Transcript of Webinar #4

# Key Priority Issue: Free Expression

# 01

Hi again. Welcome to the final webinar of this four-part webinar series on business and human rights in the digital environment. My name is Spandana Singh at New America’s Open Technology Institute.

# 02

The first two webinars in this series provided an introduction to business and human rights issues and how they intersect with the digital environment. In the previous webinar we provided a basic overview of privacy in the digital age, including how issues of privacy relate to corporations and governments, and what some of the best practices for assessing and protecting privacy in the digital sphere are. This final webinar will provide a foundational overview of another key priority issue -- free expression online. This webinar will include discussions on content regulation, transparency reporting and relevant best practice frameworks and principles for understanding how to safeguard free expression in the digital age.

# 03

Freedom of expression can broadly be defined as an individual’s right to say, write and produce content on almost anything they’d like without restrictions from the government. The rise of the digital age, in particular the growth and large-scale adoption of platforms that enable users to create and disseminate their own content has significantly increased global information flows and democratized and enhanced freedom of expression online.

# 04

On an international level, freedom of expression is enshrined in Article 19 of the Universal Declaration of Human Rights (UDHR) as well as the International Covenant on Civil and Political Rights (ICCPR). Freedom of expression is also a major component of regional conventions and charters including the European Convention for the Protection of Human Rights and Fundamental Freedoms, the American Convention on Human Rights and the African Charter on Human and Peoples’ Rights (Banjul Charter) which are also applicable to free expression online. These international treaties and agreements generally outline the obligations for governments regarding human rights, whereas frameworks such as the UN Guiding Principles provide guidance for how companies should integrate respect for human rights.

# 05

According to the UDHR: “everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.” The ICCPR states something similar, including that the right to free expression includes the “freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice”.

# 06

The principle of freedom of expression is also outlined in the GNI Principles, which are accompanied by a set of Implementation Guidelines for use by participating companies.

# 07

As with privacy in the digital age, companies and governments play different roles in protecting free expression, and they pose different threats to the exercise of this right. With regard to companies, they have become integral to creating spaces for free expression and often serving as defenders against overbroad censorship. On the other hand, companies’ practices in determining what content to allow on their platforms can undermine free expression. With regard to governments, they can provide vital legal protections to promote free expression, including rules related to intermediary liability that enable companies to permit free-ranging speech on their platforms. On the other hand, too often governments act to stifle free expression, for example, through network disruptions or by demanding that certain content be removed.

# 08

As technology platforms such as Facebook, Twitter and Google have been adopted across the globe, they have assumed the role of digital public squares, enabling for the free flow of information and enhanced free expression. They have also, however, assumed the role of de facto gatekeepers of online speech. And just as threats to speech offline have grown, so have threats to online speech, raising significant concerns about censorship and limits to free expression online.

Most American technology companies apply a First Amendment perspective on speech issues in the United States but comply with other countries’ laws when applicable. However, depending on local laws in varying countries, this may require taking specific steps in certain countries. For example, in Germany it is illegal to deny the Holocaust and as a result, platforms that host user-generated content must treat such speech as illegal in the context of that country. Many companies attempt to limit such takedowns to the applicable country through such techniques as geoblocking.

In addition, many companies develop their own standards and regulations around permissible speech on their platforms. These rules may restrict certain categories of speech beyond what would or could be prohibited by law in particular countries. For example, Facebook prohibits nudity, violence, bullying and harassment even though some speech in these categories would be protected by the First Amendment if the government sought to prohibit it.

Companies have a commitment to comply with speech-related laws around the world, but as private actors they have the right to develop their own standards for permissible content on their services. However, the overbroad censorship of speech is a significant human rights concern and a threat to free expression that needs to be assessed carefully. Some companies aid in the protection of free expression by often pushing back against requests to implement overbroad censorship regimes, however, given the amount of information and speech they engage with, they need to be held accountable in their management of online speech.

# 09

To enforce legal rules and companies’ own terms of service, companies will review content to determine whether or not to remove it from their platforms.

Content moderation can be defined as the process companies employ to review and potentially to remove, restrict or regulate content that has been identified or flagged as violating a particular law or a company’s content policies or terms of Service. Content moderation can also result in the suspension or closure of the accounts of users who violate these regulations.

# 10

Companies deploy an array of processes to remove or restrict content and accounts. There are two common categories of approaches to content moderation. The first is the use of human moderators. Human moderators include company employees and contractors who are trained and familiar with a company’s policies and the laws they need to comply with. Users and trusted organizations such as NGOs familiar with a particular community or issue can also flag violating content. The second approach is the use of algorithms or artificial intelligence to remove content. AI tools present risks of algorithmic bias since biases present in training data can be exacerbated by machine learning tools. Further, both human moderation and AI tools can have difficulty judging context and can result in mistaken takedowns and overbroad censorship. Most companies utilize a combination of these approaches to make decisions on when to remove content. Generally, human reviewers are better at judging context, such as whether a post represents terrorist recruitment or whether it aims to educate users about human rights abuses. Thus, having a human in the loop can help mitigate some risks created by AI tools.

# 11

Examples of content that is commonly removed by platforms for violating laws are content that infringe on copyright or trademark. In addition, as mentioned before, companies typically also remove content based on local regulations and laws. For example, in Germany, major platforms are required to remove terror content or face significant fines. There are similar movements to institute legislation to restrict fake news around the world. In cases where the content is only illegal in one particular jurisdiction, and is not illegal in other areas, such as content that violates Blasphemy laws in Pakistan or Holocaust denial laws in Germany, companies typically restrict content in a particular geographic location, through a process known as geo-blocking, rather than removing the content altogether. This prevents the content from being viewed using geographic specific IP addresses, but permits users globally to see it. Depending on the type of content, companies can receive requests to remove content from governments, users, companies etc.

# 12

Companies are increasingly facing pressure to remove content, including hate speech and terror content in order to make their platforms safer. However, they also have a commitment to protect free expression, including the speech of communities that may be victimized by hate speech. These tensions can come to a head in cases like when minority communities ask platforms to take down hate speech or abusive content and platforms must work to comply while also attempting to uphold and not impinge on these communities’ and other communities’ free expression.

# 13

As noted previously, companies can also remove content on their platforms based on their own content policies, which define standards of acceptability on their platforms. Typically, these policies restrict speech far beyond what speech related laws do. For example, in the United States, companies’ Terms of Service can restrict speech that the First Amendment cannot legally restrict because the First Amendment only regulates what the government can do. The expansion and development of more comprehensive Terms of Service speech restrictions raise several free expression concerns. Company content policies are continuously evolving, mostly in response to emerging forms of controversial speech. Today, for example, we are seeing more and more companies take stances and institute policies in response to hate speech, terror content, graphic violence and fake news. Because these removals do not occur on the basis of legal frameworks, there is often significant controversy around these policies, and overbroad and strict policies are often criticized for curbing free expression and limiting free speech. As a result, companies have faced greater pressure to provide transparency around their content policies and content moderation processes.

# 14

Here is an example of how Facebook has responded to these calls. In May, they released a detailed set of their Community Standards which outline what content is and isn’t permissible on their platform, and how Facebook moderators typically think about these situations.

Their “Dangerous Individuals and Organizations” section outlines that they do not allow any organizations or individuals that are engaged in the following activities to have a presence on Facebook: terrorist activity, organized hate, mass or serial murder, human trafficking and organized violence or criminal activity. They also remove content that expresses support or praise for these groups and their leaders and members.

# 15

Companies can remove content at three stages: 1) before it is published, 2) after it has been flagged for moderation and 3) actively assessing content on the platform for removal.

There are significant free expression concerns when companies attempt to remove content before it is published as this typically signals that they are attempting to act as speech gatekeepers and curb expression. One of the only categories of content that is internationally recognized as illegal and therefore can be acceptably moderated in the pre-publication stage is child pornography and child sexual abuse images. This content is identified online by expert organizations and using the algorithmic PhotoDNA technology is converted into digital hashes. These hashes can be used to identify copies of such child pornography content on platforms and remove them after they have been uploaded, or when they are still in the upload process. The pre-publication takedown of this content in this case is permissible because the content is universally considered to be illegal, and is easily identified as fitting this category. However, given that other forms of content, even terror content and fake news lie within a grey zone of acceptability and permissibility, pre-publication takedowns are considered a major free expression concern. Pre- publication takedown technology has also been applied to copyrighted content. Companies can receive flags to moderate content from a range of actors including governments and users. Many companies like YouTube have Trusted Flaggers programs in which organizations and individuals who have a high accuracy rate of flags for particular types of content, such as content that violates a company’s graphic violence or hate speech policies, are permitted to flag content to moderators in bulk, and this content receives priority review. Some governments in the EU have also set up Internet Referral Units to flag terror content to companies to remove.

Companies can also actively assess content on their platforms using trained algorithmic models in order to identify and review content that violates laws (especially when there are fines and legal consequences involved) and their own content policies. However, this technology has not yet been perfected and as a result companies are always pushed to incorporate human reviewers in order to ensure content that is permissible on the platform is not mistakenly being removed. We will talk more about the limits of AI in content moderation later on in this webinar.

# 16

Companies are receiving an increasing number of requests to moderate and remove content, as demonstrated by this map image from Twitter’s latest transparency report on content removal requests, which covers the period of January to June 2018. This poses significant concerns regarding free expression safeguards, especially when there is little transparency around the policies and practices companies use to guide their moderation decisions, the scope of requests for moderation companies are receiving and the volume of content being removed. As a result, companies have come under pressure to disclose data related to their content removals in their transparency reports. Providing such transparency is also beneficial as it highlights where companies have successfully safeguarded users’ free expression.

# 17

This is a snapshot from Google’s latest transparency report on government requests to remove content, which provides data from December 2009 to June 2018. This snapshot highlights the range of categories of content that are being targeted by government requests, and how they have become increasingly prominent.

OTI recently released a Transparency Toolkit which surveyed international and domestic technology and telecommunications companies and found that 35 companies were reporting on content removals in their transparency reports. However, there was little standardization in how and what companies were reporting, and little granular information was provided, particularly related to Terms of Service based removals and the use automation. If you’d like to take a closer look at the report, we have included it in the Supplementary Materials section.

# 18

Currently, there are a number of resources that offer best practices for promoting and protecting free expression online.

These include the Santa Clara Principles on Transparency and Accountability in Content Moderation, which were launched in May 2018 and outline minimum standards tech platforms must meet in order to provide adequate transparency and accountability around their efforts to take down user generated content or suspend accounts that violate their rules in order to ensure meaningful due process and that the enforcement of their content guidelines is fair, unbiased and respectful of users’ free expression. The Principles urge companies to disclose further data around numbers, notice and appeals.

OTI also recently released a Transparency Toolkit which surveyed international and domestic technology and telecommunications companies and found that 35 prominent companies were reporting on content removals in their transparency reports. However, there was little standardization in how and what companies were reporting, and little granular information was provided, particularly related to Terms of Service based removals and the use automation. The Toolkit also offers a set of best practices for reporting on content takedowns.

Furthermore, Ranking Digital Rights’ Corporate Accountability Index has a Freedom of Expression category which seeks evidence that companies demonstrate a strong public commitment to transparency not only in terms of how they respond to government and others’ demands, but also how they determine, communicate, and enforce private rules and commercial practices that affect users’ freedom of expression.

The Electronic Frontier Foundation also recently released a new edition of its Who’s Got Your Back, which focuses on censorship free expression for the first time.

If you’d like to take a closer look at these resources, we have included them in the Supplementary Materials section as well as other related resources from organizations such as the ACLU, BSR and the United Nations.

# 19

Governments also play an important role when it comes to free expression in the digital age. Through legislation and other actions, they can promote free expression or they can undermine it. In the United States, for example, Section 230 of the Communications Decency Act, provides immunity from liability, except for in certain cases, for providers and users of an “interactive computer service” who publish content created by third-party users. As a result, platforms do not face legal pressure to censor user content or otherwise undermine users’ speech. CDA 230 has been widely regarded as a landmark piece of legislation and for enabling the internet to become a space where free expression is permitted. However, strong intermediary liability protections such as these do not exist around the world, and in many countries, they are still being shaped. In India, for example, the government is currently working on a draft set of rules regarding intermediary liability, which have sparked significant controversy for their lack of strong protections, and the consequences these will have for free expression. Such controversial practices have also been adopted in various European countries, where companies are increasingly being pressured to remove content such as hate speech and terror content or face liability fines.

In addition, in some African nations, such as Tanzania, free expression online is being challenged by regulations that require bloggers, online streaming providers, forums and more, to pay a fee and receive approval from the government in order to operate, thus significantly hindering free expression and effectively thwarting the notion of the internet as a free marketplace of ideas. Furthermore, governments across the world, including in the Democratic Republic of Congo, Sri Lanka and the Philippines are increasingly relying on internet and network disruptions as a means of quelling opposition to existing regimes and stifling free expression, particularly around elections.

# 20

As emerging technologies develop, they are increasingly being applied to the management of online free expression. Artificial intelligence is the best example of this. As we previously discussed, many platforms deploy AI systems in order to identify and regulate content at scale. This occurs at various stages of moderation including pre-publication moderation and during active assessment of content on the platform.

AI systems vary in their ability to identify violating content. For example, some models are better trained and more accurate when identifying child sexual abuse materials on the platform. However, the models are not as reliable when identifying instances of content such as hate speech.

In addition, AI tools have resulted in some contentious takedowns. For example, some images such as the famous Vietnam War image of the little girl which carry historical value and weight have been removed by models for violating the nudity policy. In addition, human rights groups who repost images of terrorist atrocities have had their content taken down for violating the policy on terror content, despite their intent being the opposite. This happened for example in Myanmar, where Rohingya activists had their content removed and accounts shut down.

Human moderators can also demonstrate bias and will not always have a full appreciation for the context of posts. In particular, human moderators may lack the necessary language skills or familiarity with the cultural context of many posts. However, many companies are making efforts to hire humans with the necessary language skills and cultural expertise, and including humans in the decision making process can help mitigate the risks of biased or otherwise wrongful takedowns. As a result, companies are continuously pushed to consistently involve human moderators in all content decision making practices as these human moderators can add cultural context and understanding and prevent such mistakes and such instances of curtailed free expression, from taking place.

# 21

In conclusion, as platforms centered on user-generated content have proliferated and been adopted across the globe, they have become essential to global public conversations. However, in an attempt to ensure their platforms remain safe, major technology companies have developed regulations around permissible content. These often infringe on the right to free expression, however, and as a result companies have been pressed for greater transparency on these decision-making processes, especially as automated systems are further deployed. These platforms are in effect privately managing our speech, and this accountability is vital. This was the last webinar in this four-part webinar series on business and human rights in the digital age. In this webinar series we have discussed key concepts, terminologies and principles related to the business and human rights field, and the relationship between human rights and the digital age, with a focus on the UN Guiding Principles on Business and Human Rights. In addition, we took a closer look into how the concepts of privacy and free expression in the digital age interface with companies and governments, and the related best practices and frameworks for both of these issues.

# 22

Thank you for joining us for this webinar series. If you have any questions, please feel free to reach out to Richard Wingfield and Global Partners Digital at richard@gp-digital.org.