Organizational efforts to prevent harmful online communication

– Final Report –

Sabine Einwiller

Department of Communication

University of Vienna, Austria

sabine.einwiller@univie.ac.at

Sora Kim

School of Journalism and Communication

The Chinese University of Hong Kong

sorakim@cuhk.edu.hk

Research funded by

The Toyota Foundation

August 1 2018

Summary

Harmful online communication (HOC) severely threatens the dignity and safety of persons, social groups or organizations. Curbing this way of online expression that contains aggressive and destructive diction is a social responsibility of the organizations that provide platforms for online comments and discussion. This research focused on the measures taken by various types of organizations (e.g., web portals, news media, and online communities) in seven countries (USA, GBR, DEU, AUT, JPN, CHN, KOR). It included the analysis of HOC policies of 266 organizations (38 per country) and in-depth interviews with 60 representatives of organizations responsible for community and/or social media management.

Results of the policy analysis reveal that organizations share their policies mainly through terms of service (esp. Japan & China) or community guidelines/netiquettes (esp. Germany & Austria). While most policies are easy to find on the websites, those by Japanese and Chinese organizations are hardest to find. Policies buried in the terms of service document are also hardest to read. Organizations from South Korea provide most readable and educational policy documents, often containing examples and illustrations. In their policies, organizations from all countries mention to delete harmful comments without explanation as their predominant course of action. Regarding the possibilities for user actions, flagging a comment was the prevailing option.

Interviews reveal that manual inspection of comments is still the “gold standard” for identifying HOC. Before doing so, organizations with large user bases apply some form of machine filtering that often includes machine learning. While these tools are advancing, they are very far from being perfect. Organizations with smaller user bases and thus generally less amounts of HOC often rely on simple black lists and human inspection; some rely on manual inspection only. To do so, large organizations outsource inspection, others employ moderators; at online communities moderators are often volunteers. Chinese organizations are the only ones excessively using elaborated upload filters that prohibit posting certain words. Many organizations in the other countries are cautious to delete comments too quickly. Contrary to what they state in their policies, they usually try to moderate through communication and warnings of users. Deleting or hiding posts is generally done only when comments are clearly offensive, harmful for other users/persons/groups or illegal. Policies are helpful moderation instruments as they instruct and educate users. Nearly all interview partners perceive an increase in HOC in their country, which is often attributed to a polarization in society, the fact that people find likeminded others online for any opinion, and a lowered inhibition threshold for attacking others due to the distance perceived online.

Recommendations for HOC identification and management were derived and are presented at the end of this report.

Acknowledgements

We would like to sincerely thank our interview partners for taking the time to give us insights into their practices and experiences to curb harmful online communication. We furthermore thank our students and co-workers at the University of Vienna and The Chinese University of Hong Kong for their valuable support in collecting data for this research. Finally, our thanks go to the Toyota Foundation for funding this project.

Table of contents

1 Introduction

In the early days of the Internet, the online sphere was envisioned to provide a context for the development of collective values and community and to serve as an electronic forum where a plurality of voices engage in rational argument, thus fostering democratization (e.g., Rheingold 1995). Yet, this vision is severely hampered by plenty of highly emotional and quite often aggressive, hateful and thereby harmful voices uttered and disseminated online. Such harmful online communication—often debated as online "hate speech"—and ways to deal with it are a pressing social issue of intense discussion (e.g., UNESCO, OECD), which is led between various conflicting priorities—above all freedom of expression and defense of human dignity and safety.

Various approaches to address and reduce harmful online communication (HOC) in general, and hate speech in particular are being discussed (e.g., George 2015). At the center of the current debate conducted in many countries is governments' enforcement of national legislation (e.g., in Germany, Japan, South Korea). However, legislation is only one piece in the puzzle against the backdrop of the Internet's cross-national reach and the existence of mainly private organizations (e.g., Facebook, Naver, Baidu) offering platforms for public discussion (e.g., Banks 2010; Gagliardone et al. 2015). Partly pressurized by the public debate and partly in their self-interest, the private organizations providing or operating online platforms for comments and discussion have become more active in tackling HOC on their sites. As the owner of the space they are the actors who have decisive power of intervention.

Because of their central role in confining HOC, this research focuses on the role of organizations providing or operating online comments/discussion platforms and their measures, esp. comments policies and their application, to tackle this issue. Different types of private organizations that allow comments and discussions on their sites serve as research objects here. These include web portal sites, online news media sites, social network sites (SNS), blog hosting sites, online communities, e-commerce sites, recommendation portals, and non-Internet companies.

The research objectives were a) to identify the policy efforts regarding HOC b) to investigate the policies' implementation practices and effectiveness in order to c) derive recommendations from best practices and d) to ultimately help foster the new value of "online considerateness." To meet the objectives, the research comprised two empirical approaches:

1) Content analysis of comments policies of organizations that provide or operate online platforms with comments/discussion functions

2) Interviews with representatives of such organizations

Although HOC is a worldwide phenomenon, differences in culture, institutions and legislation suggest that there may be differences between regions and more specifically between different nations. Therefore this research adopts a cross-national approach to gain differentiated insights from organizations operating in different countries. For this purpose, organizations in four countries of the Western world (The United States of America USA, Great Britain GBR, Germany DEU, and Austria AUT) and three countries of the Eastern world (Japan JPN, China CHN, and South Korea KOR) were chosen as research objects.

2 Harmful online communication

2.1 Defining "harmful online communication"

The terminology "online hate speech" is frequently used in the public debate. One commonly used definition is that of the European Commission (2008). The European Commission defines illegal hate speech in its Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law and national laws as the public incitement to violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin. Accordingly, the term hate speech is primarily a political term, and as such it is also politically contested.

In this research, we favor the term "harmful online communication" (HOC) over "hate speech" as it focuses on the effects of the hateful and aggressive online communication. We define HOC in the following way:

> *Harmful online communication means ways of expression in online environments containing aggressive and destructive diction that violate social norms and aim at harming the dignity or safety of the attacked target, which can be a person, a social group or an organization.*

HOC often contains "hate speech", although the concept of HOC has a broader meaning and comprises the manner of expression (aggressive, hateful, or destructive) as well as its potential effect (harmful or hurting). Compared to offline expressions of hateful and harmful communication, HOC engenders various specific challenges including its lastingness, potential for virality, often anonymous or unidentifiable sender(s), and cross-jurisdictional characteristics.

2.2 Free speech and preventing HOC

The discussion on how to curb HOC is inextricably linked to the discussion on freedom of speech. In all of the countries in our research, freedom of expression is institutionalized. Yet, in all of the countries there are also legislative restrictions to this principle of free speech. These restrictions differ considerably between countries, and so does the freedom of expression on the net. There are also cultural differences regarding the unlimited acceptance of free speech by citizens.

2.2.1 Regional and cultural differences

Cultures differ in their support of free speech and their tolerance of offensive speech. According to a 38-nation Pew Research Center survey conducted in 2015 (Pew Research Center, 2015), Americans are the biggest supporters of freedom of expression among the 38 nations studied. Of the countries included in our research, Japanese citizens are the most open for restricting free speech in certain circumstances, for example when it comes to preventing people from voicing statements that are offensive to minority groups or people's religion/beliefs, that are critical of the government's policies, call for violent protests or that may be destabilizing the country's economy (see Fig. 1).

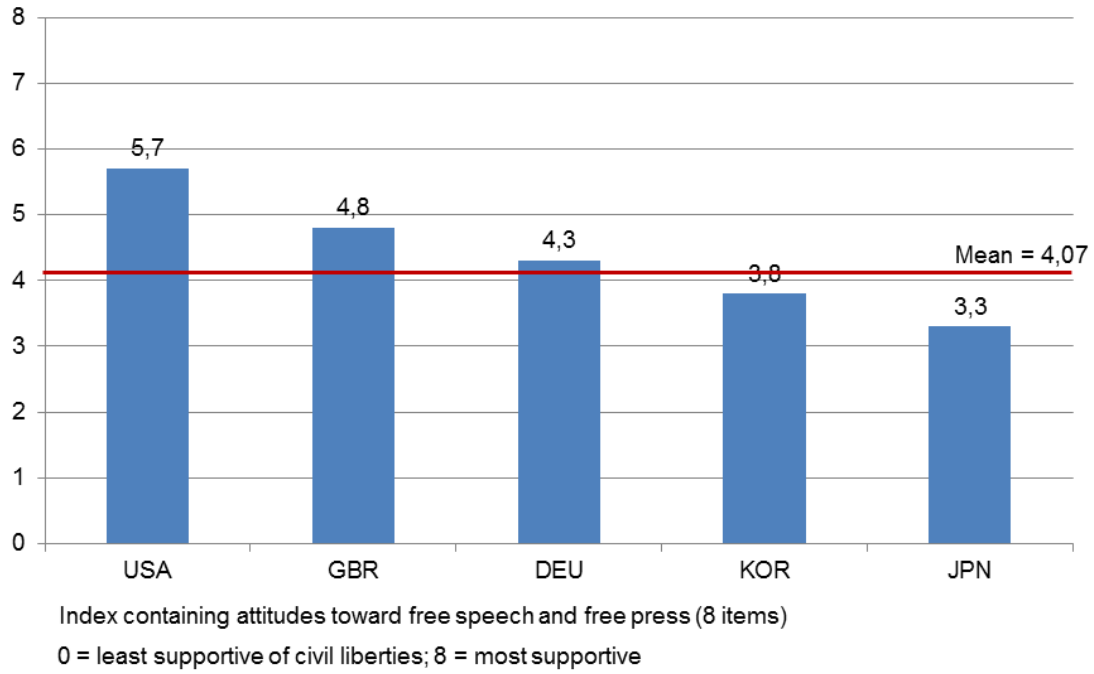

Figure 1: Attitudes towards free speech (index); Source: Pew Research Center, Global Attitudes Survey (2015); AU and CHN were not part of the study

In a comprehensive study of Internet freedom, the organization Freedom House tracks improvements and declines in government policies and practices in 65 countries each year. The score is based on examinations of laws and practices relevant to the Internet, tests of the accessibility of select websites and services, and interviews with a wide range of sources (Freedom House, 2017). Results of this study reveal the highest score, indicating least freedom on the net, for China, and the lowest scores, indicating most freedom on the net, for Germany and the USA (see Fig. 2).

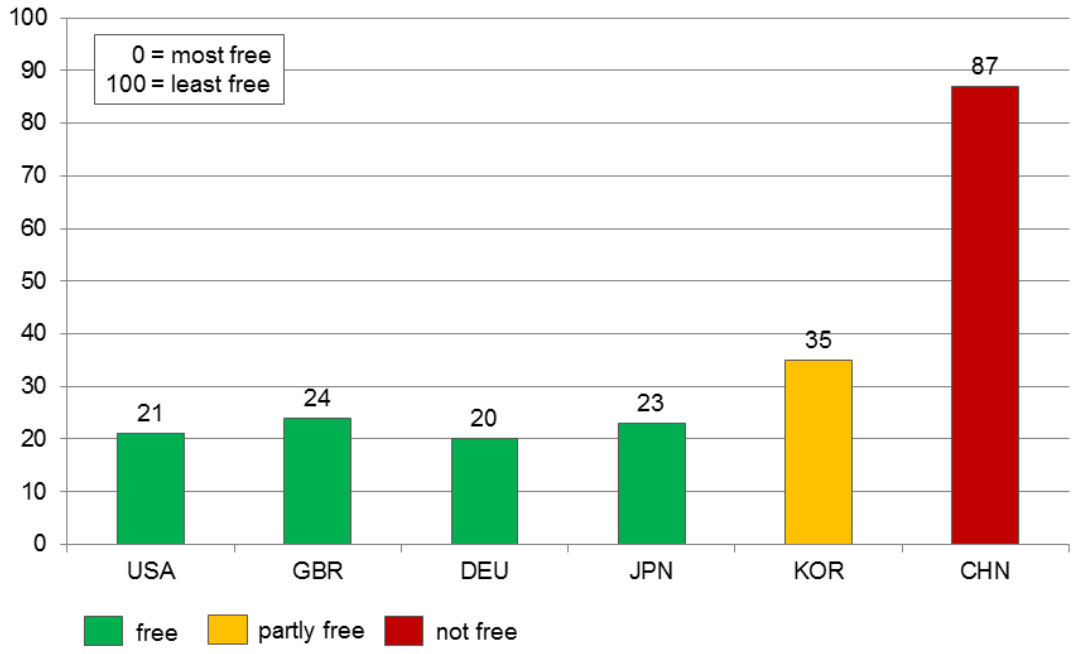

Figure 2: Freedom on the net; Source: Freedom House (2017); AU was not part of the study.

2.2.2 Legislative limitations to free speech

Countries also differ regarding their legislative limitations to free speech. In line with the results on citizens' attitudes towards free speech and freedom on the net, this principle is most strongly protected in the USA and least in China.

USA

The First Amendment - Freedom of Religion, the Press, and Expression – is part of the Bill of Rights and was intended to ensure that any ideas can be freely exchanged. Besides spoken and written words, freedom of speech encompasses all kinds of expression (incl. non-verbal communications, e.g. sit-ins, art, photographs, films and advertisements). Yet, the U.S. Supreme Court has ruled that the government may limit or ban libel (communication of false statements about a person that may harm his/her reputation), obscenity, child pornography, and fighting words (words which would incite recipients commit an act of violence), and true threats. (Ruane, 2014)

Even though freedom of expression is not unlimited, organizations providing Internet platforms are not liable when such harmful content is posted on their sites. In 1996 Congress passed the Communication Decency Act (CDA), which was meant to address a myriad of problems with regards to pornographic and illegal content on the Internet. Yet, the act's Section 230 contains a crucial aspect by stating that Internet service providers (ISPs) that host content are not legally liable for what people say on their platforms. Whether intended by Congress or not, ISPs and other interactive computer services have since used Section 230 as a complete defense against several suits brought against them. (Ehrlich, 2012)

EUROPE

In Europe Freedom of Expression is laid down in Article 10 stating that "Everyone has the right to freedom of expression". Yet, number 2 of this article codifies that states can limit freedom of speech "in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary". (European Commission, n.d.)

GREAT BRITAIN

In Great Britain, the European Convention on Human Rights is incorporated in the law, and this guarantees the freedom of expression. Yet, there are a number of exceptions. The relevant regulations regarding HOC can be found in the Public Order Act (POA) and in the Malicious Communications Act (MCA).

Section 18 of the POA prohibits the use of threatening, abusive or insulting words or behavior that intends to stir up racial hatred, whereby the offence may be committed in a public or a private place, yet inside a dwelling it has to be heard or seen also by other persons outside that dwelling. Section 4 regulates if words of behavior cause fear or provocation of violence, Section 5 harassment, alarm or distress. (Legislation.gov.uk, n.d.)

The MCA's Section 1 criminalizes sending another any article which is indecent or grossly offensive with an intent to cause distress or anxiety (which has been used to prohibit speech

of a racist or anti-religious nature). Section 127 makes it an offence to send a message that is grossly offensive or of an indecent, obscene or menacing character over a public electronic communications network. The section has controversially been widely used to prosecute users of social media. On 19 December 2012, to strike a balance between freedom of speech and criminality, the Director of Public Prosecutions issued interim guidelines, clarifying when social messaging is eligible for criminal prosecution under the law. Revisions to the interim guidelines were issued on 20 June 2013 following a public consultation. (Awan 2014)

GERMANY

While freedom of expression is constitutionally protected (Art. 5), various activities are considered punishable offenses according to the Strafgesetzbuch (StGB, Germany's criminal code). In fact, after World War II, Germany passed some of the world's toughest laws around hate speech, including prison sentences for Holocaust denial and incitement of hatred against minorities. Specifically the laws include public incitement to crime (Section 111 StGB) and incitement of popular hatred against parts of the population or to call for violent or arbitrary measures against them or to insult, maliciously slur or defame them in a manner violating their human dignity (Section 130 StGB). Furthermore punishable are the dissemination of depictions of violence (Section 131 StGB), insult, if committed by means of an assault (Section 185 StGB), defamation (Sections 186 & 187 StGB), violation of intimate privacy by taking photographs (Section 201a StGB), using threats or force to cause a person to do, suffer or omit an act (Section 240 StGB) and threatening the commission of felony (Section 241 StGB). (Bohlander, n.d.)

After criticizing the large social networks (esp. Facebook and Twitter) to insufficiently delete hate speech on their platforms, the German Bundestag passed a law – the Network Enforcement Act (NetzDG) – which came into force on January 1 2018. The law states that SNSs may be fined up to EUR 50 million if they persistently fail to remove illegal content within 24 hours that has been reported to them. The act concerns objectively criminal content, i.e. content that is clearly punishable if no in-depth examination is required to establish criminal liability. For content that requires somewhat more assessment, networks have one week to remove it. The law applies to social network operators such as Facebook, Twitter and YouTube, but not to e-mail and messenger services, professional networks, subject portals, online games and sales platforms. A limit of at least two million registered users in Germany should also prevent start-ups from being obstructed by the law in their development. (German Law Archive, n.d.) The law is controversial and tech executives and lobbyists have claimed that the NetzDG has the potential to limit freedom of expression of the country's citizens (Scott & Delcker, 2018). In preparation to the law, the networks installed additional ways to flag up offensive posts in order to report posts that contravene not just the platform's community standards but the law. It is reported that Facebook now has 1,200 people reviewing flagged content from "deletion centers" in Berlin and Essen, and Twitter is said to have hired more German-language moderators with a background for its operations in Dublin. (Oltermann, 2018)

AUSTRIA

In Austria, freedom of expression is protected by the Constitution (Article 13). Yet, various activities are considered punishable offenses according to the Strafgesetzbuch (StGB, Austria's criminal code): libel (Section 111 StGB), slander (Section 115 StGB), defamation (Section 297 StGB), coercion (Section 105 StGB), dangerous threat (Section 107 StGB), cyber mobbing and cyber stalking (Section 107 StGB) as well as public incitement (Section 283 StGB). The incitement law was tightened in 2015; now incitement needs an audience of only 30 people compared to 150 previously to be punishable (Alchinger, 2015). This was done to also make online hate speech punishable that is voiced for example in a Facebook group. Finally, the Prohibition Act (VerbotsG) punishes anyone who publicly denies the National Socialist genocide or other Nazi crimes against humanity, grossly downplays, approves or justifies it, or who works in the National Socialist sense. Austrian Media Law (MedienG, Paragraphs 6 and 7) furthermore regulates infringements regarding personality protection, libel, insult, ridicule and slander in media. (Jusline, n.d.)

In Austria, operators of online platforms where users can post and share their own content are in principle not responsible for the content posted by others and are therefore not liable for it. Yet, the platform provider becomes responsible if he learns of unlawful content, e.g. when users flag the content or send a message, and makes himself punishable if he does nothing to remove that content. This means that platform providers are obliged to delete illegal content if they receive a report about it. This is principally the same as in Germany; however, Austria does not declare such drastic punishments as proposed by the German NetzDG. (CounterAct, n.d.)

JAPAN

Article 21 of the Japanese Constitution states “freedom of assembly and association as well as speech, press and all other forms of expression are guaranteed. No censorship shall be maintained, nor shall the secrecy of any means of communication be violated” (Nihon-Koku Kenpō, 1946). In 2012, however, one sentence was added to the Article 21, amending the constitution: “notwithstanding the provisions of the preceding paragraph, engaging in activities with the purpose of harming the public interest and public order and forming associations to attain this objective shall not be recognized.” This has been receiving much criticism regarding legal limitations to free speech in Japan (Jitsuhara, 2018).

In addition, recently in 2013, the Act on the Protection of Specially Designated Secrets (SDS) has been enacted. This Designated Secrets Law allows the Japanese government to designate certain sensitive information as “special secrets” that can be free from public disclosure. Under this law, the government has a greater power to suppress free flow of information (Repeta, 2014). As a result, many are concerned about the erosion of guarantees of free speech in Japan (Jitsuhara, 2018).

SOUTH KOREA

The freedom of speech, press, petition, and gathering is protected by the South Korean Constitution. The national security law, however, prohibits the speeches or behaviors supporting the North Korean regime or communism, although recently there has been little application of this law. Identity verification law so called real-name identification law was

enacted in 2005 for online posting. However, it was revoked by the constitutional court in 2012 ruling that the identity verification law violated individuals' right to free speech (Cho & Kwon, 2015).

South Korea has relatively strict defamation laws (Statutes of the Republic of Korea, n.d.; Chapter XXXIII Crimes against reputation: Articles 307-312) under which the violators of the laws can face up to seven years imprisonments for publishing false statements. Even for publishing true statements, individuals who defame others by "publicly alleging facts" can face up to three years imprisonment (South Korea: Criminal defamation provisions, 2018). South Korea has also a specific Internet law called "Act on Promotion of Information and Communication Network Utilization and Information Protection" which governs information and communication service providers. Under this Internet law, Article 70 states similar penalties for online defamation. For instance, Article 70 (Penal Provisions) indicates "Any person who has defamed any other person by clearly and openly alleging facts with a slanderous purpose through information and communications networks shall be subject to imprisonment with prison labor for not more than 3 years or by a fine not exceeding 30 million won <Amended by Act No. 12681, May 28, 2014>" (Statutes of the Republic of Korea, n.d.; Act on Promotion of Information). These online and offline defamation laws severely undermine the freedom of expression on and offline in Korea. Although South Koreans tend to freely speak of their opinions through online and offline protests, their defamation laws lag behind, and attempts to amend the defamation laws have not been successful (Holtz, 2016).

CHINA

Although the Constitution of China (i.e., Chapter 2, Article 35) claims "citizens enjoy freedom of speech, of the press, of assembly, of associations, of procession and of demonstration" (Constitution of the People's Republic of China, n.d.), the freedom of speech has been largely limited in China. For instance, the Chinese government governs most of media organizations in China. Article 5 of the "Computer Information Network and Internet Security, Protection and Management Regulations" has also clearly prohibited inciting hatred or discriminating among nationals or harming the unity of the nation (Constitution of the People's Republic of China, n.d.). This provides the government plenty of room for limiting free speech for the purpose of national security. Many online websites and SNSs have also been banned for public use, and publications and films are often subject to active censorship in China (Jiang, 2016).

In addition, Chinese government has formulated Internet regulations and launched a series of campaigns targeting Internet content. At the end of 2011, authorities compelled microblog service operators to enforce a real-name registration program by March of 2012 for the purpose of tracking user identities (Jiang, 2016; Han, 2016; Nip & Fu, 2016). Later, the real-name registration program has been extended to all Internet platforms. In September 2013, the government launched anti-rumor and monitoring of "sensitive users" campaign together with related legislation, stating there will be penalties for publishing rumors that have been seen more than 5,000 times or shared 500 times (Jiang, 2016; Han, 2016; Nip & Fu, 2016).

3 Content analysis of online policies

Study 1 employed a quantitative content analysis method to capture the landscape of communication policy efforts made by organizations. This enabled us to identify the structures and patterns of comments policies and to empirically compare them among the different countries and types of organizations.

3.1 Sample

Documents from organizations in seven countries were collected: USA, GBR, DEU, AUT, JPN, CHN, and KOR. Because many organizations issue multiple policies, a total number of 872 policy documents from 266 organizations were included in the content analysis.

To select organizations from various categories, this study classified online platforms into eight categories: (1) web portal sites (e.g., Baidu.com, Yahoo.co.jp), (2) major online news media sites (e.g., Nytimes.com, Asahi.com) (3) social network sites (e.g., Facebook.com, Weibo.com), (4) blog hosting sites (e.g., fc2.com, blogspot.com), (5) community sites (e.g., Tianya.cn, 2ch.net), (6) e-commerce sites (e.g., Amazon.com, Taobao.com), (7) recommendation portals (e.g., Yelp.com, Tripadvisor.com), and (8) large non-Internet companies (e.g., Fortune listed or listed in a national ranking: Siemens, Toyota).

Based on the web traffic data of the selected countries the first top three organizations per each country were selected for the five categories web portal, blog hosting, community, e-commerce, and recommendation portal sites; eight online news media sites, five SNSs and ten large non-Internet companies (website or SNS) were selected. This yielded 38 organizations per country (N = 266).

3.2 Procedure

Documents including HOC-related policies were identified from the websites of the selected organizations. Documents were considered individual documents when they appeared under a separate URL or when they were under the same URL but appeared in a separate hyper link. If policies relating to HOC were part of a larger document (e.g., terms of service), they were identified in each document by drawing on the definition of HOC (see 2.1). Content addressing privacy or copyright was not considered HOC. The unit of analysis was each individual policy document, or passage on HOC within a larger document.

A code book was designed to measure the variables under investigation. The first section of the code book included variables to capture basic information about the policy document (e.g., country, type of document). The type of the policy document was classified into (1) terms of use/service, (2) community guideline/netiquette, (3) content guideline, and (4) reporting guideline. The second section focused on the accessibility and readability how the policy document, and the third section dealt with the content communicated, including how organizations handle HOC issues, what kinds of opportunities are offered for user actions, and whether it referenced related laws. At least two coders native to local languages independently coded approximately 20% of the HOC policy documents collected for each country to check intercoder reliability, which was satisfactory for all measured items.

3.3 Findings of the content analysis

3.3.1. Ways of sharing HOC policies

Of the total 872 documents analyzed, Korean organizations share significantly more policy documents with their users on their platforms (*n* = 235) than organizations from the other countries, who share similar numbers of HOC policy documents (on average *n* = 106).

About one third of the HOC policy documents are shared through terms of service or conditions of use, 30% through community guidelines/netiquette, about one quarter through reporting and 12% through content guidelines. There are some differences between the countries: Nearly half of the organizations from Japan and China communicate HOC policies through their terms of service, much more than in the other countries. Organizations from Germany and Austria, on the other hand, communicate their HOC policies mainly as community guidelines/netiquette. South Korean organizations tend to focus significantly more on reporting guidelines when communicating their HOC policies, and organizations from the USA and GBR use terms of service and community guidelines equally often to share their HOC policies.

3.3.2. Accessibility and readability of HOC policies

To assess how easy it is to find the documents, coders evaluated each document on three levels: 1) very hard to find (took more than 15 minutes to find), 2) hard to find (labelling isn't obvious or placement unexpected), and 3) easy to find (labelling is obvious and placement where one would expect it). It shows that on average 87% of the documents are easy to find. Yet, documents were easiest to find on the sites of the Korean organizations, and hardest on the sites of Chinese and Japanese organizations.

Readability was also assessed on three levels: 1) unreadable (very small font), 2) readable (medium sized font, but no illustrations or helpful color scheme), and 3) very well readable (very well designed, e.g. with illustrations, color scheme). It shows that the policy documents by Korean organizations are the most readable. They especially enhance readability of their policies by using illustrations and different colors that make the content easier to digest. HOC policies by Japanese organizations that are mainly shared through their terms of service are least readable, as terms of service documents usually use small font and technical language. SNSs put most effort into the readability of their policies, news media and E-commerce organizations the least (except for KOR).

3.3.3. Organizations' handling of HOC

The handling of HOC that is mentioned most often in organizations' policies is deleting a harmful post without explanation or comment; 85% of the organizations state that they will do so. The second most commonly mentioned HOC management method is deleting or closing user accounts in the case of HOC (80 percent). About 40% threaten with legal/judicial persecution. Only about one third mentions in their policies that they would warn a user when posting HOC and about 20% state that they would delete a post with an explanation or comment. Mentioning a committee that takes care of HOC is a specificity for the organizations from Asian countries, where about 1 in 5 mentions to use this method.

3.3.4. Opportunities for user actions

The 266 organizations provide various opportunities to users for taking action against HOC, above all marking or flagging a post (63%), followed by notifying the platform through a standardized template or email. Only 1 in 4 offers a number to make contact via telephone. The opportunity to notify an authority or government agency is provided by 20% of the organizations, almost solely from the three Asian countries. Overall, Japanese organizations state relatively limited opportunities for user actions compared to organizations from all other countries, while Korean organizations state comparatively more.

3.3.5. Content characteristics of HOC policies

A little more than half of the organizations address how users should interact with each other; German and Austrian organizations do this most often as they primarily communicate their policies through community guidelines/netiquette. Not many (14%) of the organizations educate their users with specific case examples of harmful prohibited content or behavior in their policies, yet one third of Korean organizations do so.

Three quarters of the organizations provide reference to laws and a hyperlink to a government or legal site that includes the law in their HOC policies. Only about one quarter of Japanese organizations do so, but almost all of the Korean and Chinese organizations.

4 Interviews

In the second phase of the research, an in-depth interview method was employed to investigate the organizations' HOC policy implementation practices. All 266 organizations whose policies were content analyzed in stage1 were invited for an interview. Because this did not result in a sufficient number of interviews, other relevant organizations that were not included in the content analysis stage were also contacted for the interview recruitment.

4.1 Sample and procedure

A total of 60 practitioners participated in in-depth interviews: 12 from KOR, 11 from AUT, 10 from DEU, 10 from CHN, 10 from USA, 5 from GBR, and 2 from JPN. Of these 60 practitioners, 23 were from large non-Internet companies (NI companies), 12 from news media organizations, 9 from online communities, 5 from large NGOs, 4 from social network sites (SNSs), 4 from web portal sites, and 1 from a blog hosting, e-commerce, and a recommendation portal site each. Interview partners were persons responsible for managing and/or moderating the interaction with their online users. For some organizations of interest, with whom we were not able to conduct interviews, we researched information on their handling of HOC online. This was mainly the case for the large US SNSs (above all Facebook) that are under scrutiny for their procedures regarding HOC.

Interviewers native or fluent in local languages conducted the interviews. Interviews were conducted either in-person or through Skype/phone depending on interviewees' preference. An interview guide was generated to conduct consistent interviews across countries. All interviewees gave their informed consent. Each interview lasted between 30 and 60 minutes. Interviews were recorded and then transcribed.

4.2 Findings of the qualitative interviews

The thematic analysis of the interviews revealed very similar themes across the countries, except for China. The variance of Chinese data can be explained by the difference in freedom of expression and legislation (see 2.2.1 and 2.2.2). More than between the countries (except CHN), differences with respect to handling HOC emerged between the types of organizations. Large differences regarding identifying and managing HOC were observed depending on organizations' size of user base which correlates with the number of user comments. Differences furthermore occur between organizations that manage their own proprietary platforms (on-domain management) and those whose spaces for stakeholder engagement are hosted by third party platforms like Facebook or Twitter (off-domain management). In the following, we will discuss the findings focusing on (1) identification and (2) handling of HOC.

4.2.1 Identification of HOC

There are various approaches and techniques organizations use to identify HOC; these differ depending on the size of the organization's user base and also between types of organizations. Different approaches are usually combined or applied sequentially. Generally speaking, manual inspection emerged as the "gold standard" for decisively identifying whether a comment contains HOC or not regardless of organizations' size of user base. Even when organizations apply machine learning technology, which many of the very large platform operators do, they all rely on manual inspection in a second step. Small(er) organizations often rely solely on manual inspection, often in combination with simple word filters.

In large organizations with a high volume of comments, inspection happens 24/7. Yet, medium sized platforms usually do not inspect comments during the night, but often have a window of about six to eight hours where the platform is not watched, unless they employ volunteer moderators that do not stick to business hours. Organizations with a small(er) volume of comments and a rather low frequency of HOC usually only inspect during business hours and generally not on weekends.

Next, we discuss the different approaches and techniques one by one.

Machine filtering & learning tools and word filters

The technologically most sophisticated form of identifying HOC is by using machine filtering & learning tools. Machine learning allows IT specialists to enter the offensive words and phrases to tell the computer what to look for. Then the machine learns from human decisions, which content was really inappropriate and thereby improves. Yet, technology is far from being perfect. There is no algorithm that can detect HOC with accuracy, and computers are still missing the context needed to know for sure whether a given comment is offensive or hateful. Particularly difficult is the detection of HOC in live streams (e.g., Snapchat). Nevertheless, the large Korean, Chinese and US Internet companies are using the technologies they have and are working to continuously improve it.

Chinese Internet companies use machine filtering extensively. They block many comments from being uploaded that contain words that are considered illegal or inappropriate, including

politically sensitive content. Filtering out comments that did actually not violate any laws or guidelines is accepted with approval. Since even relatively smaller sized organizations in China have very large user bases and they do not want to risk themselves for having the illegal content posted on their sites due to the strict censorship of Chinese government, they often outsource filtering to firms that offer such service.

Korean Internet companies also adopt both machine filtering and manual inspection. Most of the Internet companies with a large user base (e.g., web portals) have developed their own machine filtering system for identifying HOC which includes hate speech, obscene or sexual exploitation related content, copyright infringement content, etc. These large Internet companies tend to more strictly filter HOC than smaller organizations such as community sites with a medium size use base.

Machine filtering & learning for identifying HOC is also used by large news media organizations in the USA, Great Britain and to some extent in Austria in order to manage their user comments sections. To handle the immense moderating effort, nearly all news media open only a certain number of articles for comments. The Perspective API developed by Alphabet Jigsaw (formerly Google Ideas) is an advanced tool that helps web publishers to identify HOC (see 5.1.1).

Some online community sites also use filtering technology, which is however less sophisticated. They mainly use word filters, which supports the process of identifying HOC. However, identification of HOC is predominantly done by manual inspection with the help of moderators (see below). Facebook also offers the possibility to filter for certain words. It has a general profanity filter that is activated by default, and moderators can create their own blacklist. However, as word filters don't take context into consideration, they create wrong hits when words on the block-/blacklist are used to express positive meaning, which is seen as highly problematic by some, who therefore rather not use word filtering.

When machine or word filters are used, the content that is considered harmful or potentially harmful is filtered out and then undergoes manual inspection before it goes online; or it is hidden from the site (but still visible to the poster and his/her friends) as is the case on Facebook.

Manual inspection centers (outsourced)

It shows that manual inspection is the most important approach to identify HOC. Also the large Internet organizations that employ machine filtering all rely on manual inspection of a considerable amount of the content in a second step.

Large Korean Internet organizations have 24 hours inspection centers, of which some are abroad in countries like China and Vietnam. Chinese organizations also often outsource the inspection of their content to service providers that specialize in this kind of work.

Large SNS, like Facebook, also have such inspection centers. In response to the new law in Germany (NetzDG), Facebook and Twitter increased the number of German speaking staff in European inspection centers to identify HOC with a focus on illegal content on its German platforms (see 2.2.2). Research by Moritz Riesewieck (2017) furthermore shows that Facebook and other large SNS employ a huge number of workers in the Philippines to inspect digital content that has been flagged by users worldwide. Riesewieck found that

workers in these inspection centers work for little money and under immense time pressure. They also suffer from psychological problems, because of the cruelties they are exposed to every day combined with the strict non-disclosure agreement which prohibits them from talking about their observations (Riesewieck, 2017, p. 217f; see also Roberts, 2014).

Manual inspection by organization (inhouse)

Organizations with smaller user bases, like smaller news media, non-Internet companies and NGOs, rely heavily on manual inspection. Some use word filters to identify profanity or predefined inappropriate words (yet, these filters are deemed futile by others). Social media management tools also help to categorize comments.

To help moderators decide which content is considered harmful, many organizations rely on their external guidelines; these are sometimes supplemented by an internal handbook containing clarifications and examples. Most interview partners state that a good moderator has "the feeling" for the level of harmfulness of user comments, which is also determined by the context and the person who posts. New staff is usually trained by experienced colleagues to develop the necessary skills. In cases of uncertainty, moderators discuss the issue with their colleague(s). Exchange within the team is considered very important.

Identification by moderators and/or users

Online communities rely heavily on moderators who are recruited from the user base. Many of them moderate voluntarily and free of charge, but get some perks (e.g., invitations to events). Some communities also employ part timers as moderators, who work remotely.

Some community managers swear by moderators being selected by other moderators. The manager of a product community, which belongs to a large business enterprise, also employs user moderators; yet, he makes a point to carefully select them himself. User moderators are identified on the basis of their social skills which they prove by arbitrating as regular users. Moderators generally have a platform for exchange among one another and also with the community management team, when in doubt whether a comment is harmful or not.

Aside from user moderators, HOC is also identified by regular users in most of the organizations. On some news media and online community sites, users can send an email to the community management reporting inappropriate content. On SNSs like Facebook and Twitter, users have the possibility to flag HOC using the "report" link near the post, photo or comment, which is then sent to one of the site's inspection centers.

4.2.2 Handling and responding to HOC

The universal themes that emerged for organizations' approaches to handle and respond to HOC are (1) warning and decisively communicating with users (3) 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 HOC management methods are widely adopted by organizations, yet the degree of employing them varies by country and type of organization.

While Chinese organizations rely heavily on preventative blocking of so called illegal content, organizations in the other countries are generally rather careful to delete content or block users, and do so only if the content is severely harmful. Trolls and spam, however, are

usually deleted or blocked quickly. Having a good system for internal discussions in place, where moderators can exchange thoughts and discuss difficult decisions is widely considered important.

Warning and decisive communication

Decisive communication with users and referring to the policies is considered highly important. Interviewees state that pointing out publicly where and why comments violated the policy can help to educate the poster and other users who are observing. However, doing so is often not possible when the volume of HOC content is too large. It is also not promising when a user is clearly trolling or posts are severely harming others so that they have to be removed immediately. Leading a discussion at eye level and giving it a "human touch" (showing that there is a human being on the other side) was generally considered important. Yet, whether the moderator should comment with his or her full name was debated; some do but many don't for reasons of personal safety.

Some news media sites pin constructive comments above or next to the article to increase the user experience and to educate the community with positive examples. One news medium developed a humorous quick-witted style to respond to HOC. While this response style seems to work for this medium, others are careful to use humor for countering HOC.

Interview partners report that other users also step in to counter harmful posts or defend the person or organization attacked; they freely do so without being animated by the organization. In online communities, appointed volunteer moderators mainly assume this job (see below).

Interviewees point out that when they are commenting posts, they are generally doing so in a non-partisan way. This is particularly the case when HOC is political.

Hiding or deleting HOC posts

Some organizations (mainly on proprietary sites) replace forbidden words with an asterisk; one of the large Korean web portal sites uses music notes (♪ ♫). Yet, some make a point to delete a comment completely when it contains inappropriate words. Only one organization from the US modifies posts by taking out the offensive parts. In most countries this is problematic, as it means taking ownership which implies becoming liable for the content.

On Facebook hiding HOC is a common approach and can be done by the organization that uses the SNS as a platform. Comments containing prohibited words are automatically hidden. As the filter sometimes gets it wrong, some companies also check automatically hidden comments. Some consider it important to comment hidden posts and let the poster know that his/her behavior was not appropriate. Sometimes organizations write a public post telling that a post was deleted and for what reason in order to clarify their procedure.

Deleting comments on SNSs (Facebook or Twitter) can only be done by the SNSs. Thus, the organization needs to report comments to the SNSs when it wants them deleted (for reporting see below). Deleting is a more common option for organizations that operate their own sites. News media organizations or community sites often pre-moderate comments, i.e. do not post HOC in the first place. Yet, not everything is caught and those not adopting pre-moderation will delete HOC that has been posted on their site afterwards.

Compared to organizations from other countries, Chinese organizations rely heavily on preventative blocking of so called illegal content such as politically sensitive information, pornography, drug-related information. Interviewees emphasized the automatic blocking of HOC posts before uploading. Users usually receive a notice when a comment contains a forbidden word, so they keep trying to find workarounds by using alternative spellings.

Blocking/locking user accounts or closing discussions

Users are usually blocked only after multiple violations and several warnings. On Facebook, blocking a user is possible but rare. However, trolls or posts from spambots are blocked quickly. Community sites usually have a system of writing bans for different periods of time.

Organizations that manage their own platforms (e.g., news media) mention to sometimes close a discussion, when comments get too aggressive and discussions derail. Many news media organizations have turned to only open a certain number of discussions to begin with.

Reporting HOC and illegal content

Some interviewees mentioned that they have reported illegal content to the police, or other government authorities responsible for legal infringements, e.g. when users have voiced death threats against one another or a community moderator.

Organizations using Facebook report severe HOC to the SNS asking to delete it, because they cannot delete but only hide comments. Reporting to the platform provider or operator can also be done by users. Flagging/reporting functions on SNSs are a required feature in many countries. In the different countries there are also reporting centers, where users/ citizens can report HOC they observed or experienced.

Management by volunteer moderators (mainly community sites)

Community sites in all countries employ volunteers and some also paid moderators to deal with HOC management. These moderators usually have a range of options. Aside from intervening by communication, they can issue warnings to users, lower their user ranking (if it exists), delete HOC posts, and ban users for a period of time. Some community sites manage HOC posts through systematic punishment (yellow-cards) and rewards.

Usually, the management team of the online community steps in when a situation gets out of hand, or when the moderators are at their wit's end. One of the community managers mentioned to be reachable by phone 24/7 for moderators in case something goes wrong.

Review management system by Korean organizations (mainly large Internet companies)

Korean web portal sites, which also operate big domestic SNS, have a unique review process in place. They transparently share how they handle HOC. Through operating a reporting center, they allow users to request blocking or deleting a potentially harmful post. Once a request form is submitted, organizations immediately perform a 30 day removal of the reported content (this is common for all Korean Internet orgs.). During the following 30 day review process, the one being accused of HOC can appeal or request a formal objection through the reporting center. For gray areas, the organizations consult with the Korea Internet Self-governance Organization (KISO). The member organizations jointly review such gray cases and try to set up shared standards. Through the review process, a reported post can be permanently removed if found to be offensive, or reposted if no harm is evident.

Although apparently transparent and democratic, many pointed out shortcomings of this process. Since the reported posts have to be removed for 30 days, people abuse it for the purpose of a temporary removal of posts they are politically or otherwise opposed of. Similar scenarios have also been observed on Facebook, as FB shuts down a site when a high number of different users report it. Critics of the new German law (NetzDG) fear that scenarios like this as well as wrongful removals of comments will increase.

4.2.3 Interviewees' assessment of the legal situation in their country

Generally speaking, interviewees emphasized that laws are not the solution, but that it needs public awareness and cultural change to fight HOC. Assessment of the legal situation varied between countries, because of the different laws and limitations to free speech (see 2.2.2).

Practitioners in the USA had given least thought to the legal situation in their country. They seem to be opposed to legal interference and rather regulate HOC themselves. In Great Britain skepticism about legal restriction is also pronounced. Yet, interviewees take regulations like the Malicious Communications Act into account to justify their actions. In Germany, interviewees do not see the new law (NetzDG) to affect them, but only the big SNSs like Facebook, Twitter and Youtube. Although they are aware of and share certain criticism, some perceive the NetzDG as a helpful move to exert pressure on the big Internet players, especially after the government had tried for some years to negotiate with these organizations to take stricter actions against online hate. The majority of Austrian practitioners consider the laws sufficient, yet some state that legislation lags behind. In both, Germany and Austria, practitioners state that a solid solution requires the collaboration of the large Internet players (esp. FB, TW, Google/Youtube). Japanese practitioners do not find too much restriction by law desirable, while laws related to hate speech are seen as vague. Korean practitioners are of the opinion that laws cannot manage HOC, especially the type related to politics. They criticize that laws can be used to suppress political views, and that too much regulation can infringe autonomy and freedom of speech. Chinese interview partners were reluctant to comment on the strict laws that limit freedom of speech. Yet, they criticized that current laws governing promotions and information dissemination by competitors were lagging behind, and wished for more regulations in that area.

5 Recommendations on identifying and managing HOC

Harmful online communication (HOC), defined as ways of expression in online environments containing aggressive and destructive diction that violate social norms and aim at harming the dignity or safety of the attacked target (a person, social group or an organization), is a severe issue around the world. Our interviews with practitioners who deal with HOC every day as online moderators, community or social media managers largely affirm that this issue has been intensifying over the past years. There are various explanations for this negative trend; our interview partners mainly mentioned the following: a polarization of society in many countries, the fact that people find likeminded others online and thus support for any opinion, and a lowered inhibition threshold for attacking another person due to the distance perceived online.

Interviewed practitioners are fully aware of the problems arising through HOC and of the necessity to keep their platforms civil, be it for business and image reasons, to safeguard them from judicial persecution or because of the social responsibility they have for the well-being of their stakeholders. No organization is completely successful in dealing with HOC, if success means that comments and discussions are kept completely free from HOC and civil all the time. Yet, all are anxious to get close to this goal and some seem more successful in approaching it than others. This has a lot to do with the size of the user base that posts comments –including HOC– day by day, but also with the methods applied. From the research conducted we derived the procedures and practices that have proven particularly useful to handle HOC independent of different national and cultural environments.

In the following, we will present recommendations for organizations providing online platforms for comments and discussion, divided into the sections “Identification of HOC” and “Managing comments and discussion to curb HOC”.

5.1 How to identify HOC

The approaches to identify HOC include machine learning tools, word filters, manual inspection and user reporting.

5.1.1 Machine learning tools

For organizations with large user bases applying a machine learning tool that supports the process of identifying HOC seems useful, yet only in combination with manual inspection of still a relatively high number of comments. Organizations may develop a proprietary solution, extend a solution for their particular needs or use a tool that is offered on the market.

A publicly available tool that came up in our interviews is “Perspective” developed by Alphabet Jigsaw (formerly Google Ideas). Perspective is based on a deep-learning model trained by a very large number of manually reviewed comments. The tool provides a score from zero to 100 that indicates how similar new comments are to others previously identified as toxic. Toxicity is defined as how likely a comment will make someone leave a conversation. The developer states that Perspective can be used in a number of ways, from giving users feedback on the toxicity of their comments to offering organizations a way to filter comments (Daisuke, 2017). Although the program is far from perfect, as has been pointed out by critics (e.g., Cheatmaster30, 2017), Perspective seems to be an applicable tool in the quest for finding a way to quickly evaluate a large number of comments. Yet, a relatively high percentage of comments identified as potentially toxic still need to be inspected manually, and those that are let pass may still contain HOC; thus, attention needs to be paid to the published ones, too.

5.1.2 Word filters

For organizations with larger as well as smaller user bases the use of word filters can be helpful to flag potentially harmful comments. Organizations can install their own filters, adapted to their specific user base and needs. On Facebook, for example, the profanity filter can be set at low, medium or high, and/or a blacklist can be created individually. The

identification and moderating process can furthermore be supported to some extent by social media management tools that help categorize comments.

However, word filters often get it wrong and block comments that contain potentially harmful words used in a positive sense. This can be particularly problematic for organizations that work in controversial contexts where supportive comments may contain "inappropriate" words. Another problem is that users find workarounds by using alternative ways of writing forbidden words to bypass filters. Thus, although word filters can be helpful in flagging potentially harmful comments, manual inspection of blocked posts and of those that go online is still necessary.

5.1.3 Manual inspection

Manual inspection is still the "gold standard" to identify HOC. Large Internet organizations often outsource inspection to inspection centers operated by third parties. If this is the case, organizations must exercise their responsibility to provide good working conditions, including psychological care; this of course also applies when inspections are done in-house. When people from other cultural backgrounds inspect comments, it is essential that they consider the cultural and legal background of the comments' country of origin.

Better than outsourcing content moderation is to employ in-house moderators, or moderators that are part of the online community. In that case, moderators have a closer connection to the organization, its values and understanding of HOC. Moderators can very well work remotely, but need regular exchange with managers and other moderators and must be able to always contact someone when in doubt or trouble.

Content moderators have to be able to draw on clear guidelines to decide what is considered harmful or not. These are, first of all, the external guidelines communicated to the users. Additionally, an internal handbook containing clarifications and case examples is helpful especially for newer staff that is not yet that experienced.

5.1.4 User reporting

Regular users of a platform need to be provided with easy to see and easy to use options to report HOC. These can be buttons right next the post and simple contact forms easily found on the site. Users should also be encouraged to report when they see something, for example by regular notices from the organization to do so.

5.2 Managing comments and discussion to curb HOC

There are several approaches to handle HOC. Importantly, there need to be clear policies in place, on the basis of which users can be warned and educated. If warning and intervening by communication deems ineffective, hiding or deleting HOC posts (or words) or –in severe cases– blocking/locking user accounts is advisable. Finally, reporting HOC/illegal content to third parties (e.g., police) may be necessary.

5.2.1 HOC policies to inform and educate users

Clear communication with users about what is considered HOC and how a platform manages such harmful content is pivotal in successful HOC management. When users are not clear

about the platform's HOC guidelines and what is considered intolerable, a vicious cycle may be recurring where users keep posting HOC, and the platform keeps deleting or blocking it. Organizations have 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.

HOC policies need to be easily found and easy to read, because users generally do not actively search for policies let alone read lengthy documents in which policies are buried. When HOC policies are included somewhere in the terms of service, users will hardly become aware of them; and even if, they may not bother reading them because of the small font and technical language. Sharing HOC policies through community or content guidelines (netiquette) is advisable as these are usually much better accessible and readable.

A good practice is to install dedicated and well-designed webpages where uses are informed and educated on what is considered inappropriate. For instance, some big Korean Internet companies provide such pages under the term "Green Internet" where explanations of what is considered harmful content and what the organization does to prevent it is provided in conjunction with examples. These instructing and informing webpages related to HOC are connected to the reporting center, where user are invited to actively report potentially harmful content they have come across.

Once clear policies are shared online, organizations must strictly follow them when handling and responding to HOC. When discrepancy is observed, users tend to not follow but to bypass the guidelines. As a result, disputes between organizations and their users increase.

5.2.2 Warning and intervening by communication

Clear communication with users about what is considered inappropriate/ harmful with reference to the policies should be done whenever capacity allows. Pointing out publicly where and why comments violated the policy helps to educate the person who posted and other users who are observing.

The conversation should be lead at eye level and not with a wagging finger as this may cause renitency. Warnings and responses should furthermore be given in a non-partisan way, unless the organization wants to take a stand against a particular issue. Responding in a humorous way can work well in specific cases, as it causes surprise and takes out the aggressiveness.

Showing a "human face" can also help to appease a situation. While clear names foster this human approach, moderators may want to write under a pseudonym for safety reasons.

A helpful system, mainly for communities, is to have users give warnings to others who do not adhere to the rules. For example, a community site successfully runs a system of punishment/rewards where users can yellow-card others; those who have received 5 yellow cards in a month are consequently blocked from posting for a certain period of time.

To foster interventions against HOC by other users, moderators can encourage such counter speech by liking the counter-posts or by writing a supportive comment publicly or also privately. Users that are showing moderator behavior can be incentivized for example by bonus points and small perks, or they may be asked to take on an official moderator role.

5.2.3 Hiding or deleting HOC posts (or words)

Generally speaking, posts should only be hidden or deleted when violations against the policies and/or the law are clearly evident. Freedom of expression is a high good, which must be protected most certainly; therefore, procedures of excessively deleting posts that only give the impression of containing HOC is highly objectionable.

Yet, posts that are clearly harming others have to be removed immediately and may be hidden or deleted by means of machine or word filtering anyways.

Deleting or replacing forbidden words with symbols is a possible approach to suppress profanity. Yet, it can animate users to find workarounds and use different spellings for such words, which is counterproductive. It seems more advisable to hide or delete such posts completely, if really offensive, and point the persons who posted them to the respective guidelines. Modifying user posts (e.g., deleting objectionable content) is not advisable, as this may cause problems of liability in many countries.

When HOC posts are hidden, but the post is still visible for the one who posted and his/her friends, the person should be informed that his/her behavior was not appropriate and why. It is also advisable to leave a notice when a post was deleted/hidden briefly stating the reason, to show transparency and to educate other users. Yet, it is usually sufficient to leave such a notice from time to time and not every time.

Whether SNSs should allow organizations that use their platforms to autonomously delete posts is a contested question, as this would shift responsibility away from the platform provider. Instead, when organizations report HOC to the SNSs, there needs to be an expedited response mechanism, so the harmful content is deleted very quickly.

5.2.4 Blocking/locking user accounts

Blocking users should only be done after several warnings. However, trolls or posts generated by spambots need to be blocked quickly.

Calling out writing bans before taking the drastic step of completely banning a user is advisable.

Closing down a discussion when comments get too aggressive is an unfortunate but sometimes necessary step.

5.2.5 Reporting HOC/illegal content

Content that clearly violates a law should be reported to the respective government authorities.

Organizations using third party platforms need to report HOC to the platform provider.

Users should be encouraged to report HOC and/or illegal content to the platform operator and/or to the platform provider.

5.2.6 Delegation of management tools to users

For platform providers that host multiple users' sites (e.g., community or blog hosting sites), the delegation of HOC management tools to those who operate their personal or manage an

organization's sites is recommended. By delegating HOC management like hiding, blocking or deleting HOC, hosts empower their users who try to curb HOC on their sites, while saving resources themselves. However, responsibility is still also with the platform provider.

The delegation of power should be done with securing a clear HOC threshold of all user sites, as some users/organizations might set their HOC threshold below an acceptable level. This could negatively affect the entire platform. Thus, a clear mandatory threshold of what is not accepted on the platform should be set.

List of references

Alchinger, P. (2015). Strafgesetzbuch: Verhetzung liegt künftig schneller vor. *Die Presse*, March 12. Retrieved from: https://diepresse.com/home/politik/innenpolitik/4684036/Strafgesetzbuch_Verhetzung-liegt-kuenftig-schneller-vor

Article 19 (2018). South Korea: Criminal defamation provisions threaten freedom of expression. *Article 19,* May 10. Retrieved from: https://www.article19.org/resources/south-korea-repressive-criminal-defamation-provisions-threaten-freedom-of-expression/

Awan, I. (2014). Islamophobia and Twitter: A Typology of Online Hate against Muslims on Social Media. *Policy & Internet, 6*(2), 133-150.

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology, 24*(3), 233-239.

Bohlander, M. (n.d.). *Translation of the German Criminal Code*. https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html#p1151

Cheatmaster30 (2017). Why Alphabet's AI Cannot Identify Hate Speech. *Towards Data Science*, Aug. 21. Retrieved from: https://towardsdatascience.com/why-alphabets-ai-cannot-fix-hate-speech-8d352892cdba

Cho, D., & Kwon, K. H. (2015). The Impacts of Identity Verification and Disclosure of Social Cues on Flaming in Online User Comments. *Computers in Human Behavior, 51*, 363-371.

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

CounterAct (n.d.). *K(NO)w more!* Retrieved from: https://counteract.or.at/know-more/

Ehrlich, P. (2012). Communications Decency Act 230. *Berkeley Technology Law Journal, 17*(1), Article 23. Retrieved from: https://scholarship.law.berkeley.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1358&context=btlj

European Commission (n.d.). *Freedom of Expression*. Retrieved from: https://ec.europa.eu/europeaid/sectors/human-rights-and-democratic-governance/democracy/freedom-expression_en

European Commission (2008). *EUR-Lex. Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law*. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178

Freedom House (2017). *Freedom on the Net*. Retrieved from: https://freedomhouse.org/sites/default/files/FOTN_2017_Final.pdf

Gagliardone, I., Gal, D., Alves, T. & Martinez, G. (2015). *Countering Online Hate Speech*. Paris: Unesco Publishing.

George, C. (2015). Managing the Dangers of Online Hate Speech in South Asia. *Media Asia, 4*2(3-4), 144-156.

German Law Archive (n.d.). *Network Enforcement Act* (Netzdurchsetzunggesetz, NetzDG). Retrieved from: https://germanlawarchive.iuscomp.org/?p=1245

Han, L. E. (2016). Control and Resistance: Remembering and Forgetting in the Changing Dynamics of State, Market, and Individuals. In *Micro-blogging Memories* (pp. 51-80). Palgrave Macmillan UK.

Holtz, M. (2016). Why a presidential scandal is boosting free speech in South Korea. *The Christian Science Monitor,* December 6. Retrieved from: https://www.csmonitor.com/World/Asia-Pacific/2016/1206/Why-a-presidential-scandal-is-boosting-free-speech-in-South-Korea

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

Jitsuhara, T. (2018). Guarantee of the Right to Freedom of Speech in Japan-A Comparison with Doctrines in Germany. In Y. Nakanishi, (ed.), *Contemporary Issues in Human Rights Law* (pp. 169-191). Singapore: Springer.

Jusline (n.d.). *Strafgesetzbuch*. Retrieved from: https://www.jusline.at/gesetz/stgb

Legislation.gov.uk (n.d.). *Public Order Act 1986*. Retrieved from: http://www.legislation.gov.uk/ukpga/1986/64#commentary-c13655761

Nihon-Koku Kenpō (1946). *Postwar Constitution. Article 21, Paragraph 1*.

Nip, J. Y., & Fu, K. W. (2016). Challenging Official Propaganda? Public Opinion Leaders on Sina Weibo. *The China Quarterly*, *225*, 122-144.

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

Pew Research Center (2015). *Global Attitudes Survey*. Retrieved from: http://www.pewresearch.org/fact-tank/2016/10/12/americans-more-tolerant-of-offensive-speech-than-others-in-the-world/

Repeta, L. (2014). Japan's 2013 State Secrecy Act -- The Abe Administration's Threat to News Reporting. *The Asia-Pacific Journal,* March 3. Retrieved from https://apjjf.org/2014/12/10/Lawrence-Repeta/4086/article.html.

Rheingold, H. (1995). *The virtual community: Finding connection in a computerized world*. London: Minerva.

Riesewieck, M. (2017). *Digitale Drecksarbeit. Wie uns Facebook & Co. von dem Bösen erlösen* [Digital dirty work. How Facebook & Co. redeem us from the evil]. München: dtv.

Roberts, S. T. (2014). *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation*. Doctoral dissertation, University of Illinois at Urbana-Champaign. Retrieved from: https://www.ideals.illinois.edu/bitstream/handle/2142/50401/Sarah_Roberts.pdf?sequence=1

Ruane, K. A. (2014). Freedom of Speech and Press: Exceptions to the First Amendment. In *Congressional Research Service*. Retrieved from: https://fas.org/sgp/crs/misc/95-815.pdf

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

Statutes of the Republic of Korea (n.d.). *Chapter XXXIII Crimes against reputation: Articles 307-312*. Retrieved from: http://elaw.klri.re.kr/eng_service/lawView.do?hseq=28627&lang=ENG

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