 DOI: http://dx.doi.org/10.5216/phi.v29i2.80137

# ONLINE HATE SPEECH IN THE NEW DIGITAL PUBLIC SPHERE: FREE SPEECH AND THE FACEBOOK[1]

**Paulo Barroso**[2,3]
pbarroso1062@gmail.com

**Abstract**: This article explores the social phenomenon of online hate speech in the contemporary digital public sphere, focusing on the intersection between free speech and the proliferation of misinformation on the Facebook. Two main objectives guide the research: first, to analyse how hate speech manifests itself in the new digital public sphere, where one of the main stages is on Facebook, exploring the dynamics that amplify the dissemination of harmful content; second, evaluate Facebook's role in the digital misinformation ecosystem, considering its impact on free speech. The methodology is theoretical-conceptual, following exploratory qualitative research, with bibliographical review and documentary research. The research explores the specific case of hate speech on Facebook, involving the dissemination of discriminatory messages and content against the Rohingya, a Muslim ethnic minority in Myanmar, highlighting patterns, narratives, and impacts. The research considers Facebook's policy responses and their effectiveness in mitigating hate speech. This article seeks to contribute to the critical understanding of the tensions between free speech and digital responsibility, offering valuable insights into the challenges of digital misinformation in the era of Facebook, as well as for a deeper understanding of the dynamics between free speech, social media networks, hate speech and digital misinformation.

**Keywords**: Facebook, free speech, misinformation, new digital public sphere, online hate speech.

---

[1] Recebido: 06-08-2024/ Aceito: 24-11-2024/ Publicado on-line: 11-12-2024.
[2] É professor no Instituto Politécnico de Viseu (IPV), Viseu, Portugal.
[3] ORCID: https://orcid.org/0000-0001-7638-5064.

## Introduction

"What is freedom of expression?
Without the freedom to offend, it ceases to exist".
Salman Rushdie, *Imaginary homelands: Essays and criticism 1981-1991*.

The invention of the internet has globally transformed social uses, customs, and interactions. The digital revolution has changed the way society understands and uses the media. The most significant changes have concerned the ability to exchange information (Young & Åkerström, 2016, p. 1). The internet has become the preeminent medium of global communication. Societies are on their way to merge into one and become a virtual *agora*, an e-sphere, a virtual public space. As Pelton (2000, p. 204) claims, to understand this new age and e-sphere in which we live online, we must recognize them as marked by the interactivity and globality based on electronic culture. In recent decades, social changes are faster and more profound. The concept of "phygital" (a fusion of the words "physical" and "digital") represents the evolution of the modern day-to-day experience influenced by technology, which adapts to our changes in social behaviour (Kipper & Rampolla, 2013, p. 68).

Digital communication technologies are constantly evolving and influencing, becoming more sophisticated (Schwab, 2016, p. 9). Social media have been weaponized in multiple ways, creating a "new reality no longer limited to the perceptual horizon" (Singer & Brooking, 2018, p. 13). Social media platforms like the Facebook continue to experience user growth, with millions of users worldwide. Video

content, live streaming, and short-form videos became dominant content formats.

The research questions guiding this approach are: how does online hate speech manifest in the contemporary digital public sphere, with a specific focus on Facebook, and what is the impact of this phenomenon on the intersection between free speech and the proliferation of misinformation? Additionally, how effective are Facebook's policy responses in mitigating hate speech, specifically in the case of the Rohingya, and what insights do these findings provide for the broader understanding of the tensions between free speech and digital responsibility in the era of Facebook? This article explores the complex interaction between free speech and hate speech in the digital public sphere, focusing on the Facebook platform. The main objective is to analyse how freedom of expression is challenged by the dissemination of hate speech online, considering the social implications in contemporary society marked by digital information. As the exploratory research questions and objectives indicate, the methodology followed in this approach is theoretical-conceptual, a critical, reflective and interpretative approach of the bibliography consulted on the topic/problem, the specific case of hate speech on Facebook, involving the dissemination of discriminatory messages and content against the Rohingya.

Online hate speech refers to any expression, conduct, or communication through digital platforms that offends, threatens, or insults individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, disability, etc. (Assimakopoulos, Baider & Millar, 2017, p. 81). The rise of social media and online communication

platforms has facilitated the rapid dissemination of hate speech, leading to various social and ethical concerns like the reproduction of racism, according to Titley. In *Is free speech racist?*, Titley (2020, pp. 8-12) states that defining "racism" is difficult because, on the one hand, "racism" and "free speech" are complex and disputed keywords and, on the other, the word "racism" is used to describe many different things. According to Titley, the issue of freedom of expression is always debatable and public, focusing on what can or cannot be said in relation to race. Therefore, racism has become central to intense disputes, offline and online, over the status and mission of freedom of expression. For Titley, the principle of free speech is organized in today's multicultural and intensely mediated societies. If Titley argues that contemporary discourse on freedom of expression reveals a lot about race and racism in contemporary societies, we add that this situation is mirrored in social media and online hate speech. Therefore, various discourses, narratives, and racist expressions are disseminated on social media under the pretext of "freedom of speech", considering that some individuals exploit the concept of free speech to propagate discriminatory ideas, hate speech, and racist ideologies. This misuse of freedom of speech can contribute to the spread of harmful narratives and reinforce discriminatory attitudes, particularly in Western countries. The challenge lies in finding a balance between protecting free speech and addressing the potential harm caused by the dissemination of racist content on social media platforms.

The unchecked dissemination of hate speech online can lead to the normalisation of discrimination, intolerance,

polarization, and violence between citizens, threatening not only social cohesion and respect for fundamental European values (Pinto, Carvalho, Magalhães, Alves, Bernardo, Lopes & Carvalho, 2023), as well as universal and estimable human values. Without any discriminatory sense towards other peoples and cultures, the idea of "European values" encompasses principles such as democracy, human rights, and the rule of law, which have played a crucial role in shaping Europe's identity. However, cultural diversity must be respected.

Hate speech encourages and incites prejudice, discrimination and oppression of individuals and groups belonging to minority social segments, due to their characteristics peculiar (Ribeiro, 2021, p. 180). The spread of hate is a complex problem, it is a violence and conflict. Hate speech has the potential to escalate into physical violence and conflict. History has shown that dehumanizing language can lay the groundwork for more severe forms of discrimination, persecution, and even violence against specific communities.

The spread of hate speech can undermine the principles of free expression by creating an atmosphere of intimidation (Brown, 2015, p. 72). Hate speech has migrated to online platforms, where it can spread rapidly and reach a global audience (Ermida, 2023, p. 55). This online dimension makes it challenging to control and may contribute to real-world harm, as seen in cases of cyberbullying, doxing, or coordinated harassment campaigns.

The hate also causes the erosion of trust and has impact on democracy. When individuals or groups are targeted based on their identities, it can create a sense of distrust and

fear, hindering cooperation and understanding among different segments of society. In democratic societies, the spread of hate speech can undermine the democratic process. It may contribute to the rise of extremist ideologies, influencing public discourse and potentially swaying political decisions in ways that are detrimental to inclusivity and diversity. According to Titley (2020, p. 20), "where racism is dominantly understood in terms of ideology and ideas, invocations of free speech have become fundamental to reshaping how racism is expressed and legitimized in public culture."

With the interconnectedness of the world through digital communication, hate speech has a global impact and fuel international tensions, contribute to conflicts between nations, and exacerbate existing geopolitical issues. Regulating hate speech is challenging due to the evolving nature of online platforms and the difficulties in defining and identifying hate speech. Striking a balance between freedom of expression and the prevention of harm remains a complex and ongoing challenge.

Addressing the problem of the spread of hate requires a comprehensive approach involving education, legal measures, community engagement, and responsible online behaviour to foster a more inclusive and tolerant society. As Titley (2020, p. 9) points out in *Is free speech racist?*, "free speech is never simply a subject of law or a question of legality". Therefore, "what is most at stake here is the shape rather than limits of speech in intensively mediated, multicultural societies" (Titley, 2020, p. 19). Based on the topic of online hate speech, this article problematizes, as the Rushdie epigraph invites, the complexity of the dichotomy free speech /

hate speech on social media and in the new digital public sphere.

In *Imaginary homelands: Essays and criticism 1981-1991*, Rushdie (1991, p. 396) emphasizes a fundamental aspect of freedom of expression, which is presented as an epigraph at the beginning of this article. Rushdie is suggesting that the true essence of freedom of expression lies in the ability to express ideas, even those that may be offensive to some. In other words, if individuals are not allowed to express opinions or ideas that might be considered offensive, then the concept of freedom of expression loses its significance.

Rushdie's perspective is controversial. However, the antagonistic perspective from Rushdie, who has been the victim of hate speech and attacks, is highlighted several times, as recently happened in August 2022, to demonstrate the complexity of the dichotomy free speech / hate speech in the public sphere. The opportunity of the free speech / hate speech debate was profoundly affected by the Rushdie case following the publication of *The Satanic Verses* (Maussen & Grillo, 2015, p. 8). The dichotomy free speech / hate speech is complex, but the issue of regulating online hate speech is even more complex and problematic and it is not limited to protect ethnic minorities from discriminatory and hate speeches.

## 1. Social media, fake news and hate speech

Choosing a social network as a case study depend on the research objectives and the questions one is seeking to answer. Each platform has its own characteristics and challenges. Therefore, there are several reasons to choose the

Facebook as a case study over other social networks: a) it is one of the largest social media platforms in the world, with billions of active users, and its broad user base offers a rich source of data and insights into online behaviours; b) Facebook supports a variety of content types (text, images, videos, and external links), which allows for a comprehensive analysis of the ways hate speech can be manifested in different formats; c) it has undergone several significant changes and challenges in terms of policies, algorithms, and features over the years related to hate speech, from the spread of fake news to political polarization, offering the opportunity to examine how approaches to hate speech have evolved; d) hate speech on this popular platform can have a significant impact on society, allowing a deeper understanding of how online interactions can influence the offline world. Furthermore, the user experience on Facebook, including social interactions, sharing content, and responding to messages, is unique. Studying hate speech in this context can reveal specific patterns.

Social media platforms have become the primary arena of online public life. Users can upload publicly available content. User-generated content (including videos, texts, images, links, etc.) is made available to, and then shared by, an audience selected by the user. From a free speech perspective, the activities of social media platforms are quite similar (Koltay, 2019, p. 146). Fox and Saunders (2019, p. 3) point out that "what is new is the growing mediation of the media as more and more of us get our news through content shared by other users on platforms such as Facebook". However, there is a connection between the social media platforms and

misinformation, fake news, and hate speech.

The social platforms' structure of rewarding users for habitually sharing information is the key reason why fake news spreads on social media. A study of more than 2,400 Facebook users suggests that platforms, more than individual users, have a larger role to play in stopping the spread of misinformation online (Ceylan, Anderson & Wood, 2023). Social media platforms and fake news or the spread of fake news are linked.

Misinformation is incorrect or misleading information. It can be simply an error in reporting or purposefully exaggerated, using clickbait headlines or out-of-context details to make a story harder to look away from. Online misinformation is growing. New technologies are making it easier than ever for anyone to substantially edit texts, photos, and videos to reflect a reality that doesn't really exist: "The move towards social media as a source for news has allowed misinformation to flourish" (Micich, 2023). Anyone with a social media account like Facebook can become a "news" source and spreading false information or hateful messages across the internet, i.e. across a broad spectrum, reaching many recipients online:

> The outrageous "fact" that blasts through audiences is louder, stickier, and more interesting than a follow-up correction. In the race between the false but interesting and the true but boring, the interesting story wins (Micich, 2023).

If there is a connection between social media platforms and fake news and their spread, there is also a more complex implication between social media, fake news (as a type of

misinformation) and hate speech. This relation is intricate and multifaceted. Social media platforms serve as powerful channels for information dissemination, connecting people globally (Ermida, 2023, p. 58). However, these platforms also face challenges related to the spread of misinformation and hate speech. Social media platforms cause amplification of information, it facilitates the rapid spread of information, including fake news and hate speech, due to its viral nature. False narratives and inflammatory content can reach a wide audience quickly.

Algorithms used by social media platforms may prioritize sensational or controversial content to engage users. This can inadvertently amplify fake news and hate speech, as these types of content often generate more interactions. Social media allows users to remain anonymous or pseudonymous, enabling the dissemination of hate speech without immediate accountability. Additionally, individuals may be exposed primarily to information that aligns with their existing beliefs, creating echo chambers that reinforce extreme views.

Social media platforms face the difficult task of moderating content at scale. Identifying and removing fake news and hate speech requires a balance between protecting free speech and preventing harm, and algorithms and human moderators may struggle to keep up with the volume of content. An example of moderation is the "Tasks" platform of Facebook, in which the social network allows the maintenance of dedicated teams for service moderation, such as the Community Well-Being Team, the Integrity Team, and the Hate-Speech Engineering Team, all to supposedly discuss, jointly, which individuals, phrases, hashtags or communities

to ban from the social network (Jacob, 2021, p. 94).

From the point of view of average consumers, the most widely used applications are social networking sites like Facebook (Keipi, Näsi, Oksanen & Räsänen, 2017, p. 5). Facebook is the most well-known and globally used social networking platform and "has also branched out from its original core product into a vast number of new, and in some cases innovative, social media services" (Keipi, Näsi, Oksanen & Räsänen, 2017, p. 8). Facebook is driven by a shared core idea dedicated to connecting users to both information and other people and it is merely providing free services that cater to some of the basic needs of their users. Like most internet-based companies operating on the same playing field, Facebook has had to build an efficient and productive revenue model to grow and remain relevant and popular. Thus, Facebook base their revenue on the information they gain from users of their services. The company keeps track of the online behaviour of those using their services, including their habits, interests, consumption, and product preferences, whom they interact with online, and so on. However, the most popular online services are also the most common sources of hate material and Facebook is one of the most common sites for witnessing hate material (Keipi, Näsi, Oksanen & Räsänen, 2017, p. 137).

Social media's global reach means that fake news and hate speech can have enlarged consequences. Addressing these challenges requires a comprehensive approach, involving improved content moderation, algorithmic transparency, media literacy initiatives, and cooperation between social media platforms, governments, and civil society to mitigate

the negative impact of fake news and hate speech on society.

One alarming attribute of the functioning of social media platforms is that they provide incentives for committing criminal acts in public, some of them are evil. A criminal can use Facebook to broadcast the shooting of another person, thereby capturing an audience that would not be available to him or her without the platform (Koltay, 2019, p. 156). Even though criminal acts are not attributable to social media, the public sphere has a different quality than the one created by traditional media (printed press, radio, and television). Today, most public debates are conducted online, and major social media platforms have user numbers and power to shape public opinions. Facebook gave birth to (the public sphere of) the twenty-first century (Koltay, 2019, p. 1). There are new forms of speech and the expansion of the public sphere based on the social media platforms.

The level of publicity of content published on these platforms depends on various factors, such as the size of the platform (user numbers), the number of contacts of the user concerned and the volume of other users forwarding (sharing) it (Koltay, 2019, p. 147). The form of published opinions has also changed. Users express their opinion by pressing Facebook's "like" button (Koltay, 2019, p. 147).

Facebook's Community Standards have recently become more demanding regarding the limitation of hate speech, with are relevant to the free speech (Koltay, 2019, p. 190). According to these standards, Facebook define hate speech as a "direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender

identity, and serious disability or disease". Facebook define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation.

In 1997, the Council of Europe defines the scope of the hate speech and the principles set out that apply to hate speech, in particular hate speech disseminated through the media: "the term 'hate speech' shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance" (Council of Europe, 1997, p. 107). Therefore, hate speech is considered an instrument of incitement to racism, discrimination, and prejudice (Costa, 2021, p. 43; Koltay, 2019, p. 36).

## 2. Online hate speech: the case of weaponizing the Facebook

Facebook is a good example of a stage for weaponizing content. As a form of online hate speech, weaponizing content refers to the intentional use of media, information, or online content to achieve certain objectives, often with a negative or harmful intent. It involves manipulating or exploiting content to influence public opinion, sow discord, or achieve specific political, social, or ideological goals (Brown, 2015, p. 61). The term "weaponizing content" gained prominence in the context of digital platforms and social media, where content can spread rapidly and reach a wide audience.

Weaponizing content can take various forms, such as misinformation and propaganda, i.e. creating or spreading false or misleading information to deceive or manipulate the audience (Brown, 2015, p. 62). This can include spreading

rumours, conspiracy theories, or fabricated news stories. Other forms of weaponizing include fake accounts and bot networks, manipulative narratives, image and video manipulation (as is currently happening with the war between Israel and Hamas), or doxing and personal attacks (revealing and disseminating private or personal information about individuals or groups with the intention of causing harm, harassment, or intimidation).

The weaponization of content can have serious consequences, including undermining trust in institutions, exacerbating social divisions, manipulating elections, inciting violence, and damaging reputations. It is important for individuals to critically evaluate the content they encounter, fact-check information, and be aware of the potential for manipulation or misinformation. Words are tools to convey intent, and as such has always been employed as manipulative instruments, whether positive or negative.

Online hate speech using the Facebook is mainly expressions of negative attitudes and thoughts through oral or printed words or images. It also includes pictorial or other non-verbal manifestations of ideas. The concepts of "free speech" and "free expression" are interchangeably (Brown, 2015, p. 5), while the terms "free speech" and "hate speech" are respectively equivalent to "free expression" and "hateful expression", clearly including non-verbal as well as verbal acts (Heinze, 2016, p. 19).

In current information and communication technological societies, the access to information is easy and immediate for anyone, anywhere and at any time. Social media platforms and the internet create a digital public space for

everyone to express and participate. Under these conditions, online speeches are expressed, transmitted, and received instantly, mediated by digital media. Immersion, immanence, and immediacy are the characteristics of this digital dimension (Baudrillard, 2005, p. 31). This is a new form of "democratic" tele-citizenship in a new digital public space, where there is a profusion of speeches and intense and constant flows of information, but the quality of which is questionable, as banalities predominate and become public, as well as false information and hate speech online.

Ensuring access to reliable information is a basic prerequisite for informed debate on all the challenges societies face, according to UNESCO's report *Survey on the impact of online disinformation and hate speech*. However, social media platforms have become the preferred source of information for a growing number of citizens. Therefore, concerns have been raised about the prevalence of falsehoods and hate speech, propelled by opaque algorithms that can favour engagement over factuality, and exacerbated by active exploitation by some political leaders and other actors (UNESCO, 2023a, p. 2).

> Overt intolerance and hate toward socially vulnerable groups continue to be expressed in social media and tolerated by most users (witnesses and victims), without significant consequences for the offenders (Pinto, Carvalho, Magalhães, Alves, Bernardo, Lopes & Carvalho, 2023).

Social media contributes to the rise of hate speech by offering the capacity to reach a wider audience. Although many social media platforms have implemented control mechanisms to reduce online misbehaviour like hate speech (e.g.

using detection algorithms and report buttons), hate speech often goes undetected, and citizens perceive the inefficacy of these control mechanisms to combat hate speech (Pinto, Carvalho, Magalhães, Alves, Bernardo, Lopes & Carvalho, 2023).

Besides the false information, the phenomenon of hate speech is also widespread: 67% of internet users have encountered it online and they overwhelmingly believe that hate speech is most prevalent on Facebook (58%). "According to citizens, it is primarily LGBT+ people (33%) and ethnic or racial minorities who are victims of online hate speech in their country, although there are significant variations between countries." (UNESCO, 2023a, p. 8).

According to UNESCO's report *Survey on the impact of online disinformation and hate speech*, Facebook is the most prevalent sources of false information and/or hate speech (UNESCO, 2023a, p. 24). 54% think online platforms are not doing enough to combat hate speech (UNESCO, 2023a, p. 26).

Free speech is a problem when one offends and harms other people. Given the generic characteristics of online hate speech, which is still a use and modality of free speech, the problem takes on greater proportions (Brown, 2015, p. 243). Online hate speech can manifest in various forms, and its characteristics can vary, but some common elements include: a) targeted content, hate speech typically targets individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, disability, or other protected characteristics; b) offensive language, hate speech often involves the use of derogatory and offensive language, slurs, or

discriminatory terms to demean and dehumanize the targeted individuals or groups; c) intolerance and prejudice, hate speech reflects intolerance and prejudice, promoting negative stereotypes and reinforcing discriminatory beliefs about certain communities; d) incitement to violence or discrimination, some instances of hate speech go beyond expressing offensive views and directly call for violence, discrimination, or harm towards the targeted individuals or groups; e) online platforms, hate speech can occur on various online platforms, including social media, forums, comment sections, and other digital spaces where users can communicate; f) anonymity and impersonation, some individuals engaging in hate speech may hide behind anonymity or create fake profiles to avoid accountability for their words and actions; g) viral spread, hate speech can spread rapidly online, gaining traction through social media shares, retweets, or other forms of digital engagement; h) hate speech can be conveyed through text, memes, images, or other multimedia content, making it more visually impactful; i) trolling and harassment, hate speech is often intertwined with trolling and online harassment, where individuals intentionally provoke and intimidate others to create a hostile online environment; j) impact on real-world actions, in some cases, online hate speech has been linked to real-world violence or discrimination, emphasizing the potential harm it can cause beyond the digital realm.

However, the definition and interpretation of hate speech can vary, and what one person considers offensive may not be universally agreed upon. Platforms and communities often have their own guidelines and policies to address

hate speech and maintain a safe online environment. Hate speech is "any kind of communication in speech, writing or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are" (UNESCO, 2023b), i.e. based on religion, ethnicity, nationality, race, color, descent, gender or other identity factor. More specifically, racist hate speech includes all the specific speech forms directed against groups based on the discrimination on grounds of race, color, descent, or national or ethnic origin (Council of Europe, 2022, p. 55). Racist hate speech can take many forms and is not confined to explicitly racial remarks. It is a form of discourse directed against others, which rejects the fundamental principles of human rights, dignity and equality, and aims to weaken the position of individuals and groups in society (Council of Europe, 2022, p. 55).

The virtual world is not just new. With the advent of the internet and the new social media, there is the amplification of old and bad habits which are as challenging as the new ones (Oliveira & Leite, 2022, p. 389). This is the case, for example, of the online hate speech. This type of manifestation is not the result of new media. In the past, expressions of intolerance used to be restricted to the conversation circles of certain groups, with a limited reach, circulation, and conservation of offenses. Hate speech has always been harmful, but now what is new is that it also takes place online, predominantly on social media. Koltay refers these are "old problems in a new context". "People violate the same norms online and offline, but in different ways. Online communication has made it easier to commit various prohibited,

illegal, or outright criminal acts, thereby increasing the scale of acts" (Koltay, 2019, p. 72), like hate speech. "The problems themselves are not new; these are old ones reappearing in a new context, some of which nonetheless require new answers from the law." (Koltay, 2019, p. 72). Therefore, hate speech and the propagation and support of terrorism are often mentioned when the negative aspects of online communication are discussed, concludes Koltay.

However, with new media, hate messages arrive quickly and reach a greater number of people, reinforcing prejudices and creating an even more polarized virtual and physical environment. In the internet, there is a dual aspect of the digital environment: i) the "invisibilization" of users under the cloak of anonymity facilitates or even encourages the exposure of prejudiced thoughts, which may not would be manifested in person; and ii) the magnitude of the speech on networks, which quickly reaches thousands or millions of people and provides the extremely fast and uncontrollable spread of hate speech (Oliveira & Leite, 2022, p. 389). Since hate speech is precisely a segregating speech, its greater dissemination directly contributes to an atmosphere oppressive to minorities, which increases the harmful nature of the conduct: with each share, the more serious the damage caused for this type of violation.

## 3. Case study of hate speech on Facebook

A famous case of hate speech on Facebook involved the dissemination of discriminatory messages and content against the Rohingya, a Muslim ethnic minority in Myanmar. Hate speech on Facebook has been cited as a

contributing factor to the ethnic violence and humanitarian crisis unfolding in the region. During the conflict in 2017, Buddhist extremist groups and other Facebook users disseminated inflammatory messages, fake images and biased news, inciting hostility against the Rohingya. These messages contributed to the escalation of violence, resulting in mass displacement, killings, and widespread human rights abuses.

Facebook has faced significant criticism for its alleged inaction in effectively combating hate speech on the platform during this period. The company was accused of not taking quick enough action to remove harmful content and not having adequate policies to deal with hate speech in sensitive contexts.

Intolerance and hate can flourish on the internet, taking advantage of the characteristics of the internet and social media platforms. The internet is a virtual public space that provides users with the capacity for expressing their views and communicating without limits, and "typically (though not always) without control; the online setting makes it easy for users to hide their identity (in whole or in part) and, in some cases, even to hide their location and activity" (Assimakopoulos, Baider & Millar, 2017, p. 11). The social media platforms and the internet are generally perceived as media and tools of hate and propaganda within public sphere that is free of restrictions (Assimakopoulos, Baider & Millar, 2017, p. 56).

Even though the "terms of service" of most relevant platforms, such as Facebook, do stipulate that it is prohibited to post content that is hateful, unlawful, harmful, defamatory, obscene, tortuous, or invasive of one's privacy, the time

it usually takes to remove such content has been an issue of growing concern (Assimakopoulos, Baider & Millar, 2017, p. 12).

This case against the Rohingya highlights the challenges social media platforms face in dealing with the spread of hate speech and misinformation, especially in areas prone to ethnic and religious conflicts. It also highlights the responsibility of online platforms to monitor and moderate content to avoid harmful real-world consequences. According to the allegations against Facebook, the company's algorithms amplified hate speech against the Rohingya people, and it failed to invest in local moderators and fact checkers; it failed to take down specific posts inciting violence against Rohingya people; and it did not shut down specific accounts or delete groups and pages that were encouraging ethnic violence (Hatano, 2023, p. 136; Milmo, 2021).

Amnesty International accused Facebook's parent company Meta of having "substantially contributed" to human rights violations perpetrated against Myanmar's Rohingya ethnic group. In a recent report from 2022, Amnesty International claims that Facebook's algorithms "proactively amplified" anti-Rohingya content and that Meta ignored civilians' and activists' pleas to curb hatemongering on the social media platform while profiting from increased engagement (Guzman, 2022). Facebook's seeming inability to manage online hate speech and misinformation reveal the major problem of the spread of hate speech and discrimination around the world. Therefore, measures are needed to be implemented on social media to prevent such problems and to provide reparations to affected communities.

In addition to the inability to prevent and manage online hate speech and misinformation, Facebook even promotes these problems, asserts the Amnesty International Report, since the Rohingya have been persecuted by Myanmar's Buddhist majority for decades, but Facebook has exacerbated the situation (Guzman, 2022). The Amnesty International claims that the Tatmadaw (Myanmar's armed forces) used Facebook to boost propaganda against the Rohingya and to amass public support for a military campaign of rampant killings, rape and arson targeting the predominantly Muslim minority (Guzman, 2022). In the report, entitle "Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya", Amnesty International highlight that Meta's algorithms proactively amplified and promoted content which incited violence, hatred, and discrimination against the Rohingya – pouring fuel on the fire of long-standing discrimination and substantially increasing the risk of an outbreak of mass violence (Amnesty International, 2022, p. 74). The report concludes that Meta substantially contributed to adverse human rights impacts suffered by the Rohingya and has a responsibility to provide survivors with an effective remedy (Amnesty International, 2022, p. 9). In its report, Amnesty International concludes that Meta was made aware as early as 2012 of how its engagement-based algorithms were contributing to serious real-world harm in Myanmar (Amnesty International, 2022, p. 26).

According to the *Time*, "a U.N. fact-finding mission in 2018 determined that Facebook had been a 'useful instrument' for vilifying the Rohingya in Myanmar 'where, for most users, Facebook is the internet'" (Guzman, 2022).

Posteriorly, Meta releases a commissioned human rights impact report in which it admitted that the company was not doing enough to stop the sowing of hatred against the Rohingya on the platform, but it has invested in more Burmese-speaking content moderators and improved technology to address the problem (Guzman, 2022). "Facebook was weaponised by military leaders and nationalists to incite ethnic tensions, resulting in brutal violence against Rohingya Muslims during a campaign of ethnic cleansing in 2018" (Hatano, 2023, p. 129).

"Facebook admitted in 2018 that it had not done enough to prevent the incitement of violence and hate speech against the Rohingya, the Muslim minority in Myanmar" (Milmo, 2021). As an independent report commissioned by the company found, Facebook has become a means for those seeking to spread hate and cause harm, and posts have been linked to offline violence (Milmo, 2021).

"Facebook is struggling to respond to criticism over the leaking of users' private data and concern about the spread of fake news and hate speech on the platform" (Hogan & Safi, 2018). The *Time* adds that some of Facebook's well-intentioned measures have backfired. For instance, in 2014, Facebook supported a civil society-led campaign against hate speech by creating virtual "sticker packs" for users to post in response to violent and discriminatory content (Guzman, 2022). However, Facebook's algorithm registered the responses as engagement and further increased the visibility and spread of the harmful content.

The Amnesty International report says Meta's content moderation practices have been no match for the sheer

amount of algorithmically boosted inflammatory, anti-Rohingya sentiment (Guzman, 2022). "In mid-2014, Meta staff admitted that they only had one single Burmese-speaking content moderator devoted to Myanmar at the time, based in their Dublin office" (Amnesty International, 2022, p. 7). Amnesty International claims the Meta had only one Burmese-speaking content moderator to monitor the posts of Myanmar's 1.2 million active users at the time (Guzman, 2022). According to the Amnesty International, Rohingya refugees recalled how their reports of posts on the platform thought to violate Facebook's community standards were often ignored or rejected. An internal document from July 2019, cited by the Amnesty International report, said that action was only taken against "approximately 2% of the hate speech on the platform" (Amnesty International, 2022, p. 8).

In November 2018, Meta announced several measures against online hate speech. One of the measures is that it had onboarded 99 Myanmar language content moderators, but anti-Rohingya sentiment has nevertheless flourished on Facebook, points out the rights group (Amnesty International, 2022, p. 8). "The limitations of global approaches to content moderation become particularly pronounced in relation to the moderation of non-English language content" (Hatano, 2023, p. 147). In April 2022, Facebook supports posts in 110 languages; however, it only has the capacity to review content in 70 languages (Allen, 2022). For example, Amnesty International determined that in 2020, a video of an anti-Rohingya Buddhist monk amassed 70% of its views on Facebook through "chaining" (the automatic playing of a recommended video after one ends) even though Meta had banned

the monk's Facebook profile for hate speech in 2018 (Guzman, 2022). "Chaining" is an example of "non follower-based distribution" and refers to video content "which autoplays after a video is complete and suggests what's 'Up Next' to viewers" (Facebook, 2022). The "escalation" of the video by the leading anti-Rohingya hate figure U Wirathu shows the problem of the spread of online hate speech and discrimination on the Facebook. "The video was reported for violating community standards" and "Meta found that its algorithms had been actively promoting the video by this hate figure" (Amnesty International, 2022, p. 46). This case exemplifies Meta's active amplification of harmful content in Myanmar, and it also highlights the weaknesses of Meta's efforts at improved content moderation in Myanmar as recently as 2020, concludes the Amnesty International Report (2022, p. 46).

"Facebook's negligence facilitated the genocide of Rohingya Muslims in Myanmar after the social media network's algorithms amplified hate speech and the platform failed to take down inflammatory posts, according to legal action launched in the US and the UK" (Milmo, 2021). For Bakali (2021), the phenomenon of Islamophobia (anti-Muslim racism) contributed and is even at the origin of a genocide of Rohingya Muslims in Myanmar, whose consequences are ongoing. Such Islamophobia increasingly has come to encompass systemic racism and anti-Muslim violence in Myanmar as a "war on terror" used to sanitise more recent violence against the Rohingya. "War on Terror" logic has increasingly become normalized in contemporary public and political discourse surrounding Muslims, and was central to the state of

Myanmar's justifications for the Rohingya genocide, as Bakali and Hafez (2022) point out in the introduction of *The rise of global Islamophobia in the War on Terror*.

"To date, Meta has championed the use of artificial intelligence to improve detection of harmful content" (Guzman, 2022). However, this measure is falling short since Facebook's AI-approved advertisements containing hate speech targeting Rohingya. "An investigation by Global Witness in March 2022 found that Meta's content moderation algorithms were still failing to detect blatant anti-Rohingya and anti-Muslim content on the platform" (Amnesty International, 2022, p. 37). Nevertheless, automated tools are being used by internet platforms to shape the content we see and influence how this content is delivered to us. Paradoxically, Artificial Intelligence serves both to try to control the spread of hate speech online and for the opposite, i.e. to spread offensive speech against the dignity of people, groups or communities across a wide spectrum.

The Meta is taking more steps to address human rights issues stemming from its platform's use in Myanmar. In February of 2021, amid a military takeover of Myanmar, Meta banned the Tatmadaw and other state-sponsored entities on Facebook and Instagram and in its July 2022 Human Rights Report, the company outlined other Myanmar-specific measures it's taken, such as a "Lock your profile" feature to provide users who may be targeted for harassment or violence with greater privacy (Guzman, 2022).

It is extremely difficult, not to mention impossible, to implement effective regulation in the use of the internet in general and social networks, as this need has been talked

about for a long time and there is still no consensus or agreement and harmony with the different national policies on access to information and digital content. Different countries have diverse hate speech laws rooted in their historical, traditional, and constitutional roots (Hatano, 2023, p. 130). However, "in just the last few years, lawmakers and advocates around the world have been trying to rein in social media companies, though it's a challenging and sometimes controversial endeavor" (Guzman, 2022).

Social media platforms have the characteristic of allowing widespread access to their platforms and free uses, respecting their operating guidelines and users' privacy rights. On the other hand, social media networks have been selling a narrative that says, "if you regulate us, if you address the most harmful aspects of our business, you will fundamentally make the internet inaccessible for all the reasons that people depend on it" (Guzman, 2022), on the other. Social media networks are effective and influential, shaping how users think and act, shaping the ways of seeing and understanding others and the world in general; they shape how human society works nowadays and how we interact with each other.

Therefore, global social media companies, including Facebook, have faced criticism for their failure to effectively remove harmful content. This case of hate speech on Facebook involving the dissemination of online hate speech against the Rohingya highlight the profound threats posed by hate speech to the fundamental rights of individuals and to public goods such as peace and social stability exacerbated in the online public sphere and digital age. "As a result, the need to strike a balance between protecting freedom of expression

and privacy while regulating hate speech has come to a critical crossroads that requires urgent attention" (Hatano, 2023, p. 130). This ongoing challenge underscores the imperative for comprehensive and effective policies that address the complexities of hate speech in the digital era while safeguarding essential democratic values.

## Conclusions

The problem with online hate speech is its potential to contribute to real-world harm by fostering discrimination, prejudice, and violence. Hate speech and its spread on online platforms can have a profound impact on targeted individuals, contribute to the marginalization of specific groups, and even incite violence in extreme cases like the Rohingya. Additionally, the online nature of hate speech makes it challenging to regulate and control effectively.

The issue of online hate speech is directly associated with both free speech and the widespread, free, and unregulated use of social media The relationship between free speech and online hate speech involves a delicate balance. While free speech is a fundamental right, online hate speech can incentivize discrimination and harm. Striking a balance requires addressing harmful incentives without compromising the principles of free speech, often involving measures to curb incitement, harassment, or violence, while safeguarding diverse perspectives and fostering constructive dialogue. Achieving this balance is challenging and involves ongoing debates about the limits of free speech in the digital era. Therefore, this article concludes that hate speech finds a privileged platform to manifest itself on digital platforms,

benefiting from the difficulty of regulating speech on the internet and the inalienable right to freedom of expression.

The online spread of hate speech presents a range of distinct problems, reflecting the unique dynamics of the digital realm. One of the key issues associated with the online spread of hate speech is the global reach and amplification. Hate speech can spread rapidly and reach a global audience through online platforms. The speed and scale at which hateful content can be disseminated make it challenging to contain and counteract.

The online spread of hate speech includes anonymity and lack of accountability. Online platforms often allow users to remain anonymous or use pseudonyms, providing a shield for those who engage in hate speech. This anonymity can reduce accountability, as individuals may feel emboldened to express harmful views without facing real-world consequences. The internet allows users to remain hidden and this facilitates the production, transmission, and consumption of false, illegal, and harmful content. Online hate speech is not radically different from offline hate speech. Online hate speech is characterized by its anonymity, the speed of spread, itinerancy, permanence, and complex cross-jurisdictional character. These unique characteristics present unprecedented challenges to regulate online hate speech, particularly on social media platforms. Anonymous, pseudonymous characters or false digital identities can easily accelerate the destructive behaviour of people engaging in online activities. Online information travels easily across multiple platforms. Online haters can maintain their digital offensive behaviour even if their posts are taken down or the

social media network close their access; they simply can migrate to another social media network with less stringent regulations on hate speech. Consequently, harms of online hate speech can grow exponentially on the social media networks, counting on the aggravation of harmful content can remain online indefinitely.

One of the main pretexts behind online hate speech is that perpetrators take advantage of the perceived anonymity of the internet and therefore feel free to utter and spread insults and vexations, which are easier to do online. The current rise of online hate speech and other online misbehaviours is a result of the disinhibition that people feel online, which increases the likelihood that someone will use hate speech and decreases witnesses' accountability, leading to under-reporting. Furthermore, there is a difference between the occurrence and reporting of hate speech.

Social media algorithms may unintentionally amplify hate speech. The algorithms that determine content visibility often prioritize engagement, and provocative or controversial content tends to generate more interaction, leading to a potential feedback loop that amplifies hate speech.

Online platforms can create echo chambers, where users are exposed primarily to content that aligns with their existing beliefs. This can contribute to the polarization of society, as individuals are less likely to encounter diverse perspectives and more susceptible to the reinforcement of prejudiced views.

Hate speech often involves the spread of misinformation, further complicating efforts to address the issue. False narratives and conspiracy theories may be intertwined

with hateful content, making it challenging to distinguish between legitimate expression and harmful speech. Furthermore, the prevalence of hate speech can erode trust in online spaces. Users may become hesitant to engage in open discourse, and the fear of harassment can deter individuals from participating in online discussions, limiting the potential for constructive dialogue.

Addressing the online spread of hate speech requires a multi-faceted approach, involving collaboration between online platforms, users, policymakers, and civil society to develop effective strategies for content moderation, user education, and the promotion of positive online behaviours.

In conclusion, this approach sheds light on the complex and pressing issue of online hate speech within the contemporary digital public sphere, with a particular emphasis on the role of Facebook. The analysis revealed intricate dynamics surrounding the manifestation of hate speech, especially in the context of the Rohingya case, where discriminatory messages and content were disseminated. The findings underscore the urgency of addressing the challenges posed by digital misinformation and its impact on free speech, emphasizing the need for effective strategies in the face of evolving online communication.

The evaluation of Facebook's role in the digital misinformation ecosystem highlighted the platform's significant influence in shaping public discourse. While the study recognizes Facebook's efforts through policy responses to mitigate hate speech, it also underscores the ongoing challenges and the need for continuous adaptation to the evolving nature of online threats.

By adopting a theoretical-conceptual methodology, this research contributes to a deeper understanding of the tensions between free speech and digital responsibility. The insights gained provide valuable knowledge for policymakers, social media platforms, and scholars grappling with the multifaceted issues surrounding hate speech and misinformation in the digital era. As we navigate the complexities of the digital public sphere, this study advocates for a comprehensive approach that balances the preservation of free speech with the responsibility to curb the dissemination of harmful content. Ultimately, fostering a critical awareness of the dynamics between free speech, social media networks, hate speech, and digital misinformation is crucial for building a more informed and resilient digital society.

**Resumo:** Este artigo explora o fenômeno social do discurso de ódio online na contemporânea esfera pública digital, com foco na intersecção entre a liberdade de expressão e a proliferação de desinformação no Facebook. Dois objetivos principais orientam a investigação: primeiro, analisar como o discurso de ódio se manifesta na nova esfera pública digital, em que um dos palcos principais é o Facebook, explorando as dinâmicas que amplificam a disseminação de conteúdo prejudicial; segundo, avaliar o papel do Facebook no ecossistema da desinformação digital, considerando o seu impacto na liberdade de expressão. A metodologia é teórica-conceitual, seguindo uma pesquisa exploratória qualitativa, com revisão bibliográfica e pesquisa documental. A pesquisa qualitativa explora o caso específico de discurso de ódio no Facebook, envolvendo a disseminação de mensagens e conteúdos discriminatórios contra os Rohingya, uma minoria étnica muçulmana em Mianmar, destacando padrões, narrativas e impactos. A pesquisa considera as respostas das políticas do Facebook e a sua eficácia na mitigação do discurso de ódio. Procura-se compreender as tensões entre a liberdade de expressão e a responsabilidade digital, oferecendo insights sobre os desafios da desinformação digital, e discutir as dinâmicas entre liberdade de expressão, redes sociais, discurso de ódio e desinformação digital.

**Palavras-chave:** Desinformação, discurso de ódio online, Facebook, liberdade de expressão, nova esfera pública digital.

## References

ALLEN, Caroline. Facebook's content moderation failures in Ethiopia. *Council on Foreign Relations*, 2022. Retrieved from https://www.cfr.org/blog/facebooks-content-moderation-failures-ethiopia.

AMNESTY INTERNATIONAL. *Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya*, n. 16/5933, 2022. Retrieved from https://www.amnesty.org/en/documents/asa16/5933/2022/en/

ASSIMAKOPOULOS, Stavros., BAIDER, Fabienne & MILLAR, Sharon. *Online hate speech in the European Union: A discourse-analytic perspective*. Cham: Springer, 2017.

BAKALI, Naved. Islamophobia in Myanmar: the Rohingya genocide and the "war on terror". *Race & Class*, v. 62, n. 4, p. 53-71, 2021. Retrieved from https://journals.sagepub.com/doi/10.1177/0306396820977753

BAKALI, Naved & HAFEZ, Farid. *The rise of global Islamophobia in the War on Terror*. Manchester: Manchester University Press, 2022.

BAUDRILLARD, Jean. *The intelligence of evil or the lucidity pact*. Oxford: Berg Publishers, 2005.

BROWN, Alex. *Hate speech law:* A philosophical examination. London: Routledge, 2015.

CEYLAN, Gizem., ANDERSON, Ian & WOOD, Wendy. Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, v. 120, n. 4, 2023. Retrieved from https://www.pnas.org/doi/10.1073/pnas.2216614120

COSTA, Alessandra. *Liberdade de expressão vs. Discurso de ódio:* Uma questão de (in)tolerância. Belo Horizonte: Dialética Editora, 2021.

COUNCIL OF EUROPE. *Recommendation no R (97) 20, of the Committee of Ministers to Member States on 'Hate Speech'*, 1997.

COUNCIL OF EUROPE. *Combating hate speech in the media in the Republic of Moldova: Guide for assessing and processing hate speech cases*, 2022.

ERMIDA, Isabel. Distinguishing online hate speech from aggressive speech: a five-factor annotation model. In: ERMIDA, Isabel (ed.). *Hate speech in social media*. Palgrave Macmillan, 2023. p. 35-76.

FACEBOOK'S COMMUNITY STANDARDS (n.d.). Retrieved from https://www.facebook.com/communitystandards/

FACEBOOK, Growing your live community on Facebook, 2022. Retrieved from facebook.com/fbgaminghome/creators/growing-your-live-community

FOX, Carl; SAUNDERS, Joe. *Media ethics, free speech, and the requirements of democracy*. London: Routledge, 2019.

GUZMAN, Chad. Meta's Facebook algorithms "proactively" promoted violence against the Rohingya, new Amnesty International Report asserts. *Time*, 2022, 28 September. Retrieved from https://time.com/6217730/myanmar-meta-rohingya-facebook/.

HATANO, Ayako. Regulating online hate speech through the prism of human rights law: The potential of localised content moderation. *The Australian Year Book of International Law Online*, v. 41, n. 1, p. 127-156, 2023. Retrieved from https://doi.org/10.1163/26660229-04101017

HEINZE, Eric. *Hate speech and democratic citizenship*. Oxford: Oxford University Press, 2016.

HOGAN, Libby. & SAFI, Michael. Revealed: Facebook hate speech exploded in Myanmar during Rohingya crisis. *The Guardian*, 2018, 3 April. Retrieved from https://www.theguardian.com/world/2018/apr/03/revealed-facebook-hate-speech-exploded-in-myanmar-during-rohingya-crisis.

JACOB, Raphael. *Liberdade de expressão, internet e telecidadania:* Uma visão crítica acerca do exercício da cidadania nos meios digitais. São Paulo: Editora Literando, 2021.

KEIPI, Teo, NÄSI, Matti, OKSANEN, Atte & RÄSÄNEN, Pekka. *Online hate and harmful content:* Cross-national perspectives. London: Routledge, 2017.

KIPPER, Gregory & RAMPOLLA, Joseph. *Augmented reality: An emerging technologies guide to AR*. Syngress/Elsevier, 2013.

KOLTAY, András. *New media and freedom of expression:* Rethinking the constitutional foundations of the public sphere. Oxford: Hart Publishing, 2019.

MAUSSEN, Marcel & GRILLO, Ralph. *Regulation of speech in multicultural societies*. London: Routledge, 2015.

MICICH, Anastasia. How misinformation on social media has changed news. U.S. Pirg Education Fund, 2023, 14 August. Retrieved from https://pirg.org/edfund/articles/misinformation-on-social-media/

MILMO, Dan. Rohingya sue Facebook for £150bn over Myanmar genocide. *The Guardian*, 2021, 6 December. Retrieved from https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence.

OLIVEIRA, Júlia & LEITE, Roberta. Direito ao esquecimento e o caso Richthofen: Qual deve ser o futuro do passado? In: SCHREIBER, Anderson, MORAES, Bruno; TEFFÉ, Chiara (org.). *Direito e mídia:* Tecnologia e liberdade de expressão. São Paulo: Editora Foco, 2022. p. 381-430.

PELTON, Joseph. *E-Sphere: The rise of the world-wide mind*. London: Quorum Books, 2000.

PINTO, Isabel; CARVALHO, Catarina; MAGALHÃES, Mariana; ALVES, Sara; BERNARDO, Márcia; LOPES, Paula; CARVALHO, Cátia. Understanding the rise in

online hate speech in Portugal and Spain: a gap between occurrence and reporting. *The Social Observatory of the "la Caixa" Foundation*, 2023. Retrieved from https://oobservatoriosocial.fundacaolacaixa.pt/en/-/compreender-o-crescimento-do-discurso-de-odio-online-em-portugal-e-espanha-um-hiato-entre-a-ocorrencia-e-a-denuncia

RIBEIRO, Raisa. *Discurso de ódio, violência de gênero e pornografia:* Entre a liberdade de expressão e a igualdade. Feminismo Literário, 2021.

RUSHDIE, Salman. *Imaginary homelands:* Essays and criticism 1981-1991. New York: Granta Books, 1991.

SINGER, Peter W.; BROOKING, Emerson. *LikeWar:* The weaponization of social media. Boston: Houghton Mifflin Harcourt Publishing, 2018.

TITLEY, Gavan. *Is free speech racist?* London: Polity Press, 2020.

UNESCO. *Survey on the impact of online disinformation and hate speech*, 2023a. Retrieved from https://www.ipsos.com/sites/default/files/ct/news/documents/2023-11/unesco-ipsos-online-disinformation-hate-speech.pdf

UNESCO. What you need to know about hate speech, 2023b. Retrieved from https://www.unesco.org/en/countering-hate-speech/need-know

YOUNG, Philip; ÅKERSTRÖM, Marja. Meet the digital naturals. In: COOMBS, W. Timothy; FALKHEIMER, Jesper; HEIDE, Mats; YOUNG, Philip (eds.). *Strategic communication, social media and democracy:* The challenge of the digital naturals. London: Routledge, 2016. p. 1-10.