# Online hate speech

Introduction into motivational causes, effects and regulatory contexts

DTCT
DETECT THEN ACT

Authors: Konrad Rudnicki and Stefan Steiger
Design: Milos Petrov


This project is funded by the Rights, Equality and Citizenship Programme of the European Union (2014-2020)

# TABLE OF CONTENTS

# WHAT IS ONLINE HATE SPEECH?

## Defining a fuzzy concept

A universally accepted definition of hate speech, and its online version: cyberhate, does not exist. Different researchers and institutions use slightly different definitions [1][2], which is not unusual in social sciences, in which concepts are very often challenging to define clearly. However, the multitude of definitions of hate speech used in research have some features in common:

It is a message directed **against** an individual or a group of individuals based on their **identity**.

Based on that message the group is viewed as negative, unwelcome or undesirable which **warrants hostility** towards them.

Some researchers assume that hate speech necessarily has to involve a threat of violence [3][4], but not all scholars agree on that matter [5][6]. Some even propose abandoning attempts at strictly defining hate speech altogether [2].

The European Union in its legislation [7] defines hate speech in part as:

*"public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, colour, descent, religion or belief, or national or ethnic origin"*

This definition will be picked up again, when it comes to the political regulation of hate speech.

Online hate speech (i.e. cyberhate) is a special case of hate speech that occurs in the online environment, making the perpetrators more anonymous, which may make them seem less accountable, and as a result potentially more ruthless [8].

Further reading:
Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and violent behavior,* 40, 108-118.

# Practical definitions

Even though both legal and academic definitions of hate speech exist, the applied aspect of combating hate rarely adheres to them. To effectively fight online hate speech, non-government organizations (NGOs) aim to be more flexible than the justice system or academic systems allow.

In particular, it is increasingly common to define hate speech broadly and include messages that do not explicitly incite violence only, but instead spread prejudice, stereotypes, biases and a general sense of ostracism.

For example, the Anti-Defamation League defines online hate speech as:

*"any use of electronic communications technology to spread anti-Semitic, racist, bigoted, extremist or terrorist messages or information."* [9]

In fact, even in legal systems, the definition of hate speech deviates from being completely strict and encompasses more than merely incitement to violence. The Protocol of the Council of Europe, in its additional protocol on cybercriminality defines online hate as:

*"Racist and xenophobic material means any written material, any image or any other representation of ideas or theories, which advocates, promotes or incites hatred, discrimination or violence, against any individual or group of individuals based on race, colour, descent or national or ethnic origin, as well as religion if used as a pretext for any of these factors."* (Art. 2–1).

Further reading:
Quintel, T., & Ullrich, C. (2019). Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond. *Fundamental Rights Protection Online* (...)

# HATE - AN EMOTION

## A psychological perspective

The definitions of online hate that exist in science and law are notoriously circular and define hate speech as *"inciting hate."* That is not helpful and we have to reach out to psychological research to learn what hatred itself is. Hate speech originates from personally experienced hatred and leads to its own further perpetuation.

In psychology, hatred is a strong, negative emotional state. It is described as an *"aversive"* emotion, which promotes avoidance of the object of hatred, similarly to fear or disgust. [10]

Three elements are thought to comprise hatred:

1. **A negation of intimacy** - creating distance between yourself and the hated object.
2. **Passion** - co-occurrence of strong emotions like anger.
3. **Devaluation** - seeing the hated object as worthless.

Interestingly, the anecdote that there is a fine line between love and hate actually justifies the 'passion' aspect of hate. Love and hatred share a high intensity which is also reflected in activity with similar patterns of brain activity for both. [11]

As every other emotion, hatred serves an evolutionary function and cannot be viewed as purely *"evil."* Just like fear saves people from danger, anger allows them to fight off violence and disgust makes them avoid diseases, hatred has its purpose too.

In particular, hatred of other unknown groups served our ancestors as a defence mechanism of their own kin. The love for one's own people was frequently painted on the canvas of hatred towards the out-group or out-groups. [12]

Further reading:
Sternberg, R. J. (Ed.). (2005). *The psychology of hate.* American Psychological Association.

# Why is it everywhere?

It may appear that online hate, trolling, cyberbullying and other forms of violence that utilize new technologies are something new. However, in reality, their existence rests upon several psychological mechanisms that are as old as humanity itself.

Our mental resources (i.e. attention, memory) are limited and have to be used with caution. Using energy sparingly allowed our ancestors to survive, so they had to learn how to properly prioritize what to remember, what to pay attention to and with whom to cooperate.

Research shows that negative emotions very often have to take priority in our mental systems [140]. For most animals, stimuli evoking negative emotions are threats that have to be dealt with in order to guarantee survival. That is why people are so keen on watching and reading negative news about catastrophes or violence. News of this sort appear to us like they are important for survival. Scientists call this phenomenon "*negativity bias*" [141].

With regard to group conflicts, negativity bias is expressed in our love of gossip [142]. People favour information about the behaviour of others to learn who is trustworthy and who is not. In turn, ostracism and hate are a tool of protecting the group from people who break its norms and values. Unfortunately, merely looking differently can be perceived by our brains as *breaking the norms*.

Because we are more inclined to care about the negative and because we are so protective of our own groups, hateful rhetoric comes easy to us. It takes more effort to express positivity and it is more exciting to observe negativity. From an evolutionary perspective ostracising others is safer than including them.

Further reading:
Waller, J. E. (2004). Our ancestral shadow: Hate and human nature in evolutionary psychology. *Journal of Hate Studies, 3,* 121.

# GLOBALIZATION OF HATE

## World-wide platform of discrimination

The invention of the Internet created a brand new environment in which hateful rhetoric was able to thrive. It serves as a platform that facilitates the emergence of collective identities across the world. On the one hand, it can facilitate the creation of inclusive identities that can span across national, racial or ethnic divisions (e.g. gamers), but on the other hand, it can also consolidate previously fractured movements based around discrimination (e.g. white supremacy, militant jihad).

Because group divisions are the source and driving force of hate, it is important to highlight how the Internet facilitated the emergence of new group identities. Some researchers called them *"imagined communities,"* since their members may never even meet, but perceive themselves as members of the same group. [13]

The Internet created a virtual community fostering a *"global racist subculture."* [14] For example, the Ku Klux Klan experienced a resurgence and renewed interest in memberships thanks to the emergence of the Web. Klanwatch in 1998 wrote: *"Even lone racists, with no coreligionists nearby, feel they are part of a movement."*

In addition to spreading the reach of hate globally, the Internet also facilitated its growth through the so-called *"toxic online disinhibition effect"* [108]. People on the Web are more anonymous, there are huge distances between perpetrators and victims of hate. In these conditions, people give in to some of their darkest drives and motivations.

Because the reach of online communication is vastly larger than the jurisdiction of any country alone, the efforts to limit the spread of online hate have to be present on an international level. The newly emerged group identities fostering hate extend far beyond previously existing national prides.

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

# VICTIMS OF HATE

## Gender

The root sources of online hate speech are group identities, and even the broadest ones, like gender, can spark victimization.

Current research clearly shows that women are the primary victims of online hate speech based around gender [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. Online attacks against women may take several different forms:

1. **Misogyny** - devaluation of women based on their gender. [27]
2. **Sexual harassment.** [21]
3. **E-bile** - taking women's voices away [28, 29]
4. **Gendertrolling** - non-sexual harassment and stalking of women online [30]

A detailed legal typology of different types of violence against women online has been developed by Halder & Karupannan in 2009 [31]. The authors also discuss the reasons for an increasing amount of hate towards women online. In particular, they highlight that sexual harassment is exceptionally dangerous when social websites provide easy access to the victims' personal information.

Compared to men, women are more often harassed on the grounds of their physical attractiveness. In contrast, when men are the targets of hate, their social status, talent or achievement are attacked the most. [32]

It is critical to note that on top of being victims of online hate speech more often, women are also more emotionally susceptible to its negative effects. This calls for special care regarding the resilience of the victims [32]

Further reading:
Gender Equality Unit (2016). Background note on sexist hate speech. Council of Europe, Strasbourg.

Almost all online environments are riddled with gender-based hate speech.

Female journalists are regularly attacked and threatened based on their gender [17]. Dating apps, such as Tinder, experience a problem with female users being harassed [24]. In gaming communities women are very often dubbed as the enemy and systematically ostracised for their gender [22. 33, 34]. This is especially evident considering that female video game players are widely labelled *"girl gamers,"* to highlight the apparently unorthodox character of the phenomenon that a woman is playing [35].

The central point of hate speech addressed to women is the attempt at silencing them [16]. Researchers identified three strategies that haters use in order to silence their victims:

1. **Intimidating** - threats of physical violence (e.g. death and rape)
2. **Shaming** - attempts at changing the opinions of the bystanders about the victims. Spreading rumours, private information or unauthorized pornography of the victim.
3. **Discrediting** - attacking the reliability of the victim and the validity of their opinions.

Gendered hate speech can have dire consequences for individuals, groups and whole societies. Ostracised women suffer psychological stress but also distance themselves from the communities where they experienced hate [36]. In the worst scenario, hate may hinder women's freedom of expression online and offline.

Further reading:
Hardaker, C., & McGlashan, M. (2016). "Real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics, 91*, 80-93.

# Race/Ethnicity/Religion

**No other group divisions have larger potential for generating hate than ones that are sanctioned by institutions.**

## Muslims

Muslims are very often the targets of hate speech [16, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48]. They are typically targeted for their alleged faith-based actions and/or beliefs that are grounded in Islam [44]. The scope of the problem was highlighted by researchers who collected tweets under the hashtags: #Islam, #Muslim and #Woolwich and found that 75% of the messages expressed strong islamophobic opinions [37].

Hatred towards the Muslim community is a perfect example of the way in which online hate permeates into the real world. Among the victims of islamophobia it is hard to delineate the online threats from the intimidation, violence and abuse they experience in the real world. Instead, the hatred occurs online and offline at the same time, creating an environment entirely hostile to Muslim people [38].

Researchers have decided to explore the motivations of politicians who used anti-Muslim rhetoric [48]. They identified four main explanations that politicians had for such behaviour:

1. Presenting themselves as *only human*, justifying their racism as trivial accidents that *could happen to anyone*.
2. Painting themselves as *voices of reason*, implying that islamophobic views are founded in facts about Islam.
3. Posing as the *victims* and arguing that they are being persecuted for their opinions.
4. Presenting themselves as heroes who have the moral high-ground and defend their nation.

Other researchers have also examined the motivations behind islamophobic hate speech of regular users on Twitter [37]. They found eight distinct reasons or strategies by which people engage in such hate:

1. **Travellers** - people who visit Twitter specifically to harass Muslim individuals.
2. **Apprentices** - new users who are guided by more experienced haters.
3. **Disseminators** - people who share and spread islamophobic memes and imagery.
4. **Impersonators** - fake accounts made specifically to spread hate.
5. **Accessories** - people who join other people's conversations to target vulnerable individuals there.
6. **Reactives** - people who start spreading online hate after a large, violent incident (e.g. a terrorist attack).
7. **Movers** - people who get banned often and have to move from account to account to continue targeting their victims.
8. **Professionals** - people with large amounts of followers who start massive campaigns against Muslims.

Similar classifications based on online posts were also made in other research papers [38,42]. In particular, islamophobia is most often motivated by: anti-immigrant sentiments, internalized racist worldviews and the support of geopolitical hegemony of the United States of America [43].

It is important to highlight that very often, Islamophobic groups hold a belief that the mainstream media are corrupt and try to cover up *actual truths*, which in their opinion is that Islam is a threat that has to be dealt with [47].

Further reading:
Awan, I. (2014). Islamophobia and Twitter: A typology of online hate against Muslims on social media. *Policy & Internet, 6*(2), 133-150.

## Black people

Even though the concept of the human race has been deemed meaningless in physical anthropology [49], humans still behave as if races objectively existed. As a result, social sciences still have to be concerned with the concept of races, because humans continue to perceive visual differences between people as indicators of racial belonging.

People who are perceived to be of African descent (i.e. black) are often the targets of online hate speech [34, 50, 51]. One in three African-American adolescents experienced online hate speech based on their race [50]. In the United States, casual racism experienced by the victims often takes form of jokes related to the history of slavery in the Americas or the racial stereotypes [51].

Jokes that perpetuate racial stereotypes but may also ignite racial hate are usually spread on the internet in the form of memes. One of the most popular ones is simple and consists of merely copying the following phrase: "Despite making up only 13% of the population, blacks make up 52% of crimes." Misconstrued federal statistics are being used to imply that black people as a race are inherently more violent.

The significance of these jokes is not to be underestimated. Research shows that black people exposed to the stress of being hated online are more likely to interpret other vague social situations as racist [52].

Fortunately, other researchers have also found that elevating self-esteem and ethnic identity of black people helps to reduce the anxiety experienced as a result of racially motivated hate [50].

Further reading:
Weaver, S. (2010). Developing a rhetorical analysis of racist humour: Examining anti-black jokes on the internet. *Social Semiotics, 20*(5), 537-555.

## Latino people

Latino people are also victims of online hate speech, especially in the United States, where they experience similar resentment to African immigrants in Europe [53, 54, 55].

The terms that are often used to refer to Latino people include: *"cockroaches," "scumbags"* or *"brown invaders"* [55].

To examine the root causes of anti-Latino racism, researchers analysed user comments regarding a murder trial in which a Latino man was killed by five White assailants [54]. The analysis yielded two main mechanisms present in the comments of the haters:

**Dehumanization** - the victim of murder was described as not being truly human.

**Generalization** - the comments about the victim often addressed all of the 12 million illegal immigrants in the US.

The fact that the murdered victim was an illegal immigrant was used as justification for dehumanizing them and deeming their life worthless.

The fear of Latino illegal immigrants is so prevalent in cyberspace, that researchers have identified something called *Latino cyber-moral panic* [53]. Racist narratives and phenomena such as dehumanization and generalization affect Internet users and result in endorsements of *systematic racism*, which are institutionalized expressions of racial discrimination.

Further reading:
Loke, J. (2013). Readers' debate a local murder trial: "Race" in the online public sphere. *Communication, Culture & Critique, 6*(1), 179-200.

# Sexuality

**Hatred towards people can be based on their sexuality. In particular, any expression of sexuality that is non-heterosexual comes under scrutiny.**

LGBTQ students are harassed significantly more often than their heterosexual peers [56]. The exact number of victimized LGBTQ young people varies between 10.5% and even 71.3% [57].

Unfortunately, the victimization of LGBTQ people is already known to have tangible, adverse consequences for their health and social functioning. LGBTQ students are shown to have lower school outcomes, lower self-esteem and more suicide attempts as a result of harassment [58].

The case of LGBTQ people is a perfect example of how online hate speech cannot be considered in disconnection from the real world. Researchers have found that being harassed online as a LGBTQ person is three times more likely if you live in the southern US than in the northern US and two times more likely if you live in a rural area, compared to an urban area [59]. As a result, the location of a person in the real world has significant implications for their experiences online.

Online violence against LGBTQ people may stem from institutionalized belief systems. In particular, LGBTQ people are often presented as enemies of the nation and the society. This view is perpetuated by religious systems that classify non-heterosexuality as a sin which means that any political system that adheres to religious dogma will also struggle with combating anti-LGBTQ sentiments [60].

Further reading:
Meyer, D. (2015). Violence against queer people: Race, class, gender, and the persistence of anti-LGBT discrimination. Rutgers University Press.

# HATE FROM THE PERSPECTIVE OF THE VICTIMS

## Consequences

The first step to understanding the effects of being hated online can have on people is to ask the question: **why are people even affected by hate in the first place?**

Humans are social animals, and for any social animal, being ostracised from their group is the first step toward starvation and early death [61]. Ostracised individuals lose access to common resources and lose protection of the group. As a result, social animals developed mental systems that prevent them from engaging in behaviours that could get them thrown out of the group. Researchers call such mechanisms: *ostracism detection systems* [62].

To avoid being excluded, humans experience extreme distress at the prospect of being ostracised. Ostracism causes an emotion called *social pain* and scientists found that feeling *social pain* activates the same areas of the brain as physical pain [63]. In other words, it feels almost as if we are physically hurt, when someone ostracises us.

One of the most basic human needs is the so-called *need to belong* [64]. Frustrating that need is perceived by humans as a threat to their very existence [65] and to prevent becoming alone, humans developed negative emotions whenever someone signals that they should be excluded.

Because online hate speech is based around group belonging and very often takes the form of ostracism (e.g. *"not real Americans," "go back where you came from"*) it may activate the systems that warn us of incoming group exclusion.

Our brains evolved for thousands of years without the online environment. We are not able to easily discern if someone who messages us with hate is an actual member of our group or not. As a result, we react with social pain, stress and anxiety even if the hater is someone whom we will never meet in person.

Further reading:
Spoor, J., & Williams, K. D. (2007). The evolution of an ostracism detection system. *In Evolution and the social mind: evolutionary psychology and social cognition,* 279-292.

In line with the predictions made by evolutionary psychological science, being exposed to online hate speech causes increased stress levels [66]. However, increased stress of the victims has more far-reaching and dire consequences as well.

Women of colour who experienced online hate on the Xbox Live platform have had to distance themselves from the bigger gaming community and created their own smaller groups where they could play videogames with other women [34].

Muslim people who have been victimized by online hate speech reported increases in insecurity, fear and vulnerability [5]. That in turn led them to a decrease in the sense of belonging and the willingness to integrate into society. The fear of online hate becoming real leads Muslim people to hide their identity offline (for example, women take their headscarves off).

Most importantly, researchers have found that frequent exposure to online hate speech desensitizes bystanders to it, which decreases sympathy felt for the victims and fuels prejudice against them [67].

Hated groups become *dehumanized* - they become ostracised not only from some social groups but from humanity as a whole. Gradually, violent behaviour against these groups becomes more and more acceptable to the bystanders of hatred, as the idea that the hated out-groups do not deserve the same moral standing becomes less and less controversial [68].

Further reading:
Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior, 44*(2), 136-146.

# Moderators

**A *moderator* is a phenomenon that affects a relationship between two other concepts. For example, the relation between being subjected to hate and becoming stressed may be moderated by the resilience of the victim. More resilient individuals will become less stressed when harassed as compared to those who are less resilient [69].**

In addition to personal resistance to stress, other personality factors can also have important implications for the consequences of being victimized. For example, having high ethnic identity and self-esteem can help protect against anxiety experienced as a result of racist hate speech [50].

Building group belonging and social cohesion within the victimized groups helps them defend against external violence in the form of hate speech. However, even though group belonging may prevent from anxiety and loss of self-esteem, researchers have found that it cannot protect from depressive symptoms [50].

Another important moderator of the effects of online hate speech are the sources of hateful messages. In general, the more sources of hate and the more diverse they are, the more pronounced the stress of the victims [52].

This is not surprising, given that humans generally assess the reliability of information based on their source. The more diverse are the sources repeating the same piece of information, the more likely we are to believe it [70]. Unfortunately, this principle applies to hateful rhetoric as well.

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

# HATE FROM THE PERSPECTIVE OF THE PERPETRATORS

## Motivations

**The reasons for engaging in hate speech may be vastly different for different people. A politician sowing divisive rhetoric has different goals to a middle-class citizen who harasses others on the Internet.**

From the personality psychology point of view, different people may differ in their innate tendencies to seek social status and power over others [71]. Those who express these tendencies more seek any reason to elevate themselves and their group over others. As a result, they subscribe to stereotypes and prejudice in order to justify their view that their *ingroups* are superior to the *outgroups*. After all, beliefs that we hold, need some justification. Researchers have developed tools to measure, such as personality traits through *motivation to express prejudice* [72]. However, focusing solely on people's personalities ignores the importance of other environmental or contextual factors.

A natural guess for the roots of hate is family upbringing. Surprisingly, research so far has shown that on average, prejudiced beliefs by parents do not predict prejudice in children very strongly [73].

Social psychology points to the importance of group norms and values in explaining the emergence of hate. People perceive out-groups as potential threats that import foreign norms and values that could disrupt the local order. The motivation to protect *old ways* is directly tied to the desire to keep your own group cohesive and intact [74, 75].

Further reading:
Walters, M., Brown, R., & Wiedlitzka, S. (2016). Causes and motivations of hate crime. *Equality and Human Rights Commission research report,* 102.

Apart from the *hidden* motivations rooted in evolution, social dynamics or personality, hate has explicit, conscious motives to which perpetrators may subscribe.

Based on police reports of hate crimes, researchers have identified four types of perpetrators, differing in their motivations [76]:

1. **Thrill seekers** - people who declare that they commit hate crimes mainly for entertainment. Victimising out-groups is seen as exciting and serves as a method of group-bonding, especially among young males [77].

2. **Defensive** - people who see the victimised groups as "invaders," threats that have to be removed. They believe themselves to be defending their groups or territory from the hated group.

3. **Retaliatory** - these are people who take the defensive motivation one step further and perceive themselves to be under an immediate attack. This strategy is especially prevalent after terrorist attacks or other events that draw attention towards group conflicts. For example, passing a new legislation that allows same-sex marriage is then seen as an attack on the norms and values of some groups.

4. **Mission** - the most devoted form of hater who organises his or her life around the fight against a victimised group. *The missionaries* are driven by ideologies that compel them to ostracise others and are the most likely to commit acts of violence.

Further reading:
Walters, M., Brown, R., & Wiedlitzka, S. (2016). Causes and motivations of hate crime. *Equality and Human Rights Commission research report,* 102.

# Moderators

**What regulates how hate is going to be expressed? Researchers have reviewed 31 different studies on racist online hate speech and have found significant differences between the ways in which individuals and groups engage in hate [78].**

With regard to communication channels, individuals may express hate virtually anywhere: social media, private or public profiles, comment sections. An individual is more likely to engage in hate speech spontaneously. In contrast, group-based hate speech is mostly concentrated in the private environments of hateful groups where they develop creative ways of expressing their views (e.g. memes, videos, games).

Individual haters engage mostly in conversational rhetoric - flipping the issue, re-interpreting news in a hateful way, ad hominem attacks. Group efforts in hate are much more organised. Groups recruit members, develop hateful imagery, create new narratives and disseminate them outside of the original environment of the group.

The most important difference between individuals and groups in online hate speech lies in their goals.

Individuals seek to validate themselves - obtain approval of their in-groups, enhance their self-esteem by basking in the glory of their supposedly superior group, etc. Groups aim to gain power in society and enforce their normative systems onto others.

Further reading:
Bliuc, A. M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018).
Online networks of racial hate: A systematic review of 10 years of research on cyber-racism.
*Computers in Human Behavior, 87*, 75-86.

# HATE FROM THE PERSPECTIVE OF THE BYSTANDERS

## Perceptions

**Bystanders who witness hate incidents represent the biggest untapped potential in combating online hate speech.**

Surveys among adolescents showed that even up to 80% of them have witnessed offline or online incidents of harassment [79]. Unfortunately, research in that same age group showed that bystanders are significantly less likely to intervene if the incident takes place online [80].

Studies on bystanders' intentions to intervene revealed mixed findings. One study found that students do not want to intervene in incidents of online harassment, but would rather provide support to the victim [81]. In contrast, in another study participants declared that they would intervene if they were bystanders of online victimization [82]. All in all, the most popular response to online hate is passive behaviour (i.e. doing nothing), even though in questionnaires people often declare willingness to support the victim in hypothetical situations.

Unfortunately, only supporting the victim does not decrease the distress that they experience [79].

The online environment entails much higher perceived distance between the perpetrators and the victims of hate speech. It also provides more anonymity for the perpetrators and bystanders alike. Researchers have shown that because of properties of the Internet, inert bystanders of hate may become disinhibited and transform into brand new perpetrators of hate [83]. The ultimate goal of interventions targeted at bystanders is to reverse that phenomenon.

Further reading:
Wachs, S., & Wright, M. F. (2018). Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. *International journal of environmental research and public health, 15*(9), 2030.

# Moderators

**Motivating bystanders to intervene is essential in combating online hate. Therefore, it is crucial to identify factors that make it more likely that bystanders will intervene. Two main types of these factors can be seen in scientific literature: contextual factors and internal factors. A substantial amount of literature concerning bystander behaviour comes from research on cyberbullying. This is important, because hate speech rarely occurs between people who know each other in real life, whereas that is the case in cyberbullying.**

## Internal factors

Having prior personal experience with being victimized makes it more likely that a bystander will intervene in an incident of harassment [84, 85]. Scientists believe that this happens because experiencing victimization themselves helps bystanders in empathizing with other victims [85].

A number of personality traits of bystanders were shown to predict their intervention. People who have higher self control [81], higher self-esteem [84] and confidence in their skills [82] are more likely to intervene. Teachers who score higher on affective empathy were also found to be more engaged in several steps in their interventions against harassment [86].

Studies have also found that positive peer norms [84] and the existence of social bonds [85] predict bystander interventions, which is in line with the fact that bystanders are much more likely to act if the victim is their friend [80, 87, 88].

Most importantly, observing active bystander behaviour of peers has been shown to make it more likely that bystanders will step in to incidents of harassment [85].

Further reading:
Brody, N., & Vangelisti, A. L. (2016). Bystander intervention in cyberbullying. *Communication Monographs, 83*(1), 94-119.

## Contextual factors

The characteristics of a situation that bystanders find themselves in have an important impact on their potential interventions.

Scientists studying social behaviour in the context of harassment coined a concept called the *bystander effect*. The name of this effect alone tells the harsh truth about natural human willingness to intervene in incidents of hate. The effect itself states that the more bystanders are present, the less likely each individual is to act. This effect was observed also in the specific context of online hate [91].

However, not surprisingly, the more a bystander finds the situation disturbing, the more likely they are to intervene [85, 91]. This means that bystanders are more likely to act when hatred is expressed more aggressively [89]. Research among adolescents showed that they are more likely to defend the victim when harassment takes place online, rather than offline [90].

In line with the disturbing phenomenon of *victim blaming*, bystanders pay attention to the levels of disclosure that victims express on social media. Victims who post more personal information are blamed more for being harassed and receive less empathy from the bystanders. This blaming and lower empathy translate into lower likelihood of bystanders helping the victims [80].

As always, group norms play a key role in determining the behaviour of the group members. In some environments, bullying others is perceived as something positive. In such social environments, bystanders are likely to step in by joining in with the bully and harassing the victim together [90].

Further reading:
Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018).
Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook.
*SCM Studies in Communication and Media, 7*(4), 555-579.

# COUNTER-COMMUNICATION

## Definition

**Counter-communication is an umbrella term that covers several closely connected concepts: counter-spaces, counter-messages and counter-narratives.**

Counter-spaces are spaces where people cooperate in creating and publishing stories as well as in building personal resilience [92]. Any reply intended to oppose an incident of hate speech can be considered a counter-message or counter-speech. These messages can also take the form of counter-narratives, which are specifically designed to shift the public discourse on the matter by challenging some existing hateful rhetoric and debunking false information.

Think tanks and NGOs often organize counter-spaces and create counter-narratives [93].

People participating in counter-communication typically engage in several different types of messaging [94]:

1. **Exposing** the racist characteristics of the messages
2. **Ridiculing** these messages
3. **Debunking** misinformation
4. ***Sharing information*** about online and offline resistance to hate

For example, after the terrorist attacks in Brussels on March 2016, the hashtag #StopIslam was used to spread hateful messages directed at the Muslim community. An effort in creating a counter-narrative soon resulted in most of the tweets with the hashtag being included in supportive instead of ostracizing messages. [95].

Further reading:
Case, A. D., & Hunter, C. D. (2012). Counterspaces: A unit of analysis for understanding the role of settings in marginalized individuals' adaptive responses to oppression. *American Journal of Community Psychology, 50*(1-2), 257-270.

## Effectiveness

There are not many studies that addressed the effectiveness of counter-communication. There are even fewer studies that compiled datasets of counter-communication that could be used in future campaigns and research. Two datasets of counter-communication were created recently. One was created in collaboration with several NGOs and contains counter-narrative pairs in English, French and Italian [96]. The other consists of counter-messages, including 13,924 YouTube comments [97].

It is thought that counter-spaces provide victimized people with safety, solidarity, hope and healing [98]. The very aim of their existence is to improve the well-being of the victims of hate. However, there is almost no empirical research on the overall effectiveness of counter-communication [99].

In one study the researchers examined the reach and impact of three counter-narrative campaigns [100]. These campaigns had vast reach, totalling 378,000 video views and over 20,000 user engagements (i.e. likes, shares, replies). The engagements with the campaign were divided into two types:

1. **Sustained constructive engagement** - comments that constituted positive conversations about the content of the campaign.

2. **Sustained antagonistic engagement** - dismissive comments contesting the content of the campaign.

Reaching a lot of views does not necessarily mean that a campaign was successful. Messages that are universally disliked also reach wide audiences. Thus, it is important to differentiate positive and negative engagements with the content of the campaign.

Further reading:
Silverman, T., Stewart, C. J., Birdwell, J., & Amanullah, Z. (2016). The impact of counter-narratives. *Institute for Strategic Dialogue,* 1-54.

In a study examining the effectiveness of counter-speech in the comment section of You Tube, the researchers found that counter-messages received more likes than other comments [97]. They also found that the effectiveness of counter-speech was dependent on the category of the videos and the community surrounding them. The authors of this study provide detailed analysis of the preferences for specific types of counter-messages among LGBT, Jewish and Black communities. For example, in the LGBT community, humorous counter-messages are the most favoured, whereas warnings of consequences for the hate perpetrators are the least liked. In contrast, the Jewish community likes the warnings of consequences for the haters [97].

In a highly rigorous study, one researcher decided to check if being sanctioned (publicly shunned for racist comments) has a long-term effect on the behaviour of online haters [101]. It turned out, that being sanctioned by a high-status in-group member significantly decreases the use of racial slurs online. In particular, the author used bots which appeared to be accounts of high-status (i.e. large number of followers) white males. Being publicly messaged by a bot like that successfully altered the behaviour of the haters [101].

In response to the lack of empirical studies, other authors assessed the efficacy of counter-communication with a simple simulation model [99]. They made two important conclusions. First, counter-communication can significantly influence a given public space, although there needs to be listeners in order for it to work. Second, even a small group of counter speakers can exert an impact on the public that is a lot bigger, if there is a crucial number of listeners that do not hold extreme views (i.e. rather undecided individuals) [99].

Further reading:
Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior, 39*(3), 629-649.

# Challenges

Although promising, the paradigm of counter-communication has issues that have to be addressed.

Some researchers point out that counter-activism exposes its practitioners to harassment and hate. Messages have to be sent out by someone and that may put a target on the back of their authors [102]. Encouraging random bystanders to participate in counter-communication poses an even more challenging ethical conundrum, since they do not have access to the support system of the counter-spaces that specialize in caring for the well-being of their members.

Other authors identified three main criticisms of the counter-communication paradigm [103]:

1. A perceived lack of strategic *effectiveness* - which means overestimation of the importance of propaganda, lack of evidence for the efficacy of campaigns, required scale of the task (there is much more hateful content than counter-messaging).

2. *Normative* elements of government involvement. To what extent are interventions reasonable? Counter-communication is a form that is placed awkwardly between valid counterterrorism and publicly intolerable ideological engineering.

3. *Capability* of counter-communication: if we do not know which campaigns work and which do not, are we even able to produce messages that resonate with the public?

Further reading:
Lee, B. J. (2019). Informal Countermessaging:
The Potential and Perils of Informal Online Countermessaging.
*Studies in Conflict & Terrorism, 42*(1-2), 161-177.

# Recommendations

Because there is not enough empirical data on the effectiveness of counter-communication, any recommendations that existing literature gives are based on the expertise of the authors, NGOs and theoretical assumptions.

In general, researchers and NGOs advocate in favour of counter-communication. The Anti-Defamation League published guidelines on how to counter online hate speech [104]. The guide provides recommendations for providers as well as Internet users. For instance, it advises Internet users to: *"identify, implement and/or encourage effective strategies of counter-speech – including direct response; comedy and satire when appropriate; or simply setting the record straight."*

Some authors advise the use of arts and arts education in counter-communication [105]. Artistic freedom can enable a creative technique to navigate between freedom of expression and tackling hate speech. However, no study so far has shown that using arts education in countering hate speech is effective.

Because of the ethical problems with susceptibility of authors of counter-messages, researchers highlight that relying on Internet users for monitoring of racist content and spreading counter-narratives is not a valid response to the growing problem of online hate speech [102].

Further reading:
Anti-Defamation League. (2014).
Best Practices for Responding to Cyberhate.

Based on the analysis of content and network of populist right-wing Facebook pages and counter-speech pages in the UK, France and Italy, researchers made the following recommendations [106]:

- ***Administrators:*** post more visuals, and, with that, focus on content that has a wider reach than merely the network of people that like your page.

- ***Commenters:*** post more ''constructive'' counter-speech, instead of attacking the haters and publish more about particular policy issues.

- ***Contributors*** to counter-speech campaigns should encourage their own social networks to share counter-communication with their friends and friends of their friends, etc.

- In general, if counter-speech administrators and members are more active, and altered their messages a little bit, it would effectively enhance the reach of their content.

One of the most important recommendations that is present in the literature concerns the collaboration of different actors engaged in counter-communication. Researchers propose collaboration between formal actors (the government and NGOs) and informal actors (Web users). Alignment of messaging of formal and informal actors boosts the credibility of both [103]. Independent users posting counter-messages appear to the audiences as more authentic, while government involvement lends gravity to the points being made. Researchers studying Islamist extremism point out that the sender of the counter-message is crucial as they have to be familiar with Islam and the Muslim religion [107].

Further reading:
Bartlett, J., & Krasodomski-Jones, A. (2015). Counter speech. Examining content that challenges extremism online. *Demos*.

# CONTENT MODERATION

## Definition

Online hate speech is predominantly posted on websites that host large communities. Such websites most often have general regulations, codes of conduct for the users and moderation policies for the community managers or administrators. As a result, the ICT industry has the power to manage online content through removing it or suspending the users [93]. However, because every site has its own rules, there is no standard definition of hate speech across different platforms. Not only that, but even within one platform, the policies of moderating content can be unclear, inconsistent and ill-defined [109].

Content moderation can happen at three stages: before a post gets published; after it gets published, but before many other users can engage with it; or after it gets published and many people engaged with it, possibly reporting it to the website moderation for violating the rules [110]. After that, the power lies on the side of the moderators, be it computer algorithms or humans. Researchers have identified four types of human moderators' strategies of dealing with hate online [111]:

1. **Unconcerned gatekeepers** believe that hate does not occur a lot and is relatively unproblematic. They deem strategic manipulation attempts to be far-fetched. They mostly utilize an authoritative, non-interactive moderation process.

2. **Relaxed gate-watchers** perceive the amount of hateful comments to be high, but consider it as a standard in human interaction. They use a variety of interactive moderation processes.

3. **Alarmed guards** believe that harassment has a small but harmful impact. They use non-interactive, hierarchical moderation processes.

4. **Struggling fighters** perceive that hate occurs a lot and that it has a big impact. They use non-interactive, hierarchal moderation processes.

Further reading:
Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. *Media Studies Publications, 12.*

## Effectiveness

The effects of content moderation can be studied widely, as almost every online platform implements some form of it in its policies.

In 2015, subreddits r/fatpeoplehate and r/CoonTown were banned because they violated Reddit's anti-harassment guidelines [112]. The effectiveness of this ban was questionable. On the one hand, some researchers pointed out that the hate speech expressed by the remaining users decreased considerably. In addition, the affected members did not use hate speech in their newly found subreddits, nor did they impact the language of the members that were already there [112]. On the other hand, another study showed that during this period of unrest, banned communities on Reddit migrated to another platform called Voat, where they continued their hateful conversations [113].

Scepticism regarding the effectiveness of content moderation was also supported by the researchers who examined the utility of proactive moderation strategies on Reddit and concluded that the present tactics that are used for community-level interventions (i.e. bans and quarantines) do not influence user civility [114].

The same researchers uncovered that another factor, unrelated to content moderation predicted if people engage in hate speech. They found that structural properties of the community (i.e. which other communities the members are connected to) are the primary determinants of future community behaviour. In other words, participating in hateful subreddits significantly deteriorates member civility, even in other circumstances [114].

Further reading:
Habib, H., Musa, M. B., Zaffar, F., & Nithyanand, R. (2019). To Act or React: Investigating Proactive Strategies For Online Community Moderation.

Online content is moderated based on regulations that websites create for themselves. The way these regulations are communicated to the users can have an impact on their effectiveness.

Researchers tested if openly announcing the rules on a subreddit would influence peoples' behaviour [115]. This normative information about a community's guidelines led to an increase in norm compliance among new members of the community as well as an increase in participating in discussions among newcomers.

Openness about the rules is important, because social media sites do not explain which forbidden activities are related to which consequences [116]. In addition, the language in which policies adopted by platforms are written is complex and difficult to understand by the average social media user. Social media users whose account or content has been removed are frustrated and confused about the moderation process, which causes them to develop their own folk theories about how platform moderation works [117].

To assess how the degree of moderation affects the presence of hate, other scientists analysed discussions on two subreddits, one defined as a ''safe space'' (i.e. highly moderated) and another as the alternative, less moderated ''free speech'' space. These different moderation policies affected norms regarding style, affect and topic. In general, the safe space was characterized by positive words and discussions about leisure, whereas the language in the free speech space was characterized by negative and angry words, and topics regarding work, money and death [118].

Further reading:
Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences, 116*(20), 9785-9789.

# Challenges

NGOs are critical about the way internet service providers manage reports and deem their self-regulatory measures such as filtering and rating systems weak [93].

Some researchers argue that content moderation can never be unbiased, impartial, or non-discriminatory because websites are incentivized to segment users and treat them differently based on their value to the site [119]. They argue that content is assessed in terms of the amount of advertising revenue it will attract for the platform, above anything else [120]. As a result, policies of websites with regard to hate speech as well as their practices are guided by the corporate view of their mission [121].

**The biggest challenges for the people who are moderating content online are [110]:**

Not every hateful comment is simple to notice, understand and judge. People who moderate content have to be familiar with obscure language, memes and context of the comment.

**Moderators themselves are often victims of hate for their involvement.**

Moderators are often forced to balance the moral obligation to remove content that violates the rules, with the economic interests of the company interested in gaining attention, publicity and engagement from users.

Further reading:
de Zwart, M. (2018). Keeping the neighbourhood safe:
How does social media moderation control what we see (and think)?
*Alternative Law Journal, 43*(4), 283-288.

**Many websites and researchers attempted to tackle the problem of content moderation by increasing its effectiveness through automatic detection of hateful messages. However, there are several limitations to that approach:**

1. The variety of languages. English language is overrepresented in research on the issue compared to its share of internet content [122, 123].
2. Overrepresentation of some social media in the research (e.g. Twitter) [124].
3. Existing systems do not take into account the cultural background of the posters, social structures, context of messages and their implicit content [125].
4. Humour, irony and sarcasm are very often false positives in the hate detection systems [126].
5. Detection systems are vulnerable to polysemy (i.e. words with multiple meanings) [126].
6. Language evolves rapidly, especially among young people that use social media [125].
7. There is low consensus in hate speech categorization among humans, making it even more difficult for machines [127].
8. Automated detection can exaggerate social bias manifested in language [122].
9. Most studies provide little explanation on how messages have been annotated during categorization, making it difficult to determine intercoder reliability [122, 126].

Further reading:
Duarte, N., Llanso, E., & Loup, A. (2018). Mixed Messages? The Limits of Automated Social Media Content Analysis. In *FAT,* 106.

# Recommendations

With regard to human-based content moderation, researchers recommend a transparent moderation process that gives more detail and clarity about the process, educates its users, gives specific explanations about the removed posts or accounts; which content was in violation of a policy; how the content was recognized; who was accountable for removal and why the decision was ultimately made that a policy had been violated [121, 128, 129].

Additionally, researchers point out that including the users of the media in the process of creating policies is necessary [109]. A combination of activism and an adjustment to the current legislation could result in diminishing the authority of the social media platforms [119].

More active and nuanced intervention methods are required in order to successfully moderate hateful and dangerous communities [114]. Guidelines created by social media users could have more salience and a better chance to be adopted in the community [109].

Strategies such as bans and prohibition of certain language can backfire. Methods that function through encouraging positive social norms, may enable a better way to create online communities that are less harmful [101].

Finally, it is important to remember the well-being of human content moderators - providing days off, counselling sessions, a schedule that ensures them time to rest. Dealing with psychological abuse of others without proper support for oneself could have severe consequences [128].

Further reading:
Bengani, P., Ananny, M., & Bell, E. J. (2018). Controlling the Conversation: The Ethics of Social Platforms and Content Moderation.

**Automated detection of hate speech - recommendations for researchers:**

1. In order to better be able to compare various approaches, a main or "benchmark" dataset for automated hate speech detection is recommended [130].
2. Comparative studies are needed in order to be better able to detect the various forms of online hate speech and to gain insight on which approaches are more efficient than others [124, 130].
3. Multilingual research (i.e. using other languages besides English) [122].
4. **A more open process, in which researchers share their code, algorithms, sampling methods, annotation procedures, sources, data that has been excluded and the reason why, and individuals who may not be correctly represented in the study [122, 124, 126, 130].**
5. Provide more description of the accuracy ratings for detection systems - how were they assessed and what do they represent, to help policymakers decide which tools and methods are appropriate for which situation [122].
6. More cooperation with online platforms in order to make datasets accessible [126].
7. Humour, irony and sarcasm should be re-assessed in terms of their abusiveness [126].
8. Develop more varied datasets to address the issue of long-range dependencies [126].
9. Include user-level features (e.g. gender and geolocation) in detection systems in order to take into account the context of the messages [126].

Further reading:
Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, 80-93.

**Automated detection of hate speech - recommendations for policymakers:**

1. Using automated detection of any type of messages should never be sanctioned by law. Law should be concerned with the properties of the abolished types of speech, not with the techniques used to detect them [122].

2. Using automated detection of hate speech is burdened with the risk of rampant censorship that could affect the groups that are already marginalized the most [122].

3. **Any decision that results in effectively reducing the rights or liberties of individuals should always be made by other human agents [122].**

4. If an automated detection programme is implemented by a global institution or platform, it should take into account different contexts and cultures, in order to not risk inflicting Western standards on the rest of the world [128].

5. Any use of automatic systems in the public domain should be accompanied by human agents controlling and reviewing the functioning and the output of the system [122].

Further reading:
Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019b).
What Do We Mean When We Talk About Transparency?
Toward Meaningful Transparency in Commercial Content Moderation.
*International Journal of Communication, 13,* 18.

# PSYCHOEDUCATION

## Definition

Educational programmes are acknowledged as "soft" methods with a long-term perspective. They involve Media Literacy Education or Media and Information Literacy to teach young people the skills they need to be critical of online content, to hierarchize information and recognise troubling, hateful content, and misinformation [93].

**Media literacy programmes address the following skills [131]:**

1. **Access.** The skill to find and use media competently and to share appropriate and valuable information with others.

2. **Analysis and evaluation.** The ability to understand content and use critical thinking and comprehension to examine their quality, veracity, integrity and point of view, while keeping in mind their potential impact or consequences.

3. **Creation.** The ability to produce media content and confidently communicate this while being aware of the goal, public and composition techniques.

4. **Reflection.** The ability to adopt social responsibility and ethics to one's own identity, interaction and demeanour, to create an awareness of and to control one's media life.

5. **Action/agency.** The ability to act and participate in citizenship through media, to become political mediators based on democratic principles and views.

The ultimate goal of any educational programme concerning hate speech is to provide people with skills so that they do not back haters or endorse the hate by reacting in the same way [93].

Further reading:
Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. *Aggression and Violent Behavior, 45,* 163-172.

## Effectiveness

The research on educational programme combating hate speech is very limited. Most of the literature examined school-based anti-cyberbullying programme [132].

These programme have some effectiveness in reducing cyberbullying perpetration and victimization [132]. However, studies regarding interventions implemented within the United States do not show actual effects on behaviour, but merely changes in attitudes and intentions about cyberbullying [133]. There is a lack of evidence-based interventions.

With regard to media literacy and attitudes, a meta-analysis including 51 studies concluded that educational interventions have generally positive effects on a number of outcomes, such as media knowledge, criticism, perceived realism, behavioural beliefs, attitudes and self-efficacy [134].

For example, a media literacy programme on discrimination helped students to become more aware of media languages and how this can be used to discriminate certain social groups. In some cases, the programme improved students' political thinking. Moreover, associating critical thinking with students' own worries, preferences and identities was found to be more effective than involving them in the abstract examination of ideology [135].

Similarly, a prevention programme that aimed to facilitate critical media literacy with regard to Islamist online propaganda significantly improved pupils' awareness regarding extremist messages. However, no significant differences between the control and treatment group were found regarding their ability to adopt responsibility or act according to their newfound knowledge [136].

Further reading:
Jeong, S. H., Cho, H., & Hwang, Y. (2012). Media literacy interventions: A meta-analytic review. *Journal of Communication, 62*(3), 454-472.

# Recommendations

The first steps to take are to authorize both adults and young people to use media communication and teach them the key elements to comprehend what is going on online, to recognize hate content and to recognize the markers of hateful content and their potential influence, how they impact their victims and how to counter them [93].

Young people also need to acquire the skills to recognize conspiracy theories, revisionism, misinformation and cloaked websites as well as to interpret strategies that hate groups use to recruit new members and sympathisers [137].

It is crucial to facilitate dialogue and communication and give a sense of identity to vulnerable individuals who may seek this in online groups [93].

With regard to media literacy programmes researchers recommend developing new programmes that would involve all five main competences of media literacy (see “Psychoeducation - Definition”) [131].

In higher education, students should be offered anti-racism trainings led by institutional leaders, faculty, and administrators teaching them about racialized aggressions on social media [139].

Teacher educators ought to use an open discourse to inform their students about social media sites and encourage them to think critically about who is able to use these sites [138]. Finally, lessons about media literacy should span over longer periods of time and be seen as a part of educating people to be active citizens [135].

Further reading:
Ranieri, M., & Fabbro, F. (2016). Questioning discrimination through critical media literacy. Findings from seven European countries.
*European Educational Research Journal, 15*(4), 462-479.

# HATE SPEECH AND POLITICAL REGULATION

## What are we talking about?

There has been growing awareness of the rapid dissemination of online hate speech in many western democracies. Together with the rise of populist movements there seems to be an increasingly harsh tone to political debates with alleged effects on democratic conflict and participation. Incivility in political discourse might foster polarization and thereby challenge democratic order itself. Even more disturbingly, incivility in discourse is said to potentially spill over into other forms of violence.

Despite the current awareness, hate speech is not a new phenomenon. Approaches to tackling hate existed in pre-digital times. This is why today *existing rules may differ significantly between countries* across the world and even within Europe. While there are democratic countries where freedom of speech even protects hateful statements (i.e. the US), others (mostly authoritarian and hybrid regimes) might use hate speech as a pretext for far-reaching interferences when it comes to freedom of expression. European countries and the EU itself have mostly taken a more nuanced route as they try to balance the regulation of hate speech / the protection of the rights of individuals and groups with the right to freedom of expression.

However, fundamental difficulties remain even where common guidelines have been set. *When it comes to judging what expressions actually constitute hate speech, assessments may differ even within countries.* What some may consider to be statements acceptable or at least bearable in a heated democratic discourse, others may perceive as hurtful and punishable. Therefore, fighting online hate speech cannot avoid dealing with the fundamental tension between protecting the rights of individuals or groups and the freedom of expression. It is almost impossible to resolve and therefore constitutes a permanent trade-off.

While protections of human and minority rights have a long-lasting tradition, the current debate on hate speech is fuelled by

the growing societal importance of social media, which are often considered to be the main drivers for the dissemination of hate speech in political discussions.

Being platforms for the creation and spread of user-generated content in contrast to more classical media institutions, they pose a particular challenge for regulators. Social media are considered to be prone to spreading hate speech since content can be shared easily and rapidly with a large group of people. So the scalability and visibility is high. Furthermore, hate speech can be accessed for a long period of time if it is not effectively removed. Moreover, social media might facilitate a more explicit debate because of its relative anonymity and lack of controllability since they are run by private companies.

Europe stands at the forefront when it comes to protecting fundamental rights in cyberspace, to upholding trust of consumers in connected communication within a digital single market and to promoting social norms with regard to cyberspace. While the regulation of speech does not fall under supranational authority but is subject to member state regulation, *European institutions have however initiated important steps to harmonisation.*

Within Europe, efforts to regulate hate speech and establish sanctions have increased in the 2000s. The first legally binding definition was laid out by the Council of Europe's Additional Protocol to the Convention on Cybercrime (SEV Nr.189) which has been in force since 2006. This was followed by the above-mentioned EU framework decision in 2008 [7], which defined hate speech as ***"publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin"*** *Definition from EU framework decision 2008/913/JHA*

The EU framework decision nevertheless leaves member states the decision to punish this behaviour if they are "likely to disturb public order." But having a legal definition does not necessarily result in perfect harmonisation. There are transposition deficits in different countries as the EU Mandola Project, an ongoing research project funded by the EU, has shown [143].

Thus, instead of being completely harmonised, there are still differences in the regulation of hate speech between member states.

Since the rise of social media has fundamentally changed the media ecosystem with publication opportunities for every user, top-down regulation by states alone must be complemented by new governance arrangements including the co-regulation of private actors. In this vein, the EU established its Code of Conduct on Countering Illegal Hate Speech Online in 2016.

The code of conduct is supported by the most important internet companies (Facebook, Microsoft, Twitter, TikTok and YouTube) and therefore accounts for the important role that such intermediaries play in countering hate speech.

# MODES OF HATE SPEECH REGULATION

## How to take action?

As has become clear, regulating hate speech is not a trivial task due to different understandings of harmful content and appropriate counter-measures among democratic societies. Moreover, it is complicated by shared responsibilities between companies, states and civil society. Theoretically, there are different modes of engaging with hate speech. One would be that states enact and enforce strict laws (*statutory media regulation*). The opposite approach would be that the state does not enact any legislation and leaves regulation for example, to the market (i.e. the companies) which could identify hate speech as harmful for business since customers (i.e. users) may be repelled (*self-regulation*). A mix of both types might see laws making hate speech punishable, but leaving the responsibility to fight it to intermediaries, e.g. the major platforms of online communication (*coregulation*) [144]. For this latter approach, effectiveness depends significantly on the implementation by intermediaries. All kinds of regulation have their pros and cons.

Statutory media regulation has become less feasible with internet development, given the sheer volume of user-generated content uploaded every minute. Self-regulation risks being ineffective or not in line with the normative and legal standards it is meant to protect. Finally, coregulation is often criticised because of the authoritative rule delegated to private firms and their alleged inclination to practice "over blocking" (deleting more than would be necessary) as they fear the consequences of failing to abide by the law. Furthermore, the state then depends on intermediaries to report severe incidents in order to actually prosecute and punish them.

*But those modes of regulation are not sufficient:* There is a lot of hate speech that does not trigger legal or self-regulatory responses but that still may be harmful. Generally speaking, laws and standards set by market actors cannot be the only solution for countering hate speech, societal responses to upholding social norms are needed.

Therefore, there is an increased need for civil society to get engaged in countering hate speech (with government and civil society) [145]. This need is explicitly acknowledged by the EU's code of conduct, other initiatives or funded projects (including DeTACT).

1. Keep in mind the tension between the right to free expression and hate speech.

2. Keep in mind that you will probably not be perceived as a neutral arbiter, but as partisan and biased. *First goal is to keep up social norms in political debates, not to convince others of the falsehood of their opinions.*

The following country profiles provide a short introduction to the German, Belgian, Dutch, Irish and Hungarian laws with regard to hate speech and a glimpse at ongoing debates about future regulations. They are supposed to help reach decisions when confronted with hate speech and get a better understanding of current developments. Therefore, the profiles present the relevant legislation and, where available, practical references in order to illustrate what has been deemed illegal before. But the laws and prosecution practices in the countries differ significantly; therefore, for some countries, there are no such prior instances. For example, the Irish Prohibition of incitement to hatred act from 1989 should theoretically be applicable to online hate speech, but has only led to a handful of prosecutions for offline conduct [146, 147]. Hungarian regulations have resulted in only six convictions. Therefore, there are no clear and binding guidelines that might help to illustrate what may be considered illegal aside from the legal texts.

The profiles focus on regulations in criminal law and do not address sanctions for personal insults or defamation, for which all countries have established different statutory offences. All profiles end with a list of sources that provide further information on the respective countries.

Further reading:

Alkiviadou, N. (2018). The Legal Regulation of Hate Speech: The International and European Frameworks. *Politička misao, 55*(4), 203-229. https://doi.org/10.20901/pm.55.4.08

Aswad, E. (2016). The Role of U.S. Technology Companies as Enforcers of Europe's New Internet Hate Speech Ban. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2829175

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology, 24(3), 233-239.* https://doi.org/10.1080/13600869.2010.522323

Banks, J. (2011). European Regulation of Cross-Border Hate Speech in Cyberspace: The Limits of Legislation. *European Journal of Crime, Criminal Law and Criminal Justice, 19*(1), 1-13. https://doi.org/10.1163/157181711X553933

Bleich, E. (2013). Freedom of Expression versus Racist Hate Speech: Explaining Differences Between High Court Regulations in the USA and Europe. *Journal of Ethnic and Migration Studies, 40*(2), 283-300. https://doi.org/10.1080/1369183X.2013.851476

Brown, A. (2015). *Hate Speech Law: A Philosophical Examination (Routledge Studies in Contemporary Philosophy)*: Routledge.

Council of Europe (2017). WE CAN! Taking Action against Hate Speech through Counter and Alternative Narratives. Retrieved from https://rm.coe.int/wecan-eng-final-23052017-web/168071ba08

Kahn, R. A. (2012). Karl Loewenstein, Robert Post and the Ongoing Conversation between Europe and America Over Hate Speech Laws. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2177047

Sorial, S. (2015). Hate Speech and Distorted Communication: Rethinking the Limits of Incitement. *Law and Philosophy, 34*(3), 299-324. https://doi.org/10.1007/s10982-014-9214-9

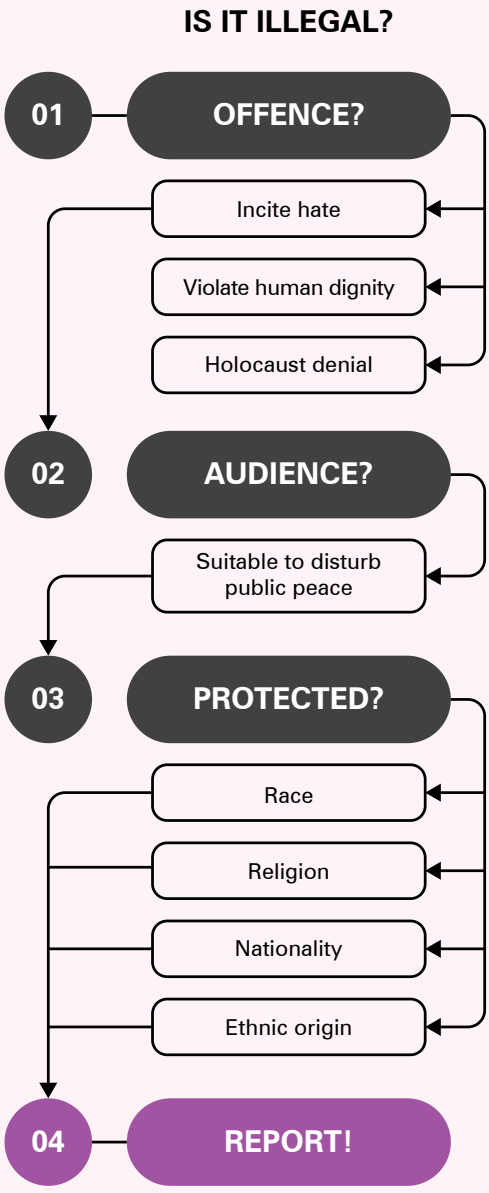
# HATE SPEECH REGULATION IN GERMANY

## Legal framework - Section 130 criminal code

The German criminal code addresses hate speech in Section 130. According to this, punishable conduct is defined as follows:

*"Whoever, in a manner which is suitable for causing a disturbance of the public peace,*

1. *incites hatred against a national, racial, religious group or a group defined by their ethnic origin, against sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population, or calls for violent or arbitrary measures against them or*
2. *violates the human dignity of others by insulting, maliciously maligning or defaming one of the aforementioned groups, sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population incurs a penalty of imprisonment for a term of between three months and five years."*



Furthermore, Section 130 of the criminal code also makes Holocaust denial a crime. Additionally, in 2017, the German government established a widely discussed law to tackle the spread of hate speech online.
The German Network Enforcement Act (Netzwerkdurchsetzungsgesetz, commonly known as NetzDG) obliges social media platforms with more than two million German users to delete “obviously” illegal content within 24 hours. In more ambiguous cases, companies have seven days to determine whether content is actually illegal. In case of noncompliance, companies can be fined with up to 5 million Euros. Heavy fines have also sparked debates whether that might lead to over blocking because companies might be inclined to delete content rather than risk a fine.

## Court rulings:

In September 2019, the Berlin district court made a controversial judgement in relation to hate speech that attracted much attention in Germany and beyond. The German politician Renate Künast, member of Bundestag for the Green Party, requested the names of a number of users from Facebook, as the respective accounts had posted comments including very offensive words and vulgar language. The court rejected the request pointing to the protection of free expression under which - according to the judges - the comments would fall given the political context of their appearance. Three months later, the judgement was partly overruled by the same court in a legal revision exerted in reaction to a complaint made by the politician [151].

As seen, the German criminal code does not sanction hate based on gender. A judgement from a court in Bonn illustrates problems following from that. The judgment can be accessed here.

Another judgment from a court in Hamm highlights the problems the global internet can create, since the defendant claimed to have written hateful comments outside of Germany. The judgment can be accessed here.

## Recent developments:

There is an ongoing debate about changing the Network Enforcement Act and obliging the affected companies to not only delete illegal content, but to further inform the Federal Criminal Police Office in order to prosecute offences like death threats or sedition [148].

Moreover, there are government ambitions to facilitate the prosecution of offenders by enabling an easier exchange of data between the police and intelligence agencies especially with regard to extremist hate speech [149]. There are also plans to strengthen the protection of politicians, who are more and more frequently confronted with hate speech [150].

## Reporting and sanctioning:

The Federal Criminal Police Office recommends reporting incidents of hate speech on a portal provided by Baden-Wuerttemberg in order to prosecute offences more efficiently. Offences violating Section 130 of the criminal code are then directly reported to the federal police [152].

According to the numbers from the Federal Criminal Police Office in 2018, there have been 1,962 cases of hate speech [153].

Further reading:
Article 19 (2018b). Hungary: Responding to 'hate speech'.
Retrieved from https://www.article19.org/wp-content/uploads/2018/03/Hungary_responding_to_HS.pdf
European Commission against Racism and Intolerance (2020b). ECRI Report on Germany (sixth monitoring cycle).
Retrieved from https://rm.coe.int/ecri-report-on-germany-sixth-monitoring-cycle-/16809ce4be

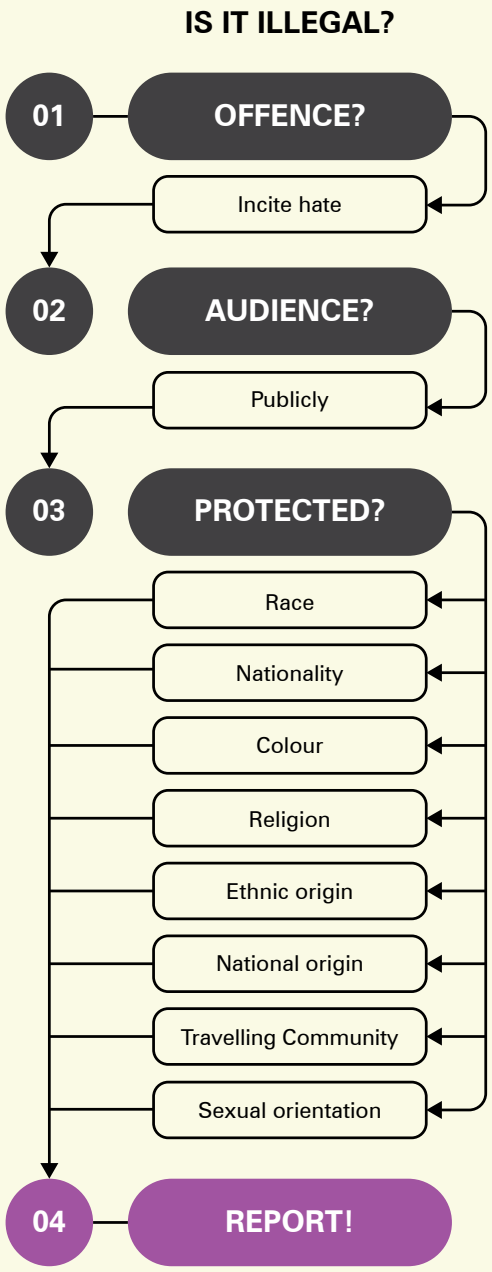
# HATE SPEECH REGULATION IN IRELAND

## Legal framework - Section 130 criminal code

The Prohibition of Incitement to Hatred Act 1989 established the following punishable offences:

*"hatred against a group of persons in the State or elsewhere on account of their race, colour, nationality, religion, ethnic or national origins, membership of the travelling community or sexual orientation;"*

1. ***(1)** It shall be an offence for a person—*
   ***(a)** to publish or distribute written material,*
   ***(b)** to use words, behave or display written material—*
   ***(i)** in any place other than inside a private residence, or*
   ***(ii)** inside a private residence so that the words, behaviour or material are heard or seen by persons outside the residence, or*
   ***(c)** to distribute, show or play a recording of visual images or sounds, if the written material, words, behaviour, visual images or sounds, as the case may be, are threatening, abusive or insulting and are intended or, having regard to all the circumstances, are likely to stir up hatred.*



## Recent developments:

In March 2019, the Minister for Communications, Climate Action and Environment proposed a new law to better tackle the spread of hate speech online. In his speech, he emphasized „that the era of self-regulation in this area is over and a new Online Safety Act is necessary" [154].

The Irish Government has thus just finished a public consultation in order to assess whether the Prohibition of Incitement to Hatred Act needs to be amended [155]. Most NGOs consider the Prohibition of Incitement to Hatred Act insufficient in dealing with online hate speech although it is argued that it theoretically would be applicable to the spread of hateful messages via the internet [156]. In the context of recent consultations on regulating hate speech, there have also been a lot of critical voices emphasising the need to guarantee free speech [157].

### Recommendations by the Irish Human Rights and Equality Commission (IHREC)

In order to assure the participation of different groups and minorities, IHREC recommended that a planned Electoral Commission should be "mandated to address the use of discriminatory rhetoric and hate speech in political campaigning" it should do so by promoting norms for political debates during elections and referendums. In a review of the Prohibition of Incitement to Hatred Act 1989 the commission concluded that the government should enact legislation that provides for a code of conduct to counter hate speech and should penalise intermediaries for failure to comply with this framework. A regulation like this would come close to the German Network Enforcement Act [158].

## Reporting and Sanctioning:

Instances of hate speech can be reported to ireport.ie. As of 2007 there have only been 44 prosecutions under the Prohibition of Incitement to Hatred Act 1989, only five of which resulted in actual convictions [159].

Further reading:

European Commission against Racism and Intolerance (2019a). ECRI Report on Ireland (fifth monitoring cycle). Retrieved from https://rm.coe.int/fifth-report-on-ireland/168094c575

Irish Human Rights and Equality Commission (2019b). Review of the Prohibition of Incitement to Hatred Act 1989. Retrieved from https://www.ihrec.ie/app/uploads/2019/12/Review-of-the-Prohibition-of-Incitement-to-Hatred-Act-1989.pdf

Schweppe, J., Haynes, A., & Carr, J. (2014). A life free from fear: Legislating for hate crime in Ireland: An NGO perspective. Retrieved from https://ulir.ul.ie/bitstream/handle/10344/4485/Schweppe_2014_crime.pdf

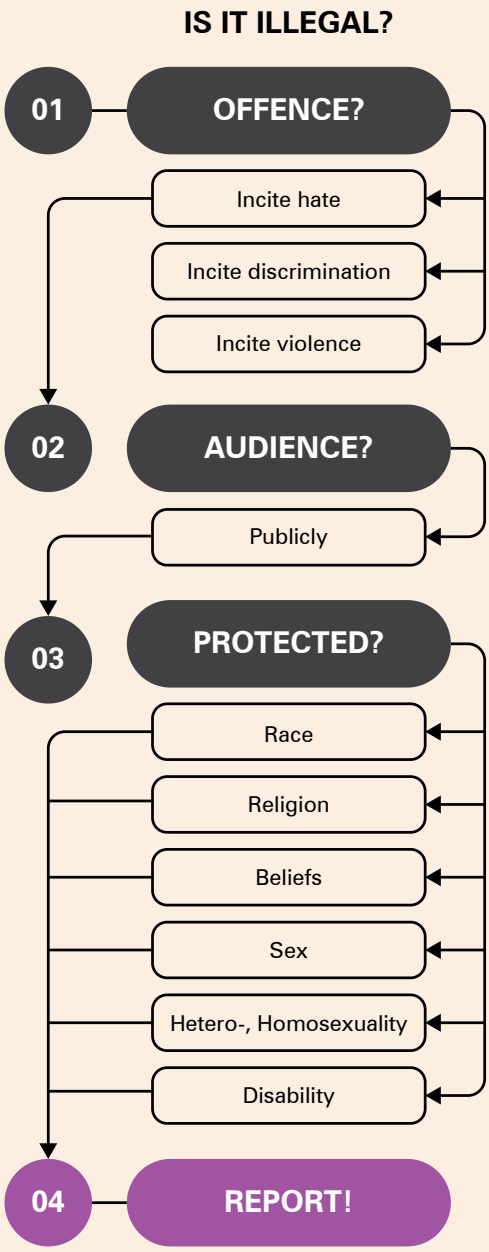
# HATE SPEECH REGULATION IN THE NETHERLANDS

## Legal framework - Section 137d of the criminal code

The Dutch criminal code regulates hate speech in Section 137d. Accordingly, punishable conduct is defined as inciting hatred, discrimination or violence against persons or property because of their race, religion or beliefs, their sex, their hetero- or homosexual orientation or their physical, mental or intellectual disability.

*"Hij die in het openbaar, mondeling of bij geschrift of afbeelding, aanzet tot haat tegen of discriminatie van mensen of geweld-dadig optreden tegen persoon of goed van mensen wegens hun ras, hun godsdienst of levensovertuiging, hun geslacht, hun hetero- of homoseksuelegerichtheid of hun lichamelijke, psychische of verstandelijke handicap, wordt gestraft met gevangenisstraf van ten hoogste twee jaren of geldboete van de vierde categorie."*

Additionally, Section 137c makes it illegal to insult groups based on the same criteria.



## Recent Developments:

The debate about hate speech regulation in the Netherlands is mostly driven by the trial against the Dutch politician Geert Wilders. In 2014, during a speech he asked whether the audience wanted “more or fewer Moroccans in the Netherlands”, to which the crowd answered “fewer”. This was considered to be an act of inciting hatred. In 2016, a court judged Wilders to be guilty of spreading hate speech, but no fine was imposed [160].

Since then there has been an ongoing debate on whether the Government should reform established rules. This is especially tricky in the Netherlands since the Dutch constitution has established rules that might not be up to the challenge in a more and more digitalised society. The constitution explicitly grants immunity for all statements a member of parliament makes in parliamentarian debates. This was a way to prevent judges from becoming involved in political matters. This, however, does not apply to statements outside of parliament. The supreme court furthermore deepened the divide between statements made in public and parliament when it ruled that politicians - because of their role as an example - should be particularly sensitive when it comes to spreading hate. This clear-cut divide is currently being criticised for being inapplicable to the modern communication environment [161].

### Court Rulings:

In May 2017, a court of appeal gave a 25-year-old a 2 month suspended sentence with a 2 year probation period for three posts on social media. This ruling might be telling for what is considered to be illegal in general. The judgment can be accessed here.

In March 2020, a defendant was sentenced for spreading racist texts. Since the accused claimed to have been drunk while writing these posts, this judgment also highlights a problem when it comes to sanctioning online hate speech - namely what can be considered mitigation. The judgment can be accessed here.

## Reporting and sanctioning:

Although denial of the Holocaust is not explicitly illegal, courts often do sanction Holocaust denial under Section 137d of the criminal code [162]. It is therefore advisable to report instances of Holocaust denial as well. Instances of discrimination can be reported via a website maintained by the Dutch government [163]. Prosecutors reported 144 discrimination offences in 2017. 19% of the incidents were conducted on the internet. Additionally, 187 instances were prosecuted under general criminal law [164].

Further reading:

European Commission against Racism and Intolerance (2019b).
ECRI Report on the Netherlands (fifth monitoring cycle).
Retrieved from https://rm.coe.int/fifth-report-on-the-netherlands/168094c577

Stam, J. (2019). The risky aspects of our hate speech laws.
Retrieved from https://leidenlawblog.nl/articles/the-risky-aspects-of-our-hate-speech-laws

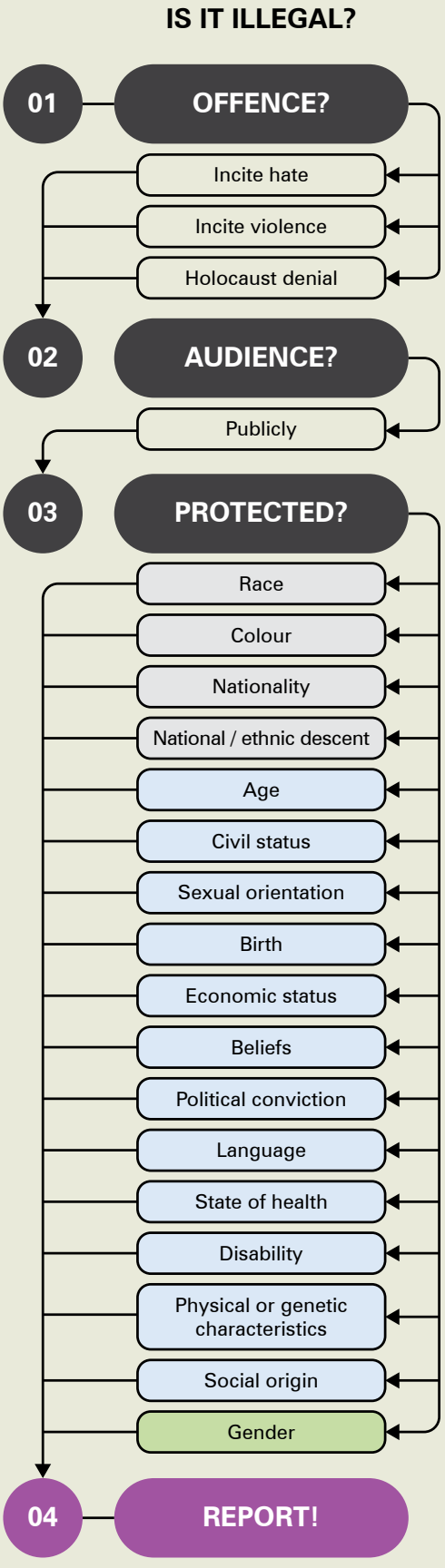
Van Noorloos, M. (2013). The Politicisation of Hate Speech Bans in the Twenty-first-century Netherlands: Law in a Changing Context. *Journal of Ethnic and Migration Studies, 40*(2), 249–265.
https://doi.org/10.1080/1369183X.2013.851474

# HATE SPEECH REGULATION IN BELGIUM

## Legal framework - Belgian Anti-Racism Law and Section 444 of the Criminal Code

The Belgian Anti-Racism Law (Law of 30 July 1981 on the punishment of certain acts inspired by racism or xenophobia) made it illegal to incite violence or hate against persons or groups on the grounds of their nationality, race, colour or national or ethnic descent (highlighted in grey). Section 444 of the Criminal Code specifies the circumstances an incident must meet. In order to be sanctionable, the incident must occur in public or in presence of several people:

*critères protégés : la nationalité, une prétendue race, la couleur de peau, l'ascendance ou l'origine nationale ou ethnique;*
*Art. 20*
*Est puni d'un emprisonnement d'un mois à un an et d'une amende de cinquante euros à mille euros, ou de l'une de ces peines seulement: [...]*



*Quiconque, dans l'une des circonstances indiquées à l'article 444 du Code pénal, incite à la haine ou à la violence à l'égard d'une personne, en raison de l'un des critères protégés, et ce, même en dehors des domaines visés à l'article 5;*

The Anti Discrimination Law made it illegal to discriminate on the grounds of age, sexual orientation, civil status, birth, economic status, beliefs, political conviction, language, current or future state of health, disability, physical or genetic characteristics or social origin (highlighted in blue).

*critères protégés : l'âge, l'orientation sexuelle, l'état civil, la naissance, la fortune, la conviction religieuse ou philosophique, la conviction politique, langue, l'état de santé actuel ou futur, un handicap, une caractéristique physique ou génétique, l'origine sociale;*

The Gender Act added gender as a protected characteristic (highlighted in green). Moreover, the Holocaust Denial Law establishes offences for publicly denying, playing down, justifying or approving of the genocide committed by the German National Socialist regime during the Second World War.

## Recent Developments:

As in the other countries, there is an ongoing debate in Belgium on the regulation of hate speech and on new regulations with regard to online hate speech. In January 2020, the Belgian Interim Prime Minister Sophie Wilmès announced that her Government intended to establish new laws to tackle the spread of hate speech online more efficiently. This debate was fuelled by a rescue operation that saved migrants on a boat and that sparked the spread of hateful messages in Belgium [165].

### Court Rulings:

Following a complaint from Unia, a court sentenced a defendant to six months in prison and a fine of 4,000 Euros for spreading racist messages on Facebook. This ruling might be telling of what is considered to be illegal in general. The judgment can be accessed here.

## Reporting and Sanctioning:

Incidents of hate speech can be reported to Unia, an independent public institution that fights discrimination [166]. Discrimination with regard to gender can be reported to the Institute for the Equality of Women and Men [167]. In 2016, Unia received 5,619 reports with regard to discrimination. These resulted in 1,907 case files. Unia publishes reports in French and Dutch that illustrate the different kinds of discrimination reported and which of the protected criteria are affected [168]. With reference to actual judgements, Unia also provides helpful advice on reporting hate speech, specifying inter alia what is considered to be public when it comes to online hate speech (e.g. an e-mail with multiple recipients) [169].


Further reading:
European Commission against Racism and Intolerance (2020a).
ECRI Report on Belgium (sixth monitoring cycle).
Retrieved from https://rm.coe.int/ecri-sixth-report-on-belgium-/16809ce9f0

Le Conseil Supérieur de l'Audiovisuel (2020).
Contenus illicites sur les réseaux sociaux et plateformes de partage vidéo : le CSA publie une note d'orientation.
Retrieved from https://www.csa.be/document/contenus-illicites-sur-les-reseaux-sociaux-et-plateformes-de-partage-video-le-csa-publie-une-note-dorientation/

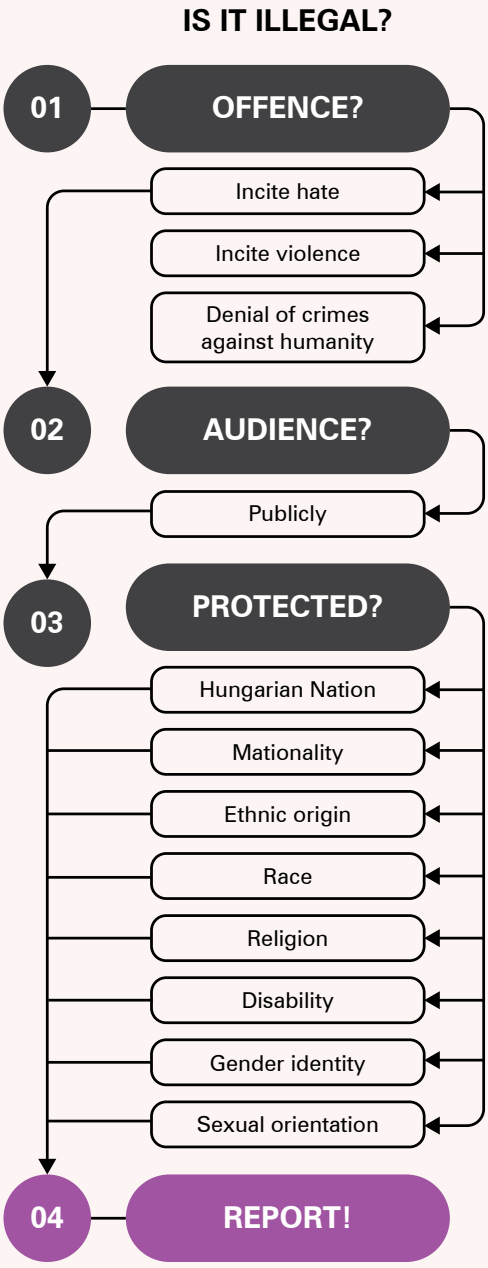
# HATE SPEECH REGULATION IN HUNGARY

## Legal framework - Section 332 of the Criminal Code

The Hungarian Criminal Law presents an encompassing set of protected characteristics in Section 332.

*Any person who, before the public at large, incites violence or hatred against: a) the Hungarian nation; b) any national, ethnic, racial or religious group or a member of such a group; or c) certain societal groups or a member of such a group, in particular on the grounds of disability, gender identity or sexual orientation is guilty of a felony punishable by imprisonment not exceeding three years.*

The Criminal Code also establishes an offence for the denial of crimes against humanity committed by National Socialists (i.e. the Holocaust) or Communists.



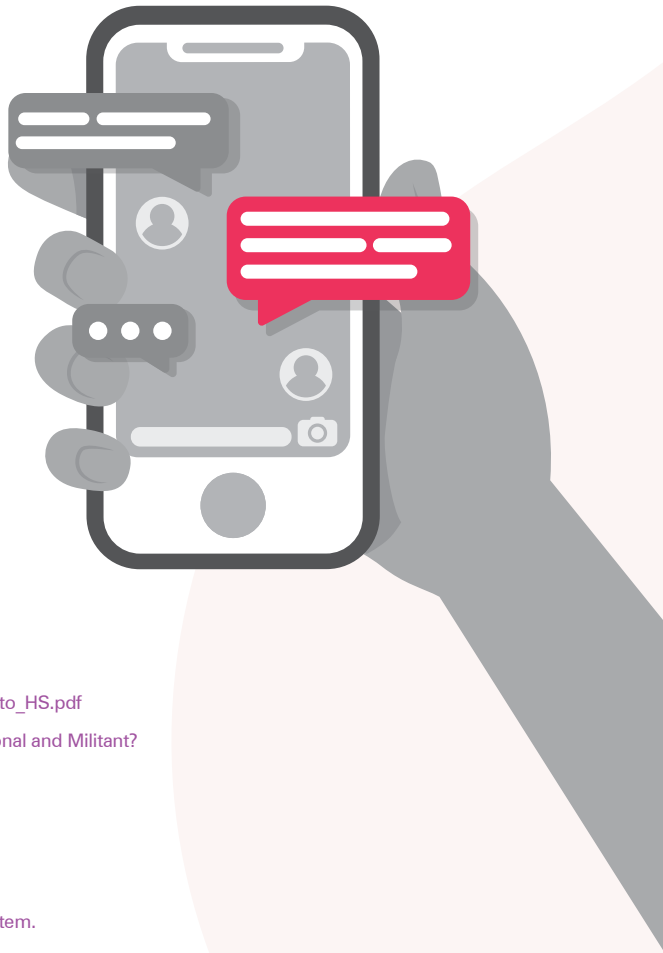
## Recent Developments:

The Hungarian government has been repeatedly criticised for the spread of hateful messages against immigrants. The UN Human Rights Committee was concerned about hate speech affecting "minorities, notably, Roma, Muslim, migrants and refugees, including in the context of government-sponsored campaigns." Therefore, it is the executive itself (and the respective supporters) that stands accused of spreading hate speech on- and offline. This is in line with governmental efforts to ban NGOs supporting migrants on the grounds of national security [170].

The established regulation on hate speech has also been criticised for its provision to prosecute hate speech against the Hungarian Nation which could be seen as a potentially inappropriate restriction of free speech that might be used to censor political opponents. It is problematic that the concept of the Hungarian Nation is not clearly defined and is too all-encompassing to constitute a protected characteristic as it might be employed arbitrarily [171]. We would therefore not recommend reporting hate speech based on this criterion.

## Reporting and Sanctioning:

Incidents can be reported to the Internet Hotline [172], although the catalogue of protected criteria has almost no prosecutions following Section 332. Between 2009 and 2013, 201 incidents were reported by the police and only six of which resulted in trials and convictions. NGOs have repeatedly criticised this restrictive practice. In 2017, the European Court also ruled that reluctant investigations by the police contravene Article 8 of the European Convention on Human Rights [171].



Further reading:

Article 19 (2018b). Hungary: Responding to 'hate speech'.
Retrieved from https://www.article19.org/wp-content/uploads/2018/03/Hungary_responding_to_HS.pdf

Belavusau, U. (2014). Hate Speech and Constitutional Democracy in Eastern Europe: Transitional and Militant? (Czech Republic, Hungary and Poland). Israel Law Review, 47(1), 27–61.
https://doi.org/10.1017/S0021223713000241

European Commission against Racism and Intolerance (2015).
ECRI Report on Hungary (fifth monitoring cycle).
Retrieved from https://rm.coe.int/fifth-report-on-hungary/16808b57e8

Koltay, A. (2013). Hate Speech and the Protection of Communities in the Hungarian Legal System.
Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2197914

# GENERAL REMARKS

When it comes to a decision on reporting hate speech, there might be some considerations that more or less apply to all countries.

Incitement to hatred or violence is central to offences in all countries. Therefore, it is helpful to clarify what might be considered to be incitement. It might include urging someone to do something (you need to...), calling for action (we should...), encouraging (someone has to...) or fueling debates (wake up...). Statements that include such phrases might qualify as incitement.

Moreover, in all countries the legal frameworks define that hate speech is punishable when it is public. But judgments differ not only between countries but also within them, when it comes to evaluating what is to be treated as public. It might be advisable to focus on instances when a post or message receives significant attention (e.g. 100+ likes).

When it comes to the different protected groups, countries differ with regard to listed characteristics. But that does not necessarily imply that a legal complaint might not be successful as it is often argued, for example, that race can include colour. Therefore, incidents might be sanctioned although a characteristic is not explicitly mentioned.

[1] Siegel, A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., ... & Tucker, J. A. (2019). Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath.

[2] Pohjonen, M. (2019). Extreme Speech - A Comparative Approach to Social Media Extreme Speech: Online Hate Speech as Media Commentary. International Journal of Communication, 13, 16.

[3] Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology*, *11*(1), 143-160.

[4] Richardson-Self, L. (2018). Woman-Hating: On Misogyny, Sexism, and Hate Speech. *Hypatia*, *33*(2), 256-272.

[5] Awan, I., & Zempi, I. (2017). 'I will blow your face OFF'—VIRTUAL and physical world anti-muslim hate crime. *The British Journal of Criminology*, *57*(2), 362-380.

[6] Erjavec, K., & Kovačič, M. P. (2012). "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society*, *15*(6), 899-920.

[7] European Commission (2008). Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law.

[8] Brown, A. (2018). What is so special about online (as compared to offline) hate speech?. *Ethnicities*, *18*(3), 297-326.

[9] Anti-Defamation League (2010). Dignity, Fairness, Respect, Solidarity, Expertise, Security, Freedom. Words Matter. New York: ADL.

[10] Sternberg, R. J. (Ed.). (2005). The psychology of hate. American Psychological Association.

[12] Waller, J. E. (2004). Our ancestral shadow: Hate and human nature in evolutionary psychology. Journal of Hate Studies, 3, 121.

[11] Zeki, S., & Romaya, J. P. (2008). Neural correlates of hate. PloS one, 3(10).

[13] Anderson, B. (2006). Imagined communities (new ed.). London: Verso.

[14] Back, L., Keith, M., & Solomes, J. (1998). Racism on the Internet: Mapping neo-fascist subcultures in space. In J. Kaplan & T. Bjørgo (Eds.), Nation and race (pp. 73–101). Boston: Northeastern University Press.

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

[16] Barlow, C., & Awan, I. (2016). "You need to be sorted out with a knife": the attempted online silencing of women and people of Muslim faith within academia. *Social Media+ Society*, *2*(4), 2056305116678896.

[17] Chen, G. M., Pain, P., Chen, V. Y., Mekelburg, M., Springer, N., & Troger, F. (2018). 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism*, 1464884918768500.

[18] Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *Journal of Intelligent & Fuzzy Systems, 36*(5), 4743-4752.

[19] Hardaker, C., & McGlashan, M. (2016). "Real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics*, *91*, 80-93.

[20] Lewis, R., Rowe, M., & Wiper, C. (2016). Online abuse of feminists as an emerging form of violence against women and girls. *British journal of criminology*, *57*(6), 1462-1481.

[21] Megarry, J. (2014). Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. In *Women's Studies International Forum* (Vol. 47, pp. 46-55). Pergamon.

[22] Salter, A., & Blodgett, B. (2012). Hypermasculinity & dickwolves: The contentious role of women in the new gaming public. *Journal of broadcasting & electronic media*, *56*(3), 401-416.

[23] Sobieraj, S. (2018). Bitch, slut, skank, cunt: Patterned resistance to women's visibility in digital publics. *Information, Communication & Society*, *21*(11), 1700-1714.

[24] Thompson, L. (2018). "I can be your Tinder nightmare": Harassment and misogyny in the online sexual marketplace. *Feminism & Psychology*, *28*(1), 69-89.

[25] Wolfe, C. (2019). Online trolls, journalism and the freedom of speech: Are the bullies taking over?. *Ethical Space: The International Journal of Communication Ethics*, *16*(1), 11-21.

[26] Wilhelm, C., & Joeckel, S. (2019). Gendered Morality and Backlash Effects in Online Discussions: An Experimental Study on How Users Respond to Hate Speech Comments Against Women and Sexual Minorities. *Sex Roles*, *80*(7-8), 381-392.

[27] Banet-Weiser, S., & Miltner, K. M. (2016). # MasculinitySoFragile: culture, structure, and networked misogyny. *Feminist Media Studies*, *16*(1), 171-174.

[28] Jane, E. A. (2012). "Your a ugly, whorish, slut" - understanding E-bile. *Feminist Media Studies*, *14*(4), 531-546. doi:10.1080/14680777.2012.741073

[29] Jane, E. A. (2014). 'Back to the kitchen, cunt': speaking the unspeakable about online misogyny. *Continuum*, *28*(4), 558-570.

[30] Mantilla, K. (2013). Gendertrolling: Misogyny adapts to new media. *Feminist Studies*, *39*(2), 563-570.

[31] Halder, D., & Karuppannan, J. (2009). Cyber socializing and victimization of women. *The Journal on Victimization*, *12*(3), 5-26.

[32] Wyckoff, J. P., Buss, D. M., & Markman, A. B. (2019). Sex differences in victimization and consequences of cyber aggression: An evolutionary perspective. *Evolutionary Behavioral Sciences*, *13*(3), 254.

[33] Consalvo, M. (2012). Confronting toxic gamer culture: A challenge for feminist game studies scholars. *Ada: A Journal of Gender, New Media, and Technology*, *1*(1), 1-6.

[34] Gray, K. L. (2012). Intersecting oppressions and online communities: Examining the experiences of women of color in Xbox Live. *Information, Communication & Society*, *15*(3), 411-428.

[35] Witkowski, E. (2018). Doing/Undoing Gender with the Girl Gamer in High-Performance Play. In *Feminism in Play* (pp. 185-203). Palgrave Macmillan, Cham.

[36] Citron, D. K. (2011). Misogynistic cyber hate speech.

[37] Awan, I. (2014). Islamophobia and Twitter: A typology of online hate against Muslims on social media. *Policy & Internet, 6*(2), 133-150.

[38] Awan, I. (2016). Islamophobia on Social Media: A Qualitative Analysis of the Facebook's Walls of Hate. *International Journal of Cyber Criminology, 10*(1).

[39] Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and violent behavior, 27*, 1-8.

[40] Awan, I., & Zempi, I. (2017). 'I will blow your face OFF'—VIRTUAL and physical world anti-muslim hate crime. *The British Journal of Criminology, 57*(2), 362-380.

[41] Cleland, J. (2014). Racism, football fans, and online message boards: How social media has added a new dimension to racist discourse in English football. *Journal of Sport and Social Issues, 38*(5), 415-431.

[42] Cleland, J., Anderson, C., & Aldridge-Deacon, J. (2018). Islamophobia, war and non-Muslims as victims: an analysis of online discourse on an English Defence League message board. *Ethnic and racial studies, 41*(9), 1541-1557.

[43] Ekman, M. (2015). Online Islamophobia and the politics of fear: manufacturing the green scare. *Ethnic and Racial Studies, 38*(11), 1986-2002.

[44] Feldman, M., & Allchorn, D. W. (2019). A working definition of anti-Muslim hatred with a focus on hate-crime work.

[45] Froio, C. (2018). Race, religion, or culture? Framing Islam between racism and neo-racism in the online network of the french far right. *Perspectives on Politics, 16*(3), 696-709.

[46] Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology, 11*(1), 143-160.

[47] Lee, B. (2015). A day in the "Swamp": Understanding discourse in the online counter-Jihad nebula. *Democracy and Security, 11*(3), 248-274.

[48] Pettersson, K. (2019). "Freedom of speech requires actions": Exploring the discourse of politicians convicted of hate-speech against Muslims. *European Journal of Social Psychology.*

[49] Lieberman, L., Kaszycka, K. A., Martinez, A. J., Yablonsky, F. L., Kirk, R. C., Štrkalj, G., ... & Sun, L. (2004). The race concept in six regions: variation without consensus. *Collegium antropologicum, 28*(2), 907-921.

[50] Tynes, B. M., Umana-Taylor, A. J., Rose, C. A., Lin, J., & Anderson, C. J. (2012). Online racial discrimination and the protective function of ethnic identity and self-esteem for African American adolescents. *Developmental psychology, 48*(2), 343.

[51] Weaver, S. (2010). Developing a rhetorical analysis of racist humour: Examining anti-black jokes on the internet. *Social Semiotics, 20*(5), 537-555.

[52] Lee-Won, R. J., White, T. N., Song, H., Lee, J. Y., & Smith, M. R. (2019). Source magnification of cyberhate: affective and cognitive effects of multiple-source hate messages on target group members. *Media Psychology*, 1-22.

[53] Flores-Yeffal, N. Y., Vidales, G., & Martinez, G. (2019). # WakeUpAmerica,# IllegalsAreCriminals: the role of the cyber public sphere in the perpetuation of the Latino cyber-moral panic in the US. *Information, Communication & Society, 22*(3), 402-419.

[54] Loke, J. (2013). Readers' debate a local murder trial: "Race" in the online public sphere. *Communication, Culture & Critique, 6*(1), 179-200.

[55] Santana, A. D. (2015). Incivility dominates online comments on immigration. *Newspaper Research Journal, 36*(1), 92-107.

[56] Finn, J. (2004). A survey of online harassment at a university campus. Journal of *Interpersonal violence, 19*(4), 468-483.

[57] Abreu, R. L., & Kenny, M. C. (2018). Cyberbullying and LGBTQ youth: A systematic literature review and recommendations for prevention and intervention. *Journal of Child & Adolescent Trauma, 11*(1), 81-97.

[58] Montoro, R., Igartua, K., & Thombs, B. D. (2016). The association of bullying with suicide ideation and attempt among adolescents with different dimensions of sexual orientation. *European Psychiatry, 33*, S71.

[59] Costello, M., Rukus, J., & Hawdon, J. (2019). We don't like your type around here: regional and residential differences in exposure to online hate material targeting sexuality. *Deviant Behavior, 40*(3), 385-401.

[60] Stakic, I. (2011). Homophobia and hate speech in Serbian public discourse: how nationalist myths and stereotypes influence prejudices against the LGBT minority (Master's thesis, Universitetet i Tromsø).

[61] Goodall, J. (1986). Social rejection, exclusion, and shunning among the Gombe chimpanzees. *Ethology and Sociobiology, 7*(3-4), 227-236.

[62] Spoor, J., & Williams, K. D. (2007). The evolution of an ostracism detection system. In *Evolution and the social mind: evolutionary psychology and social cognition*, 279-292.

[63] Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science, 302*(5643), 290-292.

[64] Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin, 117*(3), 497.

[65] MacDonald, G., Kingsbury, R., & Shaw, S. (2005). Adding insult to injury. In *The social outcast: Ostracism, social exclusion, rejection, and bullying*, 77-90.

[66] Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Hateful Speech in Online College Communities.

[67] Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior, 44*(2), 136-146.

[68] Esses, V. M., Medianu, S., & Lawson, A. S. (2013). Uncertainty, threat, and the role of the media in promoting the dehumanization of immigrants and refugees. *Journal of Social Issues, 69*(3), 518-536.

[69] Franklin, T. B., Saab, B. J., & Mansuy, I. M. (2012). Neural mechanisms of stress resilience and vulnerability. *Neuron, 75*(5), 747-761.

[70] Miller, R. L. (2015). How people judge the credibility of information: Lessons for evaluation from cognitive and information sciences. In *Credible and actionable evidence: The foundation for rigorous and influential evaluations*, 39-61.

[71] Sidanius, J. and Pratto, F. (1999) Social dominance: An intergroup theory of social hierarchy and oppression. Cambridge: Cambridge University Press.

[72] Forscher, P.S., Cox, W. T. L., Graetz, N. and Devine, P. G. (2015) 'The motivation to express prejudice', *Journal of Personality and Social Psychology, 109*, 791–812.

[73] Brown, R. (2010) Prejudice: Its social psychology. Chichester: Wiley-Blackwell.

[74] Gadd, D. (2009) 'Aggravating racism and elusive motivation', British Journal of Criminology, 49(6), 755–71.

[75] Roberts, C., Innes, M., Williams, M., Tregidga, J.and Gadd, D. (2013) Understanding who commits hate crime and why they do it, Welsh Government Social Research.

[76] Levin, J. and McDevitt, J. (1993) Hate crimes: The rising tide of bigotry and bloodshed. New York: Plenum.

[77] Steinberg, A., Brooks, J. and Remtulla, T. (2003) Youth hate crimes: Identification, prevention, and intervention. *American Journal of Psychiatry, 160*(5), 979–89.

[78] Bliuc, A. M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior, 87*, 75-86.

[79] Jones, L. M., Mitchell, K. J., & Turner, H. A. (2015). Victim reports of bystander reactions to in-person and online peer harassment: A national survey of adolescents. *Journal of youth and adolescence*, *44*(12), 2308-2320.

[80] Patterson, L. J., Allan, A., & Cross, D. (2017). Adolescent bystander behavior in the school and online environments and the implications for interventions targeting cyberbullying. *Journal of school violence*, *16*(4), 361-375.

[81] Hayes, B. E. (2019). Bystander Intervention to Abusive Behavior on Social Networking Websites. *Violence against women, 25*(4), 463-484.

[82] Henson, B., Fisher, B. S., & Reyns, B. W. (2019). There Is Virtually No Excuse: The Frequency and Predictors of College Students' Bystander Intervention Behaviors Directed at Online Victimization. *Violence against women*, 1077801219835050.

[83] Wachs, S., & Wright, M. F. (2018). Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. *International journal of environmental research and public health, 15*(9), 2030.

[84] Henson, B., Fisher, B. S., & Reyns, B. W. (2019). There Is Virtually No Excuse: The Frequency and Predictors of College Students' Bystander Intervention Behaviors Directed at Online Victimization. *Violence against women*, 1077801219835050.

[85] Costello, M., Hawdon, J., & Cross, A. (2016). Virtually standing up or standing by? Correlates of enacting social control online. *International Journal of Criminology and Sociology*, *6*, 16-28.

[86] Eldridge, M. A., & Jenkins, L. N. (2019). The Bystander Intervention Model: Teacher Intervention in Traditional and Cyber Bullying. *International Journal of Bullying Prevention*, 1-11.

[87] Brody, N., & Vangelisti, A. L. (2016). Bystander intervention in cyberbullying. *Communication Monographs, 83*(1), 94-119.

[88] Domínguez-Hernández, F., Bonell, L., & Martínez-González, A. (2018). A systematic literature review of factors that moderate bystanders' actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 12*(4).

[89] Schacter, H. L., Greenberg, S., & Juvonen, J. (2016). Who's to blame?: The effects of victim disclosure on bystander reactions to cyberbullying. *Computers in Human Behavior, 57*, 115-121.

[90] Domínguez-Hernández, F., Bonell, L., & Martínez-González, A. (2018). A systematic literature review of factors that moderate bystanders' actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 12*(4).

[91] Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *SCM Studies in Communication and Media, 7*(4), 555-579.

[92] Case, A. D., & Hunter, C. D. (2012). Counterspaces: A unit of analysis for understanding the role of settings in marginalized individuals' adaptive responses to oppression. *American Journal of Community Psychology, 50*(1-2), 257-270.

93] Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. Aggression and violent behavior, 45, 163-172.

[94] Hatakka, N. (2019). Expose, debunk, ridicule, resist! Networked civic monitoring of populist radical right online action in Finland. *Information, Communication & Society*, 1-16.

[95] Poole, E. A., Giraud, E., & de Quincey, E. (2019). Contesting# Stoplslam: The dynamics of a counter-narrative against right-wing populism. *Open Library of Humanities, 5*(1).

[96] Chung, Y. L., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019, July). CONAN-COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Conference of the Association for Computational Linguistics* (pp. 2819-2829).

[97] Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., ... & Mukherjee, A. (2019, July). Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, No. 01, pp. 369-380).

[98] Case, A. D., & Hunter, C. D. (2012). Counterspaces: A unit of analysis for understanding the role of settings in marginalized individuals' adaptive responses to oppression. *American Journal of Community Psychology*, *50*(1-2), 257-270.

[99] Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counterspeech on Facebook. In *66th ICA annual conference, at Fukuoka, Japan* (pp. 1-23).

[100] Silverman, T., Stewart, C. J., Birdwell, J., & Amanullah, Z. (2016). The impact of counter-narratives. *Institute for Strategic Dialogue*, 1-54.

[101] Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629-649.

[102] Hatakka, N. (2019). Expose, debunk, ridicule, resist! Networked civic monitoring of populist radical right online action in Finland. *Information, Communication & Society*, 1-16.

[103] Lee, B. J. (2019). Informal Countermessaging: The Potential and Perils of Informal Online Countermessaging. *Studies in Conflict & Terrorism*, *42*(1-2), 161-177.

[104] Anti-Defamation League. (2014). Best Practices for Responding to Cyberhate. Retrieved 31 July 2019, from https://www.adl.org/best-practices-for-responding-to-cyberhate

[105] Jääskeläinen, T. (2019). Countering hate speech through arts and arts education: Addressing intersections and policy implications. *Policy Futures in Education*, 1478210319848953.

[106] Bartlett, J., & Krasodomski-Jones, A. (2015). Counter speech. Examining content that challenges extremism online. *Demos.*

[107] Greenberg, K. J. (2016). Counter-radicalization via the internet. *The Annals of the American Academy of Political and Social Science, 668*(1), 165-179.

[108] Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, *7*(3), 321-326.

[109] Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, 369-374. ACM.

[110] Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. *Media Studies Publications, 12*.

[111] Frischlich, L., Boberg, S., & Quandt, T. (2018). Comment Sections as Targets of Dark Participation? Journalists' Evaluation and Moderation of Deviant User Comments. *Journalism Studies*, 1-20.

[112] Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction, 1*(CSCW), 31.

[113] Newell, E., Jurgens, D., Saleem, H. M., Vala, H., Sassine, J., Armstrong, C., & Ruths, D. (2016). User migration in online social networks: A case study on Reddit during a period of community unrest. In *Tenth International AAAI Conference on Web and Social Media*.

[114] Habib, H., Musa, M. B., Zaffar, F., & Nithyanand, R. (2019). To Act or React: Investigating Proactive Strategies For Online Community Moderation. *arXiv preprint arXiv:1906.11932*.

[115] Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences, 116*(20), 9785-9789.

[116] Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, 369-374. ACM.

[117] Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication, 13*, 18.

[118] Gibson, A. (2019). Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society, 5*(1), 2056305119832588.

[119] Nurik, C. (2019). "Men Are Scum": Self-Regulation, Hate Speech, and Gender-Based Censorship on Facebook. *International Journal of Communication, 13*(21), 2878-2898.

[120] Roberts, S. T. (2018). Digital detritus:'Error'and the logic of opacity in social media content moderation. *First Monday, 23*(3).

[121] de Zwart, M. (2018). Keeping the neighbourhood safe: How does social media moderation control what we see (and think)? *Alternative Law Journal, 43*(4), 283-288.

[122] Duarte, N., Llanso, E., & Loup, A. (2018). Mixed Messages? The Limits of Automated Social Media Content Analysis. In *FAT,* 106.

[123] Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1-10.

[124] Siegel, A. A. (2018). Online Hate Speech.

[125] Raisi, E., & Huang, B. (2016). Cyberbullying identification using participant-vocabulary consistency. *arXiv preprint* arXiv:1606.08084.

[126] Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online,* 80-93.

[127] Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence.*

[128] Bengani, P., Ananny, M., & Bell, E. J. (2018). Controlling the Conversation: The Ethics of Social Platforms and Content Moderation.

[129] Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019b). What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication, 13*, 18.

[130] Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR), 51*(4), 1-30.

[131] McDougall, J., Zezulkova, M., Van Driel, B., & Sternadel, D. (2018). Teaching media literacy in Europe: evidence of effective school practices in primary and secondary education, NESET II report.

[132] Gaffney, H., Farrington, D. P., Espelage, D. L., & Ttofi, M. M. (2019). Are cyberbullying intervention and prevention programs effective? A systematic and meta-analytical review. *Aggression and Violent Behavior, 45*, 134-153.

[133] Lancaster, M. (2018). A Systematic Research Synthesis on Cyberbullying Interventions in the United States. *Cyberpsychology, Behavior, and Social Networking, 21*(10), 593-602.

[134] Jeong, S. H., Cho, H., & Hwang, Y. (2012). Media literacy interventions: A meta-analytic review. *Journal of Communication, 62*(3), 454-472.

[135] Ranieri, M., & Fabbro, F. (2016). Questioning discrimination through critical media literacy. Findings from seven European countries. *European Educational Research Journal, 15*(4), 462-479.

[136] Schmitt, J. B., Rieger, D., Ernst, J., & Roth, H. J. (2019). Critical Media Literacy and Islamist Online Propaganda: The Feasibility, Applicability and Impact of Three Learning Arrangements. *International Journal of Conflict and Violence, 12*, 642.

[137] Meddaugh, P. M., & Kay, J. (2009). Hate speech or "reasonable racism?" the other in stormfront. *Journal of Mass Media Ethics, 24*(4), 251-268.

[138] Nagle, J. (2018). Twitter, cyber-violence, and the need for a critical social media literacy in teacher education: A review of the literature. *Teaching and Teacher Education, 76*, 86-94.

[139] Gin, K. J., Martínez-Alemán, A. M., Knight, S., Radimer, S., Lewis, J., & Rowan-Kenyon, H. T. (2016). Democratic education online: Combating racialized aggressions on social media. *Change: The Magazine of Higher Learning, 48*(3), 28-35.

[140] Forgas, J. P. (2012). *Affect in social thinking and behavior.* Psychology Press.

[141] Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review, 5*(4), 296-320.

[142] Dunbar, R. I. (2004). Gossip in evolutionary perspective. Review of general psychology, 8(2), 100-110.

[143] For further information see: http://mandola-project.eu/

[144] Puppis, M. (2010). Media Governance: A New Concept for the Analysis of Media Policy and Regulation. *Communication, Culture & Critique*, *3*(2), 134–149. https://doi.org/10.1111/j.1753-9137.2010.01063.x

[145] Jarren, O. (2007). Die Regulierung der öffentlichen Kommunikation: Medienpolitik zwischen Government und Governance. *Zeitschrift für Literaturwissenschaft und Linguistik*, *37*(2), 131–153.

[146] Schweppe, J., Haynes, A., & Carr, J. (2014). A life free from fear: Legislating for hate crime in Ireland: An NGO perspective. Retrieved from https://ulir.ul.ie/bitstream/handle/10344/4485/Schweppe_2014_crime.pdf

[147] PRISM - Preventing, Redressing and Inhibiting hate speech in new Media (2016). Hate Crime and Hate Speech in Europe: Comprehensive Analysis of International Law Principles, EU-wide Study and National Assessments. Retrieved from http://www.unicri.it/special_topics/hate_crimes/Hate_Crime_and_Hate_Speech_in_Europe_Comprehensive_Analysis_of_International_Law_Principles_EU-wide_Study_and_National_Assessments.pdf

[148] Spiegel (2019). Soziale Medien müssen Hasspostings künftig dem BKA melden. Retrieved from https://www.spiegel.de/netzwelt/netzpolitik/netzdg-soziale-medien-muessen-hasspostings-kuenftig-dem-bka-melden-a-1300007.html

[149] Der Tagesspiegel (2019). Innenminister wollen gefährliche Straftäter nach Syrien abschieben können. Retrieved from https://www.tagesspiegel.de/politik/union-und-spd-einig-innenminister-wollen-gefaehrliche-straftaeter-nach-syrien-abschieben-koennen/25302600.html

[150] Zeit (2019). Justizministerin will Politiker besser vor Hass schützen. Retrieved from https://www.zeit.de/politik/deutschland/2019-10/christine-lambrecht-justizministerin-kommunalpolitiker-schutz-anfeindungen-hatespeech-briefgeheimnis

[151] Tagesschau (2020). Gericht mit Kehrtwende im Fall Künast. Retrieved from https://www.tagesschau.de/inland/kuenast-beleidigung-103.html

152] Bundeskriminalamt (2020). Meldestelle für Hetze im Internet. Retrieved from https://www.bka.de/DE/KontaktAufnehmen/HinweisGeben/MeldestelleHetzeImInternet/meldestelle_node.html

[153] Bundeskriminalamt (2019). Fünfter Aktionstag gegen Hasspostings. Retrieved from https://www.bka.de/DE/Presse/Listenseite_Pressemitteilungen/2019/Presse2019/191106_AktionstagHasspostings.html

[154] Irish Department of Communications, Climate Action and Environment (2019). Regulation of Harmful Online Content and the Implementation of the revised Audiovisual Media Services Directive. Retrieved from https://www.dccae.gov.ie/en-ie/communications/consultations/Pages/Regulation-of-Harmful-Online-Content-and-the-Implementation-of-the-revised-Audiovisual-Media-Services-Directive.aspx

[155] Irish Department of Justice and Equality (2019). Hate Speech Public Consultation. Retrieved from http://www.justice.ie/en/JELR/Pages/Hate_Speech_Public_Consultation

[156] Schweppe, J., Haynes, A., & Carr, J. (2014). A life free from fear: Legislating for hate crime in Ireland: An NGO perspective. Retrieved from https://ulir.ul.ie/bitstream/handle/10344/4485/Schweppe_2014_crime.pdf

[157] The Irish Times (2019). Legislation against hate speech is ill-advised and counter-productive. Retrieved from https://www.irishtimes.com/opinion/legislation-against-hate-speech-is-ill-advised-and-counter-productive-1.4112340

[158] Irish Human Rights and Equality Commission (2019a). Recommendations on the Establishment of an Electoral Commission. Retrieved from https://www.ihrec.ie/app/uploads/2019/03/March-2019-IHREC-Submission-on-Establishment-of-Electoral-Commission.-1.pdf

[159] The Irish Times (2017). Courts Service reveals five convictions for hate crime since 1989. Retrieved from https://www.irishtimes.com/news/crime-and-law/courts-service-reveals-five-convictions-for-hate-crime-since-1989-1.3124352

[160] The Guardian (2016). Geert Wilders found guilty of inciting discrimination. Retrieved from https://www.theguardian.com/world/2016/dec/09/geert-wilders-found-guilty-in-hate-speech-trial-but-no-sentence-imposed

[161] Stam, J. (2019). The risky aspects of our hate speech laws. Retrieved from https://leidenlawblog.nl/articles/the-risky-aspects-of-our-hate-speech-laws

[162] Medoff, R., & Grobman, A. (2005). Holocaust Denial:A Global Survey - 2005. Retrieved from https://tandis.odihr.pl/bitstream/20.500.12389/19555/1/02249.pdf

[163] https://www.government.nl/topics/discrimination/reporting-discrimination

[164] European Commission against Racism and Intolerance (2019b). ECRI Report on the Netherlands (fifth monitoring cycle). Retrieved from https://rm.coe.int/fifth-report-on-the-netherlands/168094c577

[165] The Brussels Times (2020). 'Racism is a crime': Belgian PM announces action plan to fight online hate speech. Retrieved from https://www.brusselstimes.com/belgium/91442/racism-is-a-crime-belgian-pm-announces-action-plan-to-fight-online-hate-speech-sophie-wilmes-free-spech-online-comments-inciting-hate-xenophobia-racism-laws/

[166] https://www.unia.be/nl/actiedomeinen/samenleving/haatboodschappen/meld-het

[167] https://igvm-iefh.belgium.be/nl/activiteiten/discriminatie

[168] Unia (2017). Number of new discrimination cases at Unia rises by 20 percent. Retrieved from https://www.unia.be/en/articles/number-of-new-discrimination-cases-at-unia-rises-by-20-percent

[169] Unia (2020). The limits of free speech. Retrieved from https://www.unia.be/en/areas-of-action/media-and-internet/internet/the-limits-of-free-speech

[170] Reuters (2018). U.N. rights watchdog urges Hungary to halt hate speech, protect refugees. Retrieved from https://www.reuters.com/article/us-hungary-rights/u-n-rights-watchdog-urges-hungary-to-halt-hate-speech-protect-refugees-idUSKCN1HC1AJ

[171] Article 19 (2018b). Hungary: Responding to 'hate speech'. Retrieved from https://www.article19.org/wp-content/uploads/2018/03/Hungary_responding_to_HS.pdf

[172] http://english.nmhh.hu/article/187270/Internet_Hotline__hotline_to_report_harmful_content

DETECT THEN ACT