P&I
POLICY & INTERNET

How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation

Sabine A. Einwiller and Sora Kim

This article reports two studies conducted in the United States, Germany, South Korea, and China to examine how online content providers (OCPs) exercise their responsibility in dealing with harmful online communication (HOC) by moderating user-generated content. The first study employed content analysis of 547 HOC policy documents. In the second study, 41 representatives of OCPs were interviewed regarding the implementation of these policies. We show that HOC policies are most often communicated through user-unfriendly terms of service. Only Korean OCPs present their policies very vividly. Few organizations, mainly United States and German, encourage counter-speech. The most common organizational actions against HOC mentioned in the policies are deleting posts or blocking accounts. The interviews reveal, however, that organizations—apart from those from China—are cautious in implementing such reactive actions. They fear accusations of censorship and acknowledge the tension between free speech and their content moderation practice. What emerged as the "gold standard" for identifying HOC was manual inspection. However, organizations operating large platforms widely apply machine-learning technology or artificial intelligence. In sum, our research suggests that OCPs are not proactive enough in their communication for HOC prevention and often focus more on avoiding legal ramifications than on educating users when handling HOC.

KEY WORDS: content filtering, content moderation, harmful online communication, online abuse, online content providers

本文报道了在美国、德国、韩国和中国完成的两项研究，以检验网络内容提供商*(OCPs)*如何通过审查由用户产生的内容， 发挥职责应对不良网络传播*(HOC)*。第一项研究对*547*个*HOC*政策文件进行内容分析。第二项研究中， 就落实这些政策对*41*名*OCPs*代表进行了访谈。我们表明，*HOC*政策最常通过用户友好型服务条款进行传播。只有韩国的*OCPs*清晰地展示了他们的政策。主要源于美国和德国的小部分组织鼓励反驳言论。这些政策所提到的打击*HOC*的最常见的组织行动是删除帖子或封锁账号。 然而访谈结果显示，除去来自中国的组织，其他组织都很谨慎地执行这类反应性行动。它们担心受到审查指控，并承认自由言论与其内容审查实践之间的紧张关系。用于识别*HOC*的"黄金标准"是人工审查。然而，运营大型平台的组织广泛采用机器学习技术或人工智能。总之， 我们的研究暗示， *OCPs*在传播*HOC*预防一事上不够积极主动，并且在处理*HOC*时经常更多地聚焦于避免法律后果，而不是对用户进行教育。

关键词： 不良网络传播，内容审核，网络内容提供商，内容过滤，网络滥用

doi: 10.1002/poi3.239

Este artículo informa dos estudios realizados en los EE. UU., Alemania, Corea del Sur y China para examinar cómo los proveedores de contenido en línea (OCP) ejercen su responsabilidad en el tratamiento de la comunicación en línea dañina (HOC) al moderar el contenido generado por el usuario. El primer estudio empleó el análisis de contenido de 547 documentos de política HOC. En el segundo estudio, se entrevistó a 41 representantes de OCP con respecto a la implementación de estas políticas. Mostramos que las políticas de HOC se comunican con mayor frecuencia a través de términos de servicio amigables para el usuario. Solo los OCP coreanos presentan sus políticas de manera muy vívida. Pocas organizaciones, principalmente estadounidenses y alemanas, fomentan el contra discurso. Las acciones organizativas más comunes contra HOC mencionadas en las políticas son eliminar publicaciones o bloquear cuentas. Sin embargo, las entrevistas revelan que las organizaciones, aparte de las de China, son cautelosas al implementar tales acciones reactivas. Temen las acusaciones de censura y reconocen la tensión entre la libertad de expresión y su práctica de moderación de contenido. Lo que surgió como el "estándar de oro" para identificar HOC fue la inspección manual. Sin embargo, las organizaciones que operan plataformas grandes aplican ampliamente la tecnología de aprendizaje automático o la inteligencia artificial. En resumen, nuestra investigación sugiere que los OCP no son lo suficientemente proactivos en su comunicación para la prevención de HOC y a menudo se centran más en evitar ramificaciones legales que en educar a los usuarios cuando manejan HOC.

PALABRAS CLAVE: comunicación perjudicial en línea, moderación de contenido, proveedores de contenido en línea, filtrado de contenido, abuso en línea

Introduction

In its early days, the Internet was envisioned as a means of providing a context for the development of community and collective values. It was to serve as an electronic forum where a plurality of voices freely engaged in rational debates, thereby fostering democratization (e.g., Rheingold, 1995). Yet this vision is continually marred by highly emotional and quite often aggressive and hateful voices being disseminated online. Nearly a fifth of U.S. Americans (18 percent) have personally experienced online harassment and 62 percent consider it a major problem (Duggan, 2017). In Germany, 8 percent have been personally targeted and 40 percent have observed hate speech online (Geschke, Klaßen, Quent, & Richter, 2019). And according to the Korean National Human Rights Commission (NHRC) 84 percent of women in South Korea have come across misogynistic hate speech online (The Asian, 2017). The possible harmful consequences for those assaulted range from diminished subjective well-being (Kaakinen, Keipi, Räsänen, & Oksanen, 2018) to emotional stress, anxiety, decreased self-esteem, and depressive symptomatology (Geschke et al., 2019; Tynes, Giang, Williams, & Thompson, 2008). Thus, such harmful online communication (HOC) is a pressing social issue that endangers the health of individuals, may stimulate social unrest (Alkiviadou, 2019) and even lead to an increase in hate crimes offline (Müller & Schwarz, 2018).

In this study, the term HOC is used as an umbrella term for different forms of online communication that violate social norms and target the dignity and/or safety of others, who can be individuals, social groups or organizations. HOC encompasses the manner of expression (aggressive, hateful, or destructive) as well as its potential effect (harmful) on the targets but also on observers of the content.

It comprises different kinds of antisocial online communication practices, including online hate speech, cyberhate, and online harassment.

Various approaches to address and reduce HOC in general, and online hate speech in particular, are being debated (e.g., Cohen-Almagor, 2011; Gagliardone, Gal, Alves, & Martinez, 2015; George, 2015; Suzor et al., 2019). At the center of the debate is often governments' enforcement of national legislation. However, governments and legal institutions are in fact only one set of actors in a larger mosaic of institutions and private entities governing the Internet (DeNardis, 2012) and sharing responsibility for HOC (Helberger, Pierson, & Poell, 2018). A great deal of power lies with private entities, above all with the organizations who make money or support their business from people using their online platforms for communication. As the owners of the platforms, they are the actors who not only have decisive power of intervention but also the responsibility to provide their users with a safe environment (Taddeo & Floridi, 2016). In an open letter, the editorial staff at *WIRED Magazine* explicitly urged private companies and organizations to counteract hostility online:

> So. Companies that created the tools that let us communicate: no more passes. You have the ability to help people feel safe in their daily online lives. You have sophisticated tools to fight spam, and you take down content that infringes on copyright in the blink of an eye. This is a call to action. And a plea. You can't say "we suck at dealing with abuse," promise to do something, and then drag your feet. Because it's starting to look like you care more about your next earnings call than the people who actually use your site. (WIRED Staff, 2016)

This research focuses on the role and activities of different types of private organizations in curbing HOC, to wit all types of organizations that provide platforms for or allow comments and discussions on their websites. Thus, we include Internet intermediaries that "give access to, host, transmit and index content, products and services originated by third parties on the Internet or provide Internet-based services to third parties" (OECD, 2010, p. 9), but also such organizations that produce their own content and use Internet intermediaries for the purpose of interacting with their customers and other stakeholders. We refer to these types of organizations as online content providers (OCPs). Due to their role in Internet communication and contemporary social life in general, these private OCPs have ethical, social, and human rights responsibilities to their users and others who are affected by what is being discussed on their sites (Council of Europe, 2017; MacKinnon, Hickock, Bar, & Lim, 2014; Taddeo & Floridi, 2016).

The Internet is of course a global environment. Nevertheless, the boundaries for freedom of speech are defined in a gray zone between nationally defined legal frameworks, culturally informed expectations by the organizations' stakeholders, as well as their own norms (Jørgensen, 2017). Among 38 nations studied, according to a Pew Research Center survey conducted in 2015, Americans were the biggest supporters of freedom of speech. Citizens of other countries like Germany

(ranked 18th) and South Korea (ranked 21st) were more open to restricting free speech in certain circumstances (Wike, 2015). Because of these differences, it is worth examining HOC prevention policies and practices by OCPs in different legal and cultural environments. To do so, OCPs from two Western (the United States and Germany) and two Eastern countries (China and South Korea) are included in this research.

With this research, we contribute to the discussion on how OCPs exercise their responsibility to prevent HOC by moderating user-generated content on their platforms. The notion of content moderation refers to the processes whereby OCPs decide on the boundaries for appropriate speech on their sites (Crawford & Gillespie, 2016; Roberts, 2014). As part of the empirical analysis, we first analyze OCPs' content policies published on their sites; second, we explore how these policies are implemented in daily content moderation by means of qualitative interviews with representatives of the organizations. Thus, this study differs from and extends previous empirical research that analyzed the nature and quality of incivility and hate speech on social network sites (e.g., Awan, 2014; Oz, Zheng, & Masullo Chen, 2018; Su et al., 2018) and from research that focuses on methods for automatically detecting hate speech online (e.g., Davidson, Warmsley, Macy, & Weber, 2017; Saleem, Dillon, Benesch, & Ruths, 2017) or forecasting its likely spread (Burnap & Williams, 2015).

Our research builds on and extends theoretical discussions and frameworks on the possibilities for countering online hate speech from a civil and legal (e.g., Banks, 2010; Citron & Norton, 2011; Delgado & Stefancic, 2014) and applied ethics perspective (Cohen-Almogor, 2011) as well as contemplations on the responsibilities of Internet intermediaries regarding specific aspects of HOC like gender-based violence (Suzor et al., 2019). Methodologically, our research resembles that of Jørgensen (2017), who examined how representatives from Facebook and Google make sense of human rights such as freedom of expression and privacy. However, we take a broader approach in terms of sample and scope regarding the actual implementation of the organizations' norms in their practice of content moderation. Next, we outline the theoretical background and specific research questions before presenting the empirical studies and implications thereof.

Literature Review

HOC

Possibly due to the characteristics of the Internet such as anonymity (Cho & Kwon, 2015; Lampe, Zube, Lee, Park, & Johnston, 2014), the lack of social or personal context cues (Moor, Heuvelman, & Verleur, 2010), and the absence of compelling legal or ethical responsibilities of Internet intermediaries and OCPs (Suzor et al., 2019), the Internet has been used as a vehicle through which hostile, aggressive, offensive, abusive (and therefore harmful) communication can be widely spread. Scholars and practitioners alike have pointed it out as a pressing social and even global issue to be tackled, and called for national and global

regulations, legislation, and social responsibilities of OCPs (Alkiviadou, 2019; Cohen-Almagor, 2011; Papacharissi, 2004; Suzor et al., 2019).

In regulations and in the public debate, the term "online hate speech" is frequently used (European Commission, 2008; Parekh, 2012). In research, the terms online hate (e.g., Keipi, Näsi, Oksanen, & Räsänen, 2017) and cyberhate (e.g., Perry & Olsson, 2009) are also applied. Hate speech is defined as the speech that "expresses, encourages, stirs up, or incites *hatred* against a group of individuals distinguished by a particular feature or set of features such as race, ethnicity, gender, religion, nationality, and sexual orientation" (Parekh, 2012, p. 40). Similarly, the European Commission (2008) defines illegal hate speech in its Framework Decision 2008/913/JHA by means of criminal law and national laws as the public incitement to violence or hatred directed at groups or individuals on the basis of certain characteristics.

Since forms of online speech can be hostile and abusive without being hate speech (Parekh, 2012), this research uses a more encompassing concept, that of HOC. HOC encompasses any form of online communication that violates social norms and targets the dignity and/or safety of others. Forms of HOC include verbal attacks (e.g., discrimination, calls for physical harm and violence) and extreme cursing (e.g., obscene swearwords), as well as visual violence and obscenity (e.g., images or videos of extreme violence, pornography). Aside from the expression of hatred or degrading attitudes toward a social group or collective to which the concept of online hate is limited (Keipi et al., 2017), HOC also includes attacks on individuals (online harassment) and on organizations. Yet, the perception of what is harmful and therefore constitutes HOC for individuals, groups or organizations can differ depending on individual factors, cultural norms and regulations (Downs & Cowan, 2012; Hawdon, Oksanen, & Räsänen, 2017).

Freedom of Speech and its Restrictions in the Online Environment

Inextricably linked to any discussion on how to curb HOC is the issue of freedom of speech. In most democratic countries, freedom of speech is institutionalized, making it a constitutionally protected human right, as it is the basis of free thought and human life (Parekh, 2012). Yet freedom of speech is not the only important value to a so-called moral community where its members have a consensus on safeguarding essential human interests (Parekh, 2012). Like most other rights, it has restraints, limits, and obligations not to undermine human dignity and safety (Banks, 2010; Parekh, 2012). In most countries, there are legislative restrictions and limits to the freedom of speech. These limits vary depending on the historical and cultural contexts of a society (Parekh, 2012). For instance, freedom of speech is most strongly protected in the United States and least in China (Jiang, 2016; Mendel, 2012).

When it comes to restrictions in Internet environments, countries also differ regarding their legislative limitations. The U.S. takes a cyber-liberalist approach, providing OCPs the most robust protections for free speech and wide exemptions from liability for third-party illegal content (Yu, 2018). In the United States, based

on Section 230 of the Communication Decency Act (CDA), OCPs are not liable when harmful content is posted on their sites (Ehrlich, 2012). At the other end of the spectrum sits China; it assumes a cyber-paternalism in controlling online content (Yu, 2018). In China, free speech is largely limited, although the Chinese constitution claims that citizens enjoy their rights to free speech (Constitution of the People's Republic of China, n.d.). Article 5 of the "Computer Information Network and Internet Security, Protection and Management Regulations" clearly prohibits inciting hatred or discriminating among nationals or harming the unity of the nation (Constitution of the People's Republic of China, n.d.). These restrictions provide the Chinese government plenty of room for limiting free speech for the purpose of ideological, political, and national security.

Germany takes a middle ground between restricting illegal online content and free speech (Yu, 2018). Germany's online hate speech rules, known as the Network Enforcement Act (NetzDG), came into force on January 1, 2018 (Scott & Delcker, 2018). This new law states that large social networks (e.g., Facebook and Twitter) may be fined up to €50 million if they persistently fail to remove, within 24 hours, illegal online content that has been reported to them (Scott & Delcker, 2018). Similarly, in South Korea, there is a specific Internet law called "Act on Promotion of Information and Communication Network Utilization and Information Protection" (the Network Act hereafter), which was enacted in 2001 (Statutes of the Republic of Korea, n.d.). The Network Act imposes liability to Internet intermediaries for illegal online content shared on their services (Park, 2015). According to Article 44-2 of the Network Act, OCPs shall take at least a temporary measure on content requested to be taken down, by blocking access to it up to 30 days (Statutes of the Republic of Korea, n.d.). Despite OCPs' obligations to remove content only when someone's rights have been violated, they generally remove content—legal or illegal—after getting a request from users, simply to avoid potential liability (Gasser & Schulz, 2015). As a result, even legal content can be taken down by Internet intermediaries upon request.

Moral Responsibilities to Minimize HOC

Aside from their legal responsibilities, Internet organizations have an obligation to protect users from abuse by individuals or collectives (Suzor et al., 2019). Vedder (2001) argues that organizations should be held morally responsible, not for the user-generated content itself, but for the dissemination of offensive and harmful material. This is because OCPs are not just hosts or facilitators of communication but also actors, in that they provide the technocommercial infrastructures designed to enhance user engagement and spreading of content (Gerlitz & Helmond, 2013).

However, as Helberger et al. (2018) state, the governing of online platforms is a situation in which multiple entities contribute to a problem, or to the solution of a problem. Drawing on the concept of "multiple hands" (Thompson, 1980), they ascribe responsibility not only to the organizations that provide the platforms for discussion but also to users participating in it. In this regard, Helberger and colleagues argue that users who like and share particular harmful content,

or neglect to flag it, are also partly responsible for its circulation. They see a "cooperative responsibility" of organizations and users when it comes to solving problems.

To be able to fulfill their responsibility, however, users need to have sufficient capacity, knowledge, and freedom to exercise necessary actions. This is where OCPs come back into play, as they have to "create the *conditions* that allow individual users to comply with their responsibilities" (Helberger et al., 2018, p. 3, emphasis in original). It is the way the platform architectures are designed that shapes how users can review, flag, or counter HOC and thereby fulfill their responsibilities. Design elements to empower users include items such as examples of prohibited content, flagging mechanisms, and encouragements to engage in counter-speech. This also includes the policy documents accessible on OCPs' sites—like terms of service, content guidelines, or community guidelines—that outline the OCPs' expectations of and relationship to users (Myers West, 2018). It is in these documents where the rights and responsibilities between the parties are allocated, and where users are educated about the dos and don'ts on the platform and about the consequences they face when violating the rules. The accessibility of these documents and the way they are written (e.g., are they in technical language or easy to understand and instructive) matter for users to be able to exercise their part of the shared responsibility. Ksiazek (2015) showed that specific policies like those regarding moderation of comments are effective facilitators of civil discussion. Thus, to analyze how organizations help users fulfill their responsibility, we examine the quality of OCPs' policy documents. This leads to the first research question:

RQ1. How do OCPs educate users in their policy documents about (a) their responsibilities (dos and don'ts) with respect to HOC and (b) the consequences they may face when found in violation of the rules?

The responsibilities of OCPs regarding the circulation of HOC imply that they themselves take action against HOC. Prospective measures include the aforementioned education of users to communicate in a civil and non-harmful manner online, and to help counter HOC by flagging or by means of counter-speech (Blaya, 2019; Schieb & Preuss, 2016). Retrospective responsibility, on the contrary, requires monitoring and moderating the user-generated content. Organizations' content moderation systems are designed to set boundaries for undesirable forms of expression (Myers West, 2018). Roberts (2014) defines commercial content moderation as "the organized practice of screening user-generate content (UGC)" posted online (p. 12).

Despite public and academic interest in the practice of content moderation, it is a challenging object of study because of the many layers involved and the lack of transparency regarding organizations' moderation practices (Leetaru, 2018; Myers West, 2018). In a first step, content moderation requires a definition of HOC by the organization and, based on this, the identification of potentially harmful content. Especially for large Internet intermediaries, this is a challenging task because of the vast amount of content posted each day by users. Filters and machine-learning

tools are used to enhance the efficiency of this process, but these organizations also rely considerably on users flagging the content they consider objectionable (Crawford & Gillespie, 2016). In a second step, procedures need to be put in place to handle violations. These include hiding or deleting HOC and warning or suspending the user who violated the rules. These procedures also include commenting on infringements online to educate violators and other users and/or to demonstrate the endeavor, to maintain a civil atmosphere on the site. Although these sanctions are often codified in the policy documents (see RQ1), it remains an open question how OCPs really handle what they perceive as HOC. On one hand, they generally want to, and are expected to, keep a civil atmosphere on their sites; on the other hand, they need to be careful not to overreach their policing efforts as this may cause users to accuse them of censorship. The following research question is thus asked:

RQ2. How do OCPs (a) identify HOC and (b) handle what they perceive as violations against the content/community guidelines?

Research indicates that the occurrence and the quality of online hostilities are affected by wider societal conditions, such as political changes or threatening events like terrorist attacks (Kaakinen, Oksanen, & Räsänen, 2018). For example, race and ethnicity was found to be a major topic of HOC in Europe, triggered by the so-called refugee crisis that peaked in 2015 (Ross et al., 2016). According to an online survey, race and ethnicity are also dominant in hateful online comments in the United States (Costello, Hawdon, Ratliff, & Grantham, 2016). More empirical insights into the changes of HOC, and the wider societal-level phenomena triggering these changes, are needed in order to better understand the conditions under which HOC is more likely to appear (Kaakinen et al., 2018). Although some societal phenomena are global, differences between countries are likely due to specific national and cultural developments. This leads us to the third and final research question:

RQ3. Which developments regarding the quantity and quality of HOC do representatives of OCPs perceive, and how do countries differ in this respect?

Empirical Research

To find answers to the research questions, we conducted two studies in the United States (USA), Germany (DEU), South Korea (KOR), and China (CHN). Study 1 employed a quantitative content analysis method to analyze how OCPs educate users through their policy documents with respect to HOC (RQ1). Study 2 relied on an in-depth interview method to investigate organizations' content moderation practices (RQ2) and their perception of changes and trends regarding HOC (RQ3). To generate broad insights into organizational efforts, the sample was drawn from eight different types of OCPs: (1) web portal sites (e.g., Naver.com, Baidu. com), (2) online news media sites (e.g., nytimes.com), (3) SNSs (e.g., Facebook, Twitter), (4) blog hosting sites (e.g., blogspot.com), (5) community sites

(e.g., Tianya.cn), (6) e-commerce sites (e.g., Amazon.com), (7) recommendation sites (e.g., Yelp.com), and (8) large non-Internet companies with online platforms (e.g., Siemens, Samsung).

Study 1: Content Analysis of Policy Documents

Sample and Procedure

On the basis of the web traffic data of the selected countries, provided by Alexa Internet, Inc., the top three OCPs per country were selected for the five categories of web portal, blog hosting, community, e-commerce, and recommendation portal sites. Eight OCPs were selected for online news media sites, five for SNS and 10 for large non-Internet companies' online platforms. This yielded 38 OCPs for each country and 152 OCPs in total.

Documents containing HOC-related policies based on the study's HOC definition were downloaded from the organizations' websites and saved for further analysis. HOC-related policy contents were considered as an individual policy document when they appeared under a separate URL or when they were under the same URL but either appeared in a separate hyperlink or under a clearly separate main title within a document. This resulted in a total of 547 HOC policy documents. The unit of analysis was each individual policy document.

The goal was to answer RQ1 asking how OCPs educate users in their policy documents about their responsibilities (dos and don'ts) with respect to HOC and the consequences for violating the rules. To enable users to receive and understand these responsibilities and consequences, policies need to be communicated in an easily accessible and comprehensible manner and form. To measure the core variables regarding responsibilities and consequences, and relevant background information on manner and form, a codebook was developed drawing on previous literature (e.g., Casey, 1999; Jiang, 2016; Myers West, 2018). Where the material suggested additional categories, the codebook was extended based on the data. We followed the procedure for quantitative media content analysis with human coders as suggested by Neuendorf (2002): conceptualization and operationalization of the variables based on theory and rationale, coding scheme development, sampling, training, coding, intercoder reliability tests, and reducing and inferring from the data through statistical computation.

To capture the manner and form of communication, we first assessed the *type of policy document* containing HOC drawing on previous literature (e.g., Myers West, 2018) and the material. HOC policy documents were classified as (1) terms of use/service, (2) community guideline or community standards, (3) content guideline, and (4) reporting guideline depending on the larger type of document or the context in which they appeared. The *emphasis* given to HOC was gauged by word counts of HOC-focused content in each larger policy document; the relative proportion of HOC content to the overall length of the document assessed the degree of emphasis put on HOC. Furthermore, the accessibility and readability of HOC policies were measured. *Accessibility* of the documents was captured on

two levels: hard to find (labeling isn't obvious or placement unexpected) versus easy to find (labeling is obvious and placement where one would expect it). Based on web-design communication literature (e.g., Casey, 1999), *readability* was also assessed on two levels: readable (medium-sized font without illustrations or helpful color scheme) versus very readable (very well designed with illustrations or color scheme). We further assessed whether the policies contained *case examples* to educate the users about what is considered prohibited content or behavior, as illustrative exemplars are crucial in effective communication (Zillmann & Brosius, 2012). Whether a *reference to laws* (e.g., by means of a hyperlink to a government or legal site) was provided was also recorded, given the varying legal regulations in the four countries (e.g., Jiang, 2016). Finally, the two core variables *possibilities for user actions* and *OCPs' ways to handle HOC* were considered based on previous literature on OCPs' content moderation systems (e.g., Myers West, 2018). The specific possibilities for users to take action against HOC were derived from the material, and coders were instructed to check on the organizations' sites whether there was, for example, a possibility to flag a post or to notify the organization via email or telephone. Specific actions regarding *how OCPs handle violations* against the policies were also derived from the material (delete without explanation, warning of violators, etc.).

To secure systematic categorization and coding of the data, the codebook contained detailed instructions, definitions, and examples of the variables for coder trainings, and coding results were tested for intercoder reliability (Neuendorf, 2002). At least two coders who were native to local languages independently first coded approximately 20 percent of the HOC policy documents collected for each country to check intercoder reliability. As Krippendorff's α tests were satisfactory for all measured variables, ranging from 0.78 to 1.00 for all countries, the remainder of the sample was coded (Hayes & Krippendorff, 2007).

Results of Study 1

Of a total of 547 policy documents, Korean OCPs share twice as many policies with their users on their platforms compared with organizations from the other countries (KOR = 235, DEU = 109, CHN = 108, and USA = 95). Chinese OCPs tend to communicate HOC policies more through their terms of service (44 percent), German OCPs more through community guidelines (40 percent) and those from South Korea use reporting guidelines (34 percent) more than in any other countries, while OCPs from the United States share the policies equally through terms of service (34 percent) and community guidelines (33 percent). Overall, HOC policy documents were shared most often through the terms of service ($n = 192$, 35.1 percent), followed by community guidelines ($n = 154$, 28.2 percent), reporting guidelines ($n = 135$, 25 percent), and content guidelines ($n = 66$, 12 percent).

Most of the documents (87 percent) were easy to find. Yet, documents were easiest to find on the sites of South Korean OCPs (94 percent), and least easy on the sites of Chinese organizations (70 percent) with German (82 percent) and U.S. sites in between (83 percent). It shows that the policy documents by Korean organizations were also the

most readable (very good readability: KOR = 68 percent, DEU = 33 percent, USA = 26 percent, and CHN = 24 percent). Korean OCPs especially enhance the readability of their policies through illustrations and different colors that make the content easier to digest; 32 percent of Korean OCPs also use case examples to educate their users. Organizations from the other countries rarely provide case examples in their policy documents (DEU = 18 percent, CHN = 16 percent, and USA = 5 percent). However, the majority provide a reference to laws relating to HOC, including a hyperlink to a government or legal site; Asian organizations do this more (CHN = 95 percent and KOR = 92 percent) than those from the United States (76 percent) or Germany (66 percent). As to the average proportion of specifically discussing HOC-relevant content in the policy documents, OCPs discuss it with a relatively low proportion (22 percent on average), although Korean OCPs do so with a higher proportion (36 percent) than OCPs from other countries (DEU = 22 percent, USA = 20 percent, and CHN = 11 percent).

The 152 OCPs provide various opportunities for users to take action against HOC (see Table 1). Of the identified user actions, marking or flagging a post is most highly adopted by all organizations (77 percent), followed by offering a standard online automated template to directly submit to the organization (67 percent), which is mostly used by Korean OCPs. Korean organizations also frequently give users the opportunity to notify them via all other ways (email, telephone, or even postal mail), which organizations from China do less frequently, and those from the United States and Germany almost never. The same holds for notifying an authority or government agency directly, which is only offered in China and South Korea. Only a few organizations, mainly from the United States and Germany, call on users to counter HOC with counter-speech.

The 152 OCPs mention a variety of actions in their documents which they would take in the case of HOC (see Table 2). The most common actions are deleting an HOC post without explanation or comment (88 percent), while doing so with a comment is a less frequently stated option (19 percent). Another commonly stated practice is deleting or closing the account of someone who violated the rules

Table 1. Possibilities for User Actions Provided by OCPs

User Actions	Total $N_{total} = 152$ (%)	USA $N = 38$ (%)	DEU $N = 38$ (%)	KOR $N = 38$ (%)	CHN $N = 38$ (%)
Mark/flag the post	117 (77)	27 (71)	32 (84)	32 (84)	26 (68)
Notify provider: standard template	102 (67)	15 (40)	22 (58)	35 (92)	30 (79)
Notify provider—email	71 (47)	13 (34)	11 (29)	35 (92)	12 (32)
Notify provider—telephone	60 (40)	1 (3)	0 (0)	34 (90)	25 (66)
Notify provider—postal mail	36 (24)	0 (0)	1 (3)	32 (84)	3 (8)
Notify authority or government agency	44 (29)	0 (0)	1 (3)	21 (55)	22 (58)
Call for counter-speech	11 (7)	5 (13)	5 (13)	0 (0)	1 (3)
Other action	11 (7)	5 (13)	5 (13)	1 (3)	0 (0)

Note: absolute numbers of OCPs (percentage by country).
CHN, China; DEU, Germany; KOR, South Korea; OCP, online content providers; USA, United States.

Table 2. OCPs' Ways to Handle HOC

Handlings	Total $N_{total} = 152$ (%)	USA $N = 38$ (%)	DEU $N = 38$ (%)	KOR $N = 38$ (%)	CHN $N = 38$ (%)
Delete without explanation or comment	133 (88)	35 (92)	35 (92)	36 (95)	27 (71)
Delete or close violator's account	104 (68)	29 (76)	24 (63)	20 (53)	31 (82)
Legal/judicial persecution	67 (44)	16 (42)	15 (40)	13 (34)	23 (61)
Warning of violators	47 (31)	8 (21)	14 (37)	11 (29)	14 (37)
Delete or curtail with explanation or comment	29 (19)	6 (16)	10 (26)	11 (29)	2 (5)
Delete or close whole discussion	34 (22)	2 (5)	3 (8)	0 (0)	29 (76)
Committee	19 (13)	0 (0)	0 (0)	9 (24)	10 (26)
User (member) management system	16 (11)	0 (0)	1 (3)	2 (5)	13 (34)
Other	39 (26)	11 (29)	17 (45)	7 (18)	4 (11)

Note: absolute numbers (percentage by country)
CHN, China; DEU, Germany; KOR, South Korea; OCP, online content providers; USA, United States.

(68 percent). Overall, less than half threaten legal/judicial action, an option mostly stated in China. In China, closing a whole discussion is a frequently stated practice, a practice rarely mentioned by organizations from any of the other countries. The warning of violators is an option communicated by about one-third of OCPs. None of the organizations from the Western countries mention employing committees that would decide how to handle a violation; of the Eastern countries, approximately 20 percent of the organizations mention doing so.

Study 2: In-Depth Interviews

Sample and Procedure

The 152 OCPs analyzed in Study 1 were approached and asked for an interview with a representative in charge of or overseeing content moderation, social media or community management. This resulted in 22 interviews with representatives of OCPs (response rate: 14.5 percent; USA = 1, DEU = 5, KOR = 10, and CHN = 6). Since the number of interviews was not sufficient, more OCPs were researched. From the 120 additional organizations contacted, 19 agreed to participate in the study (response rate: 16 percent). Thus, 41 interviews with OCP representatives were realized (USA = 10, DEU = 11, KOR = 10, and CHN = 10). In addition to the interviews, we researched relevant publicly available information on the large U.S. SNSs (e.g., Facebook), which are under intense scrutiny for their procedures regarding HOC, and with whom we were not able to conduct interviews. This information, which comprised reports by NGOs (e.g., Online Civil Courage Initiative and SaferInternet), news media (e.g., The New York Times and Daily Mail), online sources (e.g., marketingland.com and netzpolitik.org) and comments/reports by the SNSs (e.g., Facebook and Twitter) on their own media, supplemented the analysis of interview data.

Of the 41 interviewees, four were from web portal sites, six from online news media sites, three from SNSs, 10 from online communities, 13 from large non-Internet companies, three from large NGOs, and one each from e-commerce and recommendation portal sites. The interviews were conducted, either in-person or through Skype/telephone, by interviewers native or fluent in the local language. Interviewees were first informed about the study's objectives and then asked to give their informed consent to participate and for us to record the interview. Interviews lasted between 30 and 60 minutes. The interview guide mainly focused on the content moderation practices, that is, identification and handling of HOC (RQ2) and the developments regarding the perceived quantity and quality of HOC (RQ3).

We used thematic analysis to analyze the qualitative interview data, a common method used for "identifying, analyzing, and reporting patterns (themes) within the data" (Braun & Clarke, 2006, p. 79). Following a multi-step process, we first prepared the data for analysis by transcribing, then identified patterns through a rigorous process of data familiarization, data coding, theme development, and revision. The supplemental information on the large United States. SNS was analyzed along the lines of the categories identified in the interviews to complement the analysis with information on these important players. The analysis was conducted by two researchers, who continuously exchanged their findings in order to identify commonalities and differences between the different countries and types of organizations.

Results of Study 2

Interviews revealed many similarities across the countries, except for China. Differences with respect to identifying and managing HOC emerged mainly between the types of organizations, depending on organizations' size of user base, which correlates with the number of user comments. We discuss below the findings focusing on the three core categories—(1) identification of HOC, (2) handling and responding to HOC, and (3) changes and trends with respect to HOC.

Identification of HOC. No universal theme emerged regarding how varying OCPs define HOC. Yet, with regard to techniques to identify HOC, (1) manual inspection emerged as a universal theme across OCPs, while (2) artificial intelligence machine learning and filtering through 24/7 inspection, and (3) simple machine filtering during business hours emerged as specific themes depending on characteristics of OCPs, as explained in more detail below.

OCPs generally do not work with a precise definition of HOC, but with a guideline describing what is to be removed. This mainly includes content that attacks, degrades, or threatens a person or group. Interview partners often mentioned that rantings against the OCP itself are generally not removed, unless they contain severe profanity and swearing, which is generally considered HOC. In addition to that, Chinese OCPs regard any politically sensitive post as HOC but also any information about their direct competitors. This shows that, in all

countries, content that can be removed can fall clearly beyond the definition of hate speech.

Approaches and techniques to identify HOC differ depending on the size of the organization's user base and also between types of organizations. In large organizations with a high volume of comments, inspection happens 24/7. Medium sized platforms usually have a window of six to eight hours during the night where the platform/site is not watched, unless they employ volunteer moderators who do not stick to business hours. OCPs with a small(er) content volume and little HOC usually inspect only during business hours.

Large Internet intermediaries apply machine-learning tools as well as simpler machine filtering. Chinese organizations use machine filtering extensively. Most of the large Korean Internet companies have developed their own machine-learning and filtering systems to identify HOC. Machine learning and filtering is also used by some large news media in order to manage their user comments sections (e.g., Perspective API used by The New York Times; Wakabayashi, 2017). Some online community sites also use filtering technology, mainly rather simple word filters. Facebook has a profanity filter installed, which moderators of organizations that use the SNS can modify and set to different levels.

Because machine learning and filtering is far from being perfect, the thereby selected content generally undergoes manual inspection in a second step. Large Korean Internet organizations have 24-hour inspection centers, some of which are located additionally in China and Vietnam, and Chinese organizations often outsource to service providers within China. Large U.S. SNSs also employ such inspection centers, for example, in the Philippines. In response to the new law in Germany (NetzDG), they also increased the number of inspection centers for German language content (Hanrahan, 2018).

Despite the use of technology, manual identification of HOC is still of great importance across all sizes of OCPs. Large organizations also rely heavily on users to flag potential HOC, which is then forwarded for inspection. Smaller OCPs, mainly online communities that manage their own platforms, employ volunteer moderators from the community to monitor the content. In non-Internet companies, where the amount of HOC is comparatively small, the community or social media manager(s) monitor the content themselves.

Handling and Responding to HOC. In organizations' approaches to handle and respond to HOC, some universal themes emerged. They include the following: (1) warning and decisively communicating with users, (2) hiding or deleting HOC posts (or words), (3) blocking/locking user accounts, and (4) reporting HOC/illegal content to third parties (e.g., police). These methods are widely adopted by OCPs, yet the degree of employing them varies by country and type of organization, as outlined below.

Interviewees consider warning and clearly communicating with users about what is considered inappropriate/harmful highly important. They stated that decisively pointing out publicly where and why comments violated the policy and referring to the respective policy could help educate the poster and those

observing. When the volume of HOC is large, however, doing so is often impossible. It is also a challenge to do this when a user is clearly trolling or posts are severely harming others so that they have to be removed immediately.

Filters such as blacklists entail hiding or deleting options. Deleting is a more common option for organizations that operate their own sites, as it is generally not possible for OCPs that use platforms by intermediaries (e.g., Facebook) to delete comments, but only to hide them. News media organizations or community sites sometimes pre-moderate comments, that is, inspect the content before posting it. Some make constructive comments more visible by elevating them to increase the user experience and to educate the community through positive examples. Some organizations, instead of deleting an entire post, replace forbidden or inappropriate words (e.g., cursing) with symbols (e.g., symbols of music notes). Many interviewees mentioned they have to be careful with deleting posts, and do so only when severe violations against their policies are evident, in order to prevent accusations of censorship. Yet interviewees from OCPs operating their own platforms (e.g., online communities) mentioned they could be rather strict in removing HOC, emphasizing their rights as the owner of the space. Chinese organizations in general rely heavily on preventive blocking of so-called illegal content to avoid the risk of having illegal content on their sites.

As noted above, reporting/flagging by users is important for HOC identification. Organizations using Facebook report severe HOC to the SNS, asking to delete it, because they cannot delete but only hide comments. Facebook also offers "social reporting," where a user who feels offended by a post can send a message directly to the poster, asking him or her to delete or change the comment. In Korea, as laid down in the Network Act, users who report a comment can invoke a 30-day ban of the post. Some interviewees mentioned reporting illegal content to the police, or other government authorities responsible for legal infringements. In China, users can also report content they considered illegal to government authorities by using reporting buttons on the sites.

Changes and Trends with Respect to HOC. In terms of HOC-related changes and trends, several universal themes emerged across OCPs and countries: (1) an increase in HOC frequency, (2) polarization of opinion, and (3) politics, gender, and LGBT as common HOC topics. A majority of our interview partners said they observed an increase in the frequency of HOC and a polarization of opinions in online media in general. Most interviewees stated that the threshold for posting offensive comments had lowered and, over time, the tone had become rougher. Some traced the increased polarization back to certain events. In the United States, the event was the 2016 election; in Germany it was mainly the increase in refugees; and in South Korea it was politics in general. Chinese organizations remarked on changes with respect to people's increasing creativity to avoid being censored by adopting homophones, special symbols and pictures.

Regarding the topics that trigger HOC, in all countries interviewees mentioned politics, as well as gender and LGBT. Race and religion were mentioned specifically by U.S. practitioners, while Germans also mentioned refugees. Koreans noted

conflicts between generations. Chinese participants noted that social issues related to the health care system and social inequality played a role in the increase in HOC.

Discussion

The findings of our content-analysis study reveal that, when it comes to educating their users, OCPs from South Korea are the best. Their policies are easiest to find, and the documents vividly present the rules with illustrations and examples. Importantly, the majority of organizations from all four countries allow users to flag posts they identify as inappropriate. Additionally, Chinese users are encouraged to report violations directly to government authorities, who maintain a tight grip on the Internet. The danger of unjust denunciations is obvious. Yet, reporting to the Internet organizations has also evoked criticism. In South Korea, it is not uncommon that users report posts simply because they express opposing viewpoints, and thereby trigger a 30-day ban on the posts (Freedom House, 2018). In fact, Article 44-2 of the Network Act has been widely criticized as unconstitutional, undermining rights to free speech in South Korea (Park, 2015). In Germany, a number of controversial deletions and suspensions in the wake of the new NetzDG have bolstered critics who say the law will impact free speech, as companies try to avoid fines (Oltermann, 2018). Thus, organizations need to act sensibly and quickly on reported posts. It is also sensible for Internet organizations to call for user action directed at the perceived violator in the form of social reporting, as done by Facebook, and encouraging counter-speech (Blaya, 2019; Schieb & Preuss, 2016). Asian OCPs, however, do not encourage counter-speech at all in their policies, and only a minority of organizations from the United States and Germany do so.

In our interview study, the representatives of the organizations emphasized the importance of clear communication with users about what is considered HOC and how a platform manages such harmful content. Interviewees furthermore mentioned the necessity to regularly evaluate their policies, whether they are still up-to-date, as new topics or tools may come up that need to be addressed. However, our findings from the content analyses reveal a somewhat different story, indicating that OCPs are not active enough in communicating or educating dos and don'ts for HOC with their users on their platforms. The fact that HOC policies are most often discussed via the terms of service is alarming, given that users rarely pay attention to the platforms' terms of service and that the readability of these legal documents is considerably low. Although German OCPs are an exception, as their most frequent mode of communication is via the use of community guidelines, one of the most common modes of communication for HOC is clearly terms of service across all of the countries. In a similar vein, our findings suggest that only a small proportion of the policy documents are dedicated to specifically discussing HOC-relevant content. Also, organizations, with the exception of those in South Korea, take a passive stance toward providing case examples or improving readability of HOC policy content. These findings imply OCPs place a relatively low emphasis on HOC-related policy communication. Moreover, the most common organizational actions against HOC mentioned in the policy documents are reactive measures,

such as deleting a post without explanation and deleting or closing the account of a violator, rather than proactive measures.

These findings show that, concerning HOC issues, OCPs are not being proactive or preventive. Scholars in the field of crisis and issue communication have long advocated the importance of taking preventive measures or of being proactive before a small issue spills over into a full-fledged social crisis (Heath & Palenchar, 2009; Penrose, 2000). The best preventive or proactive measures OCPs can take against HOC certainly include increasing user awareness of HOC-related policies through informing and educating users about the dos and don'ts of HOC. Although the organizations may not have moral responsibility for user-generated HOC itself, they are certainly responsible for the dissemination of and protecting their users from HOC content (Vedder, 2001). OCPs can indeed improve civility in online user interactions by employing a more effective mode of HOC policy communication such as community guidelines rather than a legal mode of communication (i.e., terms of service).

In addition, the Internet has brought new issues of stakeholders, adding such issues as online privacy or hostility into the inventory of corporate social responsibility agendas (Pollach, 2011). As such, organizations may also engage in or initiate corporate social responsibility projects or programs against HOC, like the Online Civil Courage Initiative (OCCI),[1] which, in partnership with Facebook, combats hate speech and extremism across Europe. Although our interviewees from large OCPs observed the necessity of corporate social responsibility initiatives to better educate users regarding HOC, organizations were not actively implementing such proactive measures. More of these initiatives related to HOC as part of OCPs' proactive measures should be considered so as to secure more civilized platforms.

As for the most common actions against HOC identified through the content analysis study—such actions as deleting posts and closing violators' accounts—our interview study revealed managerial sensitivity regarding actual implementations of such actions. Especially large platform providers remarked on having to be cautious about deleting comments or blocking users because they are afraid of alienating their users and of criticism for censorship (except for Chinese OCPs that heavily use preventive blocking of so-called illegal content). Most of the interview partners in the United States, Germany, and South Korea perceive tensions between their content moderation and freedom of speech; they also mentioned that they regularly discuss this matter internally. Yet OCPs that operate their own platforms (e.g., online communities, news media sites) emphasized their rights as the owner of the space, and that users had to play by house rules. Interestingly, this was particularly stressed by organizations from the United States, where freedom of speech is held to a very high standard.

Our interview partners stressed the necessity to keep their sites civil. As argued above, OCPs are not just hosts or facilitators of communication, but actors, which makes them morally responsible for the content they circulate (Vedder, 2001). The interviews revealed that a perception of moral responsibility indeed partly motivates organization's content moderation efforts. Yet civility was considered

just as or even more important for business reasons, so users could enjoy their experience and remain loyal members, fans, or customers. The strongest motivator, especially in the United States but also in Germany and South Korea, seemed to be the business and image case. This is why organizations in the United States, where legal requirements are the most lenient, are no less concerned about civility on their sites than their counterparts in Germany or South Korea. Nevertheless, legal requirements are the driver for the strictest forms of content moderation, which are practiced in China. However, the Network Act and respective fear of legal persecution also makes organizations in South Korea stick to the 30-day-ban rule often more strictly than necessary. In Germany, where the new NetzDG mainly targets the large Internet intermediaries, organizations also reacted and enhanced their inspecting facilities to avoid legal ramifications.

Helberger et al. (2018) see the relevant governmental institutions at the local, national, and transnational levels as the third "hand" that shares responsibility for civility on the Internet, in addition to the private organizations and the users. As this is valid, our study also shows that government regulations in the form of laws can lead to exaggerated reactions by OCPs and more emphasis on avoiding legal responsibility or liability of HOC content on their platforms, rather than proactively taking a part in preventing HOC. However, research by Hawdon et al. (2017) suggests that anti-hate speech laws may protect Internet users from being exposed to online hate. Thus, while "the law must be our last resort" (Parekh, 2012, p. 46), its intervention cannot be ruled out.

Aside from laws, implementing a code of conduct seems to be a sensible path that was taken by the European Commission. In 2016, it implemented The Code of Conduct Countering Illegal Hate Speech Online (European Commission, 2016), which was signed by several IT companies, the first ones being Facebook, Microsoft, YouTube and Twitter. Similar to the German NetzDG, the companies have to review removal notifications in less than 24 hours. The code furthermore demands, among others, that companies educate users about types of unpermitted content, inform them how to submit notices for removal and encourage the provision of notices and flagging. First experiences with the code are promising (European Commission, 2019), and the code is "a template that could be used by other countries and regions" (Alkiviadou, 2019, p. 34).

Limitations

The results of the study need to be interpreted in light of certain limitations. The analyses are based on a limited sample of 38 OCPs per country. The same applies for the interviews, which were limited to 10 or 11 OCPs per country. Because we aimed to get a broad picture of different types of organizations, the sample was drawn from a variety of OCPs. It is therefore not viable to derive conclusions for a specific type of OCP. Further in-depth research focusing, for example, on news media organizations or online communities can provide more detailed insights into content moderation practices of a specific type of OCP. The four countries differed considerably in parts regarding their contextual factors.

This also invites further in-depth analyses of moderation practices in light of the specific cultural and legal influences.

Because our investigation focused on the perspective of the organization, the impact of online policy communication and content moderation practices cannot be definitely assessed. To do so, surveys among users from the different countries investigating their perception of HOC and acceptance of organizations' moderation practices need to be conducted. Having stressed the role of users as moderators and their responsibility to curb HOC, in cooperation with OCPs (Helberger et al., 2018), user surveys seem particularly interesting as a next step.

Conclusion

Our findings delineate a rather comprehensive landscape on organizations' efforts to combat HOC, suggesting that (1) OCPs are not forceful and proactive enough in preventing HOC through communication, because they widely use "terms of service" as the mode of HOC communication, lack in providing case examples of HOC, and have a small proportion of HOC-related policy content in their overall policy documents; (2) deleting HOC without explanation is most often listed in OCPs' HOC policies as a consequence of HOC violations; (3) flagging a post is most highly adopted by OCPs for user actions, but it is also abused by individuals or OCPs to avoid legal ramifications; (4) manual inspection is universally adopted for HOC identification across OCPs, while big OCPs utilize artificial intelligence machine learning and filtering with 24/7 inspection; finally (5) OCPs observe an increase in HOC frequency and polarization of online opinion.

With political polarization on the rise in many countries, the issue of HOC is likely only intensifying. Technological advancements especially in the area of artificial intelligence will drive the development of tools to identify HOC more precisely than today. Yet, it is important to apply technology sensibly. That is, to foster civility and not to exert even more control over users, who need to be granted their right to free speech within the limits of not endangering others' rights to safety and dignity. All in all, as part of shared responsibility for HOC social issues (Helberger et al., 2018), organizations should take a more proactive stance in preventing HOC and protecting their users rather than resorting to government regulations or user civility. Importantly, this includes more effective HOC-related policy communication with users, and proactive initiatives to educate users on does and don'ts with regard to HOC.

Sabine A. Einwiller, University of Vienna, Faculty of Social Sciences, Department of Communication, Althanstrasse 14, UZA2, Wien 1090, Austria [sabine.einwiller @univie.ac.at].

Sora Kim, The Chinese University of Hong Kong - School of Journalism and Communication, Shatin, N.T., Hong Kong [sorakim@cuhk.edu.hk, sorakim91@ gmail.com].

Note

The research was funded by a grant from The Toyota Foundation.

1. https://counterspeech.fb.com/en/initiatives/online-civil-courage-initiative-occi/

References

Alkiviadou, N. 2019. "Hate Speech on Social Media Networks: Towards a Regulatory Framework." *Information & Communications Technology Law* 28 (1): 19–35.

Awan, I. 2014. "Islamophobia and Twitter: A Typology of Online Hate Against Muslims on Social Media." *Policy & Internet* 6 (2): 133–50.

Banks, J. 2010. "Regulating Hate Speech Online." *International Review of Law, Computers & Technology* 24 (3): 233–239.

Blaya, C. 2019. "Cyberhate: A Review and Content Analysis of Intervention Strategies." *Aggression and Violent Behavior* 45: 163–72.

Braun, V., and V. Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3 (2): 77–101.

Burnap, P., and M.L. Williams. 2015. "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making." *Policy & Internet* 7 (2): 223–42.

Casey, C. 1999. "Accessibility in the Virtual Library: Creating Equal Opportunity Web Sites." *Information Technology and Libraries* 18 (1): 22–5.

Cho, D., and K.H. Kwon. 2015. "The Impacts of Identity Verification and Disclosure of Social Cues on Flaming in Online User Comments." *Computers in Human Behavior* 15: 363–72.

Citron, D.K., and H.L. Norton. 2011. "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age." *Boston University Law Review* 91: 1435–84.

Cohen-Almagor, R. 2011. "Fighting Hate and Bigotry on the Internet." *Policy & Internet* 3: 3.

Constitution of the People's Republic of China. n.d. *The National People's Congress of the People's Republic of China.* http://www.npc.gov.cn/englishnpc/Constitution/node_2825.htm.

Costello, M., J. Hawdon, T. Ratliff, and T. Grantham. 2016. "Who Views Online Extremism? Individual Attributes Leading to Exposure." *Computers in Human Behavior* 63 (Suppl C): 311–20.

Council of Europe. 2017. "Recommendation CM/Rec(2017x)xx of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries." *Third revised draft* (January 12). https://rm.coe.int/recommendation-cm-rec-2017x-xx-of-the-committee-of-ministersto-member/1680731980.

Crawford, K., and T. Gillespie. 2016. "What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media & Society* 18: 410–28.

Davidson, T., D. Warmsley, M. Macy, and I. Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the Eleventh International Conference on Web and Social Media (ICWSM-2017)*. arXiv:1703.04009.

Delgado, R., and J. Stefancic. 2014. "Hate Speech in Cyberspace." *Wake Forest Law Review* 49: 319–43.

DeNardis, L. 2012. "Hidden Levers of Internet Control." *Information, Communication and Society* 15 (5): 720–38.

Downs, D.M., and G. Cowan. 2012. "Predicting the Importance of Freedom of Speech and the Perceived Harm of Hate Speech." *Journal of Applied Social Psychology* 42 (6): 1353–75.

Duggan, M. 2017. *Online Harassment 2017*. Washington, DC: Pew Research Center. http://www.pewinternet.org/2017/07/11/online-harassment-2017/.

European Commission. 2008. *EUR-Lex. Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178.

European Commission. 2016. *Code of Conduct on Countering Illegal Hate Speech Online*. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en#relatedlinks.

European Commission. 2019. *How the Code of Conduct Helped Countering Illegal Hate Speech Online*. https://ec.europa.eu/info/sites/info/files/hatespeech_infographic3_web.pdf.

Freedom House. 2018. *Freedom on the Net*. South Korea: Freedom House. https://freedomhouse.org/report/freedom-net/2018/south-korea.

Gagliardone, I., D. Gal, T. Alves, and G. Martinez. 2015. *Countering Online Hate Speech*. Paris: Unesco Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000233231.

Gasser, U., and Schulz, W. 2015. "Governance of Online Intermediaries: Observations from a Series of National Case Studies." Berkman Center Research Publication No. 2015-5. https://ssrn.com/abstract=2566364.

George, C. 2015. "Managing the Dangers of Online Hate Speech in South Asia." *Media Asia* 42 (3–4): 144–56.

Gerlitz, C., and A. Helmond. 2013. "The Like Economy: Social Buttons and the Data-Intensive Web." *New Media & Society* 15 (8): 1348–65.

Geschke, D., A. Klaßen, M. Quent, and C. Richter. 2019. #*Hass im Netz: Der schleichende Angriff auf unsere Demokratie. Eine bundesweite repräsentative Untersuchung*. Jena, Germany: Institut für Demokratie und Zivilgesellschaft. https://www.idz-jena.de/forschungsprojekte/hass-im-netz-eine-bundesweite-repraesentative-untersuchung-2019/.

Hanrahan, B. 2018. "Controlling Filth. At Facebook, the Content Police are Faceless." *Handelsblatt* (May 14). https://www.handelsblatt.com/today/companies/controlling-filth-at-facebook-the-content-police-are-faceless/23582122.html?ticket=ST-996690-WI1VB3I3sSqpg3I1MYS2-ap1.

Hawdon, J., A. Oksanen, and P. Räsänen. 2017. "Exposure to Online Hate in Four Nations: A Cross-National Consideration." *Deviant Behavior* 38 (3): 254–66.

Hayes, A.F., and K. Krippendorff. 2007. "Answering the Call for a Standard Reliability Measure for Coding data." *Communication Methods and Measures* 1 (1): 77–89.

Heath, R.L., and M.J. Palenchar. 2009. "Issue Management and Crisis Communication." In *Strategic Issues Management: Organizations and Public Policy Challenges*, 2nd ed., eds. R.L. Heath, and M.J. Palenchar. Thousand Oaks, CA: Sage Publications, 125–56.

Helberger, N., J. Pierson, and T. Poell. 2018. "Governing Online Platforms: From Contested to Cooperative Responsibility." *The Information Society* 34 (1): 1–14.

Jiang, M. 2016. "The Co-Evolution of the Internet, (Un)Civil Society and Authoritarianism in China." In *The Internet, Social Media, and a Changing China*, eds. J. deLisle, A. Goldstein, and G. Yang. Philadelphia, PA: University of Pennsylvania Press, 28–48.

Jørgensen, R.F. 2017. "What Platforms Mean When They Talk About Human Rights." *Policy & Internet* 9 (3): 280–96.

Kaakinen, M., T. Keipi, P. Räsänen, and A. Oksanen. 2018. "Cybercrime Victimization and Subjective Well-Being: An Examination of the Buffering Effect Hypothesis Among Adolescents and Young Adults." *Cyberpsychology, Behavior and Social Networking* 21 (2): 129–37.

Kaakinen, M., A. Oksanen, and P. Räsänen. 2018. "Did the Risk of Exposure to Online Hate Increase After the November 2015 Paris Attacks? A Group Relations Approach." *Computers in Human Behavior* 78: 90–7.

Keipi, T., M. Näsi, A. Oksanen, and P. Räsänen. 2017. *Online Hate and Harmful Content. Cross-National Perspectives*. New York: Routledge.

Ksiazek, T.B. 2015. "Civil Interactivity: How News Organizations' Commenting Policies Explain Civility and Hostility in User Comments." *Journal of Broadcasting & Electronic Media* 59 (4): 556–73.

Lampe, C., P. Zube, J. Lee, C.H. Park, and E. Johnston. 2014. "Crowdsourcing Civility: A Natural Experiment Examining the Effects of Distributed Moderation in Online Forums." *Government Information Quarterly* 31 (2): 317–26.

Leetaru, K. 2018. "Is Twitter Really Censoring Free Speech?" *Forbes online* (January 12). https://www.forbes.com/sites/kalevleetaru/2018/01/12/is-twitter-really-censoring-free-speech/.

MacKinnon, R., E. Hickock, A. Bar, and H. Lim. 2014. *Fostering Freedom Online: The Role of Internet Intermediaries.* Paris: UNESCO.

Mendel, T. 2012. "Does International Law Provide for Consistent Rules on Hate Speech?"." In *In The Content and Context of Hate Speech*, eds. M. Herz, and P. Molnar. New York: Cambridge University Press, 417–29.

Moor, P.J., A. Heuvelman, and R. Verleur. 2010. "Flaming on YouTube." *Computers in Human Behavior* 26: 1536–46.

Myers West, S. 2018. "Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms." *New Media & Society* 20 (11): 4366–83.

Müller, K., and Schwarz, C. 2018. "Flaming the Flames of Hate: Social Media and Hate Crime." *SSRN.* https://ssrn.com/abstract=3082972.

Neuendorf, K.A. 2002. *The Content Analysis Guidebook*. London: Sage Publications.

OECD (Organisation for Economic Cooperation and Development). 2010. *The Economic and Social Role of Internet Intermediaries*. Geneva: Organisation for Economic Co-operation and Development. https://www.oecd.org/internet/ieconomy/44949023.pdf.

Oltermann, P. 2018. "Tough New German Law Puts Tech Firms and Free Speech in Spotlight." *The Guardian* (January 5). https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight.

Oz, M., P. Zheng, and G. Masullo Chen. 2018. "Twitter Versus Facebook: Comparing Incivility, Impoliteness, and Deliberative Attributes." *New Media & Society* 20 (9): 3400–19.

Papacharissi, Z. 2004. "Democracy Online: Civility, Politeness, and the Democratic Potential of Online Political Discussion Groups." *New Media & Society* 6 (2): 259–83.

Parekh, B. 2012. "Is There a Case for Banning Hate Speech?" In *The Content and Context of Hate Speech*, eds. M. Herz, and P. Molnar. New York: Cambridge University Press, 37–56.

Park, K.S. 2015. "Intermediary Liability-Not Just Backward but Going Back." *The Global Network of Internet & Society Research Centers*, https://networkofcenters.net/research/online-intermediaries.

Penrose, J.M. 2000. "The Role of Perception in Crisis Planning." *Public Relations Review* 26 (2): 155–71.

Perry, B., and P. Olsson. 2009. "Cyberhate: The Globalization of Hate." *Information & Communications Technology Law* 18 (2): 185–99.

Pollach, I. 2011. "Online Privacy as a Corporate Social Responsibility: An Empirical Study." *Business Ethics: A European Review* 20 (1): 88–102.

Rheingold, H. 1995. *The Virtual Community: Finding Connection in a Computerized World*. London: Minerva.

Roberts, S.T. 2014. Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation (PhD thesis). University of Illinois at Urbana-Champaign.

Ross, B., M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, eds. M. Beiswenger, M. Wojatzki, and T. Zesch. Bochum: Bochumer Linguistische Arbeitsberichte, 6–9.

Saleem, H.M., K.P. Dillon, S. Benesch, and D. Ruths. 2017. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. First *Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)* at LREC 2016. arXiv:1709.10159.

Schieb, C., and M. Preuss. 2016. "Governing Hate Speech by Means of Counterspeech on Face-Book." Paper presented at the 66th Annual Conference of the International Communication Association (ICA 2016), Fukuoka, Japan, 1–23.

Scott, M., and Delcker, J. 2018. "Free Speech vs. Censorship in Germany." *Politico* (January 4). https://www.politico.eu/article/germany-hate-speech-netzdg-facebook-youtube-google-twitter-free-speech/.

Statutes of the Republic of Korea. n.d. *Act on Promotion of Information and Communication Network Utilization and Information Protection*. http://elaw.klri.re.kr/eng_service/lawView.do?hseq=38422&lang=ENG.

Su, L.Y.-F., M.A. Xenos, K.M. Rose, C. Wirz, D.A. Scheufele, and D. Brossard. 2018. "Uncivil and Personal? Comparing Patterns of Incivility in Comments on the Facebook Pages of News Outlets." *New Media & Society* 20 (10): 3678–99.

Suzor, N., M. Dragiewicz, B. Harris, R. Gillett, J. Burgess, and T. Van Geelen. 2019. "Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online." *Policy & Internet* 11 (1): 84–103.

Taddeo, M., and L. Floridi. 2016. "The Debate on the Moral Responsibilities of Online Service Providers." *Science and Engineering Ethics* 22 (6): 1575–1603.

The Asian. 2017. "84% of Women Victimized by Online Hate Speech in Korea." *The Asian* (February 20). http://www.theasian.asia/archives/98225.

Thompson, D.F. 1980. "Moral Responsibility of Public Officials: The Problem of Many Hands." *The American Political Science Review* 74 (4): 905–16.

Vedder, A. 2001. "Accountability of Internet Access and Service Providers—Strict Liability Entering Ethics?" *Ethics and Information Technology* 3 (1): 67–74.

Wakabayashi, D. 2017. Google Cousin Develops Technology to Flag Toxic Online Comments." *The New York Times* (February 13). https://www.nytimes.com/2017/02/23/technology/google-jigsaw-monitor-toxic-online-comments.html.

Wike, R. 2015. *Global Attitudes Survey.* Washington, DC: Pew Research Center. http://www.pewresearch.org/fact-tank/2016/10/12/americans-more-tolerant-of-offensive-speech-than-others-in-the-world/.

Wired Staff. 2016. "Dear Internet: It's Time to Fix This Mess You Made." *WIRED* (January 8). https://www.wired.com/2016/08/open-letter-to-the-Internet/.

Yu, W. 2018. "Internet Intermediaries' Liability for Online Illegal Hate Speech." *Frontiers of Law in China* 13 (3): 342–56.

Zillmann, D., and H.-B. Brosius. 2012. *Exemplification in Communication: The Influence of Case Reports on the Perception of Issues*. Mahwah: Lawrence Erlbaum Associates.