Hate speech in social media: a state-of-the-art review

11/12/2016

Alex Cabo Isasi and Ana García Juanatey

Table of Contents

INTRODUCTION

There has been a great deal of theorising on the deliberative and democratising potential of the Internet and, more specifically, social media. Their participatory potential is undeniable and this has generated high expectations regarding the possibilities they offer for political and social change. Social media provide broad access to information sources not controlled by governments or major corporations, thus facilitating the creation and coordination of activist networks, and they constitute an unbeatable space for deliberating and exchanging ideas. In the words of the information and communication technology expert Manuel Castells, social media are a platform for "mass self-communication", a space of communicative autonomy that is ideal for social players to foster the transition from indignation to hope. In this regard, the so-called Arab Spring, 15-M and Occupy Wall Street are examples of how social media can help to transform indignation into action and horizontal, transformative, liberating movements (Castells 2012).

However, we should avoid being overly optimistic, because it is clear that the Internet and social media can also be used for ends that are radically different to those outlined. As the cyber sceptic Evgeny Morrozov maintains, the Internet does not "*necessarily lead to universal respect for human rights.*" It is neither liberating nor democratising in itself, rather it can produce "*different political outcomes in different contexts*", so we should not be carried away by a certain "technological solutionism" or "cyber utopianism" (Morozov 2012).

Just as some liberation movements have made promising use of social media, forces of the opposite kind have demonstrated their capacity for making the most of the potential these media have for opposite ends. Thanks to Edward Snowden, it was shown how the United States uses new technologies for indiscriminate surveillance and spying. Another example, but from the dark side of the Net, is the so-called "50 Cent Army" of commentators hired by the Chinese authorities to spread propaganda messages favourable to the government in social media debates.

But perhaps the most tragic example of how the Internet and social media can be used is the crisis unleashed in Kenya following the elections in December 2007. The dissemination of messages inciting violence, via a variety of online forums, is documented as "*a decisive channel through which violence was fuelled, taking the lives of over a thousand people and displacing more than 600,000*" (Gagliardone et al. 2014: 5).

1. A DIAGNOSIS OF THE PROBLEM

These kinds of messages that incite violence have found the right channel for their dissemination on the Internet and social media. Although the real scope of the problem has not been determined quantitatively, there is a general feeling shared by journalists, jurists, NGOs, researchers and users that the problem of extreme speech in social media is becoming one of increasing concern.

A great proliferation of extremist messages is taking place all over Europe, linked to the context of the refugee crisis, with worrying "spikes" of Islamophobic hatred detected in the wake of the terrorist attacks in Paris, Brussels and Nice. For a few hours after the Paris attacks, *#matadatodoslosmusulmanes* ("kill all Muslims") became the third most used hashtag in Spain (Jubany y Roiha 2016).

According to studies carried out by the think tank Demos, Twitter has approximately 10,000 tweets a day with racist insults in English, which represents one out of every 15,000 tweets (Gagliardone et al. 2014). A different study carried out by Demos among British users of Twitter on another recurring motive for intolerance, misogyny, over three weeks in April 2016, found more than 200,000 tweets with the words "bitch" and "whore", showing that every ten seconds someone insults a woman with these words on Twitter.

The problem is no less serious in Spain. Below, some examples from 2016. Moha Gerehou, president of SOS Racismo Madrid and born in Gambia, has been the target of various attacks. In July 2016 he was "auctioned" on Twitter like a slave or hunted animal, in response to the campaign *#EstadoEspanolNoTanBlanco* ("Spanish State Not So White"). In September he received a death threat after posting a photo on Twitter of a rally against withdrawing the name of Millán Astray, founder of the Spanish Foreign Legion, from the Madrid street names[1].

In August 2016, Jordi Ballart, the Mayor of Terrassa, received homophobic threats on Twitter and Facebook, following his decision to remove the name from a street named after a member of the Blue Division[2].

In September, the transsexual Carla Antonelli, a member of the Madrid Assembly, received death threats via Twitter for defending the rights of the LGBT collective.[3] The same month, the youth football referee, Jesús Tomillero, president of the association Contra la LGTBFobia in sport, received insults and death threats on Twitter, for being the first self-declared gay official in Spain[4].

These are just some of the cases of individual harassment based on hatred and bigotry that have surfaced in the media and they represent the tip of the iceberg of a situation that has become the norm on social media. At a collective level, Romaphobia

---

[1] http://www.eldiario.es/sociedad/Nuevas-presidente-SOS-Racismo-Madrid_0_563094347.html

[2] http://www.elperiodico.com/es/noticias/sociedad/jordi-ballart-alcalde-terrassa-denuncia-amenazas-homofobas-redes-sociales-5473496

[3] http://ccaa.elpais.com/ccaa/2016/09/04/madrid/1472990074_481479.html

[4] http://www.elperiodico.com/es/noticias/deportes/arbitro-gay-jesus-tomillero-denuncia-amenazas-muerte-5376384

proliferates in the social media with impunity, while Islamophobia frequently becomes a trending topic through hashtags such as #stopIslam, #terroristaswelcome and #musulmanesterroristas. Anti-Semitism, homophobia, misogyny, etc., all intolerant ideologies have a unique space for expressing themselves on social media, which has generated a kind of *hate culture*[5] that pollutes and poisons those media with abusive, derogatory and aggressive language, largely inspired by intolerance of the immigrant population, refugees, Muslims, homosexuals and other minorities.

International and European institutions are not immune to this problem. Proof of their concern is the intense activity carried out by the European Union this year in an effort to find solutions, despite the fact that, at the same time, certain types of racism have been gaining more and more ground in European political discourse. The milestones of this dedication are the signing of the Code of Conduct with the tech companies, and the launch of a High Level Group, involving representatives of the EU member states, the European Parliament, the Council of Europe, the UN Commissioner for Refugees (UNHCR), the Organisation for Security and Cooperation in Europe (OSCE), the European Agency for Fundamental Rights (FRA), and civil society organisations such as Amnesty International, the European Network Against Racism (ENAR), the European Platform of Social NGOs and the International Lesbian, Gay, Bisexual, Trans and Intersex Association (ILGA), with the aim of preventing and combating hate crimes and hate speech, and with a special emphasis on countering online hate speech.

In recent reports, UNESCO and the European Commission Against Racism and Intolerance (ECRI) have also highlighted the growth of intolerant speech online and in social media directed towards various minorities, and the need to find effective strategies to combat it.

To sum up, the problem of online hate speech, in social media in particular, has become such a big issue that today it can be found on the agenda of many international and European bodies. Social media have become a place where people can express their anger and hatred with impunity. Sexism, homophobia, xenophobia, Islamophobia, Romaphobia, anti-Semitism and other intolerant ideologies take advantage of the Internet and social media to insult, humiliate, harass, threaten and carry out social lynching.

1.1. What is understood by hate speech? A concept subject to debate.

There is no universally accepted definition of hate speech. First of all, because the ground for definitions with ethical and legal implications is always controversial. And, secondly, because the very term “hate” makes it an emotional concept and open to subjective interpretation. It is a concept that creates confusion and, given its subjective nature, is relatively easy to manipulate.

The lowest common denominator for any definition of hate speech would be any expression of opinion or ideas based on contempt and animosity towards individuals or groups who are wished harm. However, this simple definition encompasses too broad a range of expressions for the concept to be of any use for social or legal analysis or intervention(Article 19 2015)

---

[5] http://time.com/4457110/internet-trolls/

We speak of hate speech to refer to expressions that directly incite the committing of acts of discrimination or violence motivated by racial hatred, xenophobia, sexual orientation and other types of intolerance. But some believe that the concept also extends to those expressions that foster prejudice and intolerance, as they think these types of expressions contribute indirectly towards creating a climate of hostility that can eventually lead to discriminatory acts and violent attacks (Gagliardone et al. 2015).

In everyday language, and especially in the media, the concept of hate speech is widely used to refer to a heterogeneous mass of manifestations ranging from threats to individuals and groups to cases where some people simply express their anger at the authorities, in a more or less offensive way (Gagliardone et al. 2015). Use of the term hate speech has also become widespread in the media and even in legal circles to refer to concepts such as inciting or advocating terrorism which, despite bearing some relationship to the subject of this report, require a specific focus.

Summing up, the term hate speech is open to discussion, and is the subject of political, legal and academic debate at an international level. It is a complex concept, as it threatens and confronts different values and cardinal principles of democratic systems: equality, human dignity, freedom of expression, and so on, which are not conceived in the same way in all socio-political contexts. Hate speech is therefore a concept that can be implemented politically with very different objectives, which may be more or less legitimate.

The definitions most widely accepted at an international level can be grouped together in two main trends: 1) those that define hate speech in a broader sense and include any expression that promotes or justifies hate for racist, xenophobic or religious motives, or reasons of gender, sexual orientation or disability[6]; and 2) those that define it in more restrictive and precise terms, including only those forms of expression which, in specific contexts of instability, can contribute towards unleashing violent episodes against a group of persons because they belong to one of the groups mentioned[7](Gagliardone et al. 2014).

The fundamental difference between the two could be summarised in the content-context dialectic, and the greater or lesser risk of the messages unleashing violent actions. The former defines hate speech exclusively in terms of its content, while the latter considers that what turns a particular expression into hate speech is the manifest risk that, given the historical and social context in which it is issued, it will provoke violent episodes. Context does indeed determine, to a large extent, the harm that certain expressions and statements can cause.This may range from emotional damage to provoking massive outbreaks of violence, such as the genocide in Yugoslavia and Rwanda in the 1990s.

The importance of these different ways of defining hate speech lies in the fact that, although the concept of hate speech is not a strictly legal category, these definitions are

---

[6] Recommendation No. R(97)20 of the Council of Europe, of 30/10/1997, defined it as “all forms of expression that spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism and other forms of hate based on intolerance (...)”

[7] Susan Benesch (2013) suggests the term “dangerous speech” for speech that has a reasonable probability of unleashing episodes of genocide violence.

reflected in national legislation and case law, and in international declarations, directives and agreements.

Either way, for the purposes of this report, we will take as a reference the definitions of the first group, which conceptualise hate speech in a relatively broad sense, because they include those types of expression which, given their proliferation in the European context, merit a more varied range of legal and social interventions.

1.2. The particular features of online hate speech

Online hate speech, for which the expression cyber hate has been coined, adds a number of specific features that turn it into an uncontrolled phenomenon with an even greater potential for causing harm. Firstly, the superabundance of communication. Thanks to the Internet and social media, the communication of messages previously limited to the private sphere has moved into the public sphere on an absolutely massive scale. Secondly, the decentralisation of communication, which "democratises" communications, in the sense that anybody can send a message to a potentially enormous audience. The multiplying effect of social media allows a given message to be turned into a phenomenon of exponential transmission, giving rise to what is metaphorically called "viralisation"

The permanence of the content, roaming between different platforms, the use of synonyms, anonymity and transnationality are other features of the online space that present additional challenges in relation to hate speech(Gagliardone et al. 2015). Permanence and roaming are features that can increase the potential harm of hate speech. Anonymity, the use of pseudonyms and transnationality are characteristics that can make prosecution difficult and, given the feeling of impunity they offer, encourage the expression of hate speech.

On occasions, we refer to the Internet in colloquial speech as a "virtual" space, but this adjective is misleading, inasmuch as it refers to something "which has an apparent, but not real existence"[8]. This feeling of "virtualness", as if something that happened on the Internet was not "real" and had no repercussions outside the Net, also has a disinhibiting effect on the expression of extreme speech.

In that regard, Jubany and Roiha (2016) warn of the fallacy of digital dualism, a term conceived by Nathan Jurgenson to refer to the false separation between the online and offline worlds. The question of hate speech in social media, despite the specific features of these communication spaces, is no more than the public externalisation of an underlying problem in society in general, regardless of the setting in which it manifests itself.

1.3. The consequences of hate speech

Hate speech, whether it is online or not, can have various repercussions. On the one hand, we need to consider the direct emotional and psychological damage that can be caused by threats, harassment and other attacks targeted at specific individuals and motivated by hate and intolerance. On the other hand, hate speech also causes indirect harm, eroding people's dignity and damaging their reputation.

---

[8] Definition R.A.E.

Hate speech plays a fundamental role in perpetuating discriminatory stereotypes, stigmatising groups, contributing towards their marginalisation, and so on. It constitutes a linguistic mechanism of vital importance in dehumanising certain groups, denying they belong to the citizenry on an equal footing. In that sense, hate speech sends a message that divides and segregates society. But at the same time, it also plays a cohesive role for originators of hate speech, reinforcing their sense of belonging to a group (Gagliardone et al. 2015).

The proliferation and acceptance of hate speech has the effect of reducing empathy towards the dehumanised groups and, as a result, can create fertile ground for justifying discriminatory acts, abuse and violent acts of various kinds. Although, generally speaking, a direct link cannot be established between the proliferation of hate speech and violent hate crimes, there is an ever greater awareness of the indirect link between both phenomena.

And while not all hate speech unleashes episodes of violence, what does seem obvious is that episodes of hate crime rarely occur without the prior stigmatisation and dehumanisation of the victims. It is therefore worth recalling at this point the incidents of hate crime recorded in 2015 by the Ministry of the Interior. There were 1,328 cases of hate crime reported and recorded in 2015, which includes injuries, sexual abuse, harm, degrading treatment, and so on. Movimiento contra la Intolerancia estimates put the real figure at around 4,000 cases. Bearing in mind the Ministry of Defence itself acknowledges that many hate crimes are not reported, these figures are an indication of just how wrong it would be to take hate speech lightly as a substratum justifying and inciting violent behaviour (Ministerio del Interior 2016).

1.4. Who are the perpetrators of hate speech?

There can be no doubt that much of the hate speech in social media comes from individuals directly or indirectly linked to fascist groups such as Hogar Social Madrid, which has over 15,000 followers on Twitter, groups of fanatical football supporters (*ultras*) and more or less marginal far-right political parties, such as Plataforma per Catalunya, Falange Española de las JONS, España 2000, Alianza Nacional, and so on.

In social media, these groups have found a useful tool for spreading their messages and fascist symbols, which they use to attack their victims and recruit members. Individuals associated with these groups usually act together in the social networks, producing a greater feeling of harassment among their victims. Moreover, to give free rein to their incitement to violence in the networking sites, they develop different linguistic codes and key words, which they use to avoid the computer filters of the tech companies as well as police and legal action against their behaviour.[9]

However, one thing that is clear is that the phenomenon of hate speech does not end at these groups with a marked ideological profile. There are other social media users, known as haters or trolls in Internet slang, who constitute an important part of the problem.

Individuals who are obsessed with attacking and verbally abusing specific people

[9] http://www.genbeta.com/redes-sociales-y-comunidades/asi-es-como-los-extremistas-camuflan-su-discurso-de-odio-en-las-redes-sociales

(largely famous persons), or groups they despise because of their ethnic origin, their religion, etc., are called “haters” on the Internet. Social media are plagued with users who have an obsession with disparaging and verbally attacking Muslims, Gypsies, homosexuals, women and other groups. A similar profile is that of users with a tendency to voice their demands and concerns on any matter in an extreme and aggressive manner, as a way of attracting attention and highlighting their messages.

The “troll” phenomenon is also widespread in social media but it is very risky to try to characterise such a heterogeneous phenomenon. Some act on their own to attract attention or vent their frustration. Others act together and organise to attack particular targets. Either way, there can be no doubt that part of the narcissistic and sadistic troll phenomenon springs from the simple impulse of having fun provoking other social media users. Without thinking, and often oblivious to the harm they might cause, many users spend their time spouting their hostility towards, and verbally abusing other individuals and minority groups so they can enjoy the reactions they provoke in others, with the excuse of an alleged sense of humour.

But distinguishing those who are trying to provoke for fun from those expressing their real opinions is not always easy. In tackling the phenomenon of hate speech in social media, it is therefore important to bear in mind the diversity and heterogeneity of author profiles, not only to establish the scale of the problem but also for developing suitable strategies for the different sources of extreme speech. Using these terms, which belong to Internet slang, runs the risk of excessively caricaturing the problem, and the tendency mentioned previously of considering it as something removed from offline reality, which Nathan Jurgenson calls the fallacy of digital dualism. The hostile obsessions of haters and the narcissistic sadism of trolls have their space for expression in the social media but their hostile motives and prejudices come from the offline space.

2. RESTRICTING HATE SPEECH

2.1. The dilemma between freedom of expression and suppression of hate speech

Increasing alarm at an international level at the proliferation of hate speech, and the devastating consequences it had in the genocides in Rwanda and Yugoslavia, for example, has led to big movements and pressure to restrict hate speech. As a result, states have opted for different levels of restriction, depending on the historical, social and political context in each case.

The diverse range of legal approaches to this problem and the corresponding level of suppression of hate speech, has its origin in the important debate this question generates regarding the limits of freedom of expression. Freedom of expression, despite being a fundamental pillar of the democratic system, is not an absolute right[10].

[10] The 1966 International Covenant on Civil and Political Rights established limits to freedom of expression Article 19 envisages the possibility of restrictions to “ensure respect for the rights and reputations of others” and for “the protection of national security, public order, or of public health or morals”. Article 20 bans “any advocacy or national, racial or religious hatred that constitutes incitement to discrimination, hostility or

All legal systems contain criminal categories for restricting it. For example, there is no right to insult, slander or threaten. In the case of hate speech, tension arises between the right to freedom of expression and respect for human dignity and equality. Hate speech seeks to dehumanise and socially exclude vulnerable minorities, thus depriving them of their dignity and right to equality.

At an international level, the different approaches to restricting hate speech are usually reduced to the division between a more liberal and "tolerant" American perspective and the European perspective, which is more militant and "intransigent" towards hate speech.

- The United States has traditionally been characterised by its uncompromising defence of the First Amendment[11], and its "commitment to upholding freedom, even in exceptional circumstances" (Revenga-Sánchez, 2015). Defence of freedom of expression has been framed in a conception of public debate as a free market of ideas, in which the truth always ends up gaining acceptance.
- Europe, along with other countries such as Canada, Australia and South Africa have generally taken a harder line with hate speech, resorting to the abuse of law to deny the possibility of invoking freedom of expression to accommodate Holocaust denial, advocating terrorism or openly xenophobic and racist messages.

However, this distinction between ways of understanding the limits to freedom of expression is not so straightforward. Firstly, the difference has softened in recent years, among other reasons because the traditional American respect for freedom of expression has lost ground following the security turn taken within the framework of the so-called "war on terror". And secondly, because in Europe and the US there are experts and academics who sympathise with and support a perspective and arguments contrary to those that would, in theory, correspond to them, given their legal tradition (Phillipson 2015).

On the other hand, in the Muslim world, some countries have pressed for the adoption of international laws that would prohibit blasphemy as a form of religious hatred, in response to cases such as caricatures of Muhammad in the Danish newspaper, Jyllands-Posten (Phillipson, 2015).

In the academic domain, the debate on banning hate speech and freedom of expression is one that transcends the legal sphere and has implications for public order, ethics and philosophy. The debate has been approached from various angles and it raises very varied questions.

- Are bans on hate speech a threat to the freedom of expression?
- Can they be considered a necessary evil?
- Is banning hate speech a necessary sign of the democratic commitment to respect for human dignity and equality?
- Or is legalising bans on hate speech a symptom of a weak democracy?

---

violence".

[11] "Congress shall make no law respecting an establishment of religion, (...) or abridging the freedom of expression, or of the press..."

- Are bans on hate speech effective?

The following table shows in a simplified form some of the main arguments that are usually used for and against banning hate speech.

For	Against [12]
Freedom of expression is not an absolute right. It has to be restricted to protect human dignity, equality, peace and social harmony, the right to live without being harassed and intimidated, etc. The International Covenant on Civil and Political Rights recognises this in Articles 19.3 and 20	Freedom of expression and public debate are essential for full democracy. Restrictions should not be imposed simply on the basis of how undesirable and offensive certain expressions might be. It runs the risk of an abuse of power on the part of governments that use it as a "criminal law for the enemy" to punish dissidents and political opponents.
The proliferation of extreme and offensive comments and expressions could create a social climate that leads to outbreaks of violence and, in the context of violence and divided societies, to genocide attacks.	No clear link has been demonstrated in established and prosperous democracies between the proliferation of hate speech and an increase in hate crime offences. The State can already punish "hate" as an aggravating circumstance when it is a motive for criminal acts.
Hate speech causes direct psychological damage (feeling of being threatened, humiliated, etc.) and indirect harm, in the sense that it contributes towards perpetuating situations of discrimination.	There are already legal means for prohibiting harassment, threats and individual attacks on dignity; more legitimate, proportionate and effective instruments for combating discrimination; laws against discrimination at work, education in diversity and pluralism, awareness campaigns, etc.
The most extreme declarations of hatred incite acts of violence and discrimination, and should therefore be restricted and punished.	The State can already punish acts of provocation and conspiracy, where a clear material link can be established with the commitment of a crime, but the concept of incitement gives it a mechanism for punishment without any need to demonstrate the possibility that the damage is a result of the public expression of ideas.
The State cannot remain neutral and bans on hate speech are a declaration of principles, symbolising its commitment to democratic values, equality and human dignity.	The State has more legitimate and more effective ways of positioning itself on a symbolic level. Bans have little effect and could even be counter-productive, because they run the risk of discrediting the real promotion of equality and because the can turn certain individuals into "martyrs" for freedom of expression.

[12] Table based on Nineteen arguments for hate speech bans-and against them by Eric Heinze.

With regard to the effectiveness or counter-productive character of bans, there is a very interesting study carried out in Holland by Spanje, J. and de Vreese, C. (2015), on the impact that judicialising the hate speech of certain far-right politicians has on support for their parties. These researchers found empirical evidence that the start of legal proceedings against the anti-immigration discourse of Geert Wilders strengthened his party's appeal among voters and they suggested the decision to judicialise his Islamophobic statements contributed towards his electoral take-off (Van Spanje y de Vreese 2015).

With regard to the possibility of the concept of hate speech being politically abused, South Africa under Apartheid, where laws against this type of speech were used to criminalise criticism of white domination, is a prime example (Gagliardone et al, 2015). We have also seen various examples recently of the concept of hate speech being used against dissidence and political enemies in the Spanish state. For example, the case of César Strawberry, leader of Def-Con-Dos[13], who the High Court accused of inciting hate; that of Guillermo Zapata, a Madrid city councillor, who was prosecuted for some jokes posted on Twitter when he was discussing the limits of black humour [14]; or the case of a young woman from Valencia called María Lluch, and known as "Madame Guillotine" on Twitter, who was condemned to a year in prison for making jokes about ETA's victims because the Supreme Court considered her tweets as hate speech not protected by freedom of expression[15], despite reducing the initial sentence of two years on the grounds it was disproportionate. Regardless of the ethical judgement each one might merit, these cases illustrate the problem of how the concept of hate speech might be used to justify criminal punishment of certain remarks made by political opponents.

We should therefore not lose sight of the notion that any legislation, including that related to restricting hate speech, is the expression of the dominant group which controls the content of the law (Gagliardone et al. 2015).

2.2. Legal aspects of hate online and in social media

As if this political, ethical and legal debate did not already pose enough difficulties, the problem of hate speech is made even more difficult by the Internet and social media. Firstly, because the Internet has become a kind of liberation fetish, idealistically regarded by many as "*a revolutionary force that should not be subject to any regulation*", and any attempt to regulate what happens on the Net is automatically branded as reactionary, illegitimate and undemocratic (Morozov, E., 2012).

But the Internet and social media pose a second problem as well, linked to the "border-free" character of the Net: the problem of delimiting jurisdiction. A message published through a server in the United States can have damaging consequences in Spain. Which legislation is applicable? American, which is more tolerant, or Spanish?

---

13 http://www.eldiario.es/politica/Absuelto-Cesar-Strawberry-enaltecido-Twitter_0_538946313.html

14 http://www.lavanguardia.com/politica/20160307/40269028704/archivada-causa-zapata-tercera-vez-tuits.html

15 http://www.poderjudicial.es/cgpj/es/Poder-Judicial/Noticias-Judiciales/El-Tribunal-Supremo-condena-a-un-ano-prision-a-una-joven-por-humillar-a-traves-de-twitter-a-Irene-Villa-y-a-Miguel-Angel-Blanco

This extraterritoriality also poses a challenge for judicial cooperation, inasmuch as legislative differences also affect very important questions relating to cyber crime, such as data protection and communications secrecy.

Furthermore, it poses problems that stem from the technical configuration and functionality of the Internet (server localisation, IP authentication, robot accounts, different encrypting procedures for hiding identity from attacks, etc.) that give rise to a number of uncertainties and difficulties in obtaining proof or determining responsibility(Moretón Toquero 2012).

Without entering highly complex legal terrain here, the change in the communication model which the Internet implies means that the debate on hate speech is faced with two alternatives. The first is to seek an international dialogue to achieve a "*global agreement on the criteria for restricting or tolerating online hate speech*", which is not very realistic. And the second is the "*enforcement on the Internet of the constitutional policies of each state*" (Rodríguez Izquierdo, 2015), which presents the limits and legal uncertainties outlined here superficially.

2.3. Spanish legislation

Spanish legislation covering hate speech in social media is conditioned by the need to adapt to European norms, which follow a trend increasingly widespread on a global level of introducing penal restrictions on hate speech[16].

The last reform of the Criminal Code[17] was in response to the need to adapt it to the aforementioned European regulation and the Constitutional Court's ruling, TC 235/2007, of 7 November, which imposed a restrictive interpretation of the crime of genocide denial, limiting its application to cases where this behaviour implies incitement to hate and hostility towards minorities.

The new regulation brings the various hate speech crimes together in Article 510 of the Criminal Code, and punishes them with the following penalties:

1- 1 to 4 years in prison and a fine of 6 to 12 months,
   a. Direct or indirect incitement to **hatred, hostility, discrimination and violence**, against groups or individuals for racist, anti-Semitic and other motives linked to ideology, religion, origin, sex, sexual orientation, illness or disability.
   b. Production and distribution of **material** with the same motives as those mentioned in Section a).
   c. **Denial or glorification** of the crime of **genocide**, crimes against humanity or against individuals or property in the case of armed conflict that may have been committed against the aforementioned

[16] Specifically, Framework Decision 2008/913/JAL of 28 November, on combating certain forms and expressions of racism and xenophobia by means of criminal law, of the European Union, establishes some basic guidelines that should be followed in all EU member states. Also worthy of mention in this regard is the Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems, approved by the Council of Europe in 2003

[17] L.O. 1/2015, 30 March

groups, when that promotes or fosters a **climate of violence, hostility and hatred** towards them.

2- 6 months to 2 years in prison and a fine of 6 to 12 months (and 1 to 4 years in prison when a climate of violence, hate or discrimination is promoted)

   a. Holding events, producing and disseminating material that might imply **humiliating or disparaging** one of the aforementioned groups or its members on discriminatory grounds;
   b. public **glorification or justification** of **crimes** committed against the aforementioned groups.

The regulation of Article 510 also establishes a series of aggravating circumstances and accessory consequences. For the purposes of this report, it is worth noting the aggravating circumstance for having engaged in any of the types of conduct described using "new technologies".[18] In that case, taking into account the potentially greater reach of the punishable behaviour, the terms of imprisonment will be in the upper half of the range. Withdrawal of the content is also envisaged, along with blocking access to, or interrupting delivery of the service[19].

Part of legal opinion sees the reform as correcting some of the shortcomings of the previous regulation. Its main contribution is to remove the word "provocation", which had created difficulties in applying the Article, a disparity of criteria and contradictory sentences, by interpreting it in the sense of Article 18 of the Criminal Code, as "direct incitement". Removing the term "provocation" makes it easier to punish "indirect incitement", explicitly stated in the text, as well (Gascón Cuenca 2015).

This part of legal opinion also considers the inclusion, in Section 2 of the Article, of two lesser offences for conduct "*that does not have all the characteristics for inciting hostility, discrimination and violence, but is ideal for wounding the dignity of individuals*", with shorter terms of imprisonment (6 months to 2 years), means the regulation is more in line with the principle of proportionality (Gascón Cuenca 2015).

However, in the view of other authors, this reform is technically poor and reactionary, and doubts are still being raised about its constitutionality, because it fails to adequately justify the intervention of criminal law and because it continues to suffer from a lack of proportionality. Some criminal lawyers even believe it has made most of the defects of the previous regulation worse by leaving too much to the opinion of the judges regarding whether or not to penalise conduct that apparently matches the types regulated by Article 510 (Teruel Lozano 2015).

In the view of critics of the new regulation, there are better and more proportionate responses than resorting to criminal law. The Spanish legal system has "*sufficient instruments to punish"*, for example, humiliations, threats and provocation of a certain and imminent danger of illegal acts*, "without the need to formulate an exceptional*

[18] Art. 510.3 CP

[19] Art. 510.6 CP

response by means of ad hoc types of anti-discrimination offences" (Teruel Lozano 2015: 39).

According to Rey Martínez (2015: 52), the problems arise from the fact that the approach to this question has leaned too much towards criminal law. In his view, "*an administrative sanction, and even the requirement of civil responsibility, might be much more respectful*" of the Constitution and would probably be more effective than a criminal law which, because of its rigorous and disproportionate nature, can ultimately prove to be inapplicable for the judges.

Indeed, until now, case law has applied a restrictive interpretation of hate crime offences. This has resulted in a legal panorama of very harsh prison sentences and no sanction of hate speech. In Rey Martínez's view (2015: 77), what is needed is "*more sanction of hate speech, but with a proportionate penal harshness*". The new regulation, despite the introduction of lesser offences in Section 2 of Article 510, has improved little in that regard, by establishing terms of up to four years in prison, even distancing itself from the EU's Framework Decision 2008/913/JHA, which stipulates maximum terms of three years (Gascón Cuenca, 2015).

We shall see how judges and magistrates interpret the new regulation. Either way, it is to be hoped they do not get caught in the contradictions that arose in applying the previous legislation, and which some experts, such as Miguel Ángel Aguilar, the public prosecutor coordinating hate crimes at the Public Prosecutor's Office in Barcelona, criticised harshly[20], after the Supreme Court delivered two contradictory rulings in the space of a week.

In any event, it should not be forgotten that criminal punishment of what is generically understood by hate speech is not exhausted by Article 510 of the Criminal Code The Criminal Code envisages a generic aggravating circumstance for discriminatory motives, which aggravates the criminal responsibility of any offence and, for example, can be applied to behaviour typical on social media[21], such as threats or harassment. Consideration that the aggravating circumstance of discrimination coincides means that the punishment corresponding to each offence imposed will be in the upper half of the range. Moreover, there is an offence for threatening groups,[22] which punishes threats capable of striking fear targeted at ethnic, cultural and similar groups.

Main problems for prosecuting hate speech crimes

Some of the main practical problems in prosecuting hate crime in general and hate speech crime in social media in particular are, in the opinion of the experts (Aguilar et al., 2015), as follows:

- Ignorance of the facts due to the very small number of complaints.
- The people in charge of investigations playing down the facts.
- Not enough police or judicial investigation.
- Difficulties in proving the "hate" or discriminatory motive of the crime.

---

[20] http://www.elmundo.es/elmundo/2011/06/13/barcelona/1307982347.html

[21] Art.22.4 Criminal Code

[22] Art.170.1 Criminal Code

- Lack of training on equality and non-discrimination for the institutions involved: the judiciary, judicial civil servants, forensic specialists, the police, private security, etc.
- Inherent difficulties of the "social media" scene: problems of jurisdiction, territoriality, identification and obtaining data on the aggressors, etc.

Administrative regulations

From the point of view of administrative law, there are some regulations that contemplate the possibility of fining conduct that can be understood as hate speech. In that regard, the most specific legislation is Act 19/2007, Against Violence, Racism, Xenophobia and Intolerance in Sport, which is an area where intolerant behaviour has a particular impact and where for years groups of far-right fanatical football fans found a space that allowed them to give free rein to their openly racist, homophobic and discriminatory attitudes with relative impunity.

In relation to the specific subject matter of this report, this Act regards the dissemination, via computer or technological media linked to sports information, "of content that promotes or supports violence, or which encourages and abets violent, terrorist, racist, xenophobic or intolerant behaviour for reasons of religion, ideology, sexual orientation (...)" as a very serious offence, and punishes it with big fines.[23]

Specialised institutions and police protocols against hate and discrimination.

In March 2013, the Prosecution Service launched a network of **public prosecutors who specialise in combating hate**, with one prosecutor in each province tasked with coordinating the provincial public prosecution service's action on hate and discrimination, and coordinating with the rest of the network.

This figure gave rise to the Discrimination and Hate Crime Service at the Public Prosecutors' Office in Barcelona, which was set up in 2009 as a pioneer in this area, and which has been recognised by the European Union Fundamental Rights Agency as an example of good practice that should be followed by other member states.

The challenges this service has set itself include the following (Aguilar et al. 2015):

- gather better information on incidents that could constitute punishable conduct in this area;
- encourage groups and individuals affected to communicate with the Prosecution Service, both to report incidents as well as share ideas and initiatives;
- help to unify the criteria for applying and interpreting the Criminal Code;
- tackle the complexity of investigating offences involving incitement to hatred, violence and discrimination, when they occur online and in social media;
- compile recommendations and case law published by international bodies and courts on non-discrimination;
- combat organised violent groups that spread hate speech;

[23] Art. 23.1 c) and 24.1 c) of Act 19/2007. Fines from €60,000.01 to €650,000

- tackle the tendency of some police officers, judges and prosecutors to minimise the importance of hate crime;
- improve coordination with the State's law enforcement agencies.

In that regard, it is also worth pointing out the development of **police intervention protocols**. The Mossos d'Esquadra (Catalan police) were pioneers in this area, with a "Procedure for Criminal Activities Motivated by Hate or Discrimination" in 2010. In 2014, the Ministry of Interior published the "Action Protocol of the Law Enforcement Agencies for Hate Crime and Conduct that Violates the Legal Norms on Discrimination", led by the Policía Nacional and the Guardia Civil, and based on material produced by the Spanish Racism and Xenophobia Observatory, which is attached to the Ministry of Employment and Social Security (Aguilar García 2015).

These protocols are in response to the need to train police forces and provide them with some guidelines for collecting statistical data, drawing up reports, investigating incidents to confirm the existence of racist or intolerant motives, dealing with and protecting victims, and talking to the community and NGOs that represent them.

The Ministry of Interior Protocol envisages specific recommendations for hate crime offences committed online and in social media. More specifically, and given the difficulties that the use of information technologies poses for prosecuting and identifying those responsible for offences, it recommends that investigations should be carried out by specialist units of the Judicial Police (Ministry of Interior, 2015).

Furthermore, both public institutions and some NGOs (SOS Racismo, Movimiento contra la Intolerancia, Observatorio contra la Homofobia, Fundación Rais, etc.) offer a number of services geared towards providing victims of hate crime offences in general, and hate speech offences in particular, with legal advice and assistance. In Catalonia, the city councils of Sabadell, Terrassa, and Barcelona (Non-Discrimination Office, OND) have a number of services that offer legal advice in this area.

3. THE ROLE OF SOCIAL MEDIA AS INTERMEDIARIES

There is no doubt that social media, as a communication channel for "hate" messages, also play an important role in the fight against hate speech. Just remember that the leading social networking site, Facebook, has over 1.7 billion active users[24], which is approximately equivalent to the combined populations of the European and American continents. These huge figures make social media a player, with an intermediary role but undeniable power.

Their role as intermediaries in digital communication makes them the first arbiter in determining what can and cannot be said. In this regard, it is worth recalling that the big social networking sites with the most users in Europe are based in the United States and, therefore, imbued with North American philosophy, which strongly defends a more absolute, more tolerant freedom of expression, with a certain amount of hate speech. As we pointed out in the previous section, this has legal repercussions, for example, in

[24] http://newsroom.fb.com/company-info/

terms of collaborating with the law on data requirements. But it also has an influence on their self-regulation policies and how these are implemented.

In relation to these intermediary companies, various alternatives are proposed:

- attributing responsibility to the companies,
- promoting self-regulation,
- and installing computer filters (Rodríguez Izquierdo 2015).

Each of these options poses certain dilemmas.

3.1. Attributing responsibility

While service providers in the United States are virtually free of any responsibility for published content, the attribution of responsibility to intermediary service providers in the European Union is largely based on their effective knowledge of the illicit nature of the content published through their service (Rodríguez Izquierdo 2015). In theory, they will only be held responsible when they do not remove or impede access to illicit content promptly, after having been made aware of its illicit nature by means of notification from the rights holder or an order from a competent body.

The Delfi ruling of the European Court of Human Rights established the responsibility of the Estonian news portal, Delfi, by regarding it as the content provider, for comments posted by its readers, even if it has not received a request for illicit content to be removed. However, in its ruling, the Court established that this interpretation is not applicable to social media, by establishing a difference between technical service providers, with a passive role, such as social media, and content service providers, such as a news website or an online communication medium[25].

3.2. Self-regulation and installing computer filters

The prosecution and restriction of hate speech offences by states is faced with the difficulties mentioned above in acting beyond their borders. However, and here is one of the main advantages of self-regulation strategies, companies, as private operators, have the right to establish globally applicable ethical codes and conditions of use.

All the main social networking sites, Twitter, Facebook, Instagram, Tumblr, etc., and even the media, whose comment forums are also a regular channel for hate speech, have policies and terms and conditions of use that, in theory, ban hate speech on their communication platforms.

Twitter, for example, expressly[26] prohibits threats, incitement to violence, harassment and behaviour that incites hate motivated by ethnic origin, nationality, sexual orientation, gender, religion and so on.

In its terms of service, Facebook includes clauses on not posting content that intimidates, discriminates or incites violence but expressly recognises it is unable to guarantee that Facebook is a safe place and describes its rules as user commitments. From the perspective that it constitutes a global community, subject to different national

[25]http://blog.garrigues.com/se-confirma-la-responsabilidad-de-los-portales-de-noticias-por-los-comentarios-difamatorios-de-los-lectores/

[26] https://support.twitter.com/articles/72688

legislation, Facebook establishes some standards as a form of self-regulation.[27] By means of these standards, Facebook sets out what content can be reported and removed, notably, for the purposes of this report, language that incites hatred motivated by race, ethnicity, nationality, religion, sexual orientation, sex, gender, sexual identity, disability or serious illnesses.

Most social media, including the two big social networks, Facebook and Twitter, have reporting systems that allow users to make the companies aware of the presence of abusive content that breaks the terms of use of the platforms, so the companies can assess whether to remove it. Facebook expressly warns in its community standards that not all offensive content infringes those standards and, therefore, it also offers the option of personalised blocking and filtering to hide content that users do not wish to see.

Nevertheless, despite this attitude which, theoretically, means the companies do not tolerate hate speech, it is clear that the mechanisms for reporting and removing comments are somewhat dysfunctional in practice. That is borne out in research carried out by the PRISM project (Jubany and Roiha 2016), which found that out of a total of 100 reports to Facebook, only 9 resulted in the comments being removed[28]. The company's erratic policy of removing comments, either because the algorithms they use malfunction or because the teams responsible for managing the reports are overwhelmed by the huge number of comments reported, led to messages such as "*Los perros son más discretos que los musulmanes*" (Dogs are more discrete than Muslims) being removed, while a message such as "*Matarlos a todos sin piedad*" (Kill them all without pity) was considered by Facebook to be a message that did not infringe its community standards.

One of the main criticisms that has been made of social media reporting mechanisms is the lack of transparency. It is not known for certain if the process is carried out by means of algorithms, human teams or a mixture of both. Nor do they publish data on the number of reports, the reasons for them or the percentage that lead to comments being removed and accounts being blocked (Jubany & Roiha 2016).

Facebook has also developed a powerful artificial intelligence system, called Deep Text, which is supposedly capable of analysing several thousand posts a second in more than 20 languages. However, at present, it does not appear to be working successfully. However sophisticated the algorithms this system is based on may be, or other computer filters developed by other social media, they tend to commit block errors and are not usually capable, for example, of differentiating between the critical use or reporting of a specific insulting expression. The case of Shaun King, a New York Daily News journalist and a Black Lives Matter activist, was particularly symptomatic. Despite having been invited by Facebook to be a guest speaker for this political movement against violence towards Black people months earlier, his Facebook account was suspended for publishing an email in which he was the target of racist insults. Thanks to his Facebook contacts, he managed to get the service re-established but he wondered

---

[27] https://www.facebook.com/communitystandards#

[28] http://www.elperiodico.com/es/noticias/sociedad/estudio-proyecto-prism-denuncia-banalizacion-discurso-odio-internet-5284591

how a person without privileged access to certain people in the company would handle a case like his[29].

Social media sites are aware of the problem they have with handling online harassment, threats and hate speech. In 2015, an internal message was filtered from Twitter's then CEO, Dick Costolo, in which he admitted they had a serious problem handling this issue, which was losing them users every day and limiting their targets for growth[30]. For that reason, Twitter has begun to see this as a priority and is immersed in a process of change that will give users more proactive control over their Twitter accounts. The company has sought the advice of an external committee of civil society organisations involved in preventing online abuse[31] and, among other tools, they are working on introducing quality filters to block certain words, hashtags, racist insults, etc., so they do not appear on their time line. Their aim is to find better mechanisms for blocking other users, as well as speeding up and optimising the systems for reporting abusive behaviour to the company.

A prime example of Twitter showing less tolerance to online hate occurred in July, 2016, when Milo Yiannopoulos, a symbol of the alt-right movement in the United States, was temporarily suspended for allegedly being behind the misogynist and racist harassment of the Afro-American actress, Leslie Jones. Milo's presence in the media, and the fact he had 300,000 followers on Twitter, turned the case into a debate on the freedom of expression and the hashtag #FreeMilo into a trending topic, making him a kind of martyr for freedom of expression[32].

There is no doubt that the fact the person being attacked was a Hollywood celebrity helped to highlight and resolve the case, which the company's current CEO, Jack Dorsey, publicly intervened in. Despite the importance of symbolic cases such as this, the problem is that "anonymous" people who suffer social media humiliation and harassment are not able to take their cases to such a high level. For now, we will have to wait and see if the technological changes the company is working on represent a real change in managing the problem of online hate and abuse.

Either way, for balance, and to soften the criticism usually aimed at these companies, it is only fair to remember the enormous difficulty of the task facing them. Twitter, for example, handles 300 million tweets a day, with spikes of up to 600 million[33]. Even with the help of computer filters and algorithms, the task of moderation and self-regulation is one of gigantic proportions, as well as a very sensitive one, inasmuch as they are grappling with such sensitive concepts for their image as freedom of expression and censorship.

3.3. The Code of Conduct with the European Union

In the wake of the refugee crisis in 2015, and the obvious growth in xenophobic and racist statements in social media, the European Union pressurised the tech companies

29 http://www.eldiario.es/theguardian/Facebook-temporalmente-Black-Lives-Matter_0_558544311.html

30 http://www.lavanguardia.com/tecnologia/20160901/4122751789/twitter-acoso-trolls-redes-sociales.html

31 https://about.twitter.com/es/safety/council

32 http://www.bbc.com/mundo/noticias-36847433

33 http://uk.businessinsider.com/tweets-on-twitter-is-in-serious-decline-2016-2

to assume a much more active role in the fight against hate speech.

As a result of this pressure from some states and the European Union itself, Facebook, Twitter, YouTube and Microsoft signed a Code of Conduct in May 2016 on illegal incitement to hatred on the Internet[34]. In that agreement, the tech companies made a number of commitments, notably a commitment to check requests for the removal of illegal content inciting hatred within a period of 24 hours. Furthermore, they undertook to establish clear procedures for examining reported content, and assessing the reports their services receive, in accordance with their self-regulation standards, but also taking into account, whenever necessary, the national legislative transposition of the European Framework Directive on combating certain forms and manifestations of racism and xenophobia into the criminal law that the agreement regards as its legal basis.

The agreement seeks cooperation between the signatory companies and the exchange of good practices with other online social communication services, whose adherence to the agreement will be encouraged.

The signatory companies also undertook to promote initiatives that develop an alternative "counter-speech", to support educational programmes that foster critical thinking and to collaborate with civil society organisations in training activities. This commitment, signed in the Code of Conduct, formalises and commits a line of action already being pursued by the tech companies in this field to improve the way hate and harassment on social networking sites are dealt with.

Compliance with the Code of Conduct and its results will be subject to evaluation by the recently created High Level Group on combating Racism, Xenophobia and other forms of Intolerance, comprising the EU member states, representatives of the European Parliament, the Council of Europe, the UN High Commissioner for Refugees, the Organisation for Security and Cooperation in Europe, the European Fundamental Rights Agency and civil society organisations.

Although the European Commissioner for Justice, Consumers and Gender Equality, Vera Jourová, presented this agreement as a turning point that would change the rules of the game[35], the signing of this code of conduct has not been exempt from criticism. Firstly, criticism has been expressed at the lack of transparency in the negotiation process and that the contributions of the civil society organisations have not been taken into account. There has also been criticism of the fact the code will allow the companies to take the lead on an issue, controlling online hate speech, which should be led by the public authorities, whose activity is subject to a framework of democratic responsibility. Leaving leadership on this matter in the hands of private companies poses a clear risk to freedom of expression, inasmuch as the company mechanisms for checking content are not subject to a clear, transparent framework of responsibility[36]. Ultimately, say the critics, it is a question of relying on the companies' being responsible or, rather, the interest of their shareholders coinciding in some way with social interest and respect for freedom of expression, as well as the principles of equality, dignity, etc., that are part of

34 http://europa.eu/rapid/press-release_IP-16-1937_es.htm

35 http://europa.eu/rapid/press-release_SPEECH-16-2197_en.htm

36 https://edri.org/edri-access-now-withdraw-eu-commission-forum-discussions/

democratic systems.

Another problematic question is that, while the alleged hate speech is removed by the companies without a clear collaboration procedure being in place to send that content to the public authorities, the authors of the alleged breaches of the law will not be subject to trial or criminal punishment of any kind.[37]

Either way, there can be no doubt that the European institutions are seriously concerned about online hate speech, and that they are trying to pressurise the companies into carrying out a more active moderation of social media sites

4. EXTRAJUDICIAL STRATEGIES FOR COMABTING HATE SPEECH ON SOCIAL MEDIA

Below we outline, without any claim to being exhaustive, the main "non-judicial" strategies being pursued in response to hate speech in general, and in social media in particular.

A large part of these strategies are directly or indirectly funded by public institutions and, for some authors, this poses two important questions. On the one hand, whether they are really effective. And on the other hand, the possibility of a certain breach of the alleged principle of State neutrality (Kahn 2015). However, it is clear that even most of the authors opposed to banning hate speech believe that public institutions must not remain neutral in the face of this type of extreme speech. Very few authors defend a principle of strict State neutrality. They argue that the mistaken, or excessive implementation of certain strategies could produce undesired effects, such as isolating or excluding certain speech from public debate that may be disagreeable but is still legitimate. Most do not doubt that public institutions should help to create a suitable climate for the exercise of freedom of expression that respects the principles of equality and non-discrimination (Article 19 2015).

Through the declarations of politicians and public servants, the authorities can foster empathy and encourage support for victims of hate speech, spreading the idea that this kind of speech harms community relations. Such statements are of vital importance in situations where there are disputes between ethnic communities in cities and neighbourhoods, or in contexts where identity tensions sharpen, such as electoral periods. Other initiatives that public authorities can take to ensure a suitable climate for equality and non-discrimination are training for public servants on these issues, strategic litigation and the promotion of inter-religious dialogue (Article 19 2015).

At a more concrete level, in relation to initiatives specifically targeted at online hate speech, and in social media in particular, various non-legal strategies are being implemented.

- *monitoring and research strategies*, with the intention of learning more about the scale of the hate speech problem and exploring the possibility of using early warning systems capable of identifying it automatically;

37 https://edri.org/guide-code-conduct-hate-speech/

- *strategies for pressurising social networking site operators*, which are geared towards getting them to modify their policies on the types of content that can be shared or withdraw specific content.
- *strategies focused on changing the perceptions and attitudes* that Internet users have towards hate speech in general, and on social media in particular.

Among the latter group, which this report will pay greater attention to because they are more easily applied at a local level, two broad categories stand out: on the one hand, those that consist of *campaigns,* which include activities of various kinds and whose purpose is usually to attack prejudice and intolerance as the main root of the problem of hate speech; and, on the other hand, strategies based on *training and education,* which include workshops, seminars and other types of activities designed to equip young people especially with the necessary skills for identifying and combating online hate speech.

4.1. Monitoring and research strategies

Such strategies are important for gaining an overall understanding of the phenomenon of hate speech in social media: the scale of the problem, the different types of hate, how and when more virulent situations arise, what kind of profile the perpetrators have, what type of language they use, which platforms are more propitious for propagating extreme speech, what effects it has on other users, and so on. Deepening our understanding of hate speech is valuable for the knowledge itself but, above all, for applying palliative solutions to the problem, or at least preventing its more damaging consequences.

The monitoring tasks face considerable difficulties, despite the undoubted technical progress in automating the identification of hate speech. Firstly, due a conceptual question. As pointed out in the first part of this report, what is and what is not hate speech is the subject of controversy and these conceptual differences can produce results that are not very comparable. Secondly, because of the numerous social networking sites. Thirdly, due to the various types of content (photographs, videos, audio, text, and so on) through which hate speech can manifest itself on social media. And lastly, because even when the medium is textual, there are a multitude of semantic combinations and codes for channelling hate speech, without the need to use insults and expressions which a programmer can anticipate (Ruiz et al. 2010). Strong evidence of the enormous difficulty of this task can be seen in the block errors made by the automatic filtering algorithms used by social media operators themselves, for example, Facebook, Twitter, Tumblr and so on, despite having the most powerful technical and human resources.

Although there are hardly any systematic studies on the links between hate speech and the unleashing of episodes of violence, on a more practical level, monitoring tasks can achieve a lot as early warning systems in unstable contexts and ethnically divided societies to avoid events such as the tragic wave of violence unleashed in the wake of the 2007 elections in Kenya (Gagliardone et al 2015).

On the other hand, in the field of research, there are studies and analyses in the area of social psychology that provide a basis for more effective initiatives against intolerant speech. Today there are plenty of programmes that are designed to combat intolerance

and prejudice, although very few are based on sound evidence for their effectiveness. Bearing in mind that these initiatives are directly or indirectly funded by public authorities, this field of research is of vital importance for ensuring that investment of public money is socially profitable or at least not counter-productive. (Legault et al. 2011)

One of the reference centres in the field of monitoring and research, due to the consistency of its studies, is the Centre for the Analysis of Social Media at the British think tank Demos, which carries out frequent studies not only on the spread of different types of hate speech (Islamophobia, sexism, etc.), but also on how the counter-narratives challenging hate in social media work. In this field of study, Facebook has commissioned Demos to carry out a series of studies on how content which combats hate speech is produced and shared, which is the most successful and which is potentially counter-productive[38].

4.2. Strategies for pressurising social media

Many organisations, aware of the crucial role played by social media operators as intermediaries for extremist and intolerant messages, have opted to put pressure on the companies to get them to take stronger action against hate speech and be more transparent as regards the results of their moderation policies and mechanisms, as well as removing content. These types of strategies can focus pressure directly on social media or indirectly by targeting their advertisers.

Direct pressure strategies have been carried out by both civil society organisations and public institutions. Civil society organisations put pressure on companies by means of online campaigns, collecting complaints, online petitions, etc. For their part, some states and supra-national organisations have pressurised companies to comply with national and international laws and to pledge to moderate content more actively, for fear of more costly sanctions and demands for regulation (Article19 2015). One example of the results achieved by the pressure of European institutions is that, in May 2016, Facebook, Twitter, YouTube and Microsoft signed a Code of Conduct with the European Union, which has already been referred to in this report.

Pressure campaigns on advertisers have, in turn, achieved considerable success. One example of this kind of initiative, carried out jointly in Great Britain and the United States by WAM! (Women, Action and Media Group) and the Everyday Sexism Project, called for the removal of content that was abusive to women. Following an intensive public communication campaign, which included collecting over 225,000 signatures via Chang.org, use of the hashtag #FBrape and publishing a list of companies that advertised in sexist “pages” on Facebook, they succeeded in getting 15 major firms to withdraw their advertising from Facebook, and getting the company to contact the campaign organisers in order to find ways they could cooperate, as well as publicly pledging to review their terms of service and content checking mechanisms (Gagliardone et al. 2015).

38 https://www.demos.co.uk/project/counter-speech-on-facebook-phase-2/

4.3. Strategies designed to change perceptions and attitudes

4.3.1 *Campaigns against online hate speech*

Based on their content and specific orientation, these campaigns can be classified into three types: awareness, affirmative and restrictive (Titley, Keen y Földi 2014). Before examining these types of campaigns and some specific examples of each one, it is worth looking at *Viviendo juntos online: acción y campaña contra el discurso del odio*, the No Hate campaign launched by the Council of Europe (COE) in March 2013, as this is a comprehensive campaign that includes a broad range of strategies against online hate. This campaign occupies a unique position in the fight against online hate speech because, for the last three years, it has served as an umbrella for a large number of initiatives in Europe on a national, regional and local level.

a) The Council of Europe's No Hate campaign

The No Hate campaign is part of the COE's effort to promote human rights on the Internet and targets young Europeans in particular. The idea for the campaign came from the youth representatives on the COE Advisory Youth Council following the mass killing in Utøya, in Norway, in the summer of 2012. This attack, in which 77 young people perished, highlighted the dangers associated with online hate speech which, in some cases, can turn into violence against certain social groups. However, rather than simply advocating the suppression of that kind of speech, this campaign adopts a human rights framework, thereby recognising the right to freedom of expression and choosing instead to encourage self-regulation on the part of users (Keen y Georgescu 2016).

The main goal of this campaign is to mobilise young Europeans by creating a social movement against online hate speech (No Hate Speech Movement). To achieve this, the campaign has set itself four specific goals: first, to reduce the level of acceptance of offline and online hate speech; second, to prevent and counter this problem with education on human rights; third, to raise awareness of the risks propagating hate speech poses for democracy and the well-being of young people; fourth, and last, developing and disseminating tools and mechanisms for reporting it, among them Hate Speech Watch, which allows users to send examples of hate speech they have identified on social media and other online spaces.

The scope of this campaign includes the whole Internet, but social media have a special importance, both as a target of campaign monitoring and as a tool in combating hate speech. In that regard, the campaign has launched activities focused specifically on social media. One of the campaign initiatives that have had most impact have been the Action Days, in which members of the campaign spend a day actively campaigning to raise awareness of a specific issue, for example, anti-Semitism (9 November 2016). More specifically, the Action Days envisage specific actions where campaign activists intervene in social media in four different ways: first, *expressing solidarity* with hate victims by publishing the Action Day image provided by the campaign or attaching the image to user profile photos in social networking sites; second, *reporting hate content* to Hate Speech Watch and adding a counter-argument that is then shared on the sites; third, *sharing content with counter-narratives*, provided by the campaign, in MEME, image, infograph and other formats; fourth, and last, *organising offline activities*, with the aim of raising awareness and educating young people on the Action Day issue.

These are some of the resources provided in the framework of the campaign and which are also developed on social media.

This campaign has been reproduced at a national level by means of the National Campaign Committees. These committees bring together the main interested parties in each country and follow the directives issued by the Council of Europe to maintain the general coherence of the campaign. For the most part, these committees are coordinated by the departments responsible for young people at state level, although in some cases they are only formed by NGOs and other interested parties. So far, more than 40 committees have been set up, both in member states and in other countries, such as the USA, Mexico and Morocco. These committees have launched various national campaigns, which have often echoed the initiatives launched by the COE. In Spain, the Institute of Youth (INJUVE) is the body in charge of launching and carrying out the campaign, in collaboration with the NGO Movimiento contra la Intolerancia[39]. The campaign presentation and launch were held on the International Day for the Elimination of Racial Discrimination, 21 March, 2014.

This campaign has also been carried out at a local level, through a series of activities held in a number of European cities. In addition, Strasbourg has actively collaborated with the COE, which is based there, to launch a local action plan and a campaign support group. Among other events, this plan led to the organisation of No Hate Sounds, a dance event held in Strasbourg on 22 May, 2016[40]. However, up to now there has not been a specific movement or network of cities that explicitly supports the campaign.

b) A type of campaign

Besides the COE's No Hate campaign, other campaigns of a more restricted scope against online hate speech have also been run across Europe. As pointed out at the start of this section, in terms of their content and specific orientation, these campaigns can be categorised into three types: awareness, affirmative or restrictive (Titley, Keen, y Földi 2014: 63).

Firstly, ***awareness campaigns***, which aim to make the general public aware of hate speech and its consequences. One example in Europe is the betterinternetforkids.eu campaign, launched by INSAFE (European Network of Awareness Centres) and INHOPE (Internet Association of Internet Hotlines), two networks active in the area of Internet safety and the prosecution of cyber crime that affects children. Although more focused on Internet safety, this campaign, funded by the European Commission, also created content linked to hate speech, above all through the celebration of Safer Internet Day[41]. One campaign worthy of mention in Spain is *Internet sin riesgos* (Internet with No Risks), launched by the Tenerife Island Council Youth Area with the aim of raising awareness of the propagation of online hate and sexual harassment[42].

---

[39] More information on the campaign in Spain: http://www.injuve.es/convivencia/noticia/campana-contra-la-intolerancia-en-internet

40 For further information: http://www.nohatespeechmovement.org/nohatesounds

41 For further information: https://www.saferinternetday.org/web/sid/resources/gallery

42 For further information: http://www.internetsinriesgos.com/wp-content/uploads/2013/11/folleto-ciberativista.pdf

Secondly, ***affirmative campaigns***, which aim to present minorities to the general public in a positive light in order to prevent discriminatory behaviour. These campaigns usually take the offline space into account as well, and have been launched to change the perception of various groups that are regularly the target of hate speech, for example, LGBT people, Gypsies or Muslims. Firstly, with regard to the LGBT community, it is worth noting the global All Out campaign that asserts the rights of this community by organising micro campaigns before big events such as the Sochi Olympics and actions in defence of LGBT activists, such as the #FreeTheFive campaign, which demanded the release of five Chinese LGBT activists[43]. Secondly, with regard to Roma people, another example is the international Typical Roma campaign, launched in Macedonia, Albania and Bulgaria, among other countries, in 2010, to break down prejudices against people who belong to this community[44]. In Spain, the #YoNoSoyTrapacero campaign also stands out. This was organised in protest at the negative image of Roma people spread by the Real Academia Española (RAE)[45]. Thirdly, with regard to Muslims, one campaign that stands out is the British campaign "Islam is Peace", which is based on providing information about Islam and Muslims that counters the numerous prejudices against this religion and its practitioners[46].

An interesting variation on affirmative campaigns are those based on counter-narratives. Counter-narratives are messages that offer a positive alternative to extremist propaganda and seek to deconstruct and discredit extremist narratives (Silverman et al. 2016: 15). Developing counter-narratives to online hate speech is one of the priorities of the No Hate campaign for 2016-2017. Likewise, campaigns based on counter-narratives to hate speech are being boosted from the EU by means of various funding instruments[47]. In fact, launching these kinds of campaigns and initiatives is included as one of the goals of the Code of Conduct on illegal incitement to hate online, signed by the European Commission and Facebook, Twitter, YouTube and Microsoft. As regards the potential this tool has, the British Institute for Strategic Dialogue has produced a number of studies and awareness-raising material on counter-narratives. One of these is a report, "The Impact of Counter-Narratives", which, following a year-long pilot study, maintains that "*exposure to alternative viewpoints can foster critical thinking or plant a 'seed of doubt' that later matures into a change in attitudes and behaviours* " (Silverman *et al*., 2016: 44). This institute has inspired various initiatives funded by Facebook, such as the "Online Civil Courage Initiative", in which users are encouraged to share personal stories with the aim of combating online extremism and hate speech[48], as well as a Manual and a Toolkit on how to create counter-narratives[49].

---

43 For further information: https://allout.org/es/campanas-destacadas/

44 Various campaign videos available at: https://www.youtube.com/user/typicalroma

45 This and other awareness campaigns

46 For further information: http://www.islamispeace.org.uk/

47 The annual fundamental rights colloquium organised by the European Commission in 2015, "Tolerance and respect: preventing and combating anti-Semitic and anti-Muslim hatred in Europe", stressed the importance of promoting counter-narratives that come from civil society. Further information on the colloquium at http://ec.europa.eu/justice/events/colloquium-fundamental-rights-2015/index_en.htm.

48 Further information at: https://www.facebook.com/onlinecivilcourage

49 For further information: http://www.counternarratives.org/about-us

Thirdly, ***restrictive campaigns***, which aim to gather information on pages and online actions based on intolerant content and taking action to restrict such activity. In this area, a foundation called the International Network Against Cyber Hate (INACH) stands out. It is formed by national organisations that collect complaints and reports of online discrimination, for example, the Movimiento contra la Intolerancia, in Spain. This network, particularly active on anti-Semitism, organised the first International Conference on anti-Semitism Online in Jerusalem, in April 2016[50]. Also in 2016 it published the report "Kick them back into the sea", about online hate speech against refugees[51]. National member organisations of the network have also launched campaigns with the aim of increasing complaints and reports in this area. An example of this is "Get the Trolls Out", an international campaign launched by various organisations with the backing of the COE to restrict online anti-Semitic hate speech[52]. Some interesting resources have been produced within the framework of this campaign, such as the guide "How to Counter Hate Speech on Twitter?"[53] as well as initiatives designed to stigmatise abusers like "Troll of the Month".

4.3.2 *Strategies to combat online hate speech based on education and training*

Another group of strategies for combating online hate is based on education and training. Many of these strategies come under human rights education, which includes education for diversity and interculturality, and are targeted at young people in the main.

So far, various educational initiatives have be launched at European, national and local levels. One worthy of mention in the COE's No Hate campaign is the publication of "Bookmarks", a handbook for combating online hate speech through education on human rights. This proposes a series of educational activities for countering this type of speech (Keen y Georgescu: 2014: 21-133). As well as at a European level, other initiatives have been launched against online hate speech based on education and training at national and local levels, which, in many cases, follow the recommendations in Bookmarks. Without claiming to be exhaustive, these initiatives can be classified into two large groups: on the one hand, those geared towards educating and training the general public, to develop what has been conceptualised as "digital citizenship"; and, on the other hand, initiatives aimed at equipping people who are already aware of the problem and activists with the skills needed to act more effectively.

a) Critical thinking and reflection as key tools against online hate speech

Since the Internet and, later, social media burst into the world of communication, people have gone from being mere recipients to senders of messages as well. It has therefore become clear that education strategies must adapt to this new situation, fostering the critical reception of messages but also empowering the creation of content (Hoechsmann y Poyntz 2012).

---

50 Further information on the conference and its recommendations at http://www.inach.net/detail.html?tx_news_pi1%5Bnews%5D=10&cHash=3bf353f3d766835f68a466d2f8c5b170.

51 Report available at: http://inach.zone35.net/fileadmin/user_upload/Refugee_Report20161.pdf

52 Further information at http://www.getthetrollsout.org/

53 Available at: http://stoppinghate.getthetrollsout.org/

To achieve this, the notion of "digital citizenship" has been put forward, which implies adapting the concept of education for citizenship to incorporate the knowledge and skills required for interacting in a digital medium (Gagliardone et al. 2015: 46). These skills refer to the need for critical thinking that allows received content to be filtered, as well as self-critical reflection before sharing one's own content or that of others.

Some anti-hate speech initiatives have concentrated on working on these questions, mainly with young people, because even though every sector of society generates, disseminates and receives content with hate speech, young people are particularly susceptible to heavy use of it on social media. Indeed, in a presentation in 2012, the director of research on social media at the think tank Demos maintained that a key element in combating hate speech is to equip young people with tools for critical thinking that is adapted to the digital era[54]. In this regard, questioning information that appears online and learning to look for evidence that backs up the opinions expressed is central to protecting young people against "digital populism" and various forms of hate speech.

In line with the concept of digital citizenship, various forums at a global and European level are promoting efforts to boost critical thinking and media literacy. On the one hand, the education programme of the United Nations Educational, Scientific and Cultural Organisation (UNESCO) has, as one of its strategic areas of work, education for global citizenship which, among other things, aims to foster a "sense of belonging to a common humanity, sharing values and responsibilities, empathy, solidarity and respect for differences and diversity"(Gagliardone et al. 2015). The specific goals for achieving a global citizenship include developing the necessary technical skills and abilities for using digital technologies, as well as the knowledge and abilities to search for, analyse, evaluate and interpret texts from communication media, create messages in those media and recognise their social and political influence (Hoechsmann y Poyntz 2012).

On the other hand, EU education ministries attending an informal meeting in March 2015 adopted the "Paris Declaration on promoting citizenship and the common values of freedom, tolerance and non-discrimination through education". This declaration identified a series of actions in the area of education to foster freedom of expression, social inclusion and intercultural dialogue[55]. It also defined one of its objectives as "enhancing critical thinking and media literacy, particularly in the use of the Internet and social media, so as to develop resistance to all forms of discrimination and indoctrination"[56].

As regards specific initiatives based on improving these skills to anticipate and act against online hate speech, various examples could be mentioned. First we should note "Prevention of Online Hate Speech", a project of the European Wergeland Centre in Norway, which specialises in developing education programmes for democratic citizenship. Workshops and training sessions were held under this project to give the

---

54 Carl Miller (2012), "Research perspective", speech delivered at the conference "Tackling hate speech. Living together online" on 26 and 27 November, 2012, in Budapest. Available at http://hub.coe.int/c/document_library/get _file?uuid=8544adcf-f707-4bc1-8484-c6c55311cd1a&groupId=10227.

55 https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/images/1/14/Leaflet_Paris_Declaration.pdf

56 Ibid, p. 3

participants (educators, adolescent students and parents) tools for anticipating hate speech and extremism. These sessions were based around the Bookmarks handbook mentioned previously[57].

Second, another interesting initiative is the Educational Anti-Digital Discrimination Pack, produced as part of the ICUD project[58]. This pack is designed for various groups (children, adolescents, parents, educators, teachers, activists, members of NGOs, political leaders, academics and others) and includes educational tools that came out of workshops held with young people. It is available in six languages, including Catalan and Spanish.

Thirdly, another initiative worthy of mention is the educational project developed as part of the No Hate campaign by the Belgian association Loupiote, "NO HATE - Ce qui vous regarde"[59]. Although more focused on the issue of school bullying via social networks, as part of this project a series of workshops and training sessions were held with young people and educators in the Brussels region which included developing educational tools to combat cyber hate and cyber bullying among young people[60]. The main tool to come out of the project is a short film, produced jointly with a group of adolescents, which shows the consequences of hate speech in social media through the experience of an adolescent who is the target of homophobic comments and bullying from his schoolmates. This short film is accompanied by various short videos on related topics, such as freedom of expression, online anonymity and education in the communication media, as well as a music video on propagating online hate[61].

b) Training cyber activists: some specific strategies for training groups already aware of hate speech

Besides the strategies mentioned in the previous section, initiatives have also been carried out with the specific aim of empowering people already aware of, or victims of, hate speech with specific skills that will enable them to be more effective online. The need to empower the actual victims of online hate speech especially has become more important in recent years. In fact, empowering and supporting online hate speech and hate crime victims is one of the funding priorities of the EU's Rights, Equality and Citizenship programme between 2014 and 2020[62].

Empowering activists is the main goal of the Young People Combating Hate Speech Online project, which is part of the No Hate campaign and intended to give young people and youth associations the necessary skills for recognising and acting against hate speech, as well as involving them in launching the campaign. The purpose of the training is to share experiences and successful strategies, in order to increase the

---

[57] http://eng.theewc.org/Content/What-we-do/Completed-projects/Prevention-of-online-hate-speech

58 Further information on the project and access to the pack at the link: http://digitaldiscrimination.eu/pack/

[59] http://www.culture-enseignement.cfwb.be/index.php?eID=tx_nawsecuredl&u=0&file=fileadmin/sites/cult_ens/upload/cult_ens_super_editor/cult_ens_editor/documents/News/Loupiote_NoHate.pdf&hash=ed1276c37b8dbfaa6864a9ee565d15fbc7a14b26

60 The video is available in French at the following link: https://vimeo.com/111615933

[61] https://loupioteasbl.wordpress.com/ce-qui-vous-regarde-no-hate/

[62] http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52016SC0158&from=EN

impact of bloggers' and activists' activities. Various workshops and seminars on different subjects have been held as part of the project, the last one on how to combat sexist hate speech[63].

Among initiatives following the same philosophy in the Spanish state, it is worth highlighting the workshop "How to Act Against Online Hate Speech" and the course "Human Rights and Hate Speech", which included a specific unit on strategies for combating online hate speech. These were held within the framework of the Online Project Against Xenophobia and Intolerance in Digital Media (PROXI), launched by the Human Rights Institute of Catalonia (IDHC) and United Explanations, and provided training on practical tools for identifying comments that violate human rights and intervening effectively in online media forums[64]. As part of the "Networks Against Hate" (*Redes Contra el Odio*) project, training was given to a group of cyber volunteers consisting of 15 young activists tasked with looking out for and reporting the existence of hate speech on the Internet, as well as identifying situations of cyber bullying against LGBT people[65]. At a local level, the round table "#Prourumors 2.0" on virtual tools for breaking down stereotypes stands out. This was held as part of Barcelona's Anti-Rumours Strategy and held in June 2015 in a city with various tools for combating "virtual" rumours.

CONCLUSION

Although it is complicated to quantify the problem comprehensively, there is a general feeling at an international level that hate speech on social media is worryingly widespread. Hate, anger and aggressiveness have become common currency in social media, causing emotional damage to the targets and contributing to the stigmatisation and dehumanisation of certain groups. Some say it could even trigger episodes of violence in conflict situations and ethnically divided societies. National and international public institutions have spent years trying to find solutions, or at least palliative remedies for a problem that appears to be out of control because of the particular features of the Internet and social media: the superabundance of communication, anonymity, transnationality and so on.

However, although there is a degree of consensus over the extent of the problem, the concept of hate speech is still being discussed and is the subject of debate at an international level. It is a complex concept, as it threatens some cardinal principles of democratic systems, such as equality, human dignity and freedom of expression, which is not conceived in the same way in all socio-political contexts. For this reason it is a concept that can be developed with more or less legitimate aims.

The term "hate speech" is used to refer to expressions that incite discrimination or violence due to racial hatred, xenophobia, sexual orientation and other types of intolerance but also to refer in broader terms to those expressions that foster hostility through prejudice and intolerance. In the media the concept is widely used to refer to a

63 Further information at: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667bcd

64 http://www.observatorioproxi.org/images/pdfs/Programa_taller_Proxi_003_VF.pdf

65 http://www.felgtb.com/redescontraodio/

heterogeneous mix of expressions that include manifestations as diverse, and requiring very different approaches, as mobbing famous personalities, threats to individuals or groups inspired by intolerance, offensive remarks against those in power, advocating terrorism and religious blasphemy.

Either way, the subject matter of this report, without wishing to adopt one definition or another, are those expressions which, in a broad sense, foster and incite violence, discrimination and hostility towards individuals or groups for reasons of race, xenophobia, religion, gender, sexual orientation or disability, and other forms of intolerance.

In order to systematise the different responses and initiatives being undertaken to combat this phenomenon, they have been classified as legal and non-legal strategies. Specific attention has also been paid to the role of social networking site operators, given their regulatory and restrictive capacity as intermediaries in the private sphere.

Legal strategies, which seek to sanction and suppress hate speech, pose a serious dilemma inasmuch as they affect an essential right for democracies, namely freedom of expression. The arguments in favour and against banning certain types of hate speech can be equally convincing. Consequently, even though criminal legislation emanates from the State, it is important to bear in mind the implications of this dilemma as the starting point for establishing a responsible strategy to deal with hate speech.

The last reform of the Criminal Code in 2015 touched on hate crime in attempting to adapt Spanish legislation to European regulations and correct its defects: the lack of proportionality, too much leeway for interpretation, and a failure to adapt to the principal of minimum intervention of criminal law. As usually happens in matters of this kind, there is no consensus between the experts as to whether the objectives have been achieved.

Apart from the doubts raised by resorting to legal strategies, and the conflict this creates with the exercise of freedom of expression, this approach appears to be insufficient for two reasons. Firstly, because harsh prison sentences make it a strategy that is rarely applicable, except in some extreme cases. And secondly, because it creates uncertainty, given the limited case law on this issue and the possibility it might have counter-productive effects.

It is undeniable that hate speech is a controversial field of action for the authorities. For that reason, the role of social network site operators as intermediaries could be crucial. The European Union viewed it that way and has signed a Code of Conduct with Facebook, Twitter, YouTube and Microsoft on illegal incitement to hate on the Internet, to secure a commitment for their involvement in moderating and withdrawing content that violates their terms and condition of use and national legislation. However, such initiatives have been heavily criticised for handing responsibility for what is a sensitive question for freedom of expression over to private companies, whose checking and moderating procedures are not subject to any clear and transparent framework of responsibility.

There is considerable consensus among experts that “non-legal” strategies raise fewer objections. Perhaps the issue that these kinds of strategies, which, to a large extent, are directly or indirectly funded by public institutions, still need to address is evaluating their effectiveness and social return.

A wide range of products and initiatives have been developed. Research and monitoring strategies have been implemented, not only as methods for acquiring knowledge but also as early warning systems in conflict situations. Other strategies have been developed to put pressure on social network site operators, requesting they modify their self-regulation policies or calling on them to withdraw specific content. Above all, numerous social intervention projects have been launched with the aim of changing people's perceptions of, and attitudes towards hate.

Prominent among the latter type of strategy are the campaigns, which include diverse activities designed to attack prejudice and intolerance as the root of the problem of hate speech, and which can be categorised in three types: awareness, affirmative and restrictive (Titley, Keen y Földi 2014). The No Hate campaign, promoted by the Council of Europe, deserves special mention, for the impact it has had and because it is a comprehensive campaign that includes a broad range of strategies, notably the development of counter-narratives. Countering hate speech by means of an alternative speech that discredits it is one of the strategies that public institutions and social media sites themselves have higher expectations for.

Finally, there are some other strategies that also aim to change perceptions and attitudes towards hate that have a special importance: educational and training strategies, whether they are designed to develop what is called "digital citizenship" or to train people already aware of the problem to become activists in the online space. As usual in the field of social intervention, education and training create high expectations because of their potential for combining responsibility with freedom and participation in the best possible way. The challenge, if they are not to remain at the level of good intentions, is to equip them with specific content that is capable of achieving the desired results.

BIBLIOGRAPHY

Aguilar García, Miguel Ángel, ed. 2015. *Manual Práctico Para La Investigación Y Enjuiciamiento de Delitos de Odio y Discriminación*. Barcelona: Generalitat de Catalunya Centre d'Estudis Jurídics i Formació Especialitzada.

Article 19. 2015. *'Hate Speech' Explained A Toolkit*. Article 19. London. Available at https://www.article19.org/data/files/medialibrary/38231/Hate_speech_report-ID-files--final.pdf

Bustos Gisbert, Rafael. 2015. "Libertad de expresión y discurso negacionista", in *Libertad de Expresión y discursos del odio,* Miguel Revenga Sánchez (ed.), *Cuadernos de la Cátedra de Democracia y Derechos Humanos* Nº12. Universidad de Alcalá.

Castells, Manuel. 2012. *Redes de Indignación y Esperanza□: Los Movimientos Sociales En La Era de Internet*. Alianza Editorial.

Gagliardone, Iginio, Alisha Patel, and Matti Pohjonen. 2014. "Mapping and Analysing Hate Speech Online: Opportunities and Challenges for Ethiopia". University of Oxford and Addis Ababa University. http://pcmlp.socleg.ox.ac.uk/sites/pcmlp.socleg.ox.ac.uk/files/Ethiopia%20hate%20speech.pdf

Gagliardone, Iginio, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. Countering

Online Hate Speech. Programme in Comparative Media Law and Policy, University of Oxford.

Gascón Cuenca, Andrés. 2015. "La Nueva Regulación Del Discurso Del Odio En El Ordenamiento Jurídico Español: La Modificación Del Artículo 510 CP. *Cuadernos Electrónicos de Filosofía Del Derecho,* Nº 32.

Heinze, Eric. 2014. Nineteen arguments for hate speech bans - and against them. Available at: http://freespeechdebate.com/en/discuss/nineteen-arguments-for-hate-speech-bans-and-against-them/ (Consulted 23/11/2016)

Hoechsmann, Michael, and Stuart R. Poyntz. 2012. *Media Literacies: A Critical Introduction*. Wiley-Blackwell.

Jubany, Olga, and Malin Roiha. 2016. "Backgrounds, Experiences and Responses to Online Hate Speech: A Comparative Cross-Country Analysis."

Jubany, Olga, Roiha, Malin, and Martínez, Arlette. 2016. "Online Hate Speech in Spain. Fieldwork Report. PRISM. Preventing, Redressing & Inhibiting Hate Speech in New Media". Universitat de Barcelona. Fundamental Rights and Citizenship Programme of the European Union

Kahn, Robert. 2015. Rethinking the Context of Hate Speech, in *First Amendment Law Review*; Univ. of St. Thomas (Minnesota) *Legal Studies Research Paper* Nº 15-20. https://works.bepress.com/robert_kahn/8/

Keen, Ellie, and Mara Georgescu. 2016. *Bookmarks - A Manual For Combating Hate Speech Online Through Human Rights Education*. Strasbourg: Council of Europe. http://site.ebrary.com/lib/uacj/docDetail.action?docID=10961376.

Legault, Lisa, Jennifer N. Gutsell, and Michael Inzlicht. 2011. "Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (but Also Increase) Prejudice". Psychological Science, 22, 1472-1477. DOI: 10.1177/0956797611427918

Spanish Ministry of Interior

- 2015. Protocolo de actuación de las fuerzas y cuerpos de seguridad para los delitos de odio y conductas que vulneran las normas legales sobre discriminación. Available at: http://gestionpolicialdiversidad.org/PDFdocumentos/PROTOCOLO%20ODIO.pdf

-2016. "Informe 2015 Sobre Incidentes Relacionados Con Los Delitos de Odio En España". Available at: http://datos.gob.es/catalogo/informe-2015-sobre-incidentes-relacionados-con-delitos-de-odio-espana.

Moretón Toquero, María Aranzazu. 2012. "El «ciberodio», La Nueva Cara Del Mensaje de Odio: Entre La Cibercriminalidad Y La Libertad de Expresión." *Revista Jurídica de Castilla Y León,* Nº 27.

Morozov, Evgeny. 2012. *El Desengaño de Internet: Los Mitos de La Libertad En La Red*. *Imago Mundi*. Barcelona: Ediciones Destino.

Phillipson, Gavin. 2015. "Hate Speech Laws: What they should and shouldn't try to do", Revue générale du droit (www.revuegeneraldudroit.eu), Etudes et reflexions 2015, Nº 13.

Revenga Sánchez, Miguel. 2015. "Los discursos del odio y la democracia adjetivada: tolerante, intransigente, ¿militante?", in *Libertad de Expresión y discursos del odio (op.cit)*.

Rey Martínez, Fernando. 2015. "Discurso del odio y racismo líquido", in *Libertad de Expresión y discursos del odio (op. cit).*

Rodríguez Izquierdo, Myriam. 2015 “El discurso del odio a través de Internet”, en *Libertad de Expresión y discursos del odio (op. cit).*

Ruiz, Carlos, Pere Masip, Josep Lluís Micó, Javier Díaz-Noci, and David Domingo 2010 “Convesación 2.0 y democracia Análisis de los comentarios de los lectores en la prensa digital catalana”. Comunicación y Sociedad. Vol. XXIII, Núm. 2, 2010, pp.7-39

Silverman, Tanya, Christopher J Stewart, Jonathan Birdwell, and Zahed Amanullah. 2016. “The Impact of Counter-Narratives. Insights from a Year-Long Cross-Platform Pilot Study of Counter-Narrative Curation, Targeting, Evaluation and Impact.”

Teruel Lozano, Germán M. 2015. “La libertad de expresión frente a los delitos de negacionismo y de provocación al odio y a la violencia□: Sombras Sin Luces En La Reforma Del Código Penal.” *Indret: Revista Para El Análisis Del Derecho,* Nº 4.

Titley, Gavan, Ellie Keen, and László Földi. 2014. “Starting Points for Combating Hate Speech Online. Three Studies about Online Hate Speech and Ways to Address It.”

Van Spanje, J., and C. de Vreese. 2015. “The Good, the Bad and the Voter: The Impact of Hate Speech Prosecution of a Politician on Electoral Support for His Party.” *Party Politics* 21 (1). SAGE Publications: 115 -30 256 DOI:10.1177 1354068812472553

Vázquez Alonso, Víctor J. 2015. “Libertad de expresión y religión en la cultura liberal: de la moralidad cristiana al miedo postsecular”, in *Libertad de Expresión y discursos del odio (op. cit.)*.

Vives Antón, Tomás. 2015. "Sobre la apología del terrorismo como ‘discurso’ del odio", en *Libertad de Expresión y discursos del odio (op. cit).*