Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-01917-2 OPEN

Consensus on community guidelines: an experimental study on the legitimacy of content removal in social media

Jesús C. Aguerri[1]✉, Fernando Miró-Llinares[1] & Ana B. Gómez-Bellvís[1]

The popularization of social media has led to a considerable increase in the importance of discursive expressions of violence, especially when directed at vulnerable communities. While social media platforms have created rules to regulate such expressions, little information is available on the perception of the legitimacy of these rules in the general population, regardless of the importance of the former for the latter. It is therefore the objective of this study to analyze the perception of the *seriousness* of such content and the degree to which the population has established a consensus on the withdrawal of restricted discursive behaviour on three major social media platforms (Facebook, Instagram and Twitter). For this purpose, 918 participants were immersed in an experimental paradigm in three different groups ($n_1 = 302$; $n_2 = 301$; $n_3 = 315$). Each was presented with stimuli containing discursive behaviour that is banned by community guidelines. The stimuli were presented differently to each group (i.e., description of the banned behaviour, description and accompanying example, example only). Our experimental data reveals that the degree of consensus on the need to remove content is quite high, regardless of the style of presentation. It furthermore suggests that the behaviour in question is perceived as very serious, due to the harm that our participants presume it to cause. These results have important implications for the debate on freedom of expression on the Internet and its regulation by private actors.

[1] CRÍMINA Research Centre for the Study and Prevention of Crime, University Miguel Hernández of Elche, Elche, Spain. ✉email: j.aguerri@crimina.es

Introduction

Social media play an important role in social communication, especially with regard to public and political debate. Their perhaps most remarkable innovation is that all users are granted the opportunity to share information (i.e., discourse) on a potentially global scale, turning receivers of information into spreaders of information (Balkin, 2004; Castells, 2009). However, this 'democratization' of public communication also allows actors to disseminate discourse and content that may be deemed illegal or harmful (Bilewicz et al. 2017; Tandoc et al. 2018; Wall, 2007). States have put in place administrative and criminal legislation to tackle this type of discourse, some of which were originally created for the physical world and complemented by new, specific regulations to fit the affordances of cyberspace (Heldt, 2019b). But social media companies, companies, pushed by supranational organizations (Klonick, 2020) and motivated by the need to preserve their reputation as friendly and safe places (Balkin, 2017), have added to this regulatory framework by creating their own rules, for example, *Twitter Rules* (Twitter, 2021b) and *Community Standards* (Facebook, 2021). These rules enable platforms to remove content and/or to permanently or temporarily suspend user accounts, while the discursive behaviour that leads to suspension needn't be illegal under the legislation of any particular state. This, in turn, has triggered a complex debate about the legitimacy of the private regulation of fundamental rights such as freedom of expression. (Balkin, 2004; Kaye, 2019; Keller, 2018; Suzor et al. 2018; Suzor, 2019)[1].

The debate surrounding the legitimacy of social media platform rules and community standards is important from both an academic and public perspective (Haggart and Keller, 2021), and it has intensified in recent years, particularly due to some politicians and public figures having their accounts and content limited (Floridi, 2021). Several proposals in the literature focus on ensuring legitimacy through the rule-of-law principles and human rights laws (Kaye, 2019; Suzor, 2019). However, another approach to legitimacy considers individual perceptions of the moral credibility of norms and their ability to dictate what is permissible and what is not. This approach is well-established in the criminological literature, where it is understood that norms and sanctions must be perceived as legitimate for individuals to comply voluntarily and avoid defiance towards norms and authorities (Gomez-Bellvis and Toledo, 2022; Kumm, 2004; Robinson, 2013; Tyler, 2006).

The legitimacy of user agreements, terms and conditions, or community standards doesn't rest on the same political and procedural principles that legitimize legalization in democratic states, but on the right of private companies to subject owned services to their own rules. However, social media companies do claim that *community guidelines* are not solely based on their own protective rights; instead, they are justified through the commitment to the defence of community values (Facebook, 2021). Beyond this justification, it is worth asking whether the community population supports community guidelines. In other words, is there consensus on rules limiting discursive behaviour on social media, and what elements determine such a putative consensus? Previous research suggests that citizens tend to have attitudes that are favorable towards regulations (Singhal et al. 2022). However, the perceived legitimacy of these rules has not been clearly established, nor have the elements that make users support them. Through an experimental between-group design on three independent samples of participants, the present article seeks to answer these questions about the regulation of discourse on social media, specifically with regard to Violent and Hateful Communication (Miro-Llinares, 2016) or VHC, for short.

Violent and Hateful Communication on social media

Since the inception of the Internet, crime has been transferred and adapted from the physical world to cyberspace; a process that hasn't left social media unaffected (Miró-Llinares and Johnson, 2018; Wall, 2007). Although cyberspace isn't a 'place' in the geographical sense of the word, social media platforms can be understood as places in a relational sense, meaning that they represent a 'place' where individuals and groups can converge with one another in communication (Miro-Llinares et al. 2018). The conversion of various subjects on a social media platform allows for specific types of criminal phenomena, which respond to the specific affordances of computer-mediated communication. Among the different types of crimes that can be committed on or via social media, purely discursive crime phenomena pose the greatest legislative challenge in that they require nothing more than the publication of a message to be carried out (Miró-Llinares and Gómez-Bellvís, 2020). Regulating this type of discursive behaviour is difficult, among other reasons, not all states forbid hate speech, and indeed racist expressions can be protected by freedom of speech in some legal frameworks (Barnum, 2006; Gagliardone, 2019), so it´s t is hard to considers these behaviors as "harmful" under some criminal frameworks (Feinberg, 1985). Among the different types of speech that have been deemed worthy of regulation, *hate speech* has been the most controversial, especially regarding the approaches to its regulation online (Burnap and Williams, 2016).

However, hate speech is by no means new, nor is its presence in cyberspace (Brown, 2018; Cammaerts, 2009; Erjavec and Kovacic, 2012; Quandt, 2018). First occurrences of online hate speech can be traced back to 1995 when the website *Stormfront* was created in the US (Meddaugh and Kay, 2009). However, hate speech does seem to have gained political relevance over the last few decades, and concern over its dissemination on social media has been growing (Farkas et al. 2018; Oksanen et al. 2014; Ybarra et al. 2011). This is particularly important because the rising popularity of social media has converted it into a public space which people use to participate in contemporary political culture and debate (Balkin, 2004; Klonick, 2018). This means that offensive or violent discourse disseminated on social media is public and political in nature and that the harm it generates for individual users can produce significant negative consequences for society as a whole (Bilewicz et al. 2017; Vehovar and Jontes, 2021). However, it should be noted that not all forms of VHC are considered hate speech (Miró-Llinares, 2016).

Over the last few decades, multiple definitions of *hate speech* have been provided, not all of which coincide with the criteria for its definition (Boeckmann and Turpin-Petrosino, 2002; Gagliardone, 2019; Jacobs and Potter, 1997; Matsuda, 1989; Robinson and Darley, 1995; Waldron, 2012). However, authors in the field do share one definitory element, which is also reproduced by the United Nations' definition of hate speech as

> "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor" (2020, p. 19).

The common element of the above definitions of hate speech is the discriminatory aim of the speaker. Basing violent and hateful communications on elements that are suitable to discriminate against a person on the grounds of their group identity places hate speech in the category of hate crimes and distinguishes it from other forms of VHC (Miro Llinares, 2016).

Other violent speech acts that (may) cause some kind of harm such as incitement to violence, insults, or slander can be found both on- and offline but lack the criterion of discrimination, excluding them from the category of hate speech. Under these premises, hate speech can be understood as a specific form of Violent and Hateful Communication (VHC) as a broader phenomenon (Miró-Llinares, 2016).

The regulation of VHC in social networks

Limitations on public discursive behaviour have traditionally been established by states, thereby defining the legal criteria by which public speech can be considered VHC and developing regulatory frameworks to restrict freedom of expression to a greater or lesser extent in accordance with pre-existing laws and legal principles (Heldt, 2019a; Klonick, 2018). These frameworks, therefore, vary both over time and across countries, as well as from one legal tradition to another (Brugger, 2002; Sarlet, 2019; Waldron, 2012).

But recently, examples of the limitation of freedom of expression can also be found in the context of private enterprises. The example with the most public repercussion was the recent and definitive suspension of former US President Donald Trump's various social media accounts. Following the assault on the Capitol on January 6th 2021 in Washington D.C., his accounts were suspended on the grounds of incitement to violence (Isaac and Conger, 2021; Twitter, 2021a). This decision spawned several questions regarding the establishment and enforcement of limitations to freedom of expression on social media platforms (Bensinger, 2021; Masnik, 2021; York, 2021). But it also resulted in a number of considerations on the removal of VHC from public discourse and political debate. Some of these arguments and questions have been present before the polemic around Donald Trump (Gerrard, 2018; Goldsmith and Wu, 2006; Van Loo, 2017). The central element in all of these arguments is the legitimacy of limitations to freedom of expression when they are not democratically justified but imposed on a community of users by the company that provides a service for the publication of personal expression.

Most European countries are still updating their legislation on this relatively recent issue, although some already provide operational legal mechanisms: In many countries, judges can order social media platforms to remove content that has been denounced and examined accordingly (Heldt, 2019b). Germany goes one step further and obligates social media platforms to remove content that falls into some offences of the German Criminal Code, and prescribes heavy fines for companies failing to comply Indeed, the European Union is expanding the obligations of intermediaries (Frosio, 2018). considering them responsible of tackling illegal content beyond secondary liability. The UK currently experiments with this approach but hasn't passed specific legislation yet (Wilson, 2020). On the other end of the spectrum, US courts of law have ruled that social media are covered by section 230 of the Communications Decency Acts, granting them, subject to a series of requirements, immunity from being held responsible for individual users' publications (Ardia, 2010).

Beyond this lack of universal legislation and the relative immunity offered by certain jurisdictions, the major social media companies have created their own regulatory frameworks to limit the circulation of certain types of discursive behaviour on their platforms (Citron and Norton, 2011). It cannot be overlooked that content curation is an essential part of platform service, and its economy (Gillespie, 2018). As Klonick points out, companies are incentivized to maintain a communicative climate that favours *user engagement* on the platform and therefore regulate content quite strictly. The self-imposed rules by which content is regulated are usually more restrictive than those of states and are enforced without the intervention of any legal or legislative institution (Crawford and Gillespie, 2016; Grimmelmann, 2010; Klonick, 2018; Tushnet, 2008). This leads Klonick to regard social media platforms as "the New Governors of online discourse" (Klonick, 2018, p. 1603) and thus central to the delimitation of freedom of expression, in that their relevance parallels that of state institutions.

Most popular social media sites have created content curation systems for the management of user-generated content and the enforcement of VHC protection regulations. These systems combine machine-learning algorithms, human moderation and self-reporting by users, so-called *flag systems* (Crawford and Gillespie, 2016). The use of algorithms for content management has proven useful in certain areas, such as the detection and removal of material that infringes the protection of intellectual property or the prohibition of sexually explicit graphic material (Crawford and Gillespie, 2016; Gerrard, 2018; Gillespie, 2010; West, 2018). However, in the area of discourse and VHC specifically, most moderation is done by hand (Crawford and Gillespie, 2016; Oksanen et al. 2014) because human communication exhibits a level of complexity (e.g., humour, irony, etc.) that hinders the effective use of automated content curation as the main strategy, as was also acknowledged by Facebook (Allan, 2017). However, this present limitation does not negate the possibility of an increase in the use of algorithms for the moderation of this type of content in the future. In fact, multiple authors of content curation algorithms already claim that they can perform this task (Burnap and Williams, 2016; Galan-Garcia et al. 2016; Sharma et al. 2018). Complementing the work of moderators and algorithms, platforms also have reporting systems through which users can report content in violation of regulations. While some platforms (e.g., Meta) allow users to review the status and outcome of the content review, the process itself and assessment procedures are opaque with platforms offering very little information. (Crawford and Gillespie, 2016).

Platforms apply the norms in their community guidelines to determine when messages can be "flagged", to guide moderators' decisions on content removal, and even to set the parameters of data curation algorithms, making community guidelines a crucial tool for all three types of content moderation (Carmi, 2019). These rules play a fundamental role, not only because they are the basis of the moderation and reporting system, but also because they constitute the guidelines by which companies determine the limits of freedom of expression on their social media platforms. While these rules vary from platform to platform, two of today's leading platforms, Twitter's (2021b) and Meta's (i.e., Facebook and Instagram) (2021) community guidelines share several similarities. In general, however, Meta is more restrictive than Twitter and regulates more specific areas of discursive expression than the latter (Crawford and Gillespie, 2016).

Despite differences between specific rules, the structure of community guidelines and the underlying protected values are quite similar. Facebook establishes four core values which its *community standards* seek to protect: *authenticity*, *safety*, *privacy*, and *dignity*. These *community standards* are further divided into six headings which, in turn, are divided into a series of chapters that cover more specific topics. Three of these chapters lay out rules in relation to VHC: "Violence and Incitement", "Bullying and harassment", and "Hate speech or incitement to hatred". Twitter's rules are also divided into some headings, although there are only three: *Safety*, *Privacy*, and *Authenticity*. Note that all three headings correspond to the values stated as a preamble to Facebook's community policies. While *dignity* is not explicitly mentioned in the case of Twitter, it is covered by the concept of safety, rendering the apparent difference immaterial. The regulations concerning violent communication and hate speech are contained in the section *safety*, specifically in the following chapters: “Violence”, “Terrorism/Violent Extremism”, “Abuse/Harassment”, and “Hateful Conduct”. This organization of message regulation coincides to a large extent with Facebook’s proposal.

Consensus on content seriousness and community guidelines

Social networks design their rules according to their duty to protect certain values. Thus, content removal on the grounds of non-compliance with a platform’s community guidelines is based on the assessment of a piece of content as sufficiently dangerous to seriously compromise the values protected by these guidelines. This assessment of *seriousness* should furthermore correspond to a certain level of consensus on the part of the community. That is, there should be a shared perception among users about the types of infractions deemed serious enough to result in removal or blocking. As social-psychology and criminology literature (Cialdini, 1993; Gomez-Bellvis and Toledo, 2022; Robinson, 2013) have pointed out, the alignment of rules and standards with social norms and shared moral values, as well as the perceived fairness, legitimacy, and proportionality of sanctions, can both enhance compliance with regulations.

This last point reflects two distinct but interrelated and fundamental issues for social media platforms. As stated above, consensus on the rejection of specific discursive behaviours must be established among users rather than community managers or corporate management. Such consensus would, in turn, legitimize the establishment and enforcement of regulation of discursive behaviour, under the condition that such perceptions are shared across a large enough proportion of the user base. There are therefore two major dimensions to the discussion of limitations to freedom of expression when established by private platform companies: *consensus* and *legitimacy*.

There is a growing body of literature about perception and support of community guidelines, as well as about its enforcement mechanism (Singhal et al. 2022). It has been observed that users seem to support platforms taking a proactive role in content moderation (Geeng et al. 2020; Riedl et al. 2022), However, it has also been found that users are critical of the specific application of rules (Schoenebeck et al. 2021; West, 2018), reporting that they feel their freedom of expression is limited and experience feelings of injustice, especially due to the lack of motivation behind certain decisions and the lack of transparency in execution processes (Duffy and Meisner, 2022; Jhaver et al. 2019). Recently, Rasmussen (2022) has also addressed the perceptions around the restriction of hate crimes on social media, finding remarkable agreement in favour of such restrictions. However, little is known about the factors that determine the apparent consensus around the restriction of certain expressive behaviors. In consequence, it is worth asking whether the factors that determine the seriousness attributed to a particular behavior are the same when we talk about social media as when we talk about the physical environment.

Several experimental studies investigate consensus (or a lack thereof) on the *seriousness* of criminal behaviour between the criminal system on the one hand, and its perception by citizens on the other hand (Robinson and Darley, 1995). The literature on community views and justice systems exposes study participants to different systematically varied descriptions of crime and tasks them with the assessment of the deserved punishment, which allow authors to show how people is able to distribute liability and punishment according to the seriousness of each crime (Miró-Llinares and Gómez-Bellvís, 2020; Robinson, 2013). These research suggest the existence of intuitions about justice which, in turn, lead to a generally high consensus on responsibility and punishment; these intuitions can be even cross-cultural (Robinson and Kurzban, 2007). One of Robinson’s main findings is the high sophistication of participants’ intuitions of justice. Most study participants are able to appreciate the institutions and circumstances that legal codes utilize to assess guilt, including *intentionality*, *degree of participation*, or the existence of *absolutory excuses*. Other authors, following Sellin and Wolfgang (1964), have studied directly the perceived seriousness of crime (Oconnell and Whelan, 1996; Rosenmerkel, 2001; Rossi et al. 1974). As the studies mentioned above, these experiments task participants with the assessment of seriousness for each case, either by assigning scores or by ranking different cases according to their perceived level of seriousness (Herzog and Einat, 2016).

The first relevant point to be drawn from the literature described is the importance of approaching opinions or intuitions about legal norms through concrete examples, since the abstract formulation of norms tends to generate illusory consensus. Such theoretical consensus tends to break down when participants are presented with concrete scenarios in which criminal behaviour is performed (Sellin and Wolfgang, 1964; Warr, 1989). Consequently, the first hypothesis (H_1) for our experimental paradigm is that consensus on the perceived seriousness of VHC as defined in community guidelines will be weaker when participants are presented with concrete examples of VHC rather than its abstract description.

Furthermore, the literature on the perceived seriousness of crime points to three main determining elements of seriousness: *harmfulness*, the factual assessment of the consequences of the crime for the victim; *wrongfulness*, the normative assessment of the moral reprehensibility of the crime; and the perceived *frequency* with which the conduct occurs (Warr, 1989). The study by Adriaenssen et al. (2020) also showed, in line with previous research, that participants give primacy to the element of wrongfulness over the other elements when assessing the seriousness of crime.

Concerning VHC, different communicative acts can cause physical and/or moral harm (Miro Llinares, 2016). However, this causal relationship is not necessarily direct, and one may ask whether the consequences of VHC constitute harm or the risk of harm, especially when harm is ‘only’ moral. Both the difference in the quality of harm caused by VHC and Adriaenssen et al. ‘s (2020) findings motivate two hypotheses. Firstly, we posit that behaviours resulting in direct physical harm will be rated as more serious (H_2). Secondly, we expect perceived wrongfulness to be the most determinative factor for the perception of the seriousness of VHC content (H_3). Furthermore, we have established above that agreement on content removal is linked to the assessment of seriousness, insofar as the 1atter guides the former. Consequently, our fourth hypothesis is that participants are more likely to call for content removal when they rate expressions as highly serious (H_4).

Data and method

In response to our research objectives and hypotheses, we have created an experimental paradigm, in which participants evaluate content from social media platforms in regard to their status as VHC. To immerse participants in different experimental conditions, they were placed in one of three groups, allowing for a between-group comparison of the experimental data. The experimental conditions differed from one another in the presence or absence of an example contextualizing the criminalization of discursive behaviour (Table 1).

Each of the three groups consisted of a sample of around 300 participants, all residents in Spain. The sample was recruited by non-probability snowball sampling. In order to avoid over-sampling populations that make greater use of the Internet and to

Table 1 Descriptive statistics by group and One-Way ANOVA or *z*-test between groups.

	Group 1: only standard	Group 2: standard and example	Group 3: only example	ANOVA/*z*-test	
				f-value/ x^2	*p*-value
n		301	315		
Age	32.7	33.6	34.6	3.01	0.0831
Gender				13.15	0.0014
Female	206 (68%)	191 (63 %)	171 (54%)		
Male	96 (32%)	110 (37%)	144 (46%)		
Formal Studies in law				2.6287	0.2686
No	259 (86%)	258 (86 %)	282 (90 %)		
Yes	43 (14%)	43 (14%)	33 (10%)		
Highest educational level attained (0 = "Primary studies; 5 = "University studies")	4.01	4.25	4.22	6078	0.0139
Political self-identification (1 = "far left", 7 = "far right")	3.42	3.39	3.4	0.03	0.863

minimize the bias derived from the use of non-random sampling procedures, 10 seeds (initial participants) with different population characteristics were selected according to sex and age stratum. The seeds were asked to disseminate the questionnaire among acquaintances with similar characteristics. The recruitment chains of each seed were mapped using different web links for access to the questionnaire. Table 1 shows no significant between-group differences in terms of age, political identity and legal literacy; however, there are slight differences in the level of education and gender composition of the groups. These biases will be taken into account in the discussion of the results.

All questionnaires contained a list of 13 descriptions of discursive behaviours extracted from Twitter's *Twitter rules*. Given that Meta's regulation is the most restrictive, the *Twitter rules* constitute a common substrate for three of the main social media platforms: *Twitter*, *Facebook*, and *Instagram*. Hence, the descriptions of discursive behaviour provided in the questionnaires were based on the wording used by Twitter to define these behaviours. Examples were taken from the *Twitter rules* whenever they were provided. Where no examples were given, we came up with examples to clearly and seriously illustrate the described behaviour[2] —Table 2—. Given the nature of these expressions, participants were informed prior to starting the questionnaire that they might encounter potentially offensive content. They were provided with the option to withdraw from the questionnaire if they preferred not to continue.

Following Warr's methodology (1989), and including the modifications by Adriaenssen et al. (2020), each participant had to answer six questions about each of the 13 described discursive behaviours. The questions asked to rate each behaviour according to a) the moral reproach that each behaviour deserves—wrongfulness—; b) the risk it involves and/or the harm it caused –harmfulness –; c) the frequency with which this type of behaviour occurs on social media—frequency—; d) the seriousness of the behaviour (cf. "Seriousness and Community Guidelines")—seriousness—; and e) their agreement with the statement that the type of behaviour in question should be removed from social media—consequences—. Participants were asked to rate each question on a scale of 0 to 10 with each question referring to one specific behaviour. Only in the case of perceived frequency, a Likert-type scale from 0 to 5 was used. In addition, the questionnaire contained 5 socio-demographic variables (age, gender, level of education, legal education, and political ideology).

Results

The presence of concrete examples increases the interquartile range of the mean ratings for each group (Fig. 1). This concerns both agreement with content removal (i.e., consequences) and the perceived seriousness of the behaviours. The standard deviation for both variables increases in the experimental condition that combines community standard descriptions and examples relative to the presentation of the description alone ($sd\ seriousness_{group\ 1} = 1'02$; $sd\ seriousness_{group\ 2} = 1'41$; $sd\ consequences_{group\ 1} = 1'51$; $sd\ consequences_{group\ 2} = 1'81$) and increases further in the group in which only an example is given ($sd\ seriousness_{group\ 3} = 1'46$; $sd\ consequences_{group\ 3} = 2'36$). An analysis of variance homogeneity using the Levene-test confirms that the variance changes between groups (Levane Test$_{seriousness}$: df = 2, $F = 15.883$, $p < 0.001$; *Levane Test* $_{consequences}$: df = 2, $F = 14.286$, $p < 0.001$). This analysis reveals that the presence of specific examples of VHC, both in isolation and accompanying the description of VHC, reduces the consensus on both consequences and seriousness. We thus reject the null hypothesis, namely that consensus on the perceived seriousness of VHC behaviour and its removal is independent of the type of presentation of this behaviour. We adopt the alternative hypothesis (H_1) that the presence of specific examples of VHC systematically reduces consensus on seriousness and consequences.

Furthermore, it can be observed that all sampled discursive behaviours are assigned a high level of seriousness (Fig. 2). This ceiling effect can also be observed in the rest of the variables studied (Table 3). This effect prevents us from ranking behaviours according to perceived seriousness, because differences between means of seriousness and their adjacent means are not statistically significant. However, grouping different discursive behaviours according to their reference to physical harm (A, B, C, D, F, G) or a lack thereof (E, H, I, J, K, L, M) reveals significant differences. The mean perceived seriousness of expressions of physical harm is 9.28, while the mean perceived seriousness of the other expressions is 9.08. An independent samples *t*-test reveals this difference to be statistically highly significant ($p < 0.001$). Grouping discursive behaviours according to group references, that is whether they refer to vulnerable groups (B, G, J, C, H, I, K) or whether they can be directed at any group (M, D, A, F, E, L) shows a slightly bigger difference. Here, the mean perceived seriousness of VHC against vulnerable groups equals 9.34 as opposed to 8.99 in the other cases ($p < 0.001$).

Table 3 suggests a correlational relationship between perceived seriousness and agreement on consequences. The application of a linear regression model reveals a moderate, but highly significant association between the two variables (*Pearson's* $r = 0.605$, $p < 0.001$).

Multiple linear regression with consequences as the dependent variable and seriousness in addition to socio-demographic variables as independent variables (Table 4) reveals that the latter account for 31% of the variance ($r^2 = 0.31$). Nevertheless, this

Table 2 Discursive behaviors included in the questionnaire.

	Standard description[a]	Source	Example provided
A	Threatening to inflict serious physical harm to a person or group of persons	*Violent threats T. p.*[b]	*"If you go on like this, I'm going to kill you." (Example provided by Twitter)*
B	Expressions directed at a person or group of persons because of their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious illness that threaten them with serious physical harm.	*Glorification of violence T.p.*	*"If you continue to be a whining Arab, we'll have to shoot you twice".*
C	Glorifying crimes, violent events—or their authors—where people were targeted based on their protected characteristics (including race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease)	*Glorification of violence T.p.*	Posting after a violent event directed at Jewish people: *"I'm glad this happened" (Example provided by Twitter)*
D	Promotion of terrorism or violent extremism	*Violent organizations T.p.*	*"Gora ETA. You must continue the fight".*
E	Using aggressive insults in order to threaten or intimidate others.	*Abusive behaviour T.p*	*"You're a son of a bitch who should shut the fuck up".*
F	Wishing or hoping for serious harm to a person or group of people	*Abusive behaviour T.p.*	*"I hope the next time you talk like that, you get run over by a car.*
G	Wishing or hoping for serious harm to a person or group of people basis on their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious illness	*Hateful content T.p.*	*"If this group of Muslim protesters doesn't shut up, they deserve to be shot." (Example provided by Twitter)*
H	Targeting individuals or groups with content that references forms of violence or violent events where protected categories were the primary target or victims, with the intent to harass	*Hateful content T.p*	Sending a person who professes Judaism a picture of the Auschwitz concentration camp *(Example provided by Twitter)*
I	Disseminating stereotypes that incite fear about a person or groups of people because of their race, ethnicity, nationality, sexual orientation, gender identity, religious affiliation, age, disability or serious illness.	*Hateful content T.p.*	*"All Muslims are terrorists" (Example provided by Twitter)*
J	Disseminating slurs, epithets, racist or sexist tropes, or other repeated or non-consensual content, which is used to degrade another person, when these refer to a protected group or its members.	*Hateful content T.p.*	*"Blacks are like apes" (Example provided by Facebook)*
K	Disseminating images, logos or symbols that promote hatred based on race, ethnicity, nationality, sexual orientation, gender identity, religious affiliation, age, disability or serious disease.	*Hateful content T.p.*	Posting an image of a Nazi swastika (Example provided by Twitter)
L	Disseminating content that may discourage participation or mislead people about when, where or how to participate in an electoral process.	*Civic integrity T.p.*	*"Remember that in today's general election you don't have to go to the polling station, just send an SMS with the name of the party to the following phone number".*
M	Disseminating false or altered multimedia content that could result in serious harm	*Synthetic and manipulated media policy*	*"Drinking bleach has been scientifically proven to prevent COVID-19. You should drink a 250 ml glass of bleach with every meal".*

[a] Here we present the rules as they were presented to the subjects (translated into English). We have tried to reproduce their original wording as faithfully as possible, available at: https://help.twitter.com/en/rules-and-policies#safety-and-cybercrime (version consulted on 11-11-2021, and revised on 31-03-2022 with no significant changes found).
[b] *T.p.* Twitter policy.

model reveals a significant effect of seriousness on the perception of adequate consequences for the behaviour in question. Consequently, we reject the null hypothesis of independence between perceived seriousness and perceived adequacy of consequences. We put forth the alternative hypothesis that an increase in perceived seriousness leads to an increase in the perceived adequateness of consequences (i.e., content removal).

Furthermore, we studied the impact of both demographic variables and other ratings on the perceived seriousness of discursive behaviour. We used Ordinary Least Squares (OLS) linear regression with seriousness as the dependent variable and modelled socio-demographic variables, the style of presentation (description, example, description + example), wrongfulness, harmfulness, and perceived frequency as predictor variables.

The model shows that both wrongfulness and harmfulness have a significant effect on perceived seriousness, while socio-demographic variables have very slight effects - and not statistically significant in the case of formal legal studies. Harmfulness exhibits the highest coefficient of determination. We, therefore, reject H_3 and assume that moral rejection is not the most important predictor of seriousness.

Discussion

The primary objective of this study was to explore user attitudes towards community guidelines on social media platforms. Specifically, we aimed to analyze the extent of consensus on the need for intervention and whether this consensus varies depending on how the guidelines are presented to users. We also sought to understand the factors that contribute to users perceiving content as offensive or harmful enough to warrant moderation. Through an experimental design, we tested four hypotheses and uncovered attitudes about the seriousness of content prohibited by community guidelines.

Our results give satisfactory answers to all of our hypotheses. Firstly, a comparison between standard deviations of means of

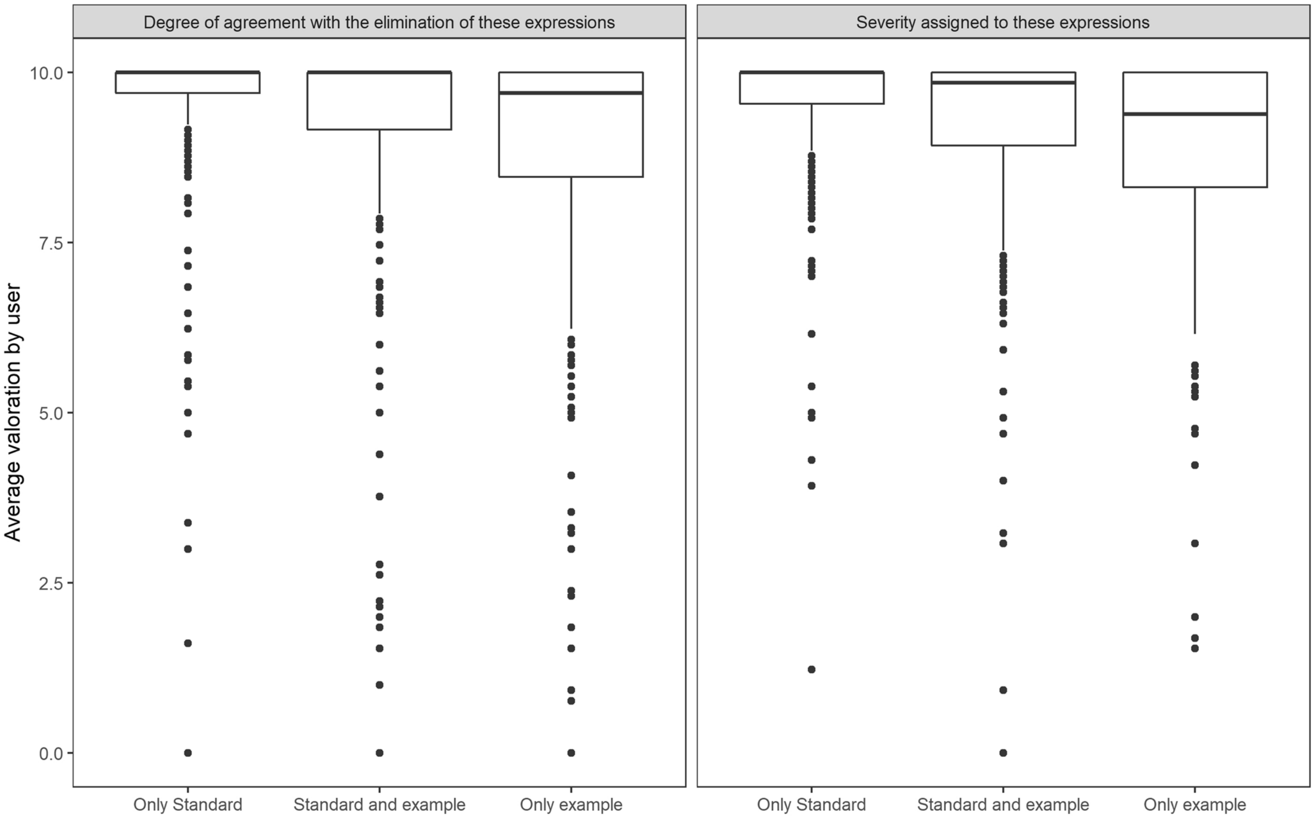

Fig. 1 Degree of consensus on seriousness and consequences.

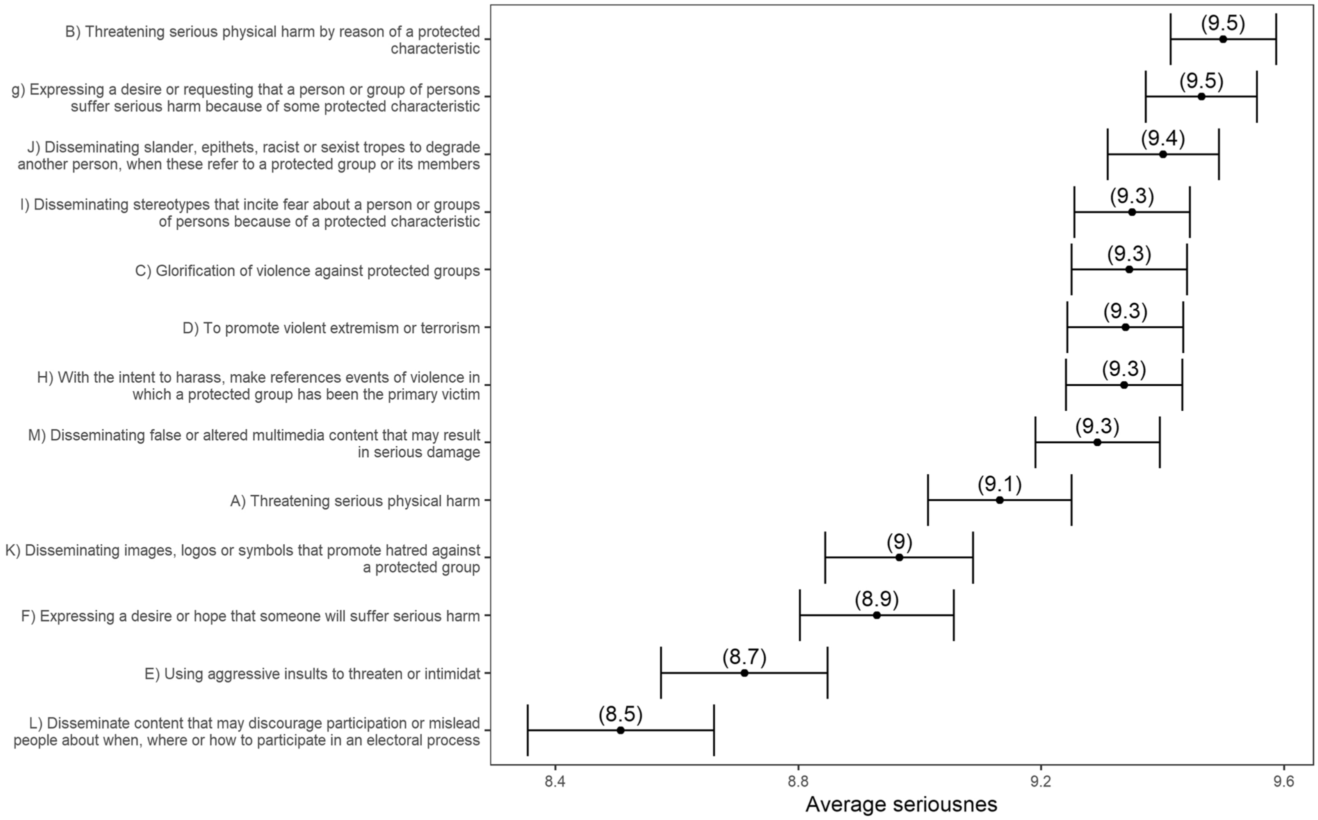

Fig. 2 Rating of expressions seriousness (95% Confidence Intervals).

Table 3 Rank of seriousness (*t*-test for adjacent ranks).

Serious.	M	sd	Conseq...	M	sd	Wrongful.	M	sd	Harmful.	M	sd	Freq.	M	sd
B	9.50*	1.34	B	9.3	2.08	B	9.76	1.05	B	9.34	1.63	B	3.58	0.99
G	9.46*	1.42	G	9.26	2.07	G	9.67	1.21	G	9.26	1.81	G	3.48	1.03
J	9.40*	1.41	J	9.28	2.03	J	9.65	1.18	J	9.23	1.81	J	3.68	1.00
I	9.35*	1.47	I	9.25	2.04	I	9.57	1.28	I	9.19	1.87	I	3.79	0.95
C	9.35*	1.47	C	9.08	2.28	C	9.64	1.26	C	9.22	1.79	C	3.60	0.97
D	9.34*	1.48	D	9.17	2.13	D	9.58	1.28	D	9.21	1.75	D	3.30	0.98
H	9.34*	1.48	H	9.19	2.10	H	9.61	1.30	H	9.21	1.82	H	3.17	1.07
M	9.29*	1.58	M	9.28	1.96	M	9.49	1.42	M	9.32	1.69	M	3.47	1.09
A	9.13*	1.82	A	8.98	2.42	A	9.41	1.59	A	9.07	1.91	A	3.64	0.93
K	8.97*	1.88	K	8.98	2.28	K	9.27	1.75	K	8.81	2.24	K	3.50	0.99
F	8.93*	1.96	F	8.81	2.54	F	9.25	1.68	F	8.83	2.24	F	3.70	0.97
E	8.71*	2.12	E	8.58	2.70	E	9.11	1.80	E	8.72	2.28	E	4.01	0.90
L	8.51	2.37	L	8.81	2.53	L	8.92	2.10	L	8.35	2.68	L	3.14	1.11

*$p < 0.05$ Significance level tested between adjacent VHC behaviors via Two-Sample *t*-Test. The asterisk signals a significant difference between the distribution sampled and the one directly below it.

perceived seriousness revealed that consensus on the latter weakens in the presence of a concrete example and, even more so, in the absence of a more general description (H_1). In other words, more concrete cases lead to less uniform answers on the seriousness of the discursive behaviour in question. This aligns with previous evidence regarding *community views* and justice systems (Robbinson and Darley, 1995; Robinson and Kurzban, 2007), and this result may contribute to explain why literature reports support for content moderation, but at the same time shows that such interventions may being perceived as unfair and generate feelings of having one's freedom of speech limited (Geeng et al. 2020; Riedl et al. 2022; Schoenebeck et al. 2021; West, 2018). It is possible that support for content moderation is based on abstract definitions of offensive or harmful content; but when applied to specific examples, it may be less clear whether those expressions, particularly those made by the subjects themselves, are truly offensive or harmful (Jhaver et al. 2019) However, consensus on the seriousness of the analyzed behaviours is quite high anyway, regardless of how the discursive behaviour was presented to participants. In addition, women are over-represented in the sample which seems to have a slight effect on the seriousness assigned to the behaviours. It is hence possible that the perceived seriousness is slightly lower in group three, not due to the style of presentation, but the disproportionate representation of the genders in this group. Despite this limitation, we claim that community rules for social media platforms generate a remarkably high consensus on the seriousness of different instances of VHC in our sample population.

Secondly, we were able to verify that discursive behaviours referencing physical harm are perceived as more serious than behaviours referencing moral harm (H_2). This difference, however, is very slight in comparison to the difference in seriousness between VHC against vulnerable groups (i.e., *hate speech*) and VCH against the general population. We, therefore, posit that the seriousness of hate crimes is assessed differently from the evaluation of traditional criminal behaviour. In particular, but contrary to what has been observed in previous literature (Rasmussen, 2022) we take the higher values of the seriousness of behaviours directed at groups that have historically been discriminated against as indicative of a notable social concern about the circulation of hate speech on social media. As far as the population of our study participants is concerned, we diagnose, as previous studies have already pointed out (Riedl et al. 2022), considerable social support for the efforts of social media companies to combat hate speech and its consequences on their platforms.

Thirdly, according to previous literature on seriousness of crime (Adriaenssen et al. 2020), we hypothesized that between harmfulness and wrongfulness, the latter would be the most determinative element for the perception of the seriousness of VHC content (H_3). However, we have observed that the strongest predictor of perceived seriousness of VHC behaviours is its perceived harmfulness. This contradicts earlier studies on criminal typologies (Adriaenssen et al. 2020), which posit moral reprehension as the most influential factor. The observation that harmfulness has greater weight than moral reproach appears to further reinforce the argument articulated above, namely that our sample population is notably concerned about the possible consequences that discursive behaviour on social media can have in the physical world. This is particularly interesting in so far as Meta and Twitter justify their policies based on the risk of hate crime dissipating from their platforms into the real world (cf. Introduction), an axiom around which scientific consensus is not unanimous.

Finally, as regards to our fourth hypothesis that posed that participants would be more likely to call for content removal

Table 4 Predictors of seriousness and consequences.

	OLS model 1 (y = seriousness)			OLS model 2 (y = consequences)		
Predictors	***coef***	**CI**	***p***	***coef***	**CI**	***p***
(Intercept)	1.50	1.32–1.68	<0.001	3.06	2.81–3.31	<0.001
Wrongfulness	0.41	0.39–0.43	<0.001			
Harmfulness	0.42	0.41–0.43	<0.001			
Perceived Incidence	0.03	0.01–0.05	<0.01			
Seriousness				0.74	0.72–0.76	<0.001
Gender (man)	−0.10	−0.14 to −0.06	<0.001	−0.60	−0.66 to −0.53	<0.001
Age	0.00	0.00–0.01	<0.001	−0.00	−0.00 to 0.00	0.092
Political self-identification (1 = “far left”, 7 = “far right”)	−0.06	−0.07 to −0.04	<0.001	−0.06	−0.08 to 0.03	<0.001
Highest educational level attained (0 = ”Primary studies; 5 = “University studies”)	0.03	0.01–0.04	<0.01	−0.09	−0.12 to −0.06	<0.001
Formal studies in law (Yes)	0.02	−0.04 to 0.07	0.605	0.04	−0.06 to 0.14	0.406
Type of questionnaire:						
Standard and example	−0.18	−0.23 to −0.14	<0.001	0.01	−0.07 to 0.09	0.790
Only example	−0.23	−0.28 to −0.18	<0.001	−0.23	−0.31 to −0.15	<0.001
Observations	11921			11921		
R^2	0.636			0.388		

when they rate expressions as highly serious (H4), our results show a palpable relationship between the perceived seriousness of an instance of VHC and desired consequences, that is the removal of this content from public discourse (H$_4$). However, the explanatory power of the corresponding model isn’t exhaustive. We, therefore, posit the existence of unknown variables beyond seriousness and the socio-demographic variables included in this study that significantly influences the perception of the response to VHC on social media. Seriousness seems to be a relevant factor, but the model constructed only accounts for 31% of the variance, with a priori important variables, such as age, not being relevant. In general terms, our research participants strongly support the removal of VHC from social media platforms, but future research will need to look more closely at the variables that determine this support.

Our findings inform the general debate on freedom of expression on social media, as they confirm the social concern about certain types of content; not so much because of its frequency, but mainly because of the harm and risks it is perceived to generate. These observations are also relevant to the more specific debate on the legitimacy of auto-regulation of social media because they confirm that the establishment and enforcement of community guidelines on social media are supported by (at least a good proportion of) the user base. This statement doesn’t settle the debate, of course. On the one hand, social support isn’t the only legitimizing factor of limitations to social behaviour. Another important aspect of legitimacy is its enforcement and the procedures established to this end. The restriction of fundamental rights (such as freedom of expression) in democratic societies is held to the standard of legal and procedural safeguards that cannot be overwritten by merge social consensus, meaning that the fundamental rights of citizens are protected from subversion even against the citizens themselves. Indeed, beyond the consensus and support of content moderation rules, previous literature has also shown that users also worry about the limitation of freedom of speech that could encompass these rules enforcement (Singhal et al. 2022), which could be strongly related not with the rules itself but with the proceedings of enforcement, future research lines should explore this later dimension of rules legitimacy.

Data availability

Data are available at Dataverse: https://doi.org/10.7910/DVN/8Y2I1V.

Received: 18 July 2022; Accepted: 5 July 2023;
Published online: 15 July 2023

Notes

[1] Adriaenssen (2020) uses different types of scales, but here it has been decided to follow the original design of Warr (1989), with scales from 0 to 10.

[2] Before conducting the survey, a pilot survey was carried out to test the questionnaires. The survey was disseminated via the researchers’ Twitter accounts, reaching a total of 404 participants. The pilot survey results allowed us to verify that the descriptions of the norms and examples were comprehensible, and that a consistent relationship existed between the selected examples and the referenced norms.

References

Adriaenssen A, Paoli L, Karstedt S, Visschers J, Greenfield VA, Pleysier S (2020) Public perceptions of the seriousness of crime: Weighing the harm and the wrong. Eur J Criminol 17(2):127–150. https://doi.org/10.1177/1477370818772768

Allan R (2017) Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community? About Facebook. https://about.fb.com/news/2017/06/hard-questions-hate-speech/

Ardia DS (2010) Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity Under Section 230 of the Communications Decency Act. Loyola Los Angeles Law Rev 43:373–506

Balkin JM (2004) Digital speech and democratic culture: A theory of freedom of expression for the information society. N Y Univ Law Rev 79(1):1–58

Balkin JM (2017) Free Speech in the Algorithmic Society: Big Data, Private Governance and New School Speech Regulation. Davis Law Rev 51:1149–1210

Barnum D (2006) The Clear and Present Danger Test in Anglo-American and European Law. San Diego Law Rev 7(2):263–292

Bensinger G (2021) Now Social Media Grows a Conscience? The New York Times. https://www.nytimes.com/2021/01/13/opinion/capitol-attack-twitter-facebook.html

Bilewicz M, Soral W, Marchlewska M, Winiewski M (2017) When Authoritarians Confront Prejudice. Differential Effects of SDO and RWA on Support for Hate-Speech Prohibition. Politic Psychol 38(1):87–99. https://doi.org/10.1111/pops.12313

Boeckmann RJ, Turpin-Petrosino C (2002) Understanding the harm of hate crime. J Soc Issues 58(2):207–225. https://doi.org/10.1111/1540-4560.00257

Brown A (2018) What is so special about online (as compared to offline) hate speech? Ethnicities 18(3):297–326. https://doi.org/10.1177/1468796817709846

Brugger W (2002) The Treatment of Hate Speech in German Constitutional Law (Part I). Ger Law J 3(12). https://doi.org/10.1017/S207183220001569

Burnap P, Williams ML (2016) Us and them: identifying cyber hate on Twitter across multiple protected characteristics. Epj Data Sci 5:11. https://doi.org/10.1140/epjds/s13688-016-0072-6

Cammaerts B (2009) Radical pluralism and free speech in online public spaces The case of North Belgian extreme right discourses. Int J Cultural Stud 12(6):555–575. https://doi.org/10.1177/1367877909342479

Carmi E (2019) The Hidden Listeners: Regulating the Line from Telephone Operators to Content Moderators. Int J Commun 13:440–458

Castells M (2009) Communication Power. Oxford University Press

Cialdini R (1993) Influence: The Psychology of Persuasion. Harpercollins

Citron DK, Norton H (2011) Intermediaries And Hate Speech: Fostering Digital Citizenship For Our Information Age. Boston Univ Law Rev 91(4):1435–1484

Crawford K, Gillespie T (2016) What is a flag for? Social media reporting tools and the vocabulary of complaint. N Media Soc 18(3):410–428. https://doi.org/10.1177/1461444814543163

Duffy BE, Meisner C (2022) Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility [Article; Early Access]. Media Culture Soc. https://doi.org/10.1177/01634437221111923

Erjavec K, Kovacic MP (2012) "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. Mass Commun Soc 15(6):899–920. https://doi.org/10.1080/15205436.2011.619679

Facebook (2021) Facebook Community Standards. https://transparency.fb.com/policies/community-standards/

Farkas J, Schou J, Neumayer C (2018) Cloaked Facebook pages: Exploring fake Islamist propaganda in social media. N Media Soc 20(5):1850–1867. https://doi.org/10.1177/1461444817707759

Feinberg J (1985) Offense to Others (The Moral Limits of Criminal Law, Vol 2). Oxford University press

Floridi L (2021) Trump, Parler, and Regulating the Infosphere as Our Commons [Editorial]. Philos Technol 34(1):1–5. https://doi.org/10.1007/s13347-021-00446-7

Frosio GF (2018) Why keep a dog and bark yourself? From intermediary liability to responsibility. Int J Law Inform Technol 26(1):1–33. https://doi.org/10.1093/ijlit/eax021

Gagliardone I (2019) Defining Online Hate and Its "Public Lives": What is the Place for "Extreme Speech"? [Article]. Int J Commun 13:3068–3086

Galan-Garcia P, de la Puerta JG, Gomez CL, Santos I, Bringas PG (2016) Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. Logic J Igpl 24(1):42–53. https://doi.org/10.1093/jigpal/jzv048

Geeng C, Francisco T, West J, Roesner F (2020) Social Media COVID-19 Misinformation Interventions Viewed Positively, But Have Limited Impact. arXiv. https://doi.org/10.48550/ARXIV.2012.11055

Gerrard Y (2018) Beyond the hashtag: Circumventing content moderation on social media. New Media Soc 20(12):4492–4511. https://doi.org/10.1177/1461444818776611

Gillespie T (2010) The politics of 'platforms' [Article]. N Media Soc 12(3):347–364. https://doi.org/10.1177/1461444809342738

Gillespie T (2018) Custodians of the internet : platforms, content moderation, and the hidden decisions that shape social media. Yale University Press

Goldsmith J, Wu T (2006) Who Controls the Internet Illusions of a Borderless World. Oxford University Press

Gomez-Bellvis AB, Toledo FJC (2022) Political speech offences in social media considering the effects of criminal sanction: Deterrent or defiance effect? [Article]. Rev Chil De Derecho Y Tecnologia 11(1):323–+. https://doi.org/10.5354/0719-2584.2022.66547

Grimmelmann J (2010) The Internet Is A Semicommons. Fordham Law Rev 78(6):2799–2842

Haggart B, Keller CI (2021) Democratic legitimacy in global platform governance [Article]. Telecommun Policy 45(6):102152. https://doi.org/10.1016/j.telpol.2021.102152

Heldt A (2019a) Reading between the lines and the numbers: an analysis of the first NetzDG reports. Internet Policy Rev 8(2). https://doi.org/10.14763/2019.2.1398

Heldt A (2019b) Sharing New Responsibilities in a Digital Age. J Inform Policy 9:336–369. https://doi.org/10.5325/jinfopoli.9.2019.0336

Herzog S, Einat T (2016) Moral Judgment, Crime Seriousness, and the Relations Between Them: An Exploratory Study. Crim Delinq 62(4):470–500. https://doi.org/10.1177/0011128712466889

Isaac M, Conger K (2021) Facebook Bars Trump Through End of His Term. The New York Times. https://www.nytimes.com/2021/01/07/technology/facebook-trump-ban.html

Jacobs JB, Potter KA (1997) Hate crimes: A critical perspective. In M. Tonry (ed), Crime and Justice: a Review of Justice (Vol. 22, pp. 1-50). https://doi.org/10.1086/449259

Jhaver S, Appling DS, Gilbert E, Bruckman A (2019) "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. 3, 1–33. https://doi.org/10.1145/3359294

Kaye D (2019) Speech Police. The global struggle to govern the Internet. Columbia Global Reports

Keller D (2018) Internet Platforms: Observations on Speech, Danger, and Money. Hoover Institution's Aegis Paper Series 1807. Available at SSRN: https://ssrn.com/abstract=3262936

Klonick K (2018) The New Governors: The People, Rules, And Processes Governing Online Speech. Harv Law Rev 131(6):598–670

Klonick K (2020) The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression [Article]. Yale Law J 129(8):2418–2499

Kumm M (2004) The legitimacy of international law: A constitutionalist framework of analysis. Eur J Int Law 15(5):907–931. https://doi.org/10.1093/ejil/chh503

Masnik M (2021) Not Easy, Not Unreasonable, Not Censorship: The Decision To Ban Trump From Twitter. TechDirt

Matsuda MJ (1989) Public response to racist speech: Considering the victim's story. Mich Law Rev 87(8):2320–2381

Meddaugh PM, Kay J (2009) Hate Speech or "Reasonable Racism?" The Other in Stormfront. J Mass Media Eth 24(4):251–268. https://doi.org/10.1080/08900520903320936

Miro Llinares F (2016) Taxonomy of violent communication and the discourse of hate on the internet. Idp-Internet Law Politic 22:82–107

Miro-Llinares F, Moneva A, Esteve M (2018) Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. Crim Sci 7(1):15. https://doi.org/10.1186/s40163-018-0089-1

Miró-Llinares F, Gómez-Bellvís AB (2020) Freedom of expression in social media and criminalization of hate speech in Spain: Evolution, impact and empirical analysis of normative compilance and self-censorship. Span J Legislative Stud 1:1–42. https://doi.org/10.21134/sjls.v0i1.1837

Miró-Llinares F, Johnson, S D (2018) Cybercrime and place: Applying environmental criminology to crimes in cyberspace. In G. Bruinsma & S. Johnson (eds), The Oxford handbooks in criminology and criminal justice (pp. 883-906). Oxford University Press

Oconnell M, Whelan A (1996) Taking wrongs seriously - Public perceptions of crime seriousness. Br J Criminol 36(2):299–318

Oksanen A, Hawdon J, Holkeri E, Nasi M, Rasanen P (2014) Exposure To Online Hate Among Young Social Media Users. Soul Soc 18:253–273. https://doi.org/10.1108/s1537-466120140000018021

Quandt T (2018) Dark Participation [Article]. Media Commun 6(4):36–48. https://doi.org/10.17645/mac.v6i4.1519

Rasmussen J (2022) When Do the Public Support Hate Speech Restrictions? Symmetries and Asymmetries across Partisans in Denmark and the United States. PsyArXiv. https://doi.org/10.31234/osf.io/j4nuc

Riedl MJ, Whipple KN, Wallace R (2022) Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others [Article]. Inform Commun Soc 25(11):1632–1649. https://doi.org/10.1080/1369118x.2021.1874040

Robinson PH (2013) Intuitions of Justice and the Utility of Desert. Oxford University Press

Robinson PH, Darley JM (1995) Justice, Liability, and Blame : Community Views and the Criminal Law. Boulder

Robinson PH, Kurzban R (2007) Concordance and conflict in intuitions of justice. Minn Law Rev 91(6):1829–1907

Rosenmerkel S (2001) Wrongfulness and harmfulness as components of seriousness of white-collar offenses. J Contemp Crim Justice 17(4):308–327

Rossi P, Waite E, Bose C, Berk R (1974) The seriousness of crimes: Normative structure and individual differences. Am Sociol Rev 39(2):224–237

Sarlet IW (2019) Freedom of expression and the problem of regulating hate speech in social networks. Rev Estud Institucionais-J Institutional Stud 5(3):1207–1233. https://doi.org/10.21783/rei.v5i3.428

Schoenebeck S, Scott CF, Hurley EG, Chang T, Selkie E (2021) Youth trust in social media companies and expectations of justice. Proc ACM Hum Comput Interact 5:1–18. https://doi.org/10.1145/3449076

Sellin T, Wolfgang ME (1964) The Measurement of Delinquency. Wiley

Sharma S, Agrawal S, Shrivastava M (2018) Degree based Classification of Harmful Speech using Twitter Data. *arXiv.* https://doi.org/10.48550/arXiv.1806.04197

Singhal M, Ling C, Paudel P, Thota P, Kumarswamy N, Stringhini G, Nilizadeh S (2022) SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. arXiv. https://doi.org/10.48550/ARXIV.2206.14855

Suzor N, Van Geelen T, West SM (2018) Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. Int Commun Gaz 80(4):385–400. https://doi.org/10.1177/1748048518757142

Suzor NP (2019) Lawless the secret rules that govern our digital lives [Book]. Cambridge University Pres. https://doi.org/10.1017/9781108666428

Tandoc Jr EC, Lim ZW, Ling R (2018) DEFINING "FAKE NEWS" A typology of scholarly definitions. Digital Journal 6(2):137–153. https://doi.org/10.1080/21670811.2017.1360143

Tushnet R (2008) Power without responsibility: Intermediaries and the First Amendment. George Wash Law Rev 76(4):986–1016

Twitter (2021a) Permanent suspension of @realDonaldTrump. Twitter Blog. https://blog.twitter.com/en_us/topics/company/2020/suspension.html

Twitter (2021b) The Twitter Rules. https://help.twitter.com/en/rules-and-policies/twitter-rules

Tyler TR (2006) Psychological perspectives on legitimacy and legitimation. In Annu Rev Psychol (Vol. 57, pp. 375-400). https://doi.org/10.1146/annurev.psych.57.102904.190038

United Nations. (2020). United Nations Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presence. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf

Van Loo R (2017) Rise Of The Digital Regulator. Duke Law J 66(6):1267–1329

Vehovar V, Jontes D (2021) Hateful and Other Negative Communication in Online Commenting Environments: Content, Structure and Targets. Acta Inform Prag 10(3):257–274. https://doi.org/10.18267/j.aip.165

Waldron J (2012) The harm in hate speech. Harvard University Press

Wall D (2007) Cybercrime: The Transformation of crime in the Information Age. Polity Press, In Cambridge

Warr M (1989) What is the perceived seriousness of crimes? Criminology 27(4):795–822

West SM (2018) Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. N Media Soc 20(11):4366–4383. https://doi.org/10.1177/1461444818773059

Wilson TJ (2020) Collaborative Justice and Harm Reduction in Cyberspace: Policing Indecent Child Images. J Crim Law 84(5):474–496. https://doi.org/10.1177/0022018320952560

Ybarra ML, Mitchell KJ, Korchmaros JD (2011) National Trends in Exposure to and Experiences of Violence on the Internet Among Children. Pediatrics 128(6):E1376–E1386. https://doi.org/10.1542/peds.2011-0118

York, J. C. (2021). Users, not tech executives, should decide what constitutes free speech online. MIT technology Review. https://www.technologyreview.com/2021/01/09/1015977/who-decides-free-speech-online/?utm_medium=tr_social&utm_campaign=site_visitor.unpaid.engagement&utm_source=Twitter#Echobox=1610385745

Acknowledgements

This publication is part of the project GamerVictim (CIPROM/2022/33) funded by the PROMETEO 2023 program of the “Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital” of the “Generalitat Valenciana”; and JCA contribution was possible thanks to Grant FJC2020-042961-I funded by MCIN/AEI/ 10.13039/501100011033 and by the “European Union NextGenerationEU/PRTR”.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: JCA, ABG-B and FM-L; analysis and interpretation of results: JCA; draft manuscript preparation: JCA, ABG-B and FM-L. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

All procedures performed in this study were in accordance with the ethical standards of the University Miguel Hernández of Elche. Ethical clearance and approval were granted by the Ethics and Integrity Committee of the University Miguel Hernandez of Elche.

Informed consent

Authors confirms that informed consent was obtained from all participants. Participants provided consent by choosing to start the survey after reading the study information.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-023-01917-2.

Correspondence and requests for materials should be addressed to Jesús C. Aguerri.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.