CENTER FOR DEMOCRACY & TECHNOLOGY

4 March, 2022

To: OHCHR

Subject: Input from the Center for Democracy & Technology on the application of the United Nations Guiding Principles on Business and Human Rights to technology companies

The Center for Democracy & Technology (CDT) respectfully submits these comments in response to OHCHR's call for input to the High Commissioner report on the practical application of the United Nations Guiding Principles on Business and Human Rights (UNGPs) to the activities of technology companies. CDT is a nonprofit organization based in Washington, D.C and Brussels that is dedicated to promoting democratic values and human rights in the digital age.

I. Technology companies' services and their regulation must be examined through a human rights lens.

Because of the vast impact that technology companies' services and their regulation can have on users' human rights, they must be considered through a human rights framework. This is especially true for intermediaries that host user-generated content.

Intermediaries' services can "implicate rights to privacy, religious freedom and belief, opinion and expression, assembly and association, and public participation, among others."[1] In some cases, they promote human rights, particularly users' rights of freedom of expression and association.[2] However, these services—and the ways users use them—can also threaten human rights by, for example, undermining privacy through intermediaries' collection and use of personal data[3] or enabling dissemination of misinformation[4] and hate speech.[5]

Intermediaries need better guidance on how to evaluate and mitigate their services' human rights risks, including how to weigh trade-offs regarding resource constraints in content moderation. For example, many intermediaries have devoted insufficient resources to content moderation in non-English languages, contributing to the spread of hate speech and misinformation in non-English speaking

[1] A/HRC/38/35 ¶ 5.
[2] A/HRC/47/52 ¶ ¶ 11-12
[3] *Id.* at IV.A.
[4] *Id.* at IV.C.
[5] *Id.* at IV.D.

1401 K Street NW, Suite 200 Washington, DC 20005

countries.[6] They have also deployed automated content moderation tools to help manage vast amounts of user-generated content, but these tools are often biased against marginalized groups.[7]

As the UNGPs recognize, governments also must ensure that businesses respect human rights. Yet government regulation of and interactions with intermediaries often lead to human rights violations. For example, government demands for users' data can interfere with users' rights to freedom of expression, association, and privacy, with especially detrimental impacts on dissidents, activists, whistleblowers, and journalists.[8] And state demands that intermediaries use their private content policies to remove lawful content allows governments to censor speech without adherence to legal process and often with little to no transparency, contrary to international human rights protections for freedom of expression.[9]

II. Third parties have applied the UNGPs to technology companies.

A. Identification and assessment of human rights impacts

UNGP Principle 18 provides that businesses should "identify and assess" actual or potential adverse human rights impacts of their activities or relationships, drawing on internal and external human rights expertise and meaningful consultation with potentially affected groups and other stakeholders.[10] Organizations like the Global Network Initiative and Ranking Digital Rights have been instrumental in assessing companies' compliance with human rights principles, including the UNGPs. GNI and RDR assessments—which focus on areas in which guidance on implementation is limited—can provide a model for the UN's attempts to operationalize the UNGPs as applied to technology companies.

GNI is a multistakeholder collaboration between ICT companies, civil society organizations, academics, and investors. Based on international human rights standards, the GNI Principles and Implementation Guidelines provide member companies with guidance on how to protect their users' rights to privacy and freedom of expression from abuse by governments.[11] GNI member companies commit to implementing the Principles on freedom of expression and privacy when faced with government pressure to hand over user data, remove content, or restrict communications.[12] GNI member companies commit to publicly disclose their commitments to the Principles and are periodically

---

[6] Nik Popli, *The 5 Most Important Revelations From the 'Facebook Papers,'* TIME (Oct. 26, 2021).
[7] Nicolas Kayser-Bril, *Automated moderation tool from Google rates People of Color and gays as "toxic,"* AlgorithmWatch (May 19, 2020); Shirin Ghaffary, *The algorithms that detect hate speech online are biased against black people*, Vox (Aug. 15, 2019).
[8] *See* A/HRC/23/40.
[9] *See Written Comments of the Ctr. for Democracy & Tech.*, In re Foreign Censorship Part 1: Policies and Practices Affecting U.S. Businesses, Investigation No. 332-585 (United States International Trade Commission) at Sec. III (hereinafter "CDT USITC Comments").
[10] HR/PUB/11/04 at 19.
[11] *GNI Principles on Freedom of Expression & Privacy*, GNI (May 2017); *Implementation Guidelines for the Principles on Freedom of Expression & Privacy*, GNI (Feb. 2017).
[12] *Global Network Initiative*, GNI (last visited Feb. 28, 2022).

independently assessed on their progress in implementing the Principles.[13] GNI also holds regular multistakeholder learning sessions and policy dialogues to further develop its guidance around protecting human rights in the ICT sector.

RDR is a non-profit research organization that evaluates 26 of the world's most powerful digital platforms and telecommunications companies on indicators directly aligned with human rights principles and, specifically, the UNGPs. The indicators drive adoption of the UNGPs around key areas including human rights due diligence, governance, transparency and accountability around systems that affect a user's fundamental rights to privacy, freedom of expression, and non-discrimination. The indicators under governance, particularly G3 - Internal implementation[14], G4 - Human rights due diligence[15], and G5 - Stakeholder engagement and accountability[16], expand upon and offer tangible measures to implement Principle 18.[17]

While participation in GNI and RDR are important steps some intermediaries have taken to implement Principle 18 and can provide a model for the operationalization of the UNGPs, very few hosts of user-generated content have conducted and published assessments of the impact their products, services, and technologies have on user rights.[18] In particular, more remains to be done around applying human rights due diligence to content moderation activities.

B. Accountability and remedy

UNGP Principle 29 provides that businesses should offer "effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted."[19] The Santa Clara Principles on Transparency and Accountability in Content Moderation can serve as a useful reference for understanding the features of operational-level grievance mechanisms that would promote users' rights.[20]

The Santa Clara Principles are a set of recommendations from academics, advocates, and civil society, intended to promote meaningful due process to impacted speakers and better ensure that the enforcement of intermediaries' content guidelines is fair, unbiased, proportional, and respectful of users' rights.[21] The newest version of the Santa Clara Principles includes both Foundational Principles and Operational Principles explicitly grounded in international human rights standards.

---

[13] *Company Assessments*, GNI (last visited Feb. 28, 2022).
[14] *2020 Indicators*, Ranking Digital Rights at G3 (last visited Feb. 28, 2022) (hereinafter "RDR Indicators").
[15] *Id.* at G4a.
[16] *Id.* at G5.
[17] HR/PUB/11/04 at 19.
[18] RDR Indicators, *supra* n.14 at G4b.
[19] HR/PUB/11/04 at 31.
[20] *The Santa Clara Principles on Transparency & Accountability in Content Moderation* (last visited Feb. 28, 2022) (hereinafter "Santa Clara Principles).
[21] *Id.*

The Operational Principles, in particular, can provide guidance on how to operationalize the UNGPs to ensure effective grievance mechanisms by larger, more mature intermediaries. They recommend that intermediaries provide notice to users whose content or accounts are removed, suspended, or otherwise acted upon and set forth minimum components and standards for notices.[22] The Operational Principles also recommend that intermediaries provide a meaningful opportunity for timely appeal of decisions to take action against a user's content or account and set forth minimum standards for appeals processes.[23]

The newest version of the Santa Clara Principles also offers guidance on governments' and state actors' compliance with their obligations under international law when regulating or interacting with intermediaries.[24] Two recommendations concerning removing barriers to company transparency and promoting government transparency can inform implementation of UNGP Chapter I.

III. The UN should clarify the application of UNGPs to government regulation of intermediaries, describe the role of transparency in implementing the UNGPs, and explain the application of human rights due diligence to content moderation.

A. State regulation of intermediaries must create an environment that enables—rather than constrains—business respect for human rights.

UNGP Principle 3(b) provides that states must ensure that laws and policies governing business enterprises "do not constrain but enable business respect for human rights." The UNGPs require governments to promote business respect for *all* human rights; states and companies should not prioritize some human rights over others, and regulation should not incentivize companies to implement only a minimum standard of human rights compliance.[25]

Intermediaries are a focal point for state efforts to control online expression and engage in surveillance. Governments have targeted intermediaries in ways that threaten users' free expression and privacy rights by, for example, direct or indirect requirements that intermediaries remove content pursuant to non-judicial notices,[26] intermediary liability regimes that lead to overbroad takedowns of speech,[27] data

[22] *Id.*
[23] *Id.*
[24] *Id.*
[25] For example, Article 26 of the European Commission's proposed Digital Services Act would require Very Large Online Platforms to engage in risk assessments concerning negative effects of their services on only certain fundamental rights, *i.e.*, those in Articles 7, 11, 21 and 24 of the Charter. CDT has urged that the DSA instead require online platforms to "focus their efforts on assessing the human rights impacts of their products and services and embed respect for human rights across the value chain," consistent with the UNGPs. Asha Allen, *European Parliament IMCO Committee Adopts DSA Report: Significant steps Forward, Leaps Still to Be Made*, Ctr. for Democracy & Tech. (Dec. 14, 2021).
[26] *See* CDT USITC Comments, *supra* n.9 at Sec. I.
[27] *See Shielding the Messengers: Protecting Platforms For Expression and Innovation*, Ctr. for Democracy & Tech. at Sec. IV (v.2 updated Dec. 2012) ("When intermediaries are liable for or obligated to police content created by

and personnel localization requirements,[28] and must-carry obligations. These obligations are incompatible with the UNGPs. The UN should consult with civil society and member states to develop a tool for benchmarking regulatory proposals against the UNGPs.

B. The UN should clarify the essential role of transparency in promoting human rights under the UNGPs.

Transparency plays a critical role in the promotion of human rights under the UNGPs. UNGP Principles 20 and 21 require businesses to track the effectiveness of their responses to adverse human rights impacts and to be prepared to communicate externally about how they address their human rights impacts, particularly when concerns are raised by or on behalf of affected stakeholders.

The UN should clarify that intermediaries must ensure that this transparency is meaningful to different stakeholders, including users, civil society organizations, and policymakers, each of which will find different kinds of transparency useful. For example, transparency reports may be informative to civil society and policymakers seeking to understand and respond to larger trends in how intermediaries respond to government demands for user data or content removals or enforce their own content policies. An individual user, however, may find more useful the transparency offered by a direct notification when a government demands their data or the intermediary takes action on their content or account under its content policies. Intermediaries should offer a variety of forms of transparency—such as transparency reports, user notifications, access to data by independent researchers, and third party audits and assessments—to help speak meaningfully to a wide variety of stakeholders.

C. The UN should explain the application of human rights due diligence to content moderation.

UNGP Principle 17 defines the parameters for human rights due diligence, with 17(a) focusing on "cover[ing] adverse human rights impacts that [a] business enterprise may cause or contribute to through its own activities, which may be directly linked to its operations, products or services…" UNGP Principles 18 through 21 elaborate on what a human rights due diligence process should look like.

The UN should clarify that, under Principle 17, technology companies should conduct human rights due diligence of content moderation processes. The increasing amount of user-generated content online has caused an expansion in research and investment in automated content analysis tools.[29] Yet due to the diversity of content across intermediaries and of linguistic and cultural expression, automated content analysis tools have limited ability to analyze content across domains.[30] They also create the risk

---

others, they will carefully screen and limit user activity in an effort to protect themselves. In doing so, they are likely to overcompensate, blocking even some lawful content out of an abundance of caution.")

[28] See CDT USITC Comments, *supra* n.9 at Sec. IV.

[29] Dhanaraj Thakur & Emma Llanso, *Do You See What I See: Capabilities and Limits of Automated Multimedia Content Analysis*, Ctr. for Democracy & Tech. (May 2021) (hereinafter "Thakur & Llanso").

[30] Natasha Duarte & Emma Llanso, *Mixed Messages: The Limits of Automated Social Media Content Analysis*, Ctr. for Democracy & Tech. (Nov. 2017).

of overbroad removal of lawful content.[31] Human rights due diligence must assess and mitigate free expression-related risks across a breadth of processes deployed for content moderation.

More explanation should be provided to ensure that due diligence efforts assessing content moderation analyze how to balance the risks posed to human rights by illegal content and "lawful but awful" content against the risks to the right to free expression posed by overbroad or indiscriminate content moderation practices. In addition, special consideration should be taken to assess the human rights impact of state demands for content removals under platform content policies. Before they enter a market, intermediaries should consider the risk of state content removal demands that will infringe on users' human rights.[32]

Automated content moderation tools also pose risks due to the limitations of the technology. Facebook transparency reporting shows that the majority of takedown decisions the company reversed came from its automated content analysis tools, which were prone to over-removing content, harming users' right to free expression.[33] CDT's study of automated content analysis tools suggest that they lack robustness and fail to identify abusive content, label trustworthy and important information as spam,[34] and are built on incomplete or poor data sets which perpetuate existing social biases.[35]

Assessing the impact of content moderation tools on fundamental rights requires flexibility in the application of human rights due diligence to take into account the diversity of content on intermediaries' platforms and new technologies deployed for automated content moderation. Greater clarity on the application of UNGPs to state regulation of intermediaries, advancing the role of transparency in the UNGPs, and explaining the application of human rights due diligence to content moderation are steps that will enable the greater adoption and implementation of the UNGPs by both states and technology companies.

---

[31] *Id.*

[32] In 2021, global civil society groups and local NGOs in the Middle East urged Google to conduct human rights due diligence before entering a new partnership with Saudi Aramco to build a data center, asserting that a data center in Saudi Arabia would harm citizens' rights to privacy and freedom of expression. *Saudi Arabia: Google Should Halt "Cloud Region"*, Human Rights Watch (May 26, 2021).

[33] *Community Standards Enforcement Report*, Facebook (last visited Feb. 28, 2022); Louise Matsakis & Paris Martineau, *Coronavirus Disrupts Social Media's First Line of Defense*, Wired (Mar. 18, 2020).

[34] *See* Rebecca Heilweil, *Facebook is flagging some coronavirus news posts as spam*, Vox (Mar. 17, 2020).

[35] Thakur & Llanso, *supra* n.29.