SYbIL | Spanish Yearbook of International Law

Some international aspects in the fight against online harmful content*

María Chiara Marullo**

Abstract: The growing prevalence of hate speech and incitement to discrimination, violent content, targeting migrants, minority, ethnic communities, and other vulnerable groups, as well as its impact on mental health to harm children, teenagers and moderators, poses significant challenges to democracy and security across the globe. States are bound by international law to combat racial discrimination, xenophobia and incitement to hatred. These standards demand that states take decisive actions against speech that incites national, racial, or religious hatred, discrimination, or violence. However, regulating those contents, especially on social media platforms, becomes increasingly complex as it involves balancing fundamental rights such as freedom of expression with the responsibilities of multinational corporations. The regulation faces considerable challenges in addressing these issues in terms of competent jurisdiction and the responsibilities of the private actors involved. This paper explores these challenges in regulating harmful content online, offering a preliminary analysis of extraterritorial measures adopted by companies, and highlighting the inadequacies of self-regulation. We will specifically examine legal action taken against Meta, emphasizing the need for an effective international legal framework to address these global issues.

Keywords: Hate speech, online harmful content, social media, self-regulation, Meta

Received: November 19, 2024 *Accepted:* December 23, 2024

(A) PREMISES

Social networks, such as Meta, provide efficient platforms for spreading users' ideas potentially rising to the level of harmful content online. Platforms have some internal policies and codes of conduct to regulate and address hate speech and violent content. While these internal policies and the codes of conduct offer a glimmer of hope for controlling hate on the internet, challenges remain due to issues of jurisdiction and technological complexities (like mirror sites), making online regulation an especially daunting task.

Hate speeches and harmful contents are most contentious issues in legislation due to the potential conflict with other fundamental rights[1]. As Professor Camarero Suárez

* The present article is being published as part of the research on hate speech in the framework of the Project: UJI-2024-02 Derecho, matrimonio y factor religioso: nuevos retos, and in the framework of the Project CIGE/2022/63: Oportunidades y desafíos en la implementación de las normas de debida diligencia empresarial en materia de derechos humanos y medio ambiente.

** Associate Professor (Profesora Contratada Doctora) of Private International Law, University of Jaume I (UJI). IP of the Project: CIGE/2022/63 Oportunidades y desafíos en la implementación de las normas de debida diligencia empresarial en materia de derechos humanos y medio ambiente, Generalitat Valenciana, coordinador of the REDHEXATA, more information at: redhexata.com. Coordinator of the research group: Grup d'Investigació en Drets Humans i Drets Fonamentals, at UJI.

1 At the supranational level, Article 10 of the European Convention on Human Rights establishes the right to freedom of expression, but this right is not absolute. It may be subject to restrictions in a democratic

notes, case law consistently emphasizes the need for balancing conflicting rights, seeking maximum protection through a proportionality test that weighs and limits these rights accordingly[2]. In this context, hate speech and violent content present a challenge for defining the boundaries of free expression[3].

While this topic cannot be fully explored here, our research starts from the premise that harmful online content includes any form of expression targeting discriminated or affected groups based on gender, sexual orientation, ethnicity, religion, or other personal or social factors, often focusing on traditionally excluded minorities. Such discourse often originates from radicalized sectors of society, fostering stigmatization and discrimination. It can also harm the mental and physical health of millions of children, teenagers, and moderators, undermining democratic coexistence, social cohesion, and intercultural integration. This study focuses on instances where hostile expressions and contents incite hate against vulnerable groups, and discrimination based on what are known as suspect categories[4]. We specifically examine social media platforms as vehicles for this harmful speech, given their rapid spread and regulatory challenges at both supranational and national levels.

Our research aims to explore the role of social media in the propagation of hate speech or violent content and assess the extraterritorial measures that companies have taken to curb its spread[5]. We will then analyze different lawsuits against Meta an emblematic case of the current challenges of regulations , arguing that the current lack of effective supranational norm and the failure of self-regulatory measures highlight the need for a more robust framework to mitigate the negative impacts of online platforms[6].

---

society for reasons such as national security, public safety, or the protection of others' rights. More information at: https://www.echr.coe.int/documents/d/echr/convention_ENG.Similarly, Article 20 of the International Covenant on Civil and Political Rights prohibits advocacy of national, racial, or religious hatred that incites violence, discrimination, or hostility. More information at: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights. Moreover, Article 13.5 of the American Convention on Human Rights bans propaganda for war and hate speech targeting individuals or groups based on race, color, religion, or national origin. More information at: https://www.oas.org/dil/treaties_b-32_american_convention_on_human_rights.pdf. Lastly, Article 17 of the European Convention addresses the abuse of rights, prohibiting any activities aimed at destroying or limiting the rights outlined in the Convention. More information at: https://www.echr.coe.int/documents/d/echr/convention_ENG.

2 On this subject, and being aware of the large volume of works related to the phenomenon of hate speech, we mention the latest article by professor M.V. Camarero Suárez, 'La protección contra la discriminación por identidad sexual en el matrimonio: una respuesta eficaz ante el impaction de la intolerancia', *ISTEL* (2024), and An interesting book coordinated by Professor Eulalia w. Petit de Gabriel, *Valores (y temores) del estado de derecho: libertad de expresión vs. delitos de opinión en derecho internacional*, (Aranzadi 2023).

3 A. Lamson Lucas de Souza Lehfeld, A. Martinez Perez Filho, freedom of speech and hate speech: an american perspective, *R. Dir. Gar. Fund., Vitória*, v. 23, n. 2, p. 31-56, jul./dez. (2022) DOI: http://dx.doi.org/10.18759/rdgf.v23i2.2029.

4 The United States Supreme Court has mentioned different criteria that may qualify a group as a suspect category, and established a judicial precedent for suspect classifications in the cases of *Hirabayashi v. United States*, 320 U.S. 81 (*Hirabayashi v. United States*, 320 U.S. 81 (1943). More information at: https://supreme.justia.com/cases/federal/us/320/81/. Os this issue see also, Jeremy Waldron, 'The Harm in Hate Speech', *The Oliver Wendell Holmes Lectures*, Volume (2009), https://doi.org/10.4159/harvard.9780674065086.

5 N. Alkiviadou, 'Platform liability, hate speech and the fundamental right to free speech', *Information & Communications Technology Law*, 1 11, (2024), at: https://doi.org/10.1080/13600834.2024.2411799.

6 N. Alkiviadou, 'Hate Speech on Social Media Networks: Towards a Regulatory Framework? ', *Information and Communications Technology Law*, 28 (1). (2019), at 19-35.

(B) DIGITAL PLATFORMS: POWERFUL VEHICLES FOR HATE SPEECH AND VIOLENT CONTENT

Social media holds significant potential by improving both the accessibility and quality of data that shape political decisions for the good of society. These platforms provide real-time access to extensive information, enabling decision-makers to act based on more comprehensive and up-to-date evidence[7]. Additionally, the interactive features of social networks allow for the incorporation of diverse viewpoints and the early detection of public concerns or trends[8], which play a crucial role in developing policies that are responsive and aligned with the needs of different communities[9]. At the same time, rapid connection around the globe and the lack of control by states or supranational regulation raise questions about their impacts on human rights[10]. Over recent years, scholars have noted the potential for social media posts to incite violence against individuals or groups[11], often with near impunity[12]. These factors make it increasingly difficult to control online speech, presenting significant risks to those targeted.

It is notorious how Facebook, and now Meta, created with the purpose of connecting people around the world, is being a vehicle for propaganda, among others, in the leakage of data or circulation of fake news that have direct consequences on state political campaigns. At the international level, the scandal became more evident after the discovery of how the platform was allowing the accumulation and use of large

---

7 A.Muna Almaududi Ausat, 'The Role of Social Media in Shaping Public Opinion and Its In-fluence on Economic Decisions', *Technology and Society Perspectives (TACIT)* Vol. 1, No. 1, (2023), at 35 44, doi:10.61100/tacit.v1i1.37.

8 S. Arshad, S. Khurram, 'Can government's presence on social media stimulate citizens' online political participation? Investigating the influence of transparency, trust, and responsiveness', *Government Information Quarterly*,(2020).

9 Casteltrione, Isidoropaolo, 'Facebook and political participation: Virtuous circle and participation intermediaries', *Interactions: Studies in Communication & Culture* 7, (2016), at: 177 96.

10 S. González-Bailón, L. Yphtach 'Do Social Media Undermine Social Cohesion? A Critical Review', *Social Issues and Policy Review* (17), (2023), 155 180.

11 A. J. F., Puerta, 'Incitación al odio y colectivos vulnerables, del Derecho internacional al Derecho español: especial referencia al delito de incitación al odio por motivos religiosos', *Revista de la Facultad de Derecho de México*, 73(285), (2023), at: 361 382. A new study has succeeded in demonstrating that it is possible to anticipate the increase of hate crimes in Spain using only social network data. The research modeled data on police complaints of hate crimes reported in Spain between 2016 and 2018, with toxic and hateful messages posted on the same dates on X (formerly Twitter) and Facebook. The results show not only a temporal correlation between the two phenomena, but it has been able to generate a series of predictive models that allow to anticipate with some accuracy when reports will increase. More information at: C. Arcila Calderón, P.Sánchez Holgado, J. Gómez, M. Barbosa, H. Qi, A. Matilla, P. Amado, A. Guzmán, D. López-Matías & T. Fernández-Villazala, 'From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and LGBT communities', *Humanities and Social Sciences Communications*, volume 11, Article number: 1369 (2024), at: https://www.nature.com/articles/s41599-024-03899-1.

12 K. Müller, C. Schwarz, *Fanning the Flames of Hate: Social Media and Hate Crime*, (2020), available at SSRN: https://ssrn.com/abstract=3082972 or http://dx.doi.org/10.2139/ssrn.3082972. In this paper the authors investigate the link between social media and hate crime. See also, C. Naganna, A. Sreejith, 'Hate speech review in the context of online social networks', *Aggression and Violent Behavior*, Volume 4, May June 2018, (2018), at: 108-118.

amounts of users' personal data by Cambridge Analytica, a British firm hired by the Trump campaign in 2016[13].

The Secretary-General of the United Nations has highlighted that using the internet to spread hateful expression represents one of the most pressing human rights challenges emerging from technological advancements[14]. Hate messages or violent content on social networks such as Facebook, Tik Tok, Instagram, Youtube, among others, are a real threat to coexistence and security[15]. Large digital platforms can be very powerful vehicles for fake news and hate campaigns[16], especially because of the speed of the internet and its ability to reach every corner of the globe. In recent years, and in the face of pressure from the international community and civil society, efforts have been intensified to minimize the impact of the messages disseminated through social platforms. These efforts have translated into the hiring of specialized teams to detect violations of the rules prohibiting hate speech, discriminatory or terrorist messages.

The Report "Promotion and protection of the right to freedom of opinion and expression" of the General Assembly of the United Nations has established as some of the most relevant current factors in the transmission harmful content online:

1. The speed of information on the Internet;
2. The lack of control of social networks;
3. and the anonymity on the networks makes it difficult to investigate and hold the company accountable[17].

It is worth mentioning that the use of a pseudonym is considered a tool to exercise freedom of expression also in the digital world[18]. Despite this, is important to highlight that research has shown that children were most likely to report having experienced anonymous trolling, which was most prevalent on Instagram, Twitter, Pinterest and Facebook. Violent content was the next most frequent impact, "occurring with highest prevalence on TikTok and YouTube respectively", the report 'childhoods: a survey of

---

13 M. Hu, 'Cambridge Analytica's black box', *Big Data & Society*, 7(2). (2020), at: https://doi.org/10.1177/2053951720938091; A. J. Brown, 'Should I Stay or Should I Leave?" *Exploring (Dis)continued Facebook Use After the Cambridge Analytica Scandal* ', *Social Media + Society*, 6(1), (2020), at:https://doi.org/10.1177/2056305120913884;

14 The Secretary-General, 'Preliminary Representation of the Secretary-General on Globalization and Its Impact on the Full Enjoyment of All Human Rights' paras 26-28, U.N. Doc A/55/342 (Aug 31 2000)

15 A. A. Siegel, *Social Media and Democracy: The State of the Field, Prospects for Reform*, (Cambridge University Press 2020), at 56-88. M. Revenga Sánchez *Libertad de expresión y discursos del odio*, (Alcalá de Henares: Universidad de Alcalá 2015), and N. Gabler, 'The Internet and Social Media Are Increasingly Divisive and Undermining of Democracy', *Alternet*, (2016).

16 Commission opens formal proceedings against Meta under the Digital Services Act related to the protection of minors on Facebook and Instagram. Facebook and Instagram were designated as Very Large Online Platforms (VLOPs), May 16, 2024. More information at: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2664.

17 J. Palmieri, 'Can Social Media Corporations be held Liable Under International Law for Human Rights Atrocities?', 34, *Pace Int'l L. Rev.* 135, (2022) at: https://digitalcommons.pace.edu/pilr/vol34/iss2/4.

18 C. Véliz, 'Online Masquerade: Redesigning the Internet for FreeSpeech Through the Use of Pseudonyms', *Journal of Applied Philosophydoi*, (2018), doi:10.1111/japp.12342, at: https://philpapers.org/archive/VLIOMR.pdf.

children and parents' said[19]. However, mechanisms should be in place that allow the identity of Internet users to be known when requested by a judge[20].

Nevertheless, in relation to the phenomenon of hate speech, one of the key challenges encountered by countries in regulating and limiting freedom of speech is the different positions of States and the lack of a unanimous consensus on the concept of hate speech in international law[21].

The spread of hate speech or violent content online has prompted initiatives to regulate digital content. However, international law lacks a clear definition. The Special Rapporteur observes that many types of hate speech do not reach the level of severity outlined in Article 20, paragraph 2, of the International Covenant, which mandates that states legally prohibit any advocacy of national, racial, or religious hatred that incites discrimination, hostility, or violence.[22].

It comes as no surprise that the media can be complicit in the commission of certain abuses[23]. Even with traditional media, such as radio, Radio, corrupt governments have used it to disseminate hate speech, as well as to justify their discourse and actions against certain ethnic groups or minorities[24]. One example is the case of the genocide in Rwanda[25], one of the most terrible episodes of recent decades, which registered more than 800,000 deaths in less than 5 months[26]. It is interesting to see how the message of hate was internalized to the point of annihilating any opposition.

---

19 Report Downloads Digital childhoods: a survey of children and parents https://www.childrenscommissioner.gov.uk/resource/digital-childhoods-a-survey-of-children-and-parents/

20 Ethnic and racially motivated hate speech has reached the Strasbourg Court on multiple occasions.in the Balázs v. Hungary case n(20 de octubre de 2015), stating emphatically that States parties to the Convention have an obligation to take all necessary measures to investigate racist motivations and to determine whether ethnic hatred or prejudice is behind the commission of any act of racism, the Strasbourg Court held that the State party to the Convention has an obligation to take all necessary measures to investigate racist motivations and to determine whether ethnic hatred or prejudice is behind the commission of any act of racism or ethnic prejudice lie behind the commission of any criminal act.

21 M. Hietanen,J. Eddebo, 'Towards a Definition of Hate Speech With a Focus on Online Contexts', *Journal of Communication Inquiry*, *47*(4), at: 440-458, (2023), at: https://doi.org/10.1177/01968599221124309 and F. Baider,), *Accountability Issues, Online Covert Hate Speech, and the Efficacy of Counter-Speech*, *Politics and governance*, Vol 11, No 2, (2023).

22 Sixty-sixth session Item 69 (b) of the provisional agenda, Promotion and protection of human rights: human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms, Sixty-sixth session. More information at: https://documents.un.org/doc/undoc/gen/n11/449/78/pdf/n1144978.pdf?OpenElement, p.9-10.

23 M. Nino 'The freedom of expression and hate speech in cyberspace', *la Comunità Internazionale*, fasc. 1/2023 pp. 33-5, Editoriale Scientifica srl, (2023).

24 *Media and Mass Atrocity: The Rwanda Genocide and Beyond*: https://www.cigionline.org/publications/media-and-mass-atrocity-rwanda-genocideand-beyond.

25 D. Rodriguez Vázquez, *El genocidio en Ruanda: análisis de los factores que influyeron en el conflicto*. Documento de Opinión, Instituto Español de Estudios Estratégicos (IEEE), (2017), at:https://www.ieee.es/Galerias/fichero/docs_opinion/2017/DIEEEO592017_Genocidio_Ruanda_DanielRguezVazquez.pdf, and W. Schabas, 'Hate speech in Rwanda. The road to genocide', in M. Lattimer, (Ed.)., *Genocide and Human Rights* (1st ed.), Routledge, 207, (2017), DOI.org/10.4324/9781351157568.

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

In recent decades, the media landcape has changed and evolved, but the spread of hate speech and violent content has not only persisted but escalated. This shift is especially concerning in terms of protecting minorities and vulnerable groups.

In this regard, since 2017, the Facebook platform has been under investigation in the case of illegal acts against the Rohingya minority[27].

Facebook's role in the Rohingya crisis serves as a case study on the dangerous intersections of social media, artificial intelligence, and hate speech. The platform, with over 1.8 billion active users worldwide, became the primary communication tool in Myanmar, where the internet is almost synonymous with Facebook. In this Country, Facebook has become "a near-ubiquitous communications tool, following the opening up of the economy". Given its far-reaching impact, the platform's misuse to disseminate dangerous speech has effectively contributed to sustaining institutionalized discrimination against the Rohingya community[28]. In this regard, this dominance allowed the platform to become a powerful vector for the dissemination of hate speech, particularly against the Rohingya minority, which exacerbated the existing ethnic tensions and served as a channel for justificatory discourses that contributed to the atrocities committed against them[29].

Myanmar's military attacks civilians on since 2017 are a considered a genocide for the control of key cities in Rakhine state[30]. The platform became a tool for government officials, the military and radical Buddhist groups to propagate misinformation and hateful ideologies. Propaganda pages linked to figures like Ashin Wirathu, referred to as the "Burmese Hitler" due to his virulent anti-Muslim rhetoric, proliferated on Facebook. These pages, such as the notorious *Kalar Beheading Gang*, spread dehumanizing messages that portrayed the Rohingya as invaders and threats to Myanmar's national identity. These false narratives fueled widespread animosity toward the Rohingya and contributed to justifying the brutal military campaigns against them.

One key issue identified in the spread of hate speech through Facebook in Myanmar is the platform's reliance on Artificial Intelligence-driven content moderation[31]. The automated systems designed to flag harmful content failed to keep pace with the volume

---

27 U.N. investigators cite Facebook role in Myanmar crisis, en: https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigatorscite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN. And J.Young, P. Swamy and D. Danks, *Beyond AI: Responses to Hate Speech and Disinformation*, at: http://jessica-young.com/research/Beyond-AI-Responses-toHate-Speech-and-Disinformation.pdf.

28 On this subject see: Social Media, Artificial Intelligence, and Hate Speech in Myanmar Case Study, This case study was utilized at an AI and Human Rights workshop, held at the Data & Society Research Institute on April 26-27, 2018. More information at: https://datasociety.net/wp-content/uploads/2018/09/Social-Media-Artificial-Intelligence-and-Hate-Speech-in-Myanmar_Case-Study_Final.pdf.

29 L. Arenal, 'Limitaciones y alcance de la responsabilidad de las empresas proveedoras de servicios en el discurso de odio online. El caso de Meta en la incitación al genocidio Rohingya', *Cuadernos de Derecho Transactional*, vol. 15, n.2, pp.141-166. (2023).

30 F. J. Zamora Cabot and M. C. Marullo, 'El conflicto rohingya y sus proyecciones jurídicas: aspectos destacados', *Ordine internazionale e diritti umani*, pp. 461-484, (2020).

31 Report Amnisty International: *Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations.* More information at: https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/.

and context-specific nature of the hate speech emerging in Myanmar[32]. Facebook's Artificial Intelligence struggled to accurately interpret content in Burmese and other local languages, allowing large amounts of inflammatory and dangerous rhetoric to go unchecked. According to Data & Society's report, this technological shortcoming highlights the risks of over-relying on Artificial Intelligence for content moderation, particularly in contexts where local linguistic and cultural nuances are critical for identifying harmful content[33].

While Facebook's role in amplifying hate speech is well documented, the platform's failure lies not only in its Artificial Intelligence systems but also in its human oversight[34]. Facebook was slow to act on repeated warnings from civil society groups and international organizations about the rise of hate speech on its platform. As the Data & Society report highlights, Facebook's reliance on under-resourced and inadequately trained human moderators exacerbated the problem, particularly in regions like Myanmar, where understanding of the local political and cultural dynamics was essential for identifying harmful content. In sum, the platform's response to these failures has been characterized as reactive rather than proactive, leading to criticism for its lack of accountability[35].

The case of Myanmar also illustrates the broader challenges posed by the global nature of platforms like Facebook, which are governed by algorithms and content moderation policies designed in one cultural context but applied universally[36]. The automated systems, which are often effective in English-speaking and Western contexts, proved woefully inadequate in Myanmar[37]. This failure underscores the importance of developing Artificial Intelligence systems that are sensitive to local languages and contexts to prevent the amplification of harmful speech in conflict zones. This negligence facilitated the spread of propaganda that dehumanized the Rohingya, labeling them as outsiders and enemies, thus justifying their mistreatment. The consequences of this unchecked spread of hate speech and violent content have led to international calls for greater regulation of social media platforms, particularly in conflict-affected regions[38].

---

32 C.Crystal, *Facebook, Telegram, and the Ongoing Struggle Against Online Hate Speech*
Case studies from Myanmar and Ethiopia show how online violence can exacerbate conflict and genocide—and what social media companies can do in response https://carnegieendowment.org/research/2023/09/facebook-telegram-and-the-ongoing-struggle-against-online-hate-speech?lang=en; J Sablosky 'Dangerous organizations: Facebook's content moderation decisions and ethnic visibility in Myanmar'. *Media, Culture & Society*, 43(6): 1017–1042, (2021).

33 *Content OR context moderation, Community-Reliant, and Industrial Approaches.* More information at: https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

34 Amnesty InternAational: Myanmar: *The social atrocity: Meta and the right to remedy for the Rohingya*, 2022. More information at: https://www.amnesty.org/en/documents/asa16/5933/2022/en/.

35 Myanmar: UN Fact-Finding Mission releases its full account of massive violations by military in Rakhine, Kachin and Shan States, 2018, https://www.ohchr.org/en/press-releases/2018/09/myanmar-un-fact-finding-mission-releases-its-full-account-massive-violations.

36 See 'From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and LGBT communities', supra note 13.

37 C. Crystal, supra note 34.

38 See, United Nations, *Hate speech and real harm*, https://www.un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm#collapseFour.

In 2019, Facebook was also implicated in the massacre of Muslims in a mosque in New Zealand by an extremist who spread the video live[39]. Or in the Molly case[40], which has also laid the groundwork for specific UK legislation to improve moderation measures on social networks and provide for more effective measures to combat child injury and suicide.

Given the social concern about the rejection of certain religions or against certain minorities or the impacts on mental and physical health, it is urgent to analyze the incidence of the so-called hate speech and violent content on social networks, and how to effective address this problem[41].

(C) SUPRANATIONAL EFFORTS TO COMBAT HARMFUL CONTENT ONLINE

The UN has launched multiple initiatives to address hate speech, including Resolution 16/18[42] and the Rabat Plan of Action, which help distinguish between blasphemy and hate speech[43]. The Rabat Plan provides a six-part test to differentiate between offensive speech and illegal hate speech, considering context, speaker, intent, content, reach, and likelihood of harm. In 2018, the UN Secretary-General introduced a strategy to combat rising global hate speech through social and political measures, without advocating for legal restrictions on speech[44]. Resolution 16/18, together with its intergovernmental

---

39 More information at: https://www.cnbc.com/2019/03/19/australias-pm-restrictsocial-media-after-christchurch-mosque-attack.html. Consultado el día 2 de abril de 2019.

40 Due to the platform's algorithm, Molly Russell, the 14-year-old girl who decided to end her life, was receiving suicide-related images. On this issue see, A. Orben, T. Dienlin, A. K. Przybylski, 'From online hate speech *Social media's enduring effect on adolescent life satisfaction*'From online hate speech, *Proceedings of the National Academy of Sciences of the United States of America*, 116(21), 10226 10228. (2019), doi: 10.1073/pnas.1902058116. C. Rodway, S. G. Tham, N. Richards, S. Ibrahim, P. Turnbull, N. Kapur and L. Appleby, 'Online harms? Suicide-related online experience: a UK-wide case series study of young people who die by suicide ', *Psychol Med*. (2023), Jul;53(10): doi: 10.1017/S0033291722001258. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10388316/.

41 B. Di Fátima (Ed.), *Hate Speech on Social Media, A Global Approach*, (LabCom Books & EdiPUCE, 2023), at: https://labcomca.ubi.pt/wp-content/uploads/2023/05/Hate-Speech-on-Social-Media.pdf.

42 Among others, Resolution adopted by the Human Rights Council* 16/18 Combating intolerance, negative stereotyping and stigmatization of, and discrimination, incitement to violence and violence against, persons based on religion or belief, Human Rights Council Sixteenth session Agenda item 9 Racism, racial discrimination, xenophobia and related form of intolerance, follow-up and implementation of the Durban Declaration and Programme of Action, more information at: https://documents.un.org/doc/resolution/gen/g11/127/27/pdf/g1112727.pdf. General Assembly of the United Nations, 20/8. The promotion, protection and enjoyment of human rights on the Internet, 16 July 2012, https://documents.un.org/doc/resolution/gen/g12/153/25/pdf/g1215325.pdf.
On this issue, see U. Kohl, 'Platform regulation of hate speech a transatlantic speech compromise?', *Journal of Media Law*, (2022), DOI: 10.1080/17577632.2022.2082520.

43 The Rabat Plan of Action, 5 October 2012, Freedom of opinion and expression, 'The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence brings together the conclusions and recommendations from several OHCHR expert workshops.", https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action.

44 The UN Strategy and Plan of Action, more information at: https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech. The initiative has two primary goals. The first is to strengthen UN efforts in addressing the root causes and underlying factors of hate speech. This aligns with the Secretary-General's prevention agenda, which aims to tackle violence, marginalization, and discrimination by

implementation mechanism, the Istanbul Process[45], serves as the primary international framework for addressing hate speech. The Council of Europe's Additional Protocol to the Convention on Cybercrime[46] is the only document specifically crafted to address online hate-related activities. Focused on the criminalization of racist and xenophobic acts committed through computer systems, the Protocol was adopted in 2003 and came into force in 2006; it addresses the criminalization of racist and xenophobic acts committed via computer systems, and acknowledges the risk of misuse or abuse of such systems to disseminate racist and xenophobic propaganda. While attentive to concerns about free expression, the Council underscores the need for regulation. Organizations like UNESCO have also further supported civil society-based action plans to prevent violent extremism and promote tolerance[47].

An interesting initiative is the development of the Santa Clara Principles on transparency and accountability in content moderation'[48]. In May 2018, a group of organizations, advocates, and academics joined forces to establish these principles in response to increasing worries about the opaque and unaccountable practices of internet platforms in developing and implementing their content moderation policies. The principles set forth baseline requirements that tech companies must follow to ensure sufficient transparency and accountability in their approaches to removing user content or suspending accounts that breach their guidelines.

> *The Principles emerged from a collaborative endeavour involving human rights organisations, advocates, and academic experts. They provide a set of standards for social media platforms, emphasising the need for meaningful transparency and accountability in content moderation, guided by a human rights-centered approach. It is notable that major social media companies have endorsed these principles*[49]

---

emphasizing early warning, early action, and preventive approaches to human rights. The second goal is to support effective UN responses to the societal impact of hate speech. To achieve this, the initiative balances two perspectives. While it adopts a broad view of what qualifies as incitement to discrimination, hostility, and violence, it focuses on fostering positive counter-narratives rather than restricting freedom of expression. The plan of action outlines 13 commitments the UN aims to undertake, such as monitoring and analyzing hate speech's root causes, providing support for its victims, using mediation strategies, improving the use of technology and education, collaborating with social media companies, enhancing UN staff skills, and engaging in advocacy to spotlight concerning hate speech trends.

45 The Istanbul Process is the dedicated mechanism for follow-up on the implementation of the action plan set out in Human Rights Council resolution 16/18 and its counterpart at the General Assembly, resolution 66/167. More information at: https://www.istanbulprocess1618.info/about/.

46 Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems Strasbourg, 28.I.2003, https://rm.coe.int/168008160f.

47 On this point see also, Organization for Security and Co-operation in Europe Vienna, August 2018, The Role of Civil Society in Preventing and Countering Violent Extremism and Radicalization that Lead to Terrorism A Guidebook for South-Eastern Europe, and the United Nations Development Programme, Preventing Violent Extremism Through Promoting Inclusive Development, Tolerance And Respect For Diversity. A development response to addressing radicalization and violent extremism, https://www.undp.org/sites/g/files/zskgke326/files/publications/Discussion%20Paper%20-%20Preventing%20Violent%20Extremism%20by%20Promoting%20Inclusive%20%20Development.pdf

48 See the Santa Clara Principles https://santaclaraprinciples.org/.

49 A. Hatano, '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*, (2023), at: https://brill.com/view/journals/auso/41/1/article-p127_6.xml#ref_FN0000068.

Other interesting initiatives are the Recommendation of the Council on Children in the Digital Environment where the principles for a safe and beneficial digital environment for children are established[50] and the G7 Digital and Technology Track Annex 3: Safety Principles[51].

At the European level, the European Commission's assessment of the Code of Conduct on hate speech online[52], launched in 2016, highlights the significant strides made by major platforms in combating hate speech, but also points to areas for improvement[53]. The Code was established to ensure faster removal of illegal content, particularly hate speech targeting various minority groups. Major tech platforms like Facebook, Twitter, Google, Microsoft, and others voluntarily signed the Code, committing to a set of guidelines designed to tackle the spread of illegal and harmful content. One of the key goals of the Code of Conduct is to enhance transparency and promote cooperation between platforms, civil society, and authorities to ensure quicker and more efficient action against hate speech. According to the 2019 assessment, platforms improved their response times significantly[54]. The removal rate of hate speech content that had been flagged by users within 24 hours rose to 72%, compared to just 28% in 2016, which constitutes a remarkable increase. However, while these numbers are promising, the report stresses that platforms need to continue refining their community standards and moderation processes.

The evaluation also emphasizes the increasing importance of artificial intelligence and automated tools in identifying and moderating hate speech. The report highlights that automated tool are becoming a more effective way to detect and act upon harmful content, with many platforms deploying such technologies to supplement human moderation efforts. Despite this progress, the report notes that there is still insufficient data on the volume of hate speech being flagged and removed. This gap in data collection impedes a more detailed understanding of the nature and scope of the problem.

Another concern raised in the assessment is the need for platforms to enhance their collaboration with trusted flaggers, which are external organizations and experts who

---

50 More information at: OECD Legal Instruments, Recommendation of the Council on Children in the Digital Environment, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0389.

51 Ministerial and Other Meetings G7/8 Digital and Technology Ministers, G7 Digital and Technology Track Annex 3: G7 Internet Safety Principles, 2021 at: https://g7.utoronto.ca/ict/2021-annex_3-internet-safety.html.

52 European Commission. Code of conduct on countering illegal hate speech online. https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985

53 About Islamophobia, N. P. Guedes, A. A Padrón, A. A., 'Herramientas jurídicas para combatir la islamofobia en la Unión Europea', *Revista Científica Universitaria Ad Hoc*, 2(5), at: 48-58 (2021).

54 Assessment of the Code of Conduct on Hate Speech on line State of Play, Brussels, 27 September 2019 (OR. en), European Commission To: Permanent Representatives Committee/Council, https://commission.europa.eu/document/download/a5c92394-8e76-434a-9f3a-3a4977d399bb_en?filename=assessment_of_the_code_of_conduct_on_hate_speech_on_line_-_state_of_play_.pdf, and the Monitoring rounds Factsheet 7th monitoring round of the Code of Conduct at: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en. See also, report IPSOS-UNESCO Study on the impact of online disinformation during election campaigns. Survey on the impact of online disinformation and hate speech September 2023, a global survey on the impact of online disinformation and hate speech, https://www.ipsos.com/sites/default/files/ct/news/documents/2023-11/unesco-ipsos-online-disinformation-hate-speech.pdf.

identify illegal content. This collaboration has proven effective, with trusted flaggers often reporting hate speech more quickly and accurately. However, the report stresses that such collaboration must be further strengthened to ensure better oversight and to improve the overall response to flagged content.

Additionally, the European Commission's assessment touches on the need for greater transparency and accountability from platforms regarding their content moderation policies. While platforms have made strides in adhering to the Code of Conduct, they must do more to provide clear information about their content removal processes and the decisions made when handling reported content. This is crucial for ensuring public trust and ensuring that moderation efforts are consistent and aligned with legal requirements. Despite these advancements, the Commission's assessment acknowledges that the current framework, while helpful, is not sufficient to fully address the challenges of online hate speech. The document calls for ongoing improvements and monitoring of the Code's implementation, with an emphasis on the need for stronger regulatory measures.

It also advocates for better coordination between national authorities, the platforms, and civil society to ensure that hate speech is effectively tackled across the EU. Looking ahead, the European Commission plans to continue its work in developing and refining online hate speech regulations. The Code of Conduct has laid the foundation for these efforts, but the growing prevalence of harmful content online means that a more robust approach is required. This includes not only technological innovations but also better alignment of legal frameworks, stronger collaboration with external stakeholders, and greater transparency in decision-making processes. These efforts aim to ensure that the EU remains a leader in the fight against online hate speech, while also preserving the core values of freedom of expression and privacy[55].

Also of major interest is The European Commission's strategy for online platforms revolves around fostering an environment that promotes fair competition, innovation, and user protection. At its core, the Commission emphasizes four guiding principles: Level Playing Field: Ensuring all digital services are subject to comparable regulations, enabling fair competition. Responsible Behavior: Platforms must act responsibly, safeguarding fundamental rights and societal values. Trust and Transparency: Platforms must be transparent about their operations, including content moderation and data use, to build user trust. Open and Non-Discriminatory Markets: Encouraging open markets while maintaining a fair, non-discriminatory approach to data use and platform access[56].

These principles are aimed at ensuring a balanced, secure, and innovative digital ecosystem in the EU, while addressing the rapid pace of technological advancement. However, the challenges lie in implementing these principles effectively, ensuring consistency across member states, and adapting to emerging digital trends. The

---

55 E. Nave, L.Lane, *'Countering online hate speech: How does human rights due diligence impact terms of service?* ', *Computer Law & Security Review*, Volume 51, (2023), at: https://doi.org/10.1016/j.clsr.2023.105884.

56 Shaping Europe's digital future. Online Platforms. "The European Commission aims to foster an environment where online platforms thrive, treat users fairly and take action to limit the spread of illegal content". More information at: https://digital-strategy.ec.europa.eu/en/policies/online-platforms.

Commission's ongoing efforts seek to establish a framework that benefits both users and businesses, while fostering innovation.

In this strategy, the Digital Service Act[57]introduced by the European Commission in December 2020plays a central role. introduced by the European Commission in December 2020. This Act constitutes regulatory proposal aimed at standardizing the definition of illegal content across platforms and establishing procedures for its removal. As a result, the decision to remove online content is delegated to each platform a private entity that, in turn, entrusts the function of censorship to individuals who must make decisions based on broad, self-regulation standards.. Furthermore, these decisions are made in a matter of seconds, despite the fact that a constitutional right is at stake: freedom of expression:

> *The EU's digital services act (DSA) helps combat propaganda, misinformation and fake news online by introducing strict requirements for online platforms: accountability for illegal content and fines for non-compliance, transparency in how algorithms work, user reporting tools and stricter ad rules, risk assessments on harmful information, crisis response to limit fake info during emergencies, independent audits of efforts against illegal content*[58].

The Digital Services Act applies to all online intermediaries in the EU[59]. Facebook and Instagram were designated as Very Large Online Platforms under the EU's Digital Services Act[60], as they both have more than 45 million monthly active users in the EU. As Very Large Online Platforms, Facebook and Instagram had to start complying with a series of obligations set out in the norm. However, there are currently no binding rules to stop online hate speech, either at the European level.

Other very relevant aspects are the regional efforts to regulate the work of moderators of digital platforms. Unfortunately, this work can lead to numerous problems due to the precarious working conditions of moderators and the effects on their mental health. In the report of the European Agency OASH, *Occupational safety and health risks of online content review work provided through digital labour platforms*[61], the risks faced by moderators are mentioned: A) Emerging risks and B) Psychosocial risks and stress.

---

57 The Digital Services Act, Ensuring a safe and accountable online environment, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en.

58 Digital Service Act, 'The EU's digital services act (DSA) aims to create a safer, more transparent internet. By making online platforms accountable for the way they manage and moderate content, the DSA helps combat the spread of harmful content online'. https://www.consilium.europa.eu/es/policies/how-the-eu-combats-harmful-content-online/

59 On the the extraterritorial implications of the Digital Services Act, see Laureline Lemoine & Mathias Vermeulen (AWO) analysis, 'As the enforcement of the Digital Services Act (DSA) is gathering speed, a number of non-EU based civil society and research organizations have wondered to what extent the DSA can have an impact on their work. This blog post provides a concise overview of the areas and provisions within the Digital Services Act that are most pertinent to the issue of extraterritorial application of the Regulation', at: https://dsa-observatory.eu/2023/11/01/the-extraterritorial-implications-of-the-digital-services-act/.
N. Ikiviadou, 'Platform liability, hate speech and the fundamental right to free speech', *Information & Communications Technology Law*, 1 11. (2024), at: https://doi.org/10.1080/13600834.2024.2411799.

60 More information at: DSA: Very large online platforms and search engines, https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops.

61 Occupational safety and health risks of online content review work provided through digital labour platforms. More information at: https://osha.europa.eu/sites/default/files/2022-03/OSH_implications_of_online_content.pdf.

Initially, content review and moderation on social media platforms were primarily managed by volunteers from the community of users. However, with the exponential growth in both the amount of content and the number of users, the task has become much more complex. Content moderators now face the challenge of reviewing vast amounts of posts, images, and videos, some of which may be live on the platform in real time. These workers must identify, categorize, verify, and validate content. This can involve tasks such as tagging objects in photos or videos and categorizing text based on keywords. Moderators are given only a few seconds to go through each step and decide whether specific content should be allowed on the platform. The content under review can include pornography, violent images, homophobic, misogynistic, or racist posts, scams, hate speech, conspiracy theories, harassment, threats, cyberbullying, and other illegal or abusive material. We can affirm that the online content review work provided through digital labour platforms is stressful, emotionally and physically demanding, and can lead to musculoskeletal disorders[62]. Digital labour platforms do not address such severe occupational safety and health (OSH) risks, or do so in a limited way. This will lead to a reconsideration of the position of European institutions regarding occupational diseases.

(D) EXTRATERRITORIAL MEASURES IN STATE REGULATION

Moving on to analyze some developments on this topic from a state's perspective, we have to start from a premise: States have different conceptions of what exactly freedom of expression on social networks should entail and the limits that can be imposed on it[63]. However, there is a growing positioning on the minimum elements for combating hate speech when the latter can have a significant impact on human rights. Furthermore, the extraterritorial nature national norms means that non-territorial companies are also subject to their provisions. This global reach poses significant challenges, as tech companies must navigate compliance with various laws across multiple jurisdictions, all of which may have conflicting standards for content moderation.

A study made in 2021 proved a link between hateful content on Facebook inciting violence against refugees and the increase in actual physical violence on migrants on EU countries[64]. From this perspective, we question whether the state measures, which are also applied beyond the territory — such as in the country where the company is based — are actually effective, or if, on the contrary, they are failing to control and mitigate the negative impacts of hate speech on social media.

---

62 M.C. Urzi Brancati, A. Pesole and E. Fernandez Macias, *New evidence on platform workers in Europe*, EUR 29958 EN, Publications Office of the European Union, Luxembourg, (2020), ISBN 978-92-76-12949-3 (online), doi:10.2760/459278 (online), JRC118570, Available at: https://publications.jrc.ec.europa.eu/repository/handle/JRC118570.

63 M. García Santos, 'El límite entre la libertad de expresión y la incitación al odio: análisis de las sentencias del Tribunal Europeo de Derechos Humanos', *Comillas Journal of International Relations*, nº. 10, (2017), R. PALOMINO, 'Libertad religiosa y libertad de expresión', *Ius Canonicum*, XLIX, nº. 98, (2009).

64 M. Cinelli, M., A. Pelicon, I. Mozetič, I. *et al. Dynamics of online hate and misinformation*, *Sci Rep* 11, 22083, (2021), https://doi.org/10.1038/s41598-021-01487-w.

The French online hate speech bill, adopted in May 2020, mandates platforms to remove illegal content such as racism and antisemitism within 24 hours of receiving a user complaint. If platforms fail to comply, they face hefty fines, potentially up to €1.25 million. While the bill aims to combat the rising tide of hate speech online, critics argue it risks over-censorship and might infringe on freedom of expression, leading to the suppression of legitimate speech. The law exemplifies France's stringent approach to online content regulation in Europe.

In Germany, the NetzDG (Netzwerkdurchsetzungsgesetz), or Facebook Act, is a law passed in 2018 aimed at enhancing the enforcement of legal accountability for social media platforms. The legislation primarily targets major platforms, including Facebook, Twitter, YouTube, and other social networks with over 2 million users within Germany[65]. It was introduced in response to growing concerns over the spread of harmful and illegal content, such as hate speech, extremist propaganda, and misinformation that were being disseminated rapidly through social media channels. The central aim of the norm is to ensure that social media platforms take immediate and effective action against illegal content.

The law imposes strict duties on these platforms to monitor, report, and remove content that breaches German laws, particularly those concerning hate speech, violent extremism, and other forms of illegal online behavior.

Under this framework, platforms are required to establish efficient reporting systems; Platforms must offer users an accessible and simple process for reporting illegal content. This applies mainly to hate speech, content that promotes violence, or terrorist content. Once a report is submitted, platforms are required to review the flagged content within 24 hours if it is clearly illegal, and remove it within 7 days. If the content is less obvious but potentially unlawful, platforms are given up to 7 days to assess and act on it. The law mandates that platforms produce detailed biannual transparency reports. These reports must outline the number of user complaints received, how many pieces of content were removed or blocked, and the platform's response to those complaints. This is intended to foster greater accountability and transparency[66]. If a platform fails to comply with the law's requirements such as not removing illegal content promptly or failing to submit transparency reports it may face heavy fines. The maximum penalty for non-compliance is €50 million[67]. The text clarified that the fines could only be levied against firms that "systematically" evaded the law.

The NetzDG primarily targets content that is explicitly illegal under German law. This includes:

---

65 T. Kasakowskij, J. Fürst, J. Fischer, K.J. Fietkiewicz, 'Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media', *Telematics and Informatics*, Volume 46, (2020)https://doi.org/10.1016/j.tele.2019.101317.

66 P. Zurth, T*he* 'German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability', *31 Fordham Intell. Prop. Media & Ent. L.J.* 1084, (2021).

67 S. Maaß, J. Wortelker, A. Rott, ' Evaluating the regulation of social media: An empirical study of the German NetzDG and Facebook', *Telecommunications Policy*, Volume 48, Issue 5, (2024), https://doi.org/10.1016/j.telpol.2024.102719.

1. Hate Speech: Content that incites discrimination, hostility, or violence against individuals or groups based on protected characteristics, such as race, ethnicity, religion, or gender.
2. Terrorist Content: Posts that promote or glorify terrorist activities or groups.
3. Child Sexual Exploitation: Content that involves the abuse or exploitation of children[68].

The law ensures that freedom of expression remains intact by excluding content that does not meet the thresholds of illegality, thus safeguarding legitimate political and social discourse[69]. Nevertheless, despite its intention to combat harmful content, the NetzDG has faced significant criticism[70]. A major concern is the potential for over-censorship[71]. Platforms, fearing the possibility of hefty fines, may adopt an overly cautious approach, leading to the removal of content that does not necessarily breach legal standards. This could result in legitimate expressions, political opinions, and controversial but lawful content being unnecessarily censored, infringing upon freedom of speech. Another concern is the operational burden placed on platforms, especially smaller ones[72]. While large social networks may have the resources to comply with the stringent requirements, smaller platforms may struggle to establish effective content moderation systems[73]. The law's scope and demands may unintentionally create a disparity in how platforms manage and enforce the law, which could also discourage new entrants to the market[74].

Though the NetzDG applies only to platforms operating in Germany, its impact has reverberated globally. The law has become a point of reference for other countries considering similar approaches to regulating harmful content online. Several nations, particularly within the European Union, have studied its provisions and effectiveness, and some have moved towards adopting their own regulatory frameworks inspired by Germany's model[75].

---

68 More information at: Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz-NetzDG) https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html.

69 H. Tworek and P. Leerssen, *An Analysis of Germany's NetzDG Law*, Transatlantic Working Group, (Amsterdam University 2019), More information at: https://pure.uva.nl/ws/files/40293503/NetzDG_Tworek_Leerssen_April_2019.pdf.

70 C. Donaldson, 'Militant Moralism: The Hegemonic Consequences of German Content Moderation', *German Law Journal*,25(3), at: 497-513, (2024), doi:10.1017/glj.2024.18.

71 Op. Cit. P. Zurth, supra note 69, And J. Ogaki, *German Content Moderation and Platform Liability Policies*, (2024), More information at: https://jsis.washington.edu/news/german-content-moderation-and-platform-liability-policies/.

72 R. Griffin, 'New School Speech Regulation and Online Hate Speech: A Case Study of Germany's NetzDG', *SSRN Electronic Journal*, 2021ff10.2139/ssrn.3920386.

73 L. M. Neudert, 'Reclaiming Digital Sovereignty: Policy and Power Dynamics Behind Germany's NetzDG', *Journal of Information Policy*, (2024), at:https://doi.org/10.5325/jinfopoli.14.2024.0013.

74 W. Echikson and O. Knodt (2018), 'Germany's NetzDG: A key test for combatting online hate', research Paper No. 2018/09 CEPS, November 2018, https://cdn.ceps.eu/wp-content/uploads/2018/11/RR%20No2018-09_Germany's%20NetzDG.pdf.

75 A. Brown, *Models of Governance of Online Hate Speech On the emergence of collaborative governance and the challenges of giving redress to targets of online hate speech within a human rights framework in Europe*, Documents and Publications, Production Department (SPDP), Council of Europe 2020. More information at: https://rm.coe.int/models-of-governance-of-online-hate-speech/168097671d.

The extraterritorial nature of the law also complicates things for companies operating internationally. Platforms must comply with the law's requirements for their German users, even if they are headquartered outside of Germany, potentially leading to challenges in reconciling conflicting regulatory standards across different jurisdictions[76]. This Act represents a critical step in regulating harmful online content and increasing the responsibility of social media platforms. By imposing clear duties on platforms to monitor and remove illegal content, it seeks to protect users from online harm while maintaining a balance with freedom of expression. However, the law is not without its challenges, particularly concerning its potential to infringe upon free speech and the burden it places on smaller platforms:

> *Supporters see the legislation as a necessary and efficient response to the threat of online hatred and extremism. Critics view it as an attempt to privatise a new 'draconian' censorship regime, forcing social media platforms to respond to this new painful liability with unnecessary takedowns. This study shows that the reality is in between these extremes. NetzDG has not provoked mass requests for takedowns. Nor has it forced internet platforms to adopt a 'take down, ask later' approach. Removal rates among the big three platforms ranged from 21.2% for Facebook to only 10.8% for Twitter. At the same time, it remains uncertain whether NetzDG has achieved significant results in reaching its stated goal of preventing hate speech. Evidence suggests that platforms are wriggling around strict compliance. Consider Facebook. The social network makes it difficult to fill out NetzDG complaints. Instead, Facebook prefers to cite their murkily defined community standards to take down vast amounts of content* [77].

As such, the NetzDG continues to be a subject of debate, both within Germany and internationally, with its outcomes likely shaping the future of online content regulation globally.

Another regulation we can mention is the Australian norm on hate speech[78] a law that imposed stringent penalties on platforms like Facebook, YouTube, and Instagram for failing to remove violent or terrorist content. In this context, the legislator has implemented some of the most progressive legal measures to address hate speech and violent content against individuals and indigenous Peoples. Among the broader population, studies reveal that approximately 14% of adults have been subjected to online hate speech[79].

---

76 O. Butler and S. Turenne 'The regulation of hate speech online and its enforcement a comparative outlook', *Journal of Media Law*, 14(1), at: 20 24, (2022), at: https://doi.org/10.1080/17577632.2022.20922. On the topic of internet jurisdiction and extraterritoriality, see the paper: M. Geist, 'Is there a there there? Toward greater certainty for internet jurisdictio' (2001), at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=266932.

77 W. Echikson and O. Knodt, 'Germany's NetzDG: A key test for combatting online hate', Supra note 76.

78 The Online Safety Act (the Act) No. 76, 2021 Compilation No. 2, Compilation date: 14 October 2024, Includes amendments: Act No. 39, 2024 https://www.legislation.gov.au/C2021A00076/latest/text. See Online Safety report. https://www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf?v=1731426825835.

79 See the eSafety Commissioner (eSafety) in Australia Report, Online hate speech, Findings from Australia, New Zealand and Europe, 2019, https://www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf?v=1731672547851. K. Gelber and L McNamara 'The Effects of Civil Hate Speech Laws: Lessons from Australia', *Law & Society Review*, 49(3):631-664. (2015), doi:10.111/lasr.12152 and M. Smith, M. Nolan, and J. Gaffey, 'Online safety and social media regulation in Australia: eSafety Commissioner v X Corp.', *Griffith Law Review*, at: 1 17, (2024). https://doi.org/10.1080/10383441.2024.2405760.

The legislation mandates fines of up to 10% of the platform's annual global revenue and potential prison sentences for executives responsible for failing to activate control mechanisms. This type of legislation points to the growing recognition that platforms must be held accountable for the content they host and the role they play in facilitating or exacerbating violence[80].

In addition, the UK Online Safety Bill[81] to address the growing concerns over the spread of harmful content on digital platforms, from social media networks to search engines. The bill targets online harm such as cyberbullying, hate speech, terrorist content, and disinformation, requiring platforms to take active measures to prevent, detect, and remove such content. Under the bill, companies would be legally obligated to protect their users from harm, and failure to comply would result in substantial penalties. The bill proposes that platforms must have a clear and accessible reporting mechanism for users, along with robust content moderation policies. A key feature of the bill is its focus on "duty of care"[82], which holds tech companies accountable for the safety of their users, especially minors. This duty is central to the bill's goal of balancing user safety with the protection of freedom of expression.

The Bill's approach is to place a duty of care on internet service providers of both user-to-user services in which users interact with each other online (CHAPTER 2). Providers of user-to-user services: duties of care). The duty of care is framed in broad terms in the Bill, but it is composed of three distinct duties to protect users from illegal content, to take additional protective measures to make their site safe and to take additional measures to protect all users from content that is harmful without being illegal, if the service is of a sufficient reach and magnitude[83].

However, critics argue that there is a risk of overreach, with the potential to lead to overzealous content moderation, which could stifle free speech[84]. Supporters of the bill argue that it holds these companies accountable for fostering safer online environments[85]. However, critics have pointed out that platforms may resort to overly

---

80 A. Brown, *Models of Governance of Online Hate Speech On the emergence of collaborative governance and the challenges of giving redress to targets of online hate speech within a human rights framework in Europe*. Supra note 77.

81 Online Safety Act 2023, Government Bill, Originated in the House of Commons, Sessions 2021-22, 2022-23. More information at: https://bills.parliament.uk/bills/3137.

82 'The Online Safety Bill extends services' duty of care to include the regulation of legal but harmful material. We argue both that this extension overburdens developers with responsibility at pain of penalty for legal content and that the specific framing of this provision risks a regulatory slippery slope toward wider censorship', Markus Trengove, Emre Kazim, Denise Almeida, Airlie Hilliard, Sara Zannone, Elizabeth Lomas, A critical review of the Online Safety Bill, Patterns, Volume 3, Issue 8, 2022, https://doi.org/10.1016/j.patter.2022.100544.

83 More information at: https://www.legislation.gov.uk/ukpga/2023/50/enacted.

84 On this topic: Peter Guest, The UK's Controversial Online Safety Act Is Now LawThe UK government says its Online Safety Act will protect people, particularly children, on the internet. Critics say it's ineffective against dangerous misinformation and may be a threat to privacy, 2023, https://www.wired.com/story/the-uks-controversial-online-safety-act-is-now-law/.

85 'National Society for the Prevention of Cruelty to Children hailed the bill's passage as 'a momentous day for children," there has been strong pushback from civil liberties groups as well as tech companies' C. Chin-Rothmann, T. Rajic and E. Brown, *Critical Questions, A New Chapter in Content Moderation: Unpacking the UK Online Safety Bill*, (2023), at: https://www.csis.org/analysis/new-chapter-content-moderation-unpacking-uk-online-safety-bill.

restrictive content moderation policies or even censorship to avoid penalties, which could unintentionally infringe on users' right to free speech[86]. Additionally, there is concern over whether platforms have the capacity and expertise to moderate complex content effectively, especially in diverse cultural and social contexts.

It has faced others critics from various stakeholders.

> *As expected, the Government's intention to show "global leadership with our groundbreaking laws to usher in a new age of accountability for tech and bring fairness and accountability to the online world" was met by support from the child protection community, but suspicion and warnings from digital rights and civil society organisations. So, is the Bill world-leading as the Government puts it, or is it introducing "state-backed censorship and monitoring on a scale never seen before in a liberal democracy", "collateral censorship, the creation of free speech martyrs, the inspiration it would provide to authoritarian regimes", "trying to legislate the impossible — a safe Internet without strong encryption"?* [87].

For one, the definition of harm under the bill is broad, and this vagueness could lead to inconsistent enforcement. Different platforms may interpret the regulations differently, leading to uneven outcomes. Some critics fear that tech companies, under the threat of hefty fines, may remove content that doesn't necessarily violate the law but could be deemed controversial or provocative and:

> *Those who think the Bill is unworkable point to its length, complexity, dependence on secondary legislation, and the operational challenges and costs of implementing its requirements — a process which is not expected to begin until mid-2024.14 It is argued that — in contrast to physical injury — there is no objective way of ascertaining that emotional or psychological harm has occurred, making it impossible to determine whether service providers have discharged their duties of care.15 At the same time, controversies of interpretation are said to be a likely consequence of relying on flexible standards and introducing categories such as "legal-but-harmful" content and "content of democratic importance"* [88].

The fine line between protecting users and over-censoring content is one of the key debates surrounding the bill[89].

The Online Safety Bill set a precedent for other countries grappling with online safety concerns. If successful, it inspired similar legislation in other jurisdictions, particularly in the United States. This could lead to a more global regulatory framework for online platforms, but it also raises questions about international jurisdiction and the differing standards in various countries regarding free speech and online content.

We end this section by mentioning the U.S. Children's Online Privacy Protection Rule, COPPA, a norm that will also have important impacts on holding large platforms accountable. This rule requires the Federal Trade Commission to create and enforce

---

86 B. Kira and L. Schertel Mendes, *A Primer on the UK Online Safety Act* (November 13, 2023). Verfassungsblog, DOI: 10.59704/2120f79b5f59e60b, at https://ssrn.com/abstract=4632326.

87 E. Harbinja, *The UK's Online Safety Bill: Safe, Harmful, Unworkable?,* (2021), at: https://verfassungsblog.de/uk-osb/.

88 https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2022/09/Policy-Brief-Online-Safety-Bill.pdf

89 See the 7 key issues from the Online Safety Bill report, may 2024, at: https://parentzone.org.uk/article/seven-key-issues-from-the-online-safety-bill-report.

regulations regarding children's online privacy and applies to operators of general audience websites or online services that have actual knowledge they are collecting, using, or disclosing personal information from children under the age of 13, as well as to websites or online services that are aware they are collecting personal information from users of another website or online service directed at children.

Operators subject to COPPA must: Post a clear and comprehensive online privacy policy detailing their practices regarding personal information collected from children and provide direct notice to parents and obtain verifiable parental consent, with limited exceptions, before collecting personal information from children online. This rule allows parents to consent to the collection and internal use of their child's information, while prohibiting the operator from disclosing that information to third parties and provides parents with access to their child's personal information so they can review it and/or request its deletion

The personal information collected online from a child is retained only for as long as necessary to fulfill the purpose for which it was collected, and deleted using reasonable measures to prevent unauthorized access or use. At the time of entering into an agreement with a customer for the provision of interactive computer services, the provider must inform the customer, in a manner it deems appropriate, that parental control tools are commercially available. These tools can help the customer restrict access to content that may be harmful to minors. The notice must either identify or give the customer access to information about the providers offering these protective services.

(E) FRAMING THE FIGHT IN LEGAL TERMS: META CASES

This section focuses on the analysis of some relevant cases against online platforms. For the sake of brevity and to maintain focus, the analysis shall be circumscribed to the cases against the platform META, an American multinational technology based California. The company owns and operates Facebook, Instagram, and WhatsApp, among other products and services.

To address these cases, we will first examine recent policies and measures established by Meta to address hate speech and violent content in the last few years. According to its website, its principles are:

> *We stand for and guide our approach to how we build technology for people and their relationships. Give People a Voice, People deserve to be heard and to have a voice — even when that means defending the right of people we disagree with. Build Connection and Community, Our services help people connect, and when they're at their best, they bring people closer together. Serve Everyone, We work to make technology accessible to everyone, and our business model is ads so our services can be free. Keep People Safe and Protect Privacy, We have a responsibility to promote the best of what people can do together by keeping people safe and preventing harm. Promote Economic Opportunity, Our tools level the playing field so businesses grow, create jobs and strengthen the economy.*[90]

[90] More information at: https://about.meta.com/company-info/.

After the events in Myanmar and the special rapporteur reports on the crimes in the country[91], which showed the correlation of the events with the activities carried out on the platform, on 2018 Meta established an Independent Assessment of the Human Rights Impact of Facebook in Myanmar:

> *Facebook stands against hate and violence, including in Myanmar, and supports justice for international crimes. We're working with the UN's Independent Investigative Mechanism for Myanmar, which has a mandate to collect evidence with appropriate safeguards in place, and assist accountability efforts. Through this work, we've begun to lawfully provide data to the IIMM that we preserved back in 2018. As these investigations proceed, we will continue to coordinate with them to provide relevant information as they investigate international crimes in Myanmar. The assessment was completed by BSR (Business for Social Responsibility) — an independent non-profit organization with expertise in human rights practices and policies — in accordance with the UN Guiding Principles on Business and Human Rights and our pledge as a member of the Global Network Initiative. The report concludes that, prior to this year, we weren't doing enough to help prevent our platform from being used to foment division and incite offline violence. We agree that we can and should do more. BSR recommends that Facebook adopt a stand-alone human rights policy, establish formalized governance structures to oversee the company's human rights strategy, and provide regular updates on progress made. BSR urges Facebook to improve enforcement of our Community Standards, the policies that outline what is and isn't allowed on Facebook. Core to this process is continued development of a team that understands the local Myanmar context and includes policy, product, and operations expertise.*[92]

Since 2018, META also established a strategy called remove, reduce, inform[93] to manage content across our platforms and created a Safety center[94]. The online safety center reflects the Facebook Community Standards and Instagram Community Guidelines and works with the support of human and technology review teams[95]. In the Facebook hate speech standards, the platform has established two levels[96].

Tier 1 content that cannot be published,

Tier 2 content to be reviewed.

Tier 1: Content aimed at an individual or group of individuals (including all groups, except those classified as non-protected for being involved in violent crimes, sexual offenses, or representing less than half of a group) based on their protected characteristic(s) or immigration status, whether in written or visual form.

Tier 2: Content targeting a person or group of people on the basis of their protected characteristic.

Related to the violent content, the platform established:

---

91 A/78/527: Report of the Special Rapporteur on the situation of human rights in Myanmar, at: https://www.ohchr.org/en/documents/country-reports/a78527-report-special-rapporteur-situation-human-rights-myanmar.

92 More information at: https://about.fb.com/news/2018/11/myanmar-hria/.

93 More information at: https://transparency.meta.com/es-es/policies/improving/prioritizing-content-review./

94 More information at: https://about.meta.com/actions/safety.

95 More information at: https://transparency.meta.com/enforcement/detecting-violations/how-review-teams-work/.

96 More information at: https://transparency.meta.com/es-es/policies/community-standards/hate-speech/.

> *To protect users from such content, we remove the most graphic content and add warning labels to other graphic content so that people are aware it may be sensitive or disturbing before they click through. We may also restrict the ability for users under 18 to view such content (or "age-gate" the content). We recognize that users may share content in order to shed light on or condemn acts such as human rights abuses or armed conflict. Our policies consider when content shared in this context and allow room for discussion and awareness raising accordingly. In ads, we provide additional protections. For example, content that has been deemed sensitive or disturbing is not eligible to run in ads. We also prohibit ads from including images and videos that are shocking, gruesome, or otherwise sensational*[97].

Something similar is established in the Instagram community standards: *We're working to remove content that has the potential to contribute to real-world harm, including through our policies prohibiting coordination of harm, sale of medical masks and related goods, hate speech, bullying and harassment and misinformation that contributes to the risk of imminent violence or physical harm*[98].

These functions are developed under the auspices of artificial intelligent systems. Each day, users upload millions of posts that undergo automated review by the artificial intelligence systems, assessing the suitability of content before it goes live. These systems are trained to detect images associated with terrorism, child sexual exploitation, and other harmful content. However, automated pre-detection is more the exception than the norm. Most content moderation relies on human agents who apply internal guidelines and extensive training to manage problematic material. Tens of thousands of these moderators work globally, typically through outsourcing and customer service firms like Teleperformance and Accenture. Due to the overwhelming volume of daily reports, moderators often have less than a minute to make decisions on flagged content. This intense pressure, combined with limited time and inadequate resources, frequently results in moderation errors. These mistakes can have two adverse outcomes: allowing harmful content to remain online or removing content that doesn't actually violate guidelines undermining both users' freedom of expression and their right to fair process.

It is worth mentioning that Meta subcontracts companies for the tasks in the different states of moderation, that is, for the search and detection of inappropriate content that may have been published by a user on its platform. In the vast majority of cases, these companies do not have specific rules on how to detect and what content to block[99]. In fact, as evidenced by the reports, [100] Facebook's measures for detecting and removing hate speech or violent content are largely ineffective. For instance, since 2018, Facebook has continued to approve advertisements containing hate speech that incite violence and

---

97 More information at: https://transparency.meta.com/es-es/policies/community-standards/violent-graphic-content/.

98 More information at: https://help.instagram.com/477434105621119?cms_id=477434105621119.

99 J. Espíndola, *Attributing Responsibility to Big Tech for Mass Atrocity: Social Media and Transitional Justice*, (Cambridge University Press 2024), doi:10.1017/S1537592724001282.

100 UN Independent International Fact-Finding Mission on Myanmar calls on UN Member States to remain vigilant in the face of the continued threat of genocide. 23 October 2019. https://www.ohchr.org/en/press-releases/2019/10/un-independent-international-fact-finding-mission-myanmar-calls-un-member?LangID=E&NewsID=25197.

genocide against the Rohingya[101]. At the same time, on meta's guidelines or on norms at companies subcontracted for moderation tasks do not contemplate specific rules to protect the mental health of moderators[102].Furthermore, the aforementioned guidelines fail to outline concrete measures or strategies to ensure that moderators can carry out their work without experiencing harmful mental repercussions.[103].

On 2021 Meta endorsed these guarantees of due diligence, vowing to "pay particular attention to the rights and needs of users from groups or populations that may be at heightened risk of becoming vulnerable or marginalized"[104].

For all these reasons, Meta, owner of Facebook, is increasingly accused of enabling human rights violations[105]. The proliferation of hate speech and violent content in its digital platforms has been in the background of recent episodes of mass atrocities. The rise of hate speech also presents multiple judicial challenges when it comes to determining Meta's responsibility for the circulation of such content, its failure to remove it, and its accountability for the mental harm caused to moderators.

The extraterritorial nature of social media platforms poses challenges to traditional judicial legal systems. Meta, based in the United States, operates globally, and the content that circulates on its platforms often has an international impact. The issue, therefore, is whether national courts can hold a foreign multinational accountable under their own laws for actions that affect citizens in other countries. In the J*ohn Doe and Jane Doe against Meta*[106], the plaintiffs argue that international human rights law applies, and courts in the USA have jurisdiction over Meta's operations, given the widespread harm caused by the company's inaction. The plaintiffs contend that Meta violated rights guaranteed under the Universal Declaration of Human Rights (UDHR) and the International Covenant on Civil and Political Rights (ICCPR), particularly concerning freedom from discrimination and violence.

However, this lawsuit also highlights a significant gap in the application of international human rights law to multinational companies. Traditionally, international human rights law focuses on state obligations to protect individuals from harm, but this case challenges the assumption that corporations, especially those operating across borders, are exempt from such standards. The central legal argument is that Meta's platforms, through their design and lack of effective moderation, allowed the dissemination of hate speech that led to tangible consequences, including violence against ethnic minorities.

---

101 Global Witness report at: https://www.globalwitness.org/en/campaigns/digital-threats/rohingya-facebook-hate-speech/.

102 Casey Newton, The Trauma Floor, The secret lives of Facebook moderators in America, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

103 Meta sued in Kenya over claims of exploitation and poor working conditions, https://edition.cnn.com/2022/05/10/tech/meta-sued-in-kenya-lgs-intl/index.html.

104 Meta 2021, Corporate Human Rights Policy, at: https:// about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf.

105 N. Hakim, 'How Social Media Companies Could Be Complicit in Incitement to Genocide', *Chicago Journal of International Law* (21)1: 83 117. (2020),

106 Superior Court Of The State Of California for the County Of San Mateo, Jane Doe, individually and on behalf of all others similarly situated, META Platforms, INC. (f/k/a Facebook, Inc.), a Delaware corporation, https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=3596&context=historical.

One of the most critical aspects of this lawsuit is the issue of extraterritorial jurisdiction. Meta's operations span across multiple countries, and the harmful content on its platforms often affects individuals worldwide. The lawsuit raises the question: can a US court hold a company based in the US accountable for actions that harm individuals in other countries?

The plaintiffs assert that the extraterritorial application of international human rights law is necessary in this case. They argue that given the global impact of Meta's platforms, international legal standards should apply regardless of where the company is based. This could have significant implications for future cases involving multinational corporations that operate across borders. If the courts accept the plaintiffs' argument, it could set a precedent for holding tech companies accountable under international law, regardless of where they are headquartered or where the harm originated.

In the *Multistate complaint against Meta*[107], the plaintiffs claim that Meta's platforms have been used to alter the psychological and social realities of a generation of young Americans. This lawsuit is not only about Meta's inability to enforce its own policies but also about how its business model exacerbates the problem, in violation of the rules protecting minors and consumers, creating irreparable damage to society. Meta, like other social media giants, uses algorithms designed to maximize user engagement, often prioritizing sensationalist content. This model, according to the plaintiffs, amplifies hateful speech and extreme content, which ultimately contributes to societal harm:

> *Meta has harnessed powerful and unprecedented technologies to entice, engage, and ultimately ensnare youth and teens. Its motive is profit, and in seeking to maximize its financial gains. Meta has repeatedly misled the public about the substantial dangers of its social media platforms. It has concealed the ways in which these platforms exploit and manipulate its most vulnerable consumers: teenager and children. And it has ignored the sweeping damage these platforms have caused to the mental and physical health of our nation's youth. In doing so, Meta engaged in, and continues to engage in, deceptive an unlawful conduct in violation of state and federal law*[108].

A significant part of the plaintiffs' argument is Meta's failure use algorithms function[109] on a user-by-user basis and to adequately moderate the content posted on its platforms. Although Meta has established community standards that prohibit hate speech and harmful content, the plaintiffs argue that these standards are not enforced consistently. In many cases, harmful content remains online for extended periods, and the moderation process, they claim, is both inefficient and biased. Facebook has options for moderating its algorithms' tendency to promote hate speech and misinformation, but it rejects those options because the production of more engaging content takes precedence. In the case Doev. Meta:

---

107 Multistate complaint against Meta. The United State District Court for the Northern district of California, case 4:23 cv-05448. More information at: https://es.scribd.com/document/679809777/Meta-Multistate-Complaint.

108 Superior Court Of The State Of California for the county Of San Mateo, Jane Doe, individually and on behalf of all others similarly situated, META PLATFORMS, INC.

109 R. Gorwa, Robert, R. Binns, and C. Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', *Big Data & Society* 7(1), (2020), at: http://doi.org/10.1177/2053951719897945

> *Facebook designed its system and the underlying algorithms and in a manner that rewarded users for posting, and thereby encouraged and trained them to post, increasingly extreme and outrageous hate speech, misinformation, and conspiracy theories attacking particular groups. The design of Facebook's algorithms and product resulted in the proliferation and intensification of hate speech, misinformation, and conspiracy theories attacking the Rohingya in Burma, radicalizing users, causing injury to Plaintiff and the Class, as described above. Accordingly, through the design of its algorithms and product, Facebook (1) contributed to the development and creation of such hate speech and misinformation and (2) radicalized users, causing them to tolerate, support, and even participate in the persecution of and ethnic violence against Plaintiff and the Class.*

The algorithms engage and then increasingly display similar material to maximize the time spent on the platform. This function applies not only to material generated by users but also to advertisements. Meta denies that its recommendation algorithms are intentionally designed to be addictive or to push emotionally distressing content. However, Meta is aware that its algorithms are structured to encourage addictive behavior and amplify such content. By misrepresenting and omitting information about how these algorithms promote harmful material, Meta prevents users, particularly parents of young users, from making fully informed decisions about their engagement with its social media platforms.:

> *Meta's Recommendation Algorithms are optimized to promote user engagement. Serving harmful or disturbing content has been shown to keep young users on the Platforms longer. Accordingly, the Recommendation Algorithms predictably and routinely present young users with psychologically and emotionally distressing content that induces them to spend increased time on the Social Media Platforms. And, once a user has interacted with such harmful content, the Recommendation Algorithm feeds that user additional similar content. […] Again, though, Meta's public statements regarding its algorithms' amplification of distressing and problematic content did not reflect Meta's true awareness of these problems*[110].

We can see that at the core of the lawsuits is the assertion that Meta's business model, which prioritizes user engagement over content moderation, exacerbates the problem. Algorithms on platforms are designed to increase user interaction by promoting content that elicits strong reactions, often amplifying sensationalist and extreme content. The algorithms reward divisive and inflammatory speech because it generates more engagement. The plaintiffs assert that this model is not only negligent but also demonstrates a deliberate indifference to the harm caused by the spread of hate speech. Furthermore, the lawsuits critique Meta's self-regulation efforts. Despite having extensive content moderation guidelines, Meta's voluntary measures have been insufficient, especially given the scale of its global operations. The plaintiffs argue that Meta has consistently failed to address harmful content and that its internal guidelines are either too vague or inconsistently enforced. This inconsistency has allowed harmful speech to flourish on the platform, contributing to real-world violence and discrimination. The plaintiffs argue that self-regulation is no longer an adequate means of addressing the issue of hate speech, and external regulatory measures are required to hold Meta accountable.

In the same line, Meta Platforms must face *a lawsuit from the state of Massachusetts*[111], which claims that the company deliberately implemented features on its Instagram

[110] 177 and 183.

[111] More information at: https://fingfx.thomsonreuters.com/gfx/legaldocs/dwvkkdqyjvm/10182024meta.pdf

platform to hook young users and misled the public regarding the risks these features posed to teenagers' mental health.

Meta was also sued in Kenya over claims of exploitation mental health and poor working conditions of moderators[112], accused the company of failing to protect them from psychological injuries resulting from their exposure to graphic and violent imagery[113]. Moderators must repeatedly review content involving terrorism, suicides, self-harm, civilian beheadings by terrorist groups, and torture tasks performed under intense time pressure that require rapid decisions to approve or remove material. The lawsuit highlights the psychosocial risks associated with these duties[114]. Recently, a Barcelona-based company subcontracted by Meta to provide content moderation services for Facebook and Instagram has been held accountable by a Spanish court for psychological harm experienced by an employee[115]. This marks the first instance in Spain where a content moderation company has been found responsible for the mental health impact on a worker.

(F) CONCLUSION: THE NEED FOR INTERNATIONAL LEGAL REVOLUTION ON THE PLATFORMS CONTENT

Meta's lawsuits represent a pivotal moment in the ongoing conversation about the role of tech companies in moderating online speech. It underscores the urgent need for an international legal framework that holds multinational corporations accountable for their actions, particularly when it comes to harmful content that spreads across borders. The case also challenges the adequacy of self-regulation in the tech industry and advocates for a more robust, external regulatory framework that can effectively address the challenges posed by social media platforms.

As the lawsuit progresses, it may set an important precedent for how courts will address the accountability of tech companies in the digital age. The outcome of this case could pave the way for stronger international regulations governing online speech, especially in cases involving racial discrimination and incitement to violence. In the long run, this lawsuit could represent a turning point in the way we understand the responsibilities of multinational corporations, and the legal obligations they bear in protecting human rights in the digital realm.

---

112 T. Meskill, *Facebook content moderator speaks about mental health impact of her job*, RTE, 12 May 2021, (2021), Available at: https://www.rte.ie/news/ireland/2021/0512/1221241-online

113 Eurofound, *Employment and Working Conditions of Selected Types of Platform Work*, Luxembourg: Publications Office of the European Union, (2018) Available online at: https://www.eurofound.europa.eu/publications/report/2018/employment-and-workingconditions-of-selected-types-of-platform-work.

114 More information at: https://www.reuters.com/world/africa/kenya-court-rules-meta-can-be-sued-over-layoffs-by-contractor-2024-09-20/#:~:text=NAIROBI%2C%20Sept%2020%20(Reuters),content%20moderators%20by%20a%20contractor.
https://web.archive.org/web/20230608141240/https://www.theguardian.com/global-development/2023/jun/07/a-watershed-meta-ordered-to-offer-mental-health-care-to-moderators-in-kenya

115 M. T.Igartua Miró, ' Sobre la Síndrome de burnout de moderador de contenidos en línea como accidente de trabajo Comentario a la Sentencia del Juzgado de lo Social n.º 28 de Barcelona 13/2024, de 12 de enero', *Revista de Trabajo y Seguridad Social CEF* N.º 480 Mayo-Junio 2024, (2024).

In our view, these lawsuits raise broader questions about the role of international law in regulating global companies. As social media platforms like Facebook, Instagram, and Twitter become integral to public discourse, the legal framework governing these companies must evolve to reflect their global impact. Traditional notions of jurisdiction and accountability must be adapted to address the challenges posed by multinational corporations and their influence on global societies. At the same time, those cases are a critical test of how international human rights law and private international law intersect. It calls for a reevaluation and a revolution of the regulatory frameworks that govern multinational corporations and offers a glimpse into the future of tech industry accountability. As the digital landscape continues to evolve, legal systems must adapt to ensure that platforms like Meta are held accountable for the impact they have on societies worldwide.