEL (35 per cent) and ML (33 per cent), and even fewer had actually used SEL (24 per cent) or ML (21 per cent) in their work with students in the last year. More positively, teachers did seem confident about the possible effectiveness of both SEL and ML approaches to work on issues related to hate speech, as illustrated in Table 5.

EFFECTIVENESS OF...	SEL (%)	ML (%)
Very effective	76 (22.8)	65 (19.5)
Effective	150 (45.1)	138 (41.4)
Neither effective nor ineffective	27 (8.1)	33 (9.9)
Ineffective	4 (1.2)	5 (1.5)
Very ineffective	3 (0.9)	0 (0.0)
I don't know/I prefer to not answer	73 (21.9)	92 (27.6)
Total	**333 (100.0)**	**333 (100.0)**

Table 5. *How effective or not do you think a SEL and ML approach would be to work on issues related to hate speech with your students?*

Finally, the correlation matrix in Table 6 gives a more explanatory hint at the systemic shift which might be needed at school level to increase teacher competence and confidence to deal with hate speech issues. Significantly, the latter correlates with having both a positive school climate and a whole-school approach to the use of digital technology and social media.

18 In the survey, SEL has been defined as "process through which children and adults acquire and effectively apply the knowledge, attitudes, and skills necessary to understand and manage emotions, set and achieve positive goals, feel and show empathy for others, establish and maintain positive relationships, and make responsible decisions" (see https://casel.org/what-is-sel/). Meanwhile, media literacy has been defined as "a set of competencies that empowers citizens to access, retrieve, understand, evaluate and use, to create as well as share information and media content in all formats, using various tools, in a critical, ethical and effective way, in order to participate and engage in personal, professional and societal activities" (see https://unesdoc.unesco.org/ark:/48223/pf0000224655).

	TEACHER CONFIDENCE IN REGARDS HATE SPEECH	POSITIVE SCHOOL CLIMATE	WHOLE-SCHOOL EDUCATION TECHNOLOGY POLICY
Teacher confidence in regards hate speech[19]			
Positive school climate[20]	+0.208 (0.0004)a		
Whole-school education technology policy[21]	+0.269 (0.0001)b	+0.335 (<0.0001)c	

***Table 6**. Spearman's rank correlation coefficients are provided (p-values in brackets) a-analysis on 284 participants; b-213 participants ; c-224 participants*

6.5. Discussion

To conclude PART II, we presented results from our quantitative survey which set a more structural point of reference for the SELMA Toolkit.

Overall, our survey data echo many of the main findings of our literature, while giving them a more concrete and contextual flavour. We were able to illustrate how our teen survey respondents indeed use media for communication and entertainment, which directly links to processes of identity formation and group dynamics. By contrast, teachers rather spend their time online for information purpose, while accessing different platforms and communication channels.

In terms of exposure, our data confirmed online hate has become an inevitable part of young people's everyday media experience. Nevertheless, one should not exaggerate the harmful impact it may have had on the individuals we surveyed.

19 This is a composite measure summing items such as "I feel confident when discussing/talking with students about hate speech", "I feel confident when discussing/talking with parents about hate speech", and "I feel confident about developing an educational or awareness-raising programme/activities in my school regarding hate speech."

20 This is a composite measure summing items such as "Students in our school get along well", "Teachers in our school actively work to create a safe and welcoming environment", and "In our school teachers, administrators, staff, students, and parents communicate with one another."

21 This is a composite measure summing items such as "In our school, we have a policy relating to the use of digital devices", "Offer opportunities for pupils to gain a more critical understanding of social media and the internet", and "Have a clear process in place to monitor/report incidents related to the use of social media".

If young people encounter it, it is mostly by accident. Meanwhile, only one in five indicated they had been targeted themselves. And, if it happened, they rarely responded with a similar hate message.

In line with what we have seen in our qualitative focus group discussions, it became clear again that the knowledge and understanding young people have of online hate is fairly limited, with some notable exceptions, as illustrated by the teenager word cloud. Teachers seem to have a somewhat better grasp of what online hate speech is about. Many of them even feel confident to discuss the topic with students and their parents. Yet, when educational or awareness-raising programmes and activities need to be developed in their school, they lose some of this self-assurance. They are also not familiar with approaches like SEL or ML, although they typically are confident about the possible effectiveness of using these approaches to work on issues related to hate speech. Finally, it is teachers who work in a positive school climate with whole-school social media policies who are most confident about their capacity to make a difference in terms of hate speech.

PART III.
THE SELMA RESPONSE TO ONLINE HATE SPEECH

7. A multiple stakeholder approach

In our literature review and empirical findings, we have seen how online hate speech – broadly defined – plays a significant role in teenager's online media experience, and how policy makers and legislator have sought, both nationally and internationally, to address and resolve increasing concerns in this regards.

The Code of Conduct on countering illegal hate speech online, which the European Commission agreed in 2016 with leading IT companies and social media platforms, should be considered against this background, with policy makers and industry stakeholders working together to regulate, monitor, or empower end users to report. As O'Neill, Staksrud and McLaughlin (2013) explain, *"regulation in the traditional sense of controlling by legislation and placing restrictions on market forces has been perceived as antithetical to the flourishing digital economy which all governments seek to support. As a consequence, a variety of voluntary and cooperative forms of regulation between industry stakeholders and government interests have sought to create the optimal conditions for innovation and development of digital opportunities while sensitive to any potentially negative" (p. 13).*

Citroen and Norton (2011) offer a range of examples of the how online companies and intermediaries might exercise their power over hate speech, some of which have also been proposed by the young people in our focus group interviews. First of all, transparency is key, with online media platforms being clear and specific in Terms of Service agreements or Community Guidelines about the harms that their hate speech policies address, as well as the consequences of policy violations. "No matter the particular definition of hate speech that intermediaries choose, an accessible and transparent policy can help users to develop a better appreciation of their responsibilities as they work, debate and connect with others online. Hard judgement calls will inevitably remain, regardless of how an intermediary chooses to define hate speech. But those decisions – however difficult – can be made in a more principled way when an intermediary grounds its policy's hate speech definition and application in terms of the specific harms it seeks to avoid" (Citroen & Norton, 2011, p. 1459). Once these principles have been established, hate speech policies should also be enforced, with companies devising strategies – drawing upon automated technological systems or human resources and judgement – to identify and remove content violating the rules. In addition, acknowledging the deletion of content can further help to support a commitment to transparent and accountable enforcement.

Secondly, even if online intermediaries deem it inappropriate to remove certain types of offensive content, for instance because it is not illegal in a certain country, they can still help with countering hate speech through different means, for instance by providing an alternative narrative. We previously elaborated on the example of www.martinlutherking.org[22] a website run by Stormfront which – until recently – figured prominently among the top results for searches of "Martin Luther King" on Google. Citroen and Norton (2011) describe the similar cases of www.jewwatch.com, a site featuring anti-Semitic content, or the image of the Michelle Obama which was altered to resemble a monkey and maliciously shared online. Both have figured prominently in Google search results and in other search engines. Rather than removing the content or changing the underlying algorithm, Google in both cases inserted an explanatory advertisement to apologise for the possibly upsetting nature of the content, while ensuring it did not endorse the views expressed, and pointing to alternative sources of information on the topic instead. More recently, Google explored a similar online advertising mechanism, using Adwords targeting tools, combing target keywords with Text Ads, Image Ads and Skippable Video Ads, to reach individuals who are sympathetic to ISIS. The underlying idea was to draw upon Google's core areas of know-how and expertise to redirect them towards curated YouTube videos debunking its terrorist recruiting themes.[23] And of course, apart from informative warning signs, social media platforms like Facebook, Instagram or YouTube are well placed to promote and campaign with more positive voices against hate, often in cooperation with civil society organisations, ideally in a language which resonates with youth audiences also.

Thirdly, the role of online intermediaries can go well beyond content removal and counter-narrative strategies, educating and empowering their users to respond to hate speech on their platforms and sites, while creating contexts in which users feel compelled to reflect on their own rights and responsibilities. As we have seen, this is in part a matter of putting in place clear Community Guidelines and reporting tools and processes. However, these kind of civic norms should also be encouraged through architectural choices which discourage speakers from fleeing responsibility for their own hateful expressions. For example, while anonymity is valuable when it enables speakers to avoid retaliation, it can also simply encourage one to avoid responsibility for socially destructive behaviour.

22 The website is currently down, but pages can still retrieved at https://web.archive.org.

23 See https://redirectmethod.org.

Online intermediaries can shape these kind of norms, for instance by permitting anonymity by default, but revoking it when a user violates the Terms of Service. Likewise, systems might be designed to slow down the posting or sharing process in certain circumstances, requiring a waiting or cool-off period, prompting the user to more carefully consider the possible impact of what is going to be communicated. Liking, sharing or comment features might be redesigned, so as to avoid controversy or polarisation becoming the implicit norm for popularity on a social media platform. Meanwhile, more respective and tolerant types of behaviours could be modelled and rewarded (Citroen & Norton, 2011).

By and large, policy makers and industry together have an essential role to put adequate regulatory frameworks and measures in place to ensure that individuals' fundamental right to human dignity is respected. This is particularly true for the most harmful end of the spectrum, with certain extreme types of content and forms of expression not to be allowed or accepted – or if this is not possible, subsequently removed and punished (Ash, 2016). Yet, even in their most effective form, top-down initiatives to regulate, monitor or report online hate speech only scrape the surface of the broader culture of online hate we have described in our review of the literature. This moves us well beyond narrow legal rules or procedures. Rather, it touches upon the nature of online hate, its causes and consequences.

In our view, this calls for a more pro-active awareness and education effort driving the multiple stakeholder approach full circle. Because indeed, to affect change, a more systemic shift is needed: one which informs and prepares children and young people for online hate by talking about it, in dialogue with their teachers, parents or other professionals or carers. Drawing upon their everyday online media experience, young people should be equipped to think critically about what hate speech is, why it occurs, which consequences is has, and what can be done about it. Alongside the critical use of reason, this calls for an approach which helps the learner to understand how he, she, we or they form(s) their own identity, opinions, attitudes and behaviours, as individuals and in group, in interplay with broader societal dynamics. Learners should be asked to step inside other people's skin, learning to empathise with the perspective of others. They should be primed for a culture of mutual respect and open debate, based on a contextual understanding of what is at stake.

8. The SELMA concept model and education strategy

The key education principles laid out above, translate into the SELMA Toolkit concept model visualised in Figure 9. The models starts from the different ecosystems in which children and young people find themselves.

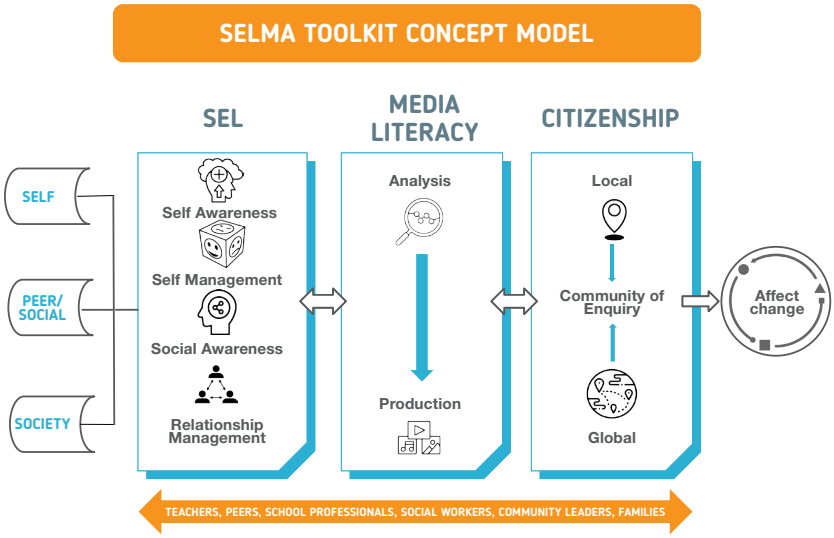

***Figure 9**. The SELMA Toolkit concept model*

The initial theme of Self focuses on how online hate speech is perceived by the individual and how it directly affects them in terms of emotional impact. It offers opportunities to explore strategies an individual can adopt to manage the personal impact of online hate speech on their wellbeing and respond in a way that benefits them. The Peer/Social theme focuses on how online hate speech is perceived by not only immediate physical peers but also within the online communities to which an individual belongs, for instance social media communities, online forums, gaming groups, and so on. It offers opportunities for young people to appraise and analyse online hate speech behaviour and develop considered strategies to respond in a way that affects positive change. The final Society theme focuses on the ethical and legal implications of hate speech; how society and industry providers determine boundaries and guidelines, and how it pertains to culture and

community. It offers opportunities to explore how to affect positive change at a wider global level through effective campaigns and calls to action.

In the SELMA Toolkit approach, each of the three master themes is broken down further into a set of subthemes. As such, the Toolkit is able to address a wide range of thematic questions, such as:

- What is hate speech? Why is there online hate speech out there? How does hate speech make me feel? What's my role and what can I do?
- Are my people really using hate speech? How can I influence my people? How can we effect change in our community?
- How can online stakeholders work together? What is needed to change the world?

Meanwhile, each subtheme or question can be approached from a number of different methodological focus angles, in particular social and emotional learning (SEL), media literacy (ML) and citizenship education (CE).

8.1. Social and emotional learning (SEL)

Emotional Intelligence is attributed to the work of Daniel Goleman, an American psychologist and science journalist, whose 1995 publication Emotional Intelligence has paved the way for educators to embrace emotional intelligence or social and emotional learning (SEL) in their schools and classrooms as an essential skill for living.

This line of thought directly taps into some of the key elements we have identified earlier on in this research report, particularly the role being played by identity formation and group dynamics. Therefore, resources within each theme of the SELMA Toolkit are led by a set of activities that develop social and emotional awareness, and management of online hate speech. This encourages the development of social and emotional skills and knowledge.

The SELMA SEL model is an adaptation of the model adopted by the Collaborative for Academic, Social, and Emotional Learning (CASEL)[24] and focuses on the following four areas:

24 www.casel.org.

SELF AWARENESS

The ability to accurately recognise one's own emotions, thoughts, and values and how they influence behaviour. The ability to accurately assess one's strengths and limitations, with a well-grounded sense of confidence, optimism, and a "growth mindset."

Identifying emotions
Accurate self-perception
Recognising strengths
Self-confidence
Self-efficacy

SELF MANAGEMENT

The ability to successfully regulate one's emotions, thoughts, and behaviours in different situations — effectively managing stress, controlling impulses, and motivating oneself. The ability to set and work toward personal and academic goals.

Impulse control
Stress management
Self-discipline
Self-motivation
Goal-setting
Organisational skills

SOCIAL AWARENESS

The ability to take the perspective of and empathise with others, including those from diverse backgrounds and cultures. The ability to understand social and ethical norms for behaviour and to recognise family, school, and community resources and supports.

Perspective-taking
Empathy
Appreciating diversity
Respect for others

RELATIONSHIP MANAGEMENT

The ability to establish and maintain healthy and rewarding relationships with diverse individuals and groups. The ability to communicate clearly, listen well, cooperate with others, resist inappropriate social pressure, negotiate conflict constructively, and seek and offer help when needed.

Communication
Social engagement
Relationship-building
Teamwork

***Figure 10**. SEL components (from CASEL, 2018)*

8.2. Media literacy

Media literacy has long been framed as a viable education strategy to address societal concerns (Martens, 2010). Defined as the ability to access, analyse, evaluate and create messages across a variety of contexts, it provides a set of perspectives from which we can expose ourselves to the media and interpret the meaning of the messages we encounter. It allows educators to start from pupils' existing understanding of the media. It uses a set of key media concepts – production, language, representation, and audience – which can be applied to the whole range of contemporary mass media. It enables children and young people to think in a more conscious and deliberate way, to understand and to analyse their own experience as content online users and creators (Buckingham, 2003; Potter, 2004).

This framework easily lends itself to the dimension of online hate we have identified in our literature review. It helps to understand how the move from offline to online media has changed the nature of hate speech and how it is spread. It provides insight into how social media contribute to the stratification of hate. It helps to look at online hate from the different perspectives of the creator, the distributor, and the audiences receiving and responding to it.

In the SELMA concept model, media literacy has been split into two discrete components: media analysis and media production.

a) Media analysis

Media analysis offers opportunities to explore how each aspect of online hate speech presents itself across the full estate of online media, and to develop critical analysis skills to understand context, drivers, cultural perspective, and so on.

The activities supporting this strand use a wide range of prepared media examples to stimulate structured discussion and provide educators with additional support with the following sections:

- Why each issue is important.
- How research indicates the issue.
- Open and closed questions to stimulate discussion.
- Activity guidance.

- Activity resource.
- Additional supporting resources curated online.
- Further professional reading.

b) Media production

Media production offers opportunities to use a wide range of media to create ways to navigate issues, raise awareness, promote counter-narratives, disrupt negative behaviours and amplify positive messages.

The activities supporting this strand explore the potential of using a range of media (in particular online technology itself) to communicate the above. Given that student's technical expertise and preferred route for communication varies depending on their own innate skill sets, each activity offers a number of multi-modal media routes. *This multi-modal approach encourages a range of benefits:*

- Opportunities for reinforcing behavioural strategies across the broader curriculum.
- Broadens ownership of SELMA across a wider staff and student body.
- Provides a variety of routes for students to express their emotional interpretations.
- Opportunities for staff to capitalise on their own individual strengths and expertise.
- Encourages a wide range of publishing platforms for students to communicate or celebrate their messages or achievement.
- Reinforces emotional interpretation through a variety of media.

Again, each media production section of the SELMA Toolkit provides educators with additional support with the following sections:

- Open and closed questions to stimulate discussion.
- Activity guidance.
- Activity resource.
- Additional supporting resources curated online.
- Further professional reading.

8.3. Citizenship in a digital world

The ultimate objective of the SELMA Toolkit is to enable teenagers to make constructive and ethical choices about personal behaviour and social interactions.

Throughout the SELMA Toolkit, learners will be encouraged to reflect and act in response to concrete online hate speech situations, which can be local, national, regional and/or global in nature. In our view, individuals need to be enabled to put online hate speech into context, starting from an awareness and critical analysis of the increasingly complex range of diverging (and often conflicting) views and perspectives in a digital society, while exploring pathways of change towards mutual tolerance and respect.

a) Global citizenship education

To this end, SELMA has adopted and adapted the key principles of UNESCO's 2015 outline document Global Citizenship Education, with key learning outcomes summarised in Figure 11.

COGNITIVE
Learners acquire knowledge and understanding of local, national and global issues and the interconnectedness and interdependence of different countries and populations.
Learners develop skills for critical thinking and analysis.
SOCIO-EMOTIONAL
Learners experience a sense of belonging to a common humanity, sharing values and responsibilities, based on human rights.
Learners develop attitudes of empathy, solidarity and respect for differences and diversity.
BEHAVIOURAL
Learners act effectively and responsibly at local, national and global levels for a more peaceful and sustainable world.
Learners develop motivation and willingness to take necessary actions.

***Figure 11**. Global citizenship education: key learning outcomes (UNESCO, 2015)*

These outcomes can further be mapped against a set of topics, across three types of learning attributes, as indicated in Figure 12. The topics provide concrete

entry points to the traits and qualities global citizenship education aims to develop. The SELMA themes have been developed against these learning attributes and topics.

INFORMED AND CRITICALLY LITERATE
Local, national and global systems and structures.
Issues affecting interaction and connectedness of communities at local, national and global levels.
Underlying assumptions and power dynamics.
SOCIALLY CONNECTED AND RESPECTFUL OF DIVERSITY
Different levels of identity.
Different communities people belong to and how these are connected.
Difference and respect for diversity.
ETHICALLY RESPONSIBLE AND ENGAGED
Actions that can be taken individually and collectively.
Ethically responsible behaviour.
Getting engaged and taking action.

***Figure 12**. Global citizenship education: from learning attributes to key topics (UNESCO, 2015)*

b) Call to action

Embedded within the citizenship element of the SELMA model is the ability to develop a "Call to action" to ultimately affect change. All activities prepare and guide SELMA Toolkit users to this point. It begs the question "So what?". It is not enough to study and understand the issue; the hacking philosophy requires us to do something about it.

Many of the activities within SELMA promote the concept of a community of inquiry within the groups using the SELMA resources. The outcomes are shaped by the collective findings of the group by creating environments for discussion, exercises to explore reasoning, opportunities for dialogue and collective decision making.

The community of inquiry is a concept first introduced by early pragmatist philosophers Charles Sander Peirce and John Dewey, concerning the nature of knowledge formation and the process of scientific inquiry. The community of inquiry is broadly defined as any group of individuals involved in a process of empirical or conceptual inquiry into problematic situations.

The community of inquiry concept emphasises that knowledge is necessarily embedded within a social context and, thus, requires intersubjective agreement among those involved in the process of inquiry for legitimacy. Applying the concept to the educational setting, Lipman (2003) argues that the classroom is a type of community of inquiry, which should lead to questioning, reasoning, connecting, deliberating, challenging, and developing problem-solving techniques.

More specifically, Lipman (2003, pp. 18-19) frames educational practice as follows:

- "Education is the outcome of participation in a teacher-guided community of inquiry, among whose goals are the achieving of understanding and good judgement.
- Students are stirred to think about the world when our knowledge of it is revealed to be ambiguous, equivocal, and mysterious.
- The teacher stance is fallibilistic (one that is ready to concede error) rather than authoritative.
- Students are thought to be reflective and increasingly reasonable and judicious.
- The focus of the education process is not on the acquisition of information but on the grasp of relationships within and among the subject matter under investigation."

9. Key lessons learned and how to engage with the SELMA journey

In line with the key findings from our literature review and empirical research, the SELMA Toolkit response to online hate speech subscribes to a number of key principles.

SELMA holds the view that children and young people can (and should) be empowered to become agents of change in their offline and online communities. Education responses to online hate speech should not preach what is "good" or "bad". Rather, they should encourage and enable children and young people to critically and creatively engage with the problem of online hate speech and its possible solutions:

- What is online hate speech?
- How does it affect my personal and social environment?
- Which role can I play – together with my peers – in addressing online hate speech and changing society for the better?

Pathways of change can only be successful when they follow a holistic plan of action. Media literacy knowledge and skills should be combined with social and emotional learning. Online hate should be considered as a pattern of behaviour, interconnected with the social and cultural contexts in which it takes place. Children and young people should be involved in a wider societal debate on how to replace the culture of online hate with tolerance and mutual respect. This should happen in dialogue with teachers and parents, professionals and carers, and a wider range of education, industry and civil society stakeholders.

The SELMA research provides a theoretical and empirical backbone for the SELMA Toolkit, which will be made available on www.hackinghate.eu in the spring of 2019. It will provide a large collection of principles, methods and activities that will enable multiple stakeholders to work on online hate speech with teenagers aged 11-16.

REFERENCES

Arbeitsgemeinschaft Kinder- und JugendschutzLandesstelle NRW e.V. (AJS). (2016). *Hate Speech/Rechtsfragen. Retrieved from* https://www.ajs.nrw.de/wp-content/uploads/2016/06/AJS-Merkblatt_Hate-Speech_Rechtsfragen.pdf.

Ash, T.G. (2016). *Free speech: Ten principles for a connected world. New Haven, CT: Yale University Press.*

Assimakopoulos, S., Baider F.B., & Millar S. (2017). *Online hate Speech in the European Union: A discourse analytic survey. Cham, Switzerland: Springer Open.*

Anti-Defamation League. (2010). *Responding to cyberhate: Toolkit for action. Retrieved from* https://www.adl.org/sites/default/files/documents/assets/pdf/combating-hate/ADL-Responding-to-Cyberhate-Toolkit.pdf.

Anti-Defamation League. (2018). *The pyramid of hate. Retrieved from* https://www.adl.org/sites/default/files/documents/pyramid-of-hate.pdf.

Bakalis, C. (2018). Rethinking cyberhate laws. *Information & Communications Technology Law, 27(1), 86-110.*

Bakuros, V. (2015). Hate speech. A diagnostic test. Grigoris publication.

Barendt, E. (2009) Freedom of expression in the United Kingdom under the Human Rights Act 1998. Indiana Law Journal, 84(3), Article 4.

Baynes, C. (2017). 'Troubling' fall in hate crime prosecutions despite spike in reports after Brexit. Evening Standard, 17 October. Retrieved from https://www.standard.co.uk/news/crime/troubling-fall-in-hate-crime-prosecutions-despite-spike-in-reports-after-brexit-a3660296.html.

BBC News. (2016). Hate crime prosecutions fall despite rise in reporting. *BBC News, 4 September. Retrieved from:* https://www.bbc.co.uk/news/uk-37266636.

BBC News. (2018a). Hate crime 'police priority' as social media cases soar. *BBC News, 17 March. Retrieved from:* https://www.bbc.co.uk/news/uk-scotland-glasgow-west-43436900.

BBC News. (2018b). Misogyny could become hate crime as legal review is announced. *BBC News, 6 September. Retrieved from* https://www.bbc.co.uk/news/uk-politics-45423789.

Benesch, S. (2013). *Dangerous speech: A proposal to prevent group violence. The Dangerous Speech Project. Retrieved from* http://dangerousspeech.org/guidelines.

Berecz, T. & Devinat, Ch. (2016a). *Quarterly Report on Cyber Hate (May-July).* INACH: Project Research - Report - Remove: Countering Cyber Hate Phenomena.

Berecz, T. & Devinat, Ch. (2016b). *Quarterly Report on Cyber Hate (August-October).* INACH: Project Research - Report - Remove: Countering Cyber Hate Phenomena.

Berecz, T. & Devinat, Ch. (2016c). *Quarterly Report on Cyber Hate (November-December).* INACH: Project Research - Report - Remove: Countering Cyber Hate Phenomena.

Berecz, T. & Devinat, Ch. (2017a). *Quarterly Report on Cyber Hate (January-March).* INACH: Project Research - Report - Remove: Countering Cyber Hate Phenomena.

Berecz, T. & Devinat, Ch. (2017b). *Quarterly Report on Cyber Hate (April-June).* INACH: Project Research - Report - Remove: Countering Cyber Hate Phenomena.

Berecz, T. & Devinat, Ch. (2017c). *Quarterly Report on Cyber Hate (July-September).* INACH: Project Research - Report - Remove: Countering Cyber Hate Phenomena.

Bowman-Grieve, L. (2009). Exploring "Stormfront": A virtual community of the Radical Right. *Studies in Conflict & Terrorism, 32(11), 989-1007.*

Bridger, S., Bachmann, C.L., & Gooch, B. (2017). LGBT in Scotland: Hate crime and discrimination. Stonewall Scotland.

Brown, A. (2017). What is Hate Speech? Part 2: Family Resemblances. *Law and Philosophy, 36(5), 561-613.*

Buckingham, D. (2003). *Media education. Literacy, learning and contemporary culture.* Cambridge: Polity Press.

Burris V., Smith E., & Strahm A., (2000). White supremacist networks on the internet. *Journal of Sociological Focus*, *33*(2), 215-235.

Cammaerts, B. (2009). Radical pluralism and free speech in online public spaces: The case of North Belgian extreme right discourses. *International journal of cultural studies, 12 (6), 555-575.*

Citron, D.K., & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review, 91*, 1435-1483.

Conway, M., & McInerney, L. (2008). Jihadi video and auto-radicalisation: Evidence from an exploratory YouTube study. In D. Ortiz-Arroyo, H.L. Larsen, D. Zeng, D.L. Hicks & G. Wagner (Eds.), *Intelligence and security informatics (pp. 108–118). Berlin, Germany: Springer-Verlag.*

Council of Europe. (1950). European Convention for the Protection of Human Rights and Fundamental Freedoms. Rome. *Retrieved from* https://www.echr.coe.int/Documents/Convention_ENG.pdf.

Council of Europe. (1997). *Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "hate speech", 30 October 1997. Retrieved from* https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680505d5b.

Council of Europe. (2003). *Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems, 28 January 2003. Retrieved from* https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168008160f.

Council of Europe: European Commission Against Racism and Intolerance (ECRI). (2016). ECRI General Policy Recommendation No. 15 on combating Hate Speech, 8 December 2015. Retrieved from https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01.

Council of the European Union. (2008). *Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law*, OJ L 328, 6 December 2008, 55-58. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=en.

Cowan L. & Hodge C. (1996). Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology, 26*(4), 355-374.

Cowan, G., Resendez, M., Marshall, E., & Quist, R. (2002). Hate speech and constitutional protection: Priming values of equality and freedom. *Journal of Social Issues, 58(2), 247-263.*

CPS. (n.d). Hate Crime. Retrieved from: https://www.cps.gov.uk/hate-crime.

CPS. (2017). *Hate Crime Annual Report 2016-17. Retrieved from* https://www.cps.gov.uk/sites/default/files/documents/publications/cps-hate-crime-report-2017_0.pdf.

CPS. (2010). *Sexual Orientation: CPS Guidance on stirring up hatred on the grounds of sexual orientation. Retrieved from* https://www.cps.gov.uk/legal-guidance/sexual-orientation-cps-guidance-stirring-hatred-grounds-sexual-orientation.

CPS. (2018). *Social Media: Guidelines on prosecuting cases involving communications sent via social media. Retrieved from* https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media.

Criminal Justice Inspection Northern Ireland. (2017). *Hate Crime: An Inspection of the Criminal Justice System's Response to Hate Crime in Northern Ireland. Retrieved from:* http://www.cjini.org/getattachment/a48b8a89-f32f-4b02-bd3c-8f77989630eb/picture.aspx.

Crown Office and Procurator Fiscal Service (2018). *Hate Crime Statistics Report 2017-18. Retrieved from* http://www.copfs.gov.uk/publications/equality-and-diversity#Hate_Crime_in_Scotland.

Daniels, J. (2008). Race, civil rights, and hate speech in the digital era. In A. Everett (Ed.), *Learning race and ethnicity: Youth and digital media* (pp. 129-154). The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, Massachusetts: The MIT Press.

Dearden, L. (2017). Hate-crime prosecutions fall despite spike in reported attacks after Brexit referendum. *The Independent, 17 October. Retrieved from* https://www.independent.co.uk/news/uk/crime/brexit-hate-crime-prosecutions-fall-attack-report-spike-cps-report-statistics-islamophobia-racist-a8003716.html.

Dearden, L. (2018). People convicted of hate crimes could get harsher sentences if they have large number of social media followers. *The Independent, 9 May. Retrieved from* https://www.independent.co.uk/news/uk/home-news/hate-crime-sentences-increased-harsher-influence-social-media-followers-a8341991.html.

Department for Digital, Culture, Media and Sport. (2018). *Policy Paper: Digital Charter. Retrieved from* https://www.gov.uk/government/publications/digital-charter/digital-charter.

Doebler, S., McAreavey, R., Shortall, S., & Shuttleworth, I. (2016*). Negativity toward immigrant out-groups among Northern Ireland's Youth – are younger cohorts becoming more tolerant? Northern Ireland Assembly. Knowledge Exchange Seminar Series 2016-17. Retrieved from* http://www.niassembly.gov.uk/globalassets/documents/raise/knowledge_exchange/briefing_papers/series6/doebler141216.pdf.

Duffy M. (2003). Web of hate: A fantasy theme analysis of the rhetorical vision of hate groups online. *Journal of Communication Inquiry, (27)*3, 291-312.

Eckstrand, N. (2018). The Ugliness of Trolls: comparing the strategies/methods of the Alt-Right and the Ku Klux Klan. *Cosmopolitan Civil Societies: an Interdisciplinary Journal, 10*(3), 41-62.

European Commission. (2016). *Code of Conduct on countering illegal hate speech online: First results on implementation*, December 2016. Retrieved from http://ec.europa.eu/information_society/newsroom/image/document/2016-50/factsheet-code-conduct-8_40573.pdf

European Commission. (2018). *Commission Recommendation on measures to effectively tackle illegal content online, 1 March 2018. Retrieved from* https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online.

Federal Agency of Civic Education. (2018). *Was ist Hate speech? [What is hate speech?]. Retrieved from* http://www.bpb.de/252396/was-ist-hate-speech.

Federal Ministry of Justice and Consumer Protection. (2017). *GesetzzurVerbesserung der Rechtsdurchsetzung in sozialenNetzwerken (Netzwerkdurchsetzungsgesetz – NetzDG). [Law on improving law enforcement in social networks (Network Enforcement Act)]. Retrieved from* https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/DE/NetzDG.html.

Finley, M.I. (1983). *Politics in the Ancient World. Cambridge, UK: Cambridge University Press.*

Foxman, A.H. & Wolf, C. (2013). Viral hate: Containing its spread on the internet. New York, N.Y.: St. Martin's Press.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech.* UNESCO Series on Internet Freedom. Paris, France: United Nations Educational, Scientific and Cultural Organisation.

Gerstenfeld P., Grant D., & Chiang C-P. (2003). Hate online: A content analysis of extremist internet sites, *Analyses of Social Issues and Public Policy*, 3(1), 29-44.

Goleman, D. (1995). *Emotional intelligence. New York, NY: Bantam Books, Inc.*

Hall, E. (2017). Why disability hate crimes are woefully under-reported. The Conversation, 25 October. Retrieved from https://theconversation.com/why-disability-hate-crimes-are-woefully-under-reported-85964.

Hall E. B. (1906). *The Friends of Voltaire*. London, UK: Smith, Elder & Co.

Hawdon J., Oksanen A., & Räsänen P. (2015). Online extremism and online hate: Exposure among adolescents and young adults in four nations. *Nordicom-Information, 37(3-4), 29-37.*

HM Government. (2018). Government response to the Internet Safety Strategy Green Paper. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/708873/Government_Response_to_the_Internet_Safety_Strategy_Green_Paper_-_Final.pdf.

Home Office. (2016). *Action Against Hate: The UK Government's plan for tackling hate crime. Retrieved from* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/543679/Action_Against_Hate_-_UK_Government_s_Plan_to_Tackle_Hate_Crime_2016.pdf.

Home Office. (2017). *Home Secretary announces new national online hate crime hub. Press release retrieved from* https://www.gov.uk/government/news/home-secretary-announces-new-national-online-hate-crime-hub.

Home Office. (2018). *Hate Crime, England and Wales, 2017/18. Statistical Bulletin 20/18. Retrieved from* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748598/hate-crime-1718-hosb2018.pdf.

Isaac, D. (2016). *Hierarchy of hate crime is undermining confidence in the law. Retrieved from* https://www.equalityhumanrights.com/en/our-work/news/hierarchy-hate-crime-undermining-confidence-law.

Jubany, O., & Roiha, M. (2015). *Backgrounds, experiences and responses to online hate speech: A comparative cross-country analysis. The PRISM Project.*

Kanter, J. (2018). Britain is coming for Silicon Valley's unruly tech giants, and it could change the way they do business forever. Business Insider, 4 October. Retrieved from https://www.businessinsider.sg/britain-will-regulate-silicon-valley-according-to-damian-collins-2018-9/.

Keen, E., & Georgescu, M. (2016). *Bookmarks - A manual for combating hate speech online through human rights education. Revised edition. Strasbourg, France: Council of Europe Publishing.*

Keipi, T. (2015). Now you see me, now you don't: A study of the relationship between Internet anonymity and Finnish young people. Doctoral dissertation: University of Turku. Retrieved from: https://www.utupub.fi/bitstream/handle/10024/113050/AnnalesB405KeipiDISS.pdf?sequence=2&isAllowed=y.

Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2017). *Online hate and harmful content: Cross-national perspectives. London, UK: Routledge.*

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist, 39(10), 1123-1134.*

LandesanstaltfürMedien NRW. (2017). *VerfolgenstattnurLöschen – Rechtsdurchsetzungim Internet. [Prosecuting instead of deleting. Law enforcement on the Internet]. Retrieved from* https://www.medienanstalt-nrw.de/service/positionen/verfolgen-statt-nur-loeschen-rechtsdurchsetzung-im-internet.html.

LandesanstaltfürMedien NRW. (2018a). *Forsa-BefragungzurWahrnehmung von Hassredeim Internet. Retrieved from:* https://www.medienanstalt-nrw.de/foerderung/forschung/abgeschlossene-projekte/forsa-befragung-zur-wahrnehmung-von-hassrede.html

LandesanstaltfürMedien NRW. (2018b). *Mehrals 130 Anzeigenbinnen 70 Tagen.* Retrieved from https://www.medienanstalt-nrw.de/service/pressemitteilungen/pressemitteilungen-2018/2018/april/verfolgen-statt-nur-loeschen-zieht-erste-bilanz.html.

Law Commission. (2014). *Hate Crime: Should the Current Offences be Extended? LawCom No 348, Chapter 7: Extending the stirring up offences. Retrieved from* https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2015/03/lc348_hate_crime.pdf.

Law Commission. (2018). *Offensive online communications. Retrieved from* https://www.lawcom.gov.uk/project/offensive-online-communications/.

Leets L. (2002). Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of Social Issues, 58*(2), 341-361.

Leets, L., & Giles, H. (1999). Harmful speech in intergroup encounters: An organizational framework for communication research. In M. Roloff (Ed.), *Communication Yearbook (Vol. 22, pp. 91–37). Thousand Oaks, CA: Sage Publications.*

Levin, B *(2002). Cyberhate: A legal and historical analysis of extremists' use of computer networks in America. American Behavioral Scientist, 45(6), 958-988.*

Livingstone, S., & Brake, D.R. (2010). On the rapid rise of social networking sites: new findings and policy implications. Children & Society, 24(1), 75-83.

Livingstone, S, & Haddon, L. (2009). EU Kids Online: Final report. LSE, London: EU Kids Online.

Lipman, M. (2003). Thinking in education. Cambridge: Cambridge University press.

Lord Bracadale. (2018). Independent Review of Hate Crime Legislation in Scotland: Final Report. Scottish Government, Justice Directorate. Retrieved from http://www.gov.scot/Publications/2018/05/2988/downloads.

Martens, H. (2010). Evaluating media literacy education: Concepts, theories and future directions. *Journal of Media Literacy Education*, *2*(1), 1-22.

Marwick, A. E., & Miller, R.W. (2014). *Online harassment, defamation, and hateful speech: A Primer of the legal landscape.* Fordham Center on Law and Information Policy Report.

Matsuda, M. J., Lawrence, C. R., Delgado, R., & Crenshaw, K. W. (1993). *Words that wound: Critical race theory, assaultive speech, and the first amendment. Boulder, Colorado: Westview Press.*

McDonald, H. (2017). Racially motivated crimes now exceed sectarian ones in Northern Ireland. The Guardian, 12 November. Retrieved from https://www.theguardian.com/uk-news/2017/nov/12/racially-motivated-crimes-now-exceed-sectarian-ones-in-northern-ireland.

McGonagle, T. (2013). *The Council of Europe against online hate speech: Conundrums and challenges. (MCM; No. 2013(005)). Belgrade: Republic of Serbia, Ministry of Culture and Information.*

McVeigh, R. (2018). Incitement to hatred in Northern Ireland. Equality Coalition. Retrieved from https://caj.org.uk/2018/04/27/incitement-to-hatred-in-northern-ireland-research-report-by-dr-robbie-mcveigh-for-the-equality-coalition.

McVeigh, R. and Rolston, B. (2007). From Good Friday to Good Relations: sectarianism, racism and the Northern Ireland state. *Race & Class, 48(4), 1-23.*

Montague, R. and Shirlow, P. (2014). Challenging Racism: Ending Hate. Unite Against Hate. Retrieved from http://democracyandpeace.org/wp-content/uploads/2015/12/Challenging_Racism_Ending_Hate.pdf.

Morley, N. (2017). Hate crime prosecutions are down despite huge rise in reports. *Metro, 17 October. Retrieved from* https://metro.co.uk/2017/10/17/hate-crime-prosecutions-are-down-despite-huge-rise-in-reports-7005263/.

No-Hate-Speech-Movement. (NHM) Germany. (2018). *Was ist eigentlich Hate Speech? [What is hate speech actually?]. Retrieved from* https://no-hate-speech.de/de/wissen/.

Northern Ireland Human Rights Commission. (2013). *Racist Hate Crime: Human Rights and the Criminal Justice System in Northern Ireland. Retrieved from* http://fra.europa.eu/sites/default/files/frc-2013-g-sauberli-investigation_report_full_en.pdf

Oksanen, A., Hawdon, J., Holkeri, E., Näsi, M., & Räsänen, P. (2014). Exposure to online hate among young social media users. *Sociological Studies of Children & Youth, 18(1), 253-273.*

Oksanen, A., Kaakinen, M., Minkkinen, J., Räsänen, P., Enjolras, B., & Steen-Johnsen, K. (2018). Perceived societal fear and cyberhate after the November 2015 Paris terrorist attacks. *Terrorism and Political Violence, 1-20.*

Olterman, P. (2018). Tough new German law puts tech firms and free speech in spotlight. The Guardian, 5 January. Retrieved from https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight.

O'Neill, B., Staksrud, E. & McLaughlin, S. (2013). *Towards a Better Internet for Children? Policy pillars, players and paradoxes. Göteborg, Sweden: Nordic*om.

Perry B. & Olsson P. (2009). Cyberhate: the globalization of hate. *Information & Communications Technology Law*, *18*(2), 185-199.

Potok, M. (2016). *The year in hate and extremism. Intelligence report.* Southern Poverty Law Center, US. Retrieved from https://www.splcenter.org/fighting-hate/intelligence-report/2016/year-hate-and-extremism.

Police Service of Northern Ireland. (2018). *Incidents and Crime with a Hate Motivation Recorded by the Police in Northern Ireland: Quarterly Update to 31 March 2018. Available at:* https://www.psni.police.uk/globalassets/inside-the-psni/our-statistics/hate-motivation-statistics/2017-18/quarterly-hate-motivations-bulletin-period-ending-mar18.pdf.

Potter, W. J. (2004). *Theory of media literacy: A cognitive approach.* Thoasand Oaks, CA: Sage.

Rogers, R. (2013). *Right-wing formations in Europe and their counter-measures: An online mapping. Amsterdam, Netherlands: Digital Methods Initiative. Retrieved from* http://govcom.org/populism/GCO_DMI_Populism_final_6May2013.pdf.

Sedler, R. (1992). The unconstitutionality of campus bans on "racist speech": The view from without and within. *University of Pittsburgh Law Review, 53, 631-683.*

Sentencing Council. (2018). Public Order Offences Consultation 2018. Annex C: Draft Guidelines. Retrieved from https://www.sentencingcouncil.org.uk/wp-content/uploads/Annex-C-Public-Order-offences-guidelines_Consultation-web.pdf.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016* (pp. 687-690). AAAI Press.

Smolla, R. A. (1990). Academic freedom, hate speech, and the idea of a university. *Law and Contemporary Problems, 53*(3), 195-205.

Strafgesetzbuch (StGB). (1998). Strafgesetzbuch der Bundesrepublik Deutschland in der Fassungvom 13.11.1998. Retrieved from https://www.gesetze-im-internet.de/stgb/.

Stray, M. (2017). *Online Hate Crime Report 2017. Galop. Retrieved from* http://www.galop.org.uk/wp-content/uploads/2017/08/Online-hate-report.pdf.

Thiesmeyer, L. (1999). Racism on the Web: Its rhetoric and marketing. *Ethics and Information Technology, 1*(2); 117-125.

Timmermann, W. (2008). Counteracting hate speech as a way of preventing genocidal violence. *Genocide Studies and Prevention: An International Journal, 3(3), 353-374.*

Timofeeva Y. (2003). Hate speech online: Restricted or protected - Comparison of regulations in the United States and Germany. Transnat'l L. & Pol'y, 12(2), 253-286.

Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the Rwandan genocide. *The Quarterly Journal of Economics*, 129(4), 1947-1994.

UNESCO. (2015). Global citizenship education: topics and learning objectives. Paris, France: United Nations Educational, Scientific and Cultural Organization.

UN General Assembly. (1948). *Universal declaration of human rights (217 [III] A). Paris. Retrieved from* http://www.un.org/en/universal-declaration-human-rights.

UN Human Rights Council. (UNHRC). (2015). *Report of the Special Rapporteur on Minority Issues, Izsák, R.,* A/HRC/28/64.

Walters, M.A., Brown, R., & Wiedlitzka, S. (2016). *Causes and motivations of hate crime. Equality and Human Rights Commission. Research report 102. Retrieved from* https://www.equalityhumanrights.com/en/publication-download/research-report-102-causes-and-motivations-hate-crime.

Walters, M.A., Wiedlitzka, S., Owusu-Bempah, A., & Goodall, K. (2017). *Hate Crime and the Legal Process: Options for Law Reform – Final Report. Retrieved from* https://www.sussex.ac.uk/webteam/gateway/file.php?name=final-report---hate-crime-and-the-legal-process.pdf&site=539.

White, M. H. II, & Crandall, C. S. (2017). Freedom of racist speech: Ego and expressive threats. *Journal of Personality and Social Psychology, 113*(3), 413-429.

Williams, M. & Burnap, P. (2016). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and Big Data. *The British Journal of Criminology*, *56*(2), 211-238.

Zuleta, L. & Rasmus, B. (2017). *Hate speech in the public online debate. Copenhagen, Denmark: The Danish Institute for Human Rights.*

SELMA
HACKING HATE

www.europeanschoolnet.org

www.youth-life.gr

www.swgfl.org.uk

www.diana-award.org.uk

www.lmk-online.de

www.cfdp.dk

CONTACT US
www.hackinghate.eu | info@hackinghate.eu

FOLLOW US
#SELMA_eu | @hackinghate

This project is funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020). The contents of this communication are the sole responsibility of the author and can in no way be taken to reflect the views of the European Commission.