Human Rights Baseline Assessment for Small and Medium Sized Technology Companies:

Assessing risks relating to freedom of expression

January 2020

GLOBAL PARTNERS DIGITAL

OPEN TECHNOLOGY INSTITUTE

Contents

Introduction 01

Acknowledgements 02

Assessing risks to privacy 03

- Terms of Service and Community Standards 03
- Legal Requests for Content Regulation 08
- Domestic Legislation Related to Freedom of Expression 10
- Providing Adequate Notice to Users 13

Introduction

This section of the tool is designed to help you assess how your company's operations and practices may pose risks to the right to freedom of expression. Although you have already answered some high-level questions on this topic, these questions are designed to help you evaluate these risks at a deeper level. Please note that some questions may be similar to questions you answered in Part 1 of this assessment. This is so you have all the relevant information on a given topic in one place. The topics covered by this section of the tool include Terms of Service and Community Standards, content regulation, domestic legislation related to freedom of expression, transparency and accountability mechanisms, and content policy development.

Acknowledgements

We would like to thank Michaela Lee from Business for Social Responsibility and Evelyn Asward from the University of Oklahoma College of Law for their support in reviewing and creating this human rights impact assessment tool.

This tool was developed by New America's Open Technology Institute and Ranking Digital Rights as part of a consortium of organizations working on promoting business and human rights in the tech sector to advance internet freedom.

Assessing risks to freedom of expression

Terms of Service and Community Standards

Generally, technology companies that enable users to create and share user-generated content on their platforms should have Terms of Service (also known as Community Standards or Community Guidelines). These policies define standards and norms regarding acceptable speech on a platform. Typically, these content guidelines are stricter than legal restrictions on speech. For example, a company's Terms of Service may prohibit forms of speech such as bullying or graphic violence, which are not considered illegal. In this way, internet platforms assume a significant amount of power and control and become gatekeepers of online speech. It is therefore vital that internet platforms construct and enforce these speech-related policies in a rights-respecting manner. Failure to do so can result in overbroad censorship, infringements on users' freedom of expression, and the stifling of free flows of information.

Today, internet platforms large and small typically rely on automated tools as well as a large body of human content moderators to enforce their content policies. Automated tools that are deployed for content moderation purposes are imperfect, and often not able to adequately parse subjective human speech. As a result, the use of automated tools can cause the erroneous removal of content or accounts, and can also cause over-censorship.

For human moderators, the content moderation process requires frequent engagement with sensitive and often graphic forms of content, including child pornography, graphic violence, and hate speech. As a result, the enforcement of content policies can touch on and pose a risk to a range of other human rights, including ones related to labor. This assessment does not touch on these other human rights risks in depth. However, if your company employs human content moderators, we strongly encourage you to take a step further to ensure that your company's treatment of these staff members is rights-respecting as well.

The following questions are designed to help you assess the current state of your policies and practices related to Terms of Service implementation and enforcement.

1. **If your company hosts or otherwise facilitates online content, does your company have established policies or rules in place that define unacceptable content or that outline what types of content will be removed?**

2. **If yes, are these policies or rules publicly available?**

a) Where on your website or platform are these policies or rules listed?

b) Do you specifically highlight or communicate changes to these policies or rules to users (e.g. through blog posts or email updates)?

c) Do you offer a public archive where users can see how these policies and rules have changed over time and reference old policies?

d) Do these policies or rules include examples of permissible and impermissible content?

3. **Does your company have a team responsible for managing Terms of Service- and Community Standards-related content removal or restriction? If so, describe how many people, where they are situated in your organization, and whether and to what extent they have had human rights training with respect to international freedom of expression standards.**

4. **Does your company use human moderators to review, remove, and/or restrict content (in addition to or instead of algorithmic or other automated processes)?**

a) How many moderators does your company employ?

b) Are these moderators direct employees or contractors?

c) Please list all countries in which these moderators are based.

d) Please describe how these moderators are trained, including how long their training is, what their training covers, and whether they receive recurring training.

5. **Does your company publicly disclose the full set of content guidelines that are used by your content moderators when making decisions on content removal based on your company's Terms of Service or Community Standards?**

6. **Does your company use automated tools to review, remove, restrict, and generally moderate content?**

a) What categories of content are these automated tools used for (e.g. hate speech, nudity, copyright, child pornography, etc.)

b) Are these tools used to proactively flag and identify content?

c) Are these tools used to remove content?

d) What role do human moderators play in monitoring and reviewing content flagged and/or removed by automated tools?

7. Do you publicly disclose how automated tools are used during the content moderation process? If yes, please detail how.

8. Which personnel in your company are engaged in the creation of new content policies regarding what content is permissible on the platform? To what extent have they had human rights training with respect to international freedom of expression standards?

9. Which personnel in your company are engaged in the implementation of new content policies?

10. Please describe the process your company goes through to create and implement new policies regarding what content is permissible on your site.

11. **Do you consult external stakeholders when creating new content policies? If yes, please select what types of experts or organizations are consulted. If there are other experts or organizations that are not covered by the suggested categories, please list them in the "other" category.**

- Domestic government ☐
- Foreign government ☐
- Domestic civil society ☐
- International civil society ☐
- National human rights institutions ☐
- International human rights institutions ☐
- Multi-stakeholder initiatives ☐
- Academics ☐
- Consultants ☐
- Business partners: ☐
- Other: ☐

Legal Requests for Content Regulation

Data collection directly impacts user privacy, and users
Over the past decade we have seen just how much power online speech has had around the world—both for positive and negative purposes. At the same time, many internet platforms have begun receiving legal requests to remove or restrict content. Legal requests can include requests from law enforcement or other government agencies, requests asserting copyright or trademark infringement, and requests related to regional laws such as the "Right to be Forgotten"). When a company operates in a particular jurisdiction it is required to comply with local laws. This includes speech-related laws that may prohibit certain forms of content. For example, in Germany it is illegal to deny the Holocaust. As a result, internet platforms that operate in Germany must restrict (also known as geoblock) any Holocaust denial content in the nation.

However, not all legal requests for content removal or restriction are based on clearly-defined laws. In fact, many governments around the world have demanded that companies remove content that they find contentious or unfavorable, despite the fact that this content is not illegal or prohibited by the company's own Terms of Service. In situations like these, companies are often torn between upholding the freedom of expression of their users while also maintaining good relationships with the government and ensuring they can continue operating in this jurisdiction. In order to ensure that the right to freedom of expression is respected in such scenarios, companies should establish and enforce clear policies for legal requests for content regulation.

The following questions are designed to help you assess the current state of your policies and practices related to legal requests for content removal and restriction.

1. **Does your company have a team responsible for managing legal and government requests for content removal or restriction? If so, describe how many people, where they are situated in your organization, and whether and to what extent they have had human rights training with respect to international freedom of expression standards.**

2. **Does your company have established policies or processes to guide how your company responds to legal requests for content removal or restriction?**

a) Please provide an outline of any relevant policies or processes.

b) Were these policies or processes created in consultation with legal counsel?

c) Were these policies or processes created in consultation with outside stakeholders, including civil society?

d) Are these policies publicly available?

e) Where on your website or platform are these policies or rules listed?

f) Do you specifically highlight or communicate changes to these policies or rules to users (e.g. through blog posts or email updates)?

3. When legal requests are applicable to a particular jurisdiction or country, do you completely remove the content from the platform or geoblock or restrict it based on location?

4. Under what circumstances, if any, have you challenged legal requests to remove or restrict content? For example, have you challenged any such requests on the grounds that they are not clear, they are not legal under the relevant domestic laws of the country concerned, that they are not legal given that country's international human rights law obligations, that they are not appropriate or in line with your policies, or on any other ground?

Domestic Legislation Related to Freedom of Expression

As your company grows and expands operations into different countries, it is important that you are aware of the different speech-related and intermediary liability-related legal frameworks that may impact your platform. Every country has different restrictions on speech and different provisions related to intermediary liability. In some jurisdictions, companies that fail to comply with such provisions can face fines or a ban on operations.

The following questions are designed to help you assess and understand the current speech-related and intermediary liability-related legal frameworks in the countries in which you currently operate. We recommend referring to Stanford University's Center for Internet and Society's World Intermediary Liability Map, Columbia University's Global Freedom of Expression Project, and the Global Network Initiative's Country Legal Frameworks Resource as a starting point.

1. **Are you aware of established protections to limit the liability of companies that host or otherwise facilitate user-generated content (often termed protections for "intermediary liability") in any of the countries in which you operate?**

a) If yes, please list any relevant laws and their scope.

b) Do these protections have any exceptions?

2. **Are you aware of any of the countries in which you operate criminalizing or otherwise banning any form of speech or content that users might generate or share? If yes, please describe.**

3. **Are you aware of any laws that mandate that your company regulate certain types of content in any of the countries in which you operate?**

a) If yes, please list any relevant laws and their scope (including what specific types of content they focus on, such as disinformation, terror propaganda, hate speech, blasphemy, etc.).

b) Do any of these laws include provisions mandating how quickly content has to be removed?

If yes:

What are the time requirements, and are there fines associated with failing to meet these time requirements?

Does your company currently have the capacity to moderate content at the pace required by these laws?

4. In the event that your company is unable to moderate content as per the timelines outlined in relevant laws, does your company have the financial resources to pay these fines?

5. How does your company comply with these laws while respecting human rights?

6. Does your company feel that time pressure and potential fines cause you to err on the side of censorship when regulating content?

Providing Adequate Notice to Users

Providing adequate notice to users during the content moderation process is a vital transparency and accountability mechanism for any technology platform. In particular, companies should provide detailed notices to users who have had their content removed or restricted. Additionally, companies should provide detailed notices and updates to users who have flagged content for removal and restriction. This ensures that both sets of users have adequate insight into the ongoing content moderation process and understand how and why speech on the platform is being regulated.

The following questions are designed to help you assess and understand the current state of your policies and practices related to providing adequate notice to users during the content moderation process.

*

1. Does your company provide a notice to users who have had their content or accounts removed, restricted, or suspended?

a) Are notices available in a durable form that is accessible even if a user's account is suspended or terminated (e.g. through an email)?

b) Do notices include an excerpt of the content in question?

c) Do notices include the specific clause of your Community Standards that the content was found to violate?

d) Do notices include how the content was detected and removed (e.g. flagged by another user, by a government or law enforcement agency, by a trusted flagger, by an automated tool, etc.)?

e) Do notices explain how a user can appeal the decision?

f) Do notices include any other information?

2. Does your company provide a notice to users who have flagged content or accounts for removal, restriction, or suspension regarding the company's decision on how to treat the flagged content or account?

a) Do these users have a durable log of all reports they have submitted and their outcomes?

b) Do these users receive specific notifications or updates when a report has been resolved? If yes, what information do these notifications or updates include?

GLOBAL PARTNERS DIGITAL

Second Home
68 Hanbury St
London E1 5JL

+44 203 818 3258