Private Governance of Freedom of Expression on Social Media Platforms

EU content regulation through the lens of human rights standards

Rikke Frank Jørgensen & Lumi Zuleta

The Danish Institute for Human Rights, Copenhagen, Denmark

Abstract

For years, social media platforms have been perceived as a democratic gain, facilitating freedom of expression, easy access to a variety of information, and new means of public participation. At the same time, social media have enabled the dissemination of illegal content and incitement to discrimination, hostility, or violence, fuelling several content regulation initiatives. From the perspective of freedom of expression, this development embraces two challenges: first, private actors govern freedom of expression, without human rights safeguards; second, this privatised governance of human rights is encouraged and legitimised by a broad range of EU policy initiatives. Informed by an analysis of Danish Facebook users' attitudes toward public debate on Facebook, we pose the question: How do social media companies such as Facebook balance various human rights considerations on their platforms, particularly in relation to freedom of expression? We analyse the abovementioned challenges through a human rights lens, which serves as the analytical framework for this article. Further, we suggest some strategies for moving forward, drawing on recent recommendations from the UN human rights system.

Keywords: human rights, social media, content regulation, freedom of expression, EU

Introduction

For years, social media platforms such as Facebook have been perceived as a democratic gain, not least due to the potential of allowing everyone to exercise freedom of expression, including voicing opinions, reaching diverse audiences, sharing information from a variety of sources, locating likeminded people across borders, and mobilising around specific interests. However, with the swift growth and intense use of social media, new challenges emerge. The widespread use of social media platforms has enabled the dissemination of illegal content, incitement to discrimination, hostility, or violence, and a broad range of potentially harmful content. All of these can have damaging consequences not only for the targeted individuals, but for public debate as well (see Dangerous Speech

Jørgensen, R. F., & Zuleta, L. (2020). Private governance of freedom of expression on social media platforms: EU content regulation through the lens of human rights standards. *Nordicom Review, 41*(1), 51-67. https://doi.org/10.2478/nor-2020-0003

Project, 2020; DIHR, 2017, 2019). In response to these challenges, EU policymakers increasingly call upon social media platforms to regulate content. This policy development has led to a growing concern for the human rights implications of private actors governing the online public sphere.

From the perspective of freedom of expression, particularly two challenges are at stake. First, individual expression, public debate, and so forth are governed by private actors operating outside the direct reach of human rights law, placing freedom of expression in a vulnerable position. Second, EU policy initiatives combatting illegal content on social media platforms encourage and legitimise this private regime of content regulation – without adequate human rights safeguards. In fact, the EU policies create a regulatory incentive for over-removal that runs counter to the "strictly necessary" and "proportionate" principles embedded in human rights law.

Part of the human rights challenge with social media platforms like Facebook occurs because of their dual role as both a private company and a public space playing a pivotal role as access points to information. The difficulty in establishing the appropriate metaphor for what Facebook *is* makes it equally challenging to find the right regulatory response to their human rights impact (Jørgensen, 2013). Facebook is not a media corporation with an editor-in-chief subject to media regulation; however, its widespread use makes it as powerful as traditional media companies in many cases. Scholars have referred to Facebook as a public infrastructure or utility, essential for social and political participation in the twenty-first century and accessible for all (Balkin 2017; Plantin et al., 2016; Van Dijck et al., 2018), but it is a privately governed sphere – and legally a commercial service – free to define what is allowed and what is not. While Facebook refers to itself as a global community, it is effectively governed by commercially defined rules and norms largely inaccessible to its community (Gillespie, 2018; Klonick, 2018; Suzor, 2019).

In this article, we analyse these challenges through the lens of human rights standards and suggest a way forward. When evaluating state regulation and company practices, a human rights-based approach is used to determine to which extent users' human rights are protected within a given social domain and to ensure that state regulation and company practices adhere to and protect human rights standards. This perspective in the study of internet policy is not new (see, e.g., Wagner et al., 2019; Kerr et al., 2019)[1]; however, the contribution of this article is to situate the EU governance model towards social media companies within an international human rights context and to highlight its deficits in relation to protecting freedom of expression. Arguably, the EU governance model is one of several (contrasted with, e.g., the American "hands-off" and the Chinese "hands-on" models); yet, the EU model is an interesting case due to the EU's strong commitment to human rights, stipulated both in the EU Charter for Fundamental Rights and Freedoms and in the European Convention on Human Rights[2].

We begin by presenting some key findings from a 2018 survey of Danish social media users, highlighting how the respondents perceive the role of social media platforms like Facebook vis-à-vis their ability to enjoy freedom of expression. On that foundation, we next address the human rights framework and the regulatory challenges involved in protecting freedom of expression, as well as the boundaries of freedom of expression on social media platforms. The analysis is informed by recent EU policy initiatives in the field of content regulation and by international human rights law, including soft

law. We conclude with some recommendations for moving forward, drawing upon the recommendations of UN Special Rapporteur on freedom of expression David Kaye.

Danish survey on social media and freedom of expression

In 2018, the Danish Institute for Human Rights commissioned a YouGov survey to examine Danish Facebook users' attitudes towards social media, freedom of expression, and content moderation. Following a previous study, "Hate speech in the online public debate" (DIHR, 2017), the Institute was keen to understand in more detail how Danish social media users perceived Facebook's role as a space for exercising freedom of expression and what expectations they had for governance of the platform. In addition to general questions related to usage patterns and perceptions, the survey posed questions about participation in the online public debate, seeking concrete experiences of encounters with, for example, harassment and offensive behaviour when using Facebook.

Method and data

The survey was internet-based and based on answers from 2,305 Danish Facebook users aged 18 and older. It focused exclusively on Facebook since it is the most commonly used social media platform by the Danish population (DIHR, 2019). In fact, a recent study shows that 63 per cent of Danes use Facebook daily and it plays a vital role as a source of news and information, particularly among those aged 18–24 (DR Medieforskning, 2018).

To identify potential respondents, e-mail invitations were sent to those meeting the relevant criteria in the YouGov panel[3]. In order to ensure that the survey captured respondents who used Facebook actively, respondents had to have a Facebook profile and must have posted a comment on Facebook at some point. A comparison of the respondents with the Danish population in general indicates that the respondents are representative of the population when it comes to gender and age; however, there is a slight overrepresentation of both respondents aged 50–59 years and of those highly educated.

Key results

The growing use of social media platforms as forums for public debate implies new conditions – as well as challenges – for freedom of expression. On the one hand, the ease of sharing opinions with a broader public is an advancement for freedom of expression; on the other hand, the ease of expressing hostile and discriminating attitudes can deter others from freely expressing their views. This duality is a recurring theme in the survey, according to which 48 per cent perceive social media to be a gain for freedom of expression. Moreover, nearly one third (28%) of the respondents indicate that social media have had a positive impact on their exercise of freedom of expression (DIHR, 2019).

The survey confirms Facebook's dominant position in Denmark – almost half (48%) of the respondents found it to be "an important platform for the public debate in Denmark". But the question remains: How representative and pluralistic is the public debate unfolding on Facebook? According to the survey, gender and age significantly influence whether an individual is likely to participate in the public debate on Facebook: among the respondents, men participated far more frequently that women, and those over 50

were overrepresented (64%) while 18–29-year-olds were underrepresented (8%). The findings also suggest that the tone of the debate has a significant chilling effect on civic engagement: 59 per cent refrain from posting a comment on Facebook because of the tone, suggesting a strong connection between the tone of the debate and self-censorship in public participation. The fact that some refrain from voicing their opinion in online debates was seen as a problem for freedom of expression by 63 per cent of the respondents. But at the same time, 62 per cent found it important to safeguard freedom of expression despite offensive comments.

Derogatory and offensive language was identified as the most prevalent type of offensive behaviour on Facebook, with half of the respondents observing this type of behaviour often or from time to time. One out of five witnessed sexually offensive comments often or from time to time, and the same number witnessed threats against others often or from time to time. Women taking part in the public debate on Facebook experienced derogatory and offensive comments based on their gender three times as often as men; contrarily, men primarily experienced derogatory and offensive comments about their political opinions.

The survey also examined attitudes towards content moderation. More than half (53%) believe that Facebook should ensure a healthy environment for public debate by moderating user-generated content, potentially removing content conflicting with Facebook's community standards[4]. This indicates that most users do not recognise content removal as an intervention in their freedom of expression – in fact, three out of four do not perceive it as a freedom of expression issue at all. The respondents are more concerned with Facebook keeping the platform free from harmful content than they are with the potential human rights implications of Facebook's content moderation practices. The nature of these human rights implications is further addressed below.

The human rights framework

Human rights are legally codified norms applying to all human beings, irrespective of national borders. International human rights law obligates states to act in certain ways or refrain from certain acts in order to protect the human rights of individuals. Since 2012, UN resolutions have iterated that human rights, including freedom of expression, must be protected online as well as offline (UNHRC, 2012). According to article 19 of the International Covenant on Civil and Political Rights (ICCPR, 1966), states must ensure an enabling environment for freedom of expression and protect the exercise thereof (para. 1). Freedom of expression is not absolute, but any restriction *must* meet the criteria: "provided by law and are necessary" for protecting individual rights or reputations, public order, or public health and morals (para. 3). According to article 20 of the ICCPR, a legitimate restriction to freedom of expression is "national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" (para. 2). Likewise, the Covenant obliges states to implement and enforce appropriate and effective measures to prevent and protect against acts of discrimination on several grounds, including sex, race, colour, descent, or national origin (article 2, para. 1).

As part of their human rights obligations, states must ensure that human rights are protected in the realm of non-state actors. As such, states incur responsibility not only for human rights abuses inflicted by themselves, but also those caused by third parties

that they fail to prevent, punish, and remediate (UNHRCom, 2004: para. 8). In an online domain dominated by privately owned platforms, the ability to enjoy human rights is thus closely related to whether states have transposed them into national regulation applicable to companies, and the willingness of companies to voluntarily undertake human rights due diligence. UN Special Rapporteur on violence against women Dubravka Šimonovic (2018: para. 115) notes:

> Internet intermediaries should uphold the principle that human rights are protected online, and voluntary [*sic*] accept and apply all core international human rights and women's rights instruments with a view to contributing to universal human rights protection and achieving the empowerment of women, and the elimination of discrimination and violence against them in digital space.

In recent years, a variety of initiatives have been introduced providing guidance to companies for ensuring compliance with human rights, most notably the Guiding Principles on Business and Human Rights adopted by the UN Human Rights Council in 2011. According to these Guiding Principles, any business entity has a responsibility to respect human rights. As part of this, they must avoid causing or contributing to adverse human rights impacts and seek to prevent or mitigate such impacts directly linked to their operations, products, or services by their business relationships – even if they have not contributed to those impacts. Moreover, the Guiding Principles stipulate that businesses should be prepared to communicate how they address their human rights impacts externally, particularly when concerns are raised by, or on behalf of, affected stakeholders.

While the Guiding Principles are nonbinding, the overwhelming role of social media companies in public life globally provides a strong argument for their adoption and implementation (Kaye, 2018). The Human Rights Council stresses that a company's responsibility to respect human rights is a global standard that "exists independently of states' abilities and/or willingness to fulfil their own human rights obligations and does not diminish those obligations" (UNHRC, 2011: para. 11). Former UN High Commissioner for Human Rights Navi Pillay (2014: para. 43) reiterates this: "The responsibility to respect human rights applies throughout a company's global operations regardless of where its users are located and exists independently of whether the state meets its own human rights obligations".

Since the Guiding Principles are the prevailing (and minimum) standard for defining and assessing the responsibility of social media platforms in relation to human rights, it is important to bear in mind the expectations to companies highlighted by these principles. Drawing on Kaye (2018: para. 11), we group the expectations into three themes relating to policy commitment, human rights due diligence, and remedy mechanisms:

1. Policy commitment: The company shall ensure high-level policy commitments to respect human rights.

2. Human rights due diligence:
   a. The company shall identify, address, and account for actual and potential human rights impacts of business activities, including through regular risk and impact assessments; meaningful consultation with potentially affected groups and other stakeholders; and appropriate follow-up action that mitigates or prevents these impacts.

b. The company shall engage in prevention and mitigation strategies that respect principles of internationally recognised human rights to the greatest extent possible when faced with conflicting local law requirements.
c. The company shall conduct ongoing review of efforts to respect rights, including through regular consultation with stakeholders and frequent, accessible, and effective communication with affected groups and the public.

3. Remedy mechanisms: The company shall provide appropriate remediation, including through operational-level grievance mechanisms that users may access without worsening their "sense of disempowerment".

We return to the governance themes below, but first we take a closer look at some of the challenges that occur when trying to determine the human rights impact of social media platforms.

The human rights impact of social media platforms

Over the past years, social media have contributed positively to individuals' ability to enjoy a broad range of human rights beyond freedom of expression, having a transformative impact on individuals' ability to assemble, mobilise, learn, educate, and so forth around the globe. A growing awareness exists, however, that the digital domain also entails negative human rights implications and might facilitate new instances of violence, hate, and discrimination.

The fact that social media platforms provide modalities for a broad range of processes related to public life and participation implies that there are additional intersections between business activities and human rights other than the traditionally well-known examples, such as human rights harm related to working conditions or impact on a local community. In addition to having obligations towards their employees and the communities in which they operate, companies may negatively affect the human rights of billions of users as part of the services and platforms they provide (BSR, 2014). This reality presents significant challenges for clarifying the human rights responsibilities of these companies. While they may contribute to a range of more well-known human rights abuses, the reach and impact on their users worldwide is unique to the sector. A specific content regulation policy may impact billion of users' ability to express themselves and seek information; yet, it is not clear if and when such a policy would amount to human rights abuse (Land, 2019).

As mentioned above, states incur responsibility not only for human rights abuses inflicted by themselves, but also those caused by third parties which they fail to prevent, punish, and remediate. In relation to freedom of expression, state action has traditionally been an essential element of alleged human rights violations. For example, if a state orders a private platform to remove content, this constitutes a violation of the right to freedom of expression under human rights law unless the order is provided by law and necessary to pursue a legitimate aim; however, when a platform decides to remove content because it violates its terms of service, this is private action outside the direct reach of human rights law (Jørgensen, 2018). Legally speaking, the relationship between the platform and the user is governed by the terms of service (contract law), rather than human rights law. While Facebook's practices may affect individuals' ability to exercise

freedom of expression, it does not have a legal obligation to protect this right. Scholarship has only recently begun to address the broader societal implications of having the online public sphere based on privately owned platforms – and the challenges raised by this governance gap (Callamard, 2019; Gillespie, 2018; Jørgensen, 2019; Klonick, 2018; Laidlaw, 2015; Suzor, 2019; Zuboff, 2019). Effectively, private actors with strong human rights impacts steer in the soft regime of guidelines and corporate social responsibility.

The UN Guiding Principles on Business and Human Rights (UNHRC, 2011), as previously discussed, requires businesses to assess their human rights impacts as part of a due diligence obligation. This requirement applies to all companies, but contrary to many other sectors, human rights impact assessment is still a relatively new concept for social media companies (Jørgensen et al., 2019). The prevailing industry initiative is the Global Network Initiative (GNI), established in 2008 to guide companies when states make requests that may violate international human rights standards of freedom of expression and privacy (Maclay, 2014). The GNI's approach has been to help companies enact policies that anticipate and respond to situations in which host country law and practice differ from international human rights standards. As part of this effort, the companies publish annual transparency reports in which they reveal aggregate numbers about state requests for interference in user communication. Moreover, the participating companies commit to undergo periodic assessment by an independent third party to evaluate their compliance with the GNI principles (these assessments are not publicly available except for a summary report).

Research on Facebook's human rights approach suggests that the company tends to focus on its role vis-à-vis suppressive states and less on the human rights impacts of its own business practices; for example, its enforcement of community standards (Jørgensen, 2017; Ranking Digital Rights, 2019). The focus on state overreach is not surprising, as these cases have attracted much attention in public debate. Moreover, the emphasis on state overreach provides the company with an element of discretion when deciding which internal processes to include or exclude in its human rights impact assessment.

In the following, state initiatives aiming to remove content from the online domain are referred to as *content regulation* (Cooke, 2007; Frydman et al., 2009; Jørgensen & Pedersen, 2017), whereas the companies' enforcement of their community standards is termed *content moderation* (Gillespie, 2018; Klonick, 2018; Roberts, 2019; York & Zuckerman, 2019). While content regulation is largely concerned with removal of illegal content – thus enforcing the boundaries for freedom of expression – content moderation typically involves both legal and illegal content, as defined by companies in their terms of service. Since human rights law provides legal standards for the former, and limited guidance for the latter, the distinction is important to understand. Moreover, as you shall see below, the two are increasingly blurred.

The EU model – from limited liability to proactive measures

Content regulation has been on the EU policy agenda since the mid-1990s, and the ramifications for freedom of expression are addressed in several studies and reports (Jørgensen et al., 2015; Keller, 2018; Tambini et al., 2008). However, it has gained new momentum recently, not least as a state response to counter illegal content on social media platforms.

For the past 20 years, the Directive on Electronic Commerce (EC, 2000) has provided the basis for the EU regime for intermediary liability in situations of "mere conduit" (article 12), "caching" (article 13), and hosting (article 14). These articles exempt intermediaries from liability in cases where the users of the platform, network, and so forth infringe the rights of others (Riis & Schwemer, 2019), while article 15 establishes that states shall not impose a general monitoring obligation on providers. The exemption from intermediary liability is conditioned on a) the provider having no actual knowledge of illegal activity or b) the provider, when obtaining such knowledge, acting expeditiously to remove or disable access to the information. In other words, a platform is lifted from liability provided it removes illegal content quickly when notified, and it shall not monitor its users' activities. However, it remains unclear how quickly a platform is expected to react to illegal content to be exempt from liability (Riis & Schwemer, 2019).

As part of the increased policy attention towards illegal content – not least on social media platforms – regulation is now introduced that supplements the limited liability regime of the Directive on Electronic Commerce, with expectations of "proactive measures" for user-generated content (Riis & Schwemer, 2019) to make the takedown regime more efficient. In 2017, the European Commission issued a Communication, "Tackling illegal content online", that aimed to establish an "enhanced responsibility of online platforms" (EC, 2017). This was followed, in 2018, by a Recommendation "on measures to effectively tackle illegal content online" (EC, 2018a), which stipulates that "hosting service providers should be encouraged to take, where appropriate, proportionate and specific proactive measures in respect of illegal content" (Ch. II, point 18). In line with this, the Terrorist Content Regulation currently being negotiated (EC, 2018b) mentions proactive measures, including by using "automated means" to effectively identify and remove "terrorist content" (EC, 2018b: 17). Likewise, the recently adopted Audiovisual Media Services Directive (EPCO, 2018) demands that video platforms use "appropriate means" to ensure that their services do not contain any incitement to violence or hatred. The combination of a limited liability regime and the call for proactive measures effectively demands social media companies to operate within a blurred mix of expectations and demands. On the one hand, they are expected not to interfere with content and to keep their status as mere conduit, caching, or host; on the other hand, they are expected to proactively detect, identify, and remove content (we shall return to this below).

Adding to this complexity, the European Commission and Facebook, Microsoft, Twitter, and YouTube have agreed on the "EU Code of Conduct on countering illegal hate speech online" (EC, 2016). The agreement includes the development of internal procedures to guarantee that the companies review notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content if necessary. It also includes partnerships with civil society organisations (so-called "trusted reporters") who should help flag content promoting incitement to violence and hateful conduct. The Code of Conduct defines illegal hate speech according to European Council's Framework Decision on combating certain forms and expressions of racism and xenophobia: as all conduct "publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent, or national or ethnic origin" (European Council, 2008; quoted in EC, 2016: para. 2). Currently, there is no uniform definition of what constitutes hate speech around the world, and the Framework Decision has been criticised for lack of compliance

with international standards on freedom of expression, as pointed out by the UK-based organisation ARTICLE 19 (2016).

Since its adoption, the Code of Conduct has been supplemented with various national initiatives. In 2017, Germany introduced the Network Enforcement Act in order to tackle "hate speech" on social media platforms (see German Law Review, 2017). The Act obliges owners of social media platforms with more than two million German users to remove illegal content within 24 hours or risk sanctions with fines up to 50 million euros. In the UK, former Prime Minister Theresa May has encouraged industry to "go further and faster" in removing terror-related content, including by developing automated filters to detect and suppress it automatically (Hope & McCann, 2017). Most recently, an Online Harms White Paper sets out the British government's plans for a "world-leading package of online safety measures" comprising legislative and non-legislative measures to make companies more responsible for their users' safety online (GOV.UK, 2019).

Scholars and commentators have repeatedly warned that the EU approach to content regulation uses intermediaries (in this case social media platforms) to implement public policy with limited oversight and with severe implications for freedom of expression (Brown, 2010; Jørgensen et al., 2015; Keller, 2018; Korff, 2014; Mackinnon et al., 2014). When companies may be sanctioned for not rapidly identifying and removing illegal content, this creates an incentive for over-removal ("better safe than sorry") and may lead to a disproportionate takedown of legal content in order to target a smaller amount of illegal material, which contradicts the "strictly necessary" and "proportionate" principles embedded in human rights law. In response to the EU draft of Terrorist Content Regulation (EC, 2018b), several UN special rapporteurs stated their concern for the short amount of time platforms have "to comply with the sub-contracted human rights responsibilities that fall to them by virtue of State mandates on takedown" (Kaye et al., 2018: 6). The rapporteurs note that the short timeframe and the threat of penalties are likely to "incentivize platforms to err on the side of caution and remove content that is legitimate or lawful" (Kaye et al., 2018: 6), profoundly effecting users' human rights and undermining the potential for meaningful remedies to be quickly activated. Likewise, proactive measures such as upload filters would enable the blocking of content without any form of due process even before it is published (Kaye et al., 2018). Such practice would reverse the human rights standard that states – not individuals – bear the burden of justifying restrictions on freedom of expression, and it would make it practically impossible to uphold the "strictly necessary" and "proportionate" principles of international human rights law. Moreover, such proactive measures seem to conflict with the obligations of the Directive on Electronic Commerce – to not interfere with content nor monitor it (EC, 2000). Anticipating this potential conflict, the Explanatory Memorandum to the draft Terrorist Content Regulation states that "any measures taken by the hosting service provider in compliance with this Regulation, including any proactive measures" will not lead to the provider losing the liability exemption under the Directive for Electronic Commerce. (Kaye et al., 2018: 9). However, as such recitals are not binding, this may lead to legal uncertainty, impacting both platforms and individuals, and potentially undermining the protection of human rights (Kaye et al., 2018).

Effectively, social media platforms operate in a legal grey-zone with conflicting expectations related to their role vis-à-vis content regulation. In practice, they are asked

to navigate between three set of norms. First, limited liability schemes that expect them to not monitor content but remove illegal content when notified in order to benefit from "safe harbours" provisions (EC, 2000). Second, expectations of enhanced responsibility and proactive measures (Code of Conduct, Terrorist Content Regulation, Audiovisual Media Services Directive). And third, conducting human rights impact assessments to mitigate negative human rights impacts, as stipulated in the UN Guiding Principles on Business and Human Rights.

This zone of unclear expectations, norms, and liability provisions is partly due to the character of the online domain. With private companies in control of social media platforms, it is no surprise that EU regulators and member states have turned to these actors to regulate content, as it is outside their direct sphere of control. Looking through the prism of the right to freedom of expression, however, this practice is problematic and calls for standards from EU regulators to ensure that fundamental rights are protected when regulatory action is delegated to private actors. In the absence of such standards, the legal grey-zone presented by regulation and codes of conduct are transposed to national level in the EU member states. Consequently, social media platforms are left with self-devised standards while carrying out practices that affect users' human rights. This privatised law enforcement challenges international human rights standards whereby states, as duty bearers, have an obligation to respect and protect individuals' human rights.

The company model – community standards rule

While content regulation is concerned with removal of illegal content mandated by states, content moderation refers to company practices involving both legal and illegal content, as defined in the terms of service and community standards (Gillespie, 2018 ; Roberts, 2019; York & Zuckerman, 2019). In recent years, the content moderation practices of social media companies increasingly evoke attention. Social media companies such as Facebook are subject to continuous criticism for not doing enough in terms of policing their platforms, for example in relation to hate speech, and for doing too much, such as removing legal content. In relation to the criticism of "not doing enough", this perception is illustrated by the Danish YouGov survey, where more than half (53%) of respondents believe that Facebook should take responsibility for the public debate and moderate user-generated content (see *Key results* above).

In contrast to this position, several UN special rapporteurs (Kaye et al., 2018: 7) have noted that social media platforms' terms of service and community standards frequently impose limitations beyond what states could do in compliance with their obligations under international human rights law:

> Such standards are commonly drafted in terms that lack sufficient clarity and fail to provide adequate guidance on the circumstances under which content may be blocked, removed or restricted, or access to a service may be restricted or terminated, thereby falling short of the legality requirement under international human rights law.

These shortcomings become particularly problematic when states expect companies to take on quasi-regulative and quasi-enforcement functions (as discussed above); hence,

it is crucial to ensure that such sub-contracting of state obligations are compliant with human rights (Kaye et al., 2018).

In practice, companies' content moderation policies are continuously revised to reflect legal standards (in particular US law) and ever-evolving company norms for the types of expressions allowed within their services. As such, a diverse mix of legal and non-legal standards guide the numerous decisions taken on content each day. The justifications for content removal (and account deactivation) range from content illegal under US law (e.g., child exploitation, terrorism, copyright violations, fraud, and criminal activity) to content that is legal but outlawed by the community norms (e.g., pseudo-identity, harassment of others, harmful content, nudity and sexually explicit content, and certain categories of graphic content). The content categories are not clearly demarcated, and their enforcement continues to evoke public debate in specific cases as illustrated, for example, by the removal of the "napalm girl" photo posted by the Norwegian newspaper *Aftenposten* (BBC News, 2016). While the companies' handling of state requests for takedown of content is governed by human rights standards, enforcement of their community standards is not. In practice, the incentive to uphold freedom of expression standards is countered by the mixed set of norms that make up community standards. Regarding enforcement, Facebook relies on a combination of software and users to flag "inappropriate" content, which is then reviewed and decided upon by a globally distributed team of content reviewers. In short, content moderation implies that social media companies set the boundaries for their users' freedom of expression based on their community standards rather than the criteria prescribed by human rights law.

Transparency, or lack thereof, is therefore an important theme in relation to content moderation practices; for example, in relation to numbers and types of content removed and the decisions and practices informing those removals. The powerful role Facebook has in governing public debate has prompted calls for an increased level of transparency to enable some level of public oversight. In 2018, for example, a group of scholars, freedom of expression advocates, and platform representatives gathered at the first "Content Moderation & Removal at Scale" conference to examine the operational challenges of content moderation and how companies are addressing them. As a result, the participants agreed on the "Santa Clara principles on transparency and accountability in content moderation", which demand that companies should publish the *number* of posts removed and accounts suspended; provide *notice* and explanation to each user affected; and provide a *meaningful opportunity for appeal* of any content removal or account suspension (Santa Clara Principles, 2018).

In 2018, Facebook – for the first time – decided to publish its internal guidelines on how it enforces its community standards (Bickert, 2018), as well as a preliminary transparency report related to community standards enforcement (Facebook transparency, 2020). The company has also started an appeal process to enable users to object to individual content decisions. Despite such initiatives, it remains an open question how freedom of expression concerns raised by corporate policy, design, and engineering choices should be reconciled with the freedom of private entities to design and customise their platforms as they choose (Kaye, 2016: para. 55).

Protecting freedom of expression on Facebook

Part of the regulatory challenge with social media platforms like Facebook occurs because their services resemble a public sphere (Jørgensen, 2018; Moore, 2016; York & Zuckerman, 2019) yet operate purely within the remit of private law. As illustrated by the Danish YouGov survey, Facebook has established itself as a significant venue for public debate in the country, with almost half (48%) of the respondents stating that they find Facebook to be "an important platform for the public debate in Denmark". However, while it provides an open, widely accessible space for debate and news consumption, this space suffers from some of the same flaws as its normative ideal[5]; for example, by enabling public debate in which men participate far more frequently than women and with an age bias towards the "older" (50+ years old), while the "younger" (18–29-year-olds) are underrepresented. Further, the survey findings suggest that it is a public sphere in which the tone of the debate has a significant chilling effect on its participants, since it keeps more than half (59%) of the respondents from sharing their opinions, in particular due to derogatory and offensive language. The chilling effect is gender biased since women in the survey experienced derogatory and offensive comments based on their gender three times as often as men. The male respondents, to the contrary, primarily experienced derogatory and offensive comments based on their political opinion.

Despite Facebook's seemingly public functions, its legal status as a private service provider affords it the freedom to design, conduct, and govern this public sphere on the basis of commercial priorities rather than public interest. In terms of freedom of expression, Facebook has no legal obligation to protect its exercise, since such an obligation would require national regulation, which is not the case in Denmark, nor elsewhere. The challenge is thus to devise a way forward, whereby users' human rights are protected, considering recent developments both regarding content regulation (state initiatives) and content moderation (company practices). In the final section, we propose three recommendations for moving forward, drawing upon the governance themes of the UN Guiding Principles on Business and Human Rights and recommendations from the UN's authoritative source on this matter, the UN Special Rapporteur on freedom of expression David Kaye.

First, states must ensure that any content regulation measure is in accordance with human rights standards

As part of their human rights obligations, states should ensure that standards of legality, necessity, and proportionality are adhered to in any content regulation measure they introduce or suggest, and in order to avoid a chilling effect on freedom of expression, they should refrain from imposing disproportionate sanctions on intermediaries (Kaye, 2018: para. 66). Any automated measure must be specific and proportionate to ensure that the tackling of illegal content does not violate users' right to freedom of expression. Also, states should avoid delegating responsibility to companies as adjudicators of content, which empowers corporate judgment over human rights values to the detriment of users (Kaye, 2018: para. 68). Finally, they should publish detailed transparency reports on all content-related requests (Kaye, 2018: para. 69) and ensure efficient complaint mechanisms.

Second, social media platforms must demonstrate a commitment to human rights

While the UN Guiding Principles (UNHCR, 2011) are not legally binding, they represent the authoritative global standard for business and human rights and should set the direction of legal obligations as soft law norms that may crystallise to hard law obligations over time (Kaye et al., 2018). Until hard law obligations are in place, states should require companies to demonstrate a top-level policy commitment to human rights grounded in the Guiding Principles. As part of this, companies must demonstrate that human rights standards form the baseline for their terms of service, community standards, and other policies governing the use of their platforms (see below). Moreover, given their impact on the public sphere, social media platforms must open themselves up to public accountability. This entails, among others, transparency reporting including granular data on the volume and types of requests the company receives, actions taken, the volume and types of users' appeals, response times, and the rate at which such appeals are granted (Kaye, 2018: para. 72).

Third, social media platforms must prove human rights due diligence across their operations

Human rights due diligence begins with "rules rooted in rights", continues with rigorous human rights impact assessments for product and policy development, and moves through operations with ongoing assessment and follow-up action that prevents or mitigates identified negative impacts including meaningful consultation with affected groups and stakeholders (Kaye, 2018: para. 70). One element in human rights due diligence is ensuring that enforcement of the platform's content rules is based on international standards of freedom of expression (ICCPR, 1966: article 19) and providing meaningful due process. As part of due process, companies should provide notice to users whose content is taken down or account is suspended, revealing the reason for the removal or suspension and providing an effective opportunity for appealing any such decision, as iterated by the Santa Clara Principles (2018).

Conclusion

The position of social media platforms as public sphere, social infrastructure, and governors of public debate has resulted in a range of policy challenges, leading EU legislators and member states to propose various forms of content regulation to tackle illegal content. Tackling and hindering the spread of illegal content online is important and necessary; however, these legislative responses effectively encourage privatised law enforcement and support self-regulatory practices based on company community standards rather than human rights standards. Moreover, company enforcement of community standards effectively provides for a public sphere governed by commercial priorities rather than public interest and human rights law. Currently, the ecosystem of social media platforms represents a governance gap in human rights protection, and it falls upon the states – as human rights duty bearers – to secure legislative responses firmly anchored in human rights law. With its declared commitment to human rights, the EU has a special responsibility to ensure that the policy and governance models it suggests

provide for effective human rights protection – including in the private realm. This is especially urgent, as platforms like Facebook increasingly constitute the primary, if not exclusive, point of access to information for many people. Unfortunately, the debate on content regulation within the EU continues to show limited attention to the freedom of expression issues evoked by these arrangements, thus leaving it to private companies to set the boundaries for freedom of expression online.

Notes

1. For a summary of the human rights and technology literature, please refer to Jørgensen, 2019: xxiv-xxvii.
2. The European Convention of Human Rights is anchored within the Council of Europe, which has for the past 20 years been instrumental in devising human rights standards related to internet governance (see Jørgensen 2013: 53–56).
3. The YouGov panel is a user panel with more than 90,000 self-registered Danish respondents.
4. These findings are part of the survey but not included in the report (DIHR, 2019).
5. Jürgen Habermas's normative and influential ideal of the "public sphere" has been widely criticised for its lack of attention to structures of inequality; for example, the absence of women, and private matters of public concern in the public sphere.

References

ARTICLE 19. (2016). *EU: European Commission's code of conduct for countering illegal hate speech online and the framework decision*. (Legal Analysis). London. https://www.article19.org/data/files/medialibrary/38430/EU-Code-of-conduct-analysis-FINAL.pdf

Balkin, J. M. (2017). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UC Davies Law Review 51*, 1149.

BBC News. (2016). Facebook u-turn over "napalm girl" photograph. *BBC News*. https://www.bbc.com/news/technology-37318040

Bickert, M. (2018, April 24). Publishing our internal enforcement guidelines and expanding our appeals process. *Facebook News*. Retrieved January 23, 2020, from https://about.fb.com/news/2018/04/comprehensive-community-standards/

Brown, I. (2010). Internet self-regulation and fundamental rights. *Index on Censorship 1*, 98–106.

BSR (2014). Legitimate and meaningful: Stakeholder engagement in human rights due diligence. Retrieved February 4, 2020, from https://www.bsr.org/en/our-insights/report-view/engaging-with-rights-holders

Callamard, A. (2019). The human rights obligations of non-state actors. In R. F. Jørgensen (Ed.), *Human rights in the age of platforms* (pp. 191–225). Cambridge, Massachusetts: MIT Press.

Cooke, L. (2007). Controlling the net: European approaches to content and access regulation. *Journal of Information Science, 33*(3), 360–376. https://doi.org/10.1177/0165551506072163

Dangerous Speech Project. (2020). *We study dangerous speech and ways to counteract it.* Retrieved February 4, 2020, from https://dangerousspeech.org

Danish Institute for Human Rights (DIHR). (2017). Hadefulde ytringer i den offentlige debat [Hate speech in the online public debate]. (In Danish – includes English summary). Retrieved February 4, 2020, from https://menneskeret.dk/sites/menneskeret.dk/files/media/dokumenter/udgivelser/ligebehandling_2017/rapport_hadefulde_ytringer_online_2017.pdf

Danish Institute for Human Rights (DIHR). (2019). Demokratisk deltagelse på Facebook [Democratic participation on Facebook]. (In Danish – includes English summary). Retrieved February 4, 2020, from https://menneskeret.dk/sites/menneskeret.dk/files/04_april_19/Rapport%20om%20demokratisk%20deltagelse.pdf

DR Medieforskning. (2018). Medieudvikling 2018 [Media development 2018]. Retrieved February 4, 2020, from https://www.dr.dk/om-dr/fakta-om-dr/medieforskning/medieudviklingen/2018

European Parliament, Council of the European Union (EPCO). (2018). *Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU (Audiovisual Media Services Directive)*. https://eur-lex.europa.eu/eli/dir/2018/1808/oj

European Commission (EC). (2000). *Directive on electronic commerce (2000/31/EC)*. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32000L0031:en:HTML

European Commission (EC). (2016). *EU code of conduct on countering illegal hate speech online*. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en

European Commission (EC). (2017). *Tackling illegal content online: Towards an enhanced responsibility of online platforms* (Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions). https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms

European Commission (EC). (2018a). *Commission recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online* (*Official Journal of the European Union* L63/50). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018H0334&from=FR

European Commission (EC). (2018b). *Proposal for a regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online. COM (2018) 640 final.* https://eur-lex.europa.eu/resource.html?uri=cellar:dc0b5b0f-b65f-11e8-99ee-01aa75ed71a1.0001.02/DOC_1&format=PDF

European Council. (2008). *Council framework decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law* (*Official Journal of the European Union* L 328/55). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=EN

Facebook transparency. (2020). Community standards enforcement report. Retrieved January 23, 2020, from https://transparency.facebook.com/community-standards-enforcement

Frydman, B., Hennebel, L., & Lewkowicz, G. (2009). Public strategies for Internet co-regulation in the United States, Europe and China. In E. Brousseau, M. Marzouki, & C. Meadel (Eds.), *Governance, regulations and powers on the internet* (pp. 133–150). Cambridge: Cambridge University Press.

German Law Review. (2017). Network enforcement act. Retrieved January 23, 2020, from https://germanlawarchive.iuscomp.org/?p=1245

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, Connecticut: Yale University Press.

GOV.UK. (2019, April 8). Closed consultation: Online harms white paper. Retrieved January 23, 2020, from https://www.gov.uk/government/consultations/online-harms-white-paper

Hope, C., & McCann, K. (2017, September 19). Google, Facebook and Twitter told to take down terror content within two hours or face fines. *The Telegraph.* https://www.telegraph.co.uk/news/2017/09/19/google-facebook-twitter-told-take-terror-content-within-two/

ICCPR (1966, December 16). *International covenant on civil and political rights*. United Nations Human Rights Office of the High Commissioner. https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx

Jørgensen, R. F. (2013). *Framing the net: The Internet and human rights*. Cheltenham, UK: Edward Elgar Publishing.

Jørgensen, R. F. (2017). Framing human rights: Exploring storytelling within internet companies. *Information, Communication & Society 21*(3), 340–355. https://doi.org/10.1080/1369118X.2017.1289233

Jørgensen, R. F. (2018). Human rights and private actors in the online domain. In M. K. Land, & J. D. Aronson (Eds.), *New technologies for human rights law and practice* (pp. 243–269). Cambridge: Cambridge University Press.

Jørgensen, R. F. (Ed.). (2019) *Human Rights in the Age of Platforms*. Cambridge, Massachusetts: MIT Press.

Jørgensen, R. F., & Pedersen, A. M. (2017). Online service providers as human rights arbiters. In M. Taddeo, & L. Floridi (Eds.), *The responsibilities of online service providers* (pp. 179–199). Oxford: Oxford University Press.

Jørgensen, R. F., Pedersen, A. M., Benedek, W., & Nindler, R. (2015, November 30). *Case study on ICT and human rights (policies of EU)* (Work package no. 2 – deliverable no. 3*)*. European Commission. http://www.fp7-frame.eu

Jørgensen, R. F., Veiberg, C. B., & ten Oever, N. (2019). Exploring the role of HRIA in the information and communication technologies (ICT) sector. In N. Götzmann (Ed.), *Handbook on human rights impact assessment* (pp. 205–218). Cheltenham, UK: Edward Elgar Publishing.

Kaye, D. (2016). *Report of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression*. (A/HRC/32/38). United Nations Human Rights Council. https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/32/38

Kaye, D. (2018). *Report of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression* (A/HRC/38/35). United Nations Human Rights Council. https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ContentRegulation.aspx

Kaye, D., Cannataci, J., & Ní Aoláin, F. (2018, December 7). *Mandates of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression; the special rapporteur on the right*

to privacy and the special rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism (Reference: OL OTH 71/2018). United Nations Human Rights Council. https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24234

Keller, D. (2018). *Internet platforms: Observations on speech, danger, and money* (Aegis series paper no. 1807). A Hoover Institution essay, Stanford University, California. Retrieved February 5, 2020, from https://cyberlaw.stanford.edu/files/publication/files/381732092-internet-platforms-observations-on-speech-danger-and-money.pdf

Kerr, A., Musiani, F., & Pohle, J. (2019). Communication and internet policy: A critical rights-based history and future. *Internet Policy Review, 8*(1). https://ssrn.com/abstract=3434969

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review, 131*(6), 1598–1670.

Korff, D. (2014). *The rule of law on the Internet and in the wider digital world* (Issue paper). Council of Europe Commissioner for Human Rights. Retrieved February 5, 2020, from https://www.sbs.ox.ac.uk/cybersecurity-capacity/system/files/70114_Rule%20of%20Law%20on%20the%20Internet_web.pdf

Laidlaw, E. (2015). *Regulating speech in cyberspace: Gatekeepers, human rights and corporate responsibility*. Cambridge University Press. https://doi.org/10.1017/CBO9781107278721

Land, M. K. (2019). Regulating private harms online: Content regulation under human rights law. In R. F. Jørgensen (Ed.), *Human rights in the age of platforms* (pp. 285–316). Cambridge, Massachusetts: MIT Press.

MacKinnon, R., Hickok, E., Bar, A., & Lim, H. (2014). *Fostering freedom online: The role of Internet intermediaries*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000231162

Maclay, C. M. (2014). *An improbable coalition: How businesses, non-governmental organizations, investors and academics formed the global network initiative to promote privacy and free expression online.* (A dissertation presented in partial fulfilment of the requirements for the degree Doctor of Philosophy). Northeastern University, Boston, Massachusetts. Retrieved February 5, 2020, from https://repository.library.northeastern.edu/files/neu:336802/fulltext.pdf

Moore, M. (2016, April). *Tech giants and civic power*. Centre for the study of Media, Communication & Power, King's College London. Retrieved February 5, 2020, from https://www.kcl.ac.uk/policy-institute/assets/cmcp/tech-giants-and-civic-power.pdf

Pillay, N. (2014). *The right to privacy in the digital age* (A/HRC/27/37). The Office of the United Nations High Commissioner for Human Rights.

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2016). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society, 15*(7), 751–760.

Ranking Digital Rights. (2019). *2019 Ranking Digital Rights Corporate Accountability Index*. Retrieved February 4, 2020, from https://rankingdigitalrights.org/index2019/

Riis, T., & Schwemer, S. F. (2019). Leaving the European safe harbor, sailing toward algoritmic content regulation. *Journal of Internet Law, 22*(7), 11–21.

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. New Haven, Connecticut: Yale University Press.

Santa Clara Principles. (2018). The Santa Clara principles on transparency and accountability in content moderation. Retrieved January 23, 2020, from https://santaclaraprinciples.org

Šimonovic, D. (2018). *Report of the special rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective* (A/HRC/38/47). United Nations, Human Rights Council. https://digitallibrary.un.org/record/1641160?ln=en

Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge: Cambridge University Press.

Tambini, D., Leonardi, D., & Marsden, C. (2008). *Codifying cyberspace: Communications self-regulation in the age of Internet convergence*. New York: Routledge.

United Nations Human Rights Committee (UNHRCom). (2004, March 29). *General comment no. 31 [80]: The nature of the general legal obligation imposed on state parties to the covenant* (CCPR/C/21/Rev.1/Add. 13). https://undocs.org/CCPR/C/21/Rev.1/Add.13

United Nations Human Rights Council (UNHRC). (2011, March 21). *Guiding principles on business and human rights: Implementing the United Nations "protect, respect and remedy" framework*. (Report of the special representative of the secretary-general on the issue of human rights and transnational corporations and other business enterprises, John Ruggie). https://www.ohchr.org/documents/issues/business/A.HRC.17.31.pdf

United Nations Human Rights Council (UNHRC). (2012, July 16). *The promotion, protection and enjoyment of human rights on the internet (*A/HRC/20/8). https://digitallibrary.un.org/record/731540?ln=en

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. New York: Oxford University Press.

Wagner B., Kettemann, M. C., & Vieth, K. (Eds.). (2019). *Research handbook on human rights and digital technology: Global politics, law and international relations*. Cheltenham, UK: Edward Elgar Publishing.

York, J. C., & Zuckerman, E. (2019). Moderating the public sphere. In R. F. Jørgensen (Ed.), *Human Rights in the Age of Platforms* (pp. 137–162). Cambridge, Massachusetts: MIT Press.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: PublicAffairs.