PAPER

BSR

A Human Rights-Based Approach to Content Governance

MARCH 2021

INTRODUCTION

The question of how social media platforms can respect the freedom of expression rights of users while also protecting others from harm is one of the most pressing challenges of our time. Taking a human rights-based approach to this challenge will help ensure alignment with internationally agreed norms and consistency across borders—but what does a human rights-based approach to content mean in practice?

We believe that the elements described in this paper constitute the foundations of a human rights-based approach to content governance. We have arrived at these elements by combining the UN Guiding Principles on Business and Human Rights (UNGPs) with a consideration of the various human rights principles, standards, and methodologies upon which the UNGPs were built.

We believe that a human rights-based approach to content governance can be segmented into four parts:

1. **Content policy**—statements about what content is and is not allowed on a social media platform, as well as about the visibility of content.
2. **Content policy implementation**—how content decisions are executed in practice.
3. **Product development**—how new features, services, and functionalities are introduced and evolve.
4. **Tracking and transparency**—how the outcomes and effectiveness of a human rights-based approach is measured and communicated.

Further, we believe that a special focus on engagement with affected stakeholders and the needs of vulnerable groups is essential across all four parts.

There are two important features to highlight about these four parts taken in combination.

First, these four parts constitute a robust framework of ongoing human rights due diligence that enable content decisions to be made thoughtfully, deliberately, and grounded upon rights-based analysis, rather than “on the go” or according to the whim of the moment. They emphasize that process matters as much as the decision itself—and that while different companies may reach different conclusions, content decisions should be intellectually consistent, defensible on human rights grounds, and conveyed transparently.

Second, these four parts encompass more than just what content is and is not allowed on a platform—our approach assumes that international human rights law provides an overall framework for decision-making and action, not simply a “copy and paste” set of content rules for companies to follow.

The paper has been written to inform discussion and debate, and we welcome comments to amend, improve, and build on this approach. A summary of this human rights-based approach structured according to some of the most relevant UNGP principles is provided in the annex.

This paper draws upon and is inspired by the work of the current and previous UN Special Rapporteurs on the promotion and protection of the right to freedom of opinion and expression; organizations such as Center for Democracy and Technology, Ranking Digital Rights, AccessNow, Article 19, Global Partners Digital, Global Network Initiative, Just Peace Labs, and the Dangerous Speech Project; initiatives such as the Santa Clara Principles; and experts such as Rebecca MacKinnon, Jillian York, Emma Llanso, Evelyn Aswad, Evelyn Douek, Jenny Domino, Jonathan Zittrain, Alex Stamos, and Barrie Sander. The perspectives shared in this paper are also informed by our engagements with social media companies in the BSR membership.

This paper does not attempt to analyze the merits of specific content decisions or repeat complex content policy dilemmas that are being well debated elsewhere, such as whether the speech of government officials should be held to a higher or lower standard than other users. Rather, this paper is primarily focused on describing in a clear and succinct manner how a UNGPs-based framework can be applied to the overall system of content governance at social media companies.

This paper mainly focuses on the role of social media companies that provide platforms for user-generated content and facilitate networks for sharing content; it does not consider the role of companies in other industry segments, such as app stores, advertising intermediaries, content delivery networks, or internet service providers. We recognize that the relative roles and responsibilities of companies across the entire technology industry value chain merits deeper consideration, and we intend to address this in due course.

A ROLE FOR GOVERNMENTS?

Before describing how companies can take a human rights-based approach to content governance, it is important to address the question of whether companies should be making content decisions in the first place. The recent controversy around user @realDonaldTrump has led many to argue that decisions about content—especially content shared by elected officials—should be a role for governments, not companies, and that therefore any discussion around human rights-based approaches deployed by companies would be moot.

This paper is primarily focused on practical approaches to content governance that can be taken by companies, rather than different public policy and regulatory options. However, there are three reasons why we believe companies should play a role and why therefore setting out a company-based human rights-based approach to content governance remains essential.

First, many of the most significant public policy proposals on content governance themselves envision a very important role for companies. For example, the UK Online Harms White Paper proposes a statutory duty of care on companies to take responsibility for harm associated with content on their platforms, the EU Digital Services Act proposes new responsibilities for online platforms to prevent misuse, and various proposals to reform US Section 230 (such as this document from the NYU Stern Center for Business and Human Rights) similarly envision a responsibility for companies to moderate content. We believe it is essential that companies adopting this responsibility deploy approaches based on international human rights standards when doing so, and in that context, we hope this paper provides useful insights into what policy makers and regulators could reasonably expect of social media companies—be that voluntary, regulated, or a mix of the two.

Second, these public policy proposals relate to specific jurisdictions, whereas the internet is global. Human rights-based approaches enable consistent approaches to be taken across international borders, including jurisdictions with no regulations whatsoever and jurisdictions where laws and regulations conflict with international rights standards. Indeed, a global approach based upon international human rights standards will be an essential counterweight against governments making bad policy, and it will provide a strong foundation from which companies, civil society organizations, and others can push back against the actions of governments seeking to suppress freedom of expression and other rights.

Third, the UNGPs clearly state that companies have a responsibility to address the adverse human rights impacts with which they are involved, including the responsibility to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products, or services.[1] User-generated content clearly has a connection to adverse human rights impacts, and therefore a human rights-based approach to content governance is essential to meet the responsibility companies have to address this connection.

PART ONE: CONTENT POLICY

Every social media company will establish unique policies for what content is and is not allowed on their platforms. Different companies may choose to be more or less permissive on different issues or in different contexts, and companies may also choose to have policies that alter the visibility and distribution of content, such as through warning labels, interstitials, or downranking.

Moreover, different content standards may be suitable for different types of service. Private messaging services (such as WhatsApp or Signal) may be more permissive than content-sharing platforms (such as Facebook, Twitter, or YouTube), and different standards may be needed according to whether the content being shared is predominantly text, images, or video, or for different parts of a single platform, such profile pages, the content feed, and messaging.

In addition, while this paper focusses mainly on social media platforms, it is important to consider the different content policies that may be appropriate for other layers in "the stack," such as app stores, cloud services, content delivery networks, domain registrars, and internet service providers.[2]

We believe that international human rights standards and instruments can shape what is and is not allowed on a social media platform in four main ways.

- **Content policy should be founded upon human rights standards and instruments.** Principle 16 of the UNGPs states that, as the basis for embedding their responsibility to respect human rights, companies should express their commitment to human rights through a statement of policy. In the social media context, we interpret this to mean that the "rules," "community standards," or "community guidelines" that determine what users can and cannot post on a social media platform should reference the International Bill of Human Rights, consisting of the Universal Declaration of Human Rights and the main instruments through which it has been codified, the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights. Other relevant international human rights instruments, such as the Convention on

[1] See Principles 11 and 13 of the UN Guiding Principles on Business and Human Rights.

[2] For further analysis about "the tech stack" and content moderation, see Navigating the Tech Stack: When, Where and How Should We Moderate Content? by Joan Donovan, and A Framework for Moderation, by Ben Thompson.

the Elimination of Discrimination against Women, the Convention on the Elimination of All Forms of Racial Discrimination, and the Convention on the Rights of the Child should also be used.

- **Content policy should encompass all human rights.** The highly diverse mix of content posted on social media platforms mean that any human rights contained in international human rights instruments can be impacted by user-generated content, not simply freedom of expression and personal safety. The platform policies used by social media companies are constantly evolving as new cases and threats arise, but we believe there is significant benefit to preempting some of this evolution by making sure that platform policies are inclusive of ***all human rights*** in the first place. Specifically, we believe that companies can undertake a systematic gap analysis between their platform policies and international human rights instruments to make sure that less obvious rights in a social media context are appropriately covered. Human rights, such as the right to a fair trial, freedom of movement, right to enjoy the benefits of scientific progress and its applications, indigenous or protected cultural heritage, and labor standards, are often missing in content policies today.

- **Content policy should be informed by stakeholder engagement.** Rather than being created in a vacuum, content policy should be informed by the perspectives of affected stakeholders—both users of the platform and non-users who may be impacted by its content—and by relevant experts. Facebook's stakeholder engagement principles and publication of the minutes from meetings where content policy is made are an example of this. Similarly, Twitter solicited input from more than 6,500 individuals worldwide to develop its policy on synthetic and manipulated media and provided users with the opportunity to comment on an initial draft of the policy.

- **Content policy should distinguish between organic and paid speech.** Specifically, two essential features of paid speech—that it can be amplified and targeted—present different risks and increase both the connection and potential causal relationship between the social media company and human rights harm. Organic speech and paid speech have very different characteristics and status, and more restrictive content policies should be applied to paid speech—indeed, greater restrictions are already common today in areas such as political speech, misinformation, and public health.

On content policy, there is a counterargument that the inclusion of such "complex" human rights terminology would reduce the effectiveness of these platform policies by making them less accessible to billions of users. We believe this can be solved by using stand-alone "policy wonk" text that is published separately from—but is very clearly connected to—more accessible content policy language designed for a wider audience. After all, human rights belong to all of us by virtue of being alive and have been translated into over 500 languages. Furthermore, regardless of public communications, key human rights principles can play an essential role in the implementation of content policy, which we turn to next.

PART TWO: CONTENT POLICY IMPLEMENTATION

Social media companies face monumental challenges when implementing content policies in practice. These challenges may be summarized as scale (i.e., the sheer volume of content to be reviewed), speed (i.e., the rapid nature in which decision about harmful content needs to be made), and context (i.e., different nuances and interpretations across variables such as language, culture, politics, and conflict). There are also gaps in the underlying international human rights framework. Even the highest functional and most well-resourced social media company will face difficult dilemmas and make mistakes; the test of a human rights-based approach is addressing dilemmas thoughtfully, minimizing errors, and taking the appropriate response when mistakes happen.

- **Policy implementation should be informed by engagement with affected stakeholders and experts that understand the relevant context.** Principle 18 of the UNGPs emphasizes meaningful consultation with potentially affected groups and other relevant stakeholders when identifying and assessing human rights impacts, while Principle 20 emphasizes feedback from both internal and external sources, including affected stakeholders, when tracking the effectiveness of policy implementation. In the case of social media companies, this means structured engagement with those with real lived experiences who can help the company interpret the relevant context, understand the link between online content and offline harm, and make informed decisions that are less likely to result in adverse human rights impacts. This consideration of context (and not simply the content itself) was a feature of decisions made relating not just to user @realDonaldTrump's posts, but also during recent elections in Myanmar too.

- **Content policy implementation should be based on international human rights principles.** Rather than just providing a long list of restricted content, social media companies have found it helpful to set out the values and principles that justify these restrictions. We believe that four key principles based on international human rights law (see UN General Comment 34) can usefully shape and inform the implementation of content policy—namely, that decisions to restrict freedom of expression should take into account whether they are: necessary (i.e., the same goal cannot be achieved by other means), proportionate (i.e., restrictions are not overbroad and are the least intrusive to achieve the legitimate purpose), legitimate (i.e., the precise nature of the threat to human rights is clear), and nondiscriminatory (i.e., restrictions are implemented in a nondiscriminatory manner). International human rights law does not provide a "copy and paste list" of what should or should not be allowed; it does provide principles that can inform challenging decisions, especially when two rights are in conflict and neither can be achieved in its entirety. The Facebook's Oversight Board's first five case decisions provided analysis against each of these principles, demonstrating how they can be applied in practice.

- **The cumulative impact of individual content decisions requires special attention.** The concept of cumulative impacts is the notion that one case taken in isolation may not have significant human rights impacts but, when combined with thousands of similar cases, may result in severe human rights impacts. This can be especially challenging in the case of platform policies: One case alone may not violate them, but a combination of cases may, for example, create a hostile environment for users and lead to human rights violations. Google's White Paper on Information Quality and Content Moderation describes this dynamic well.

- **Companies should seek ways to honor the principles of internationally recognized human rights when faced with conflicting requirements.** At the time of writing, social media platforms are facing complex and uncertain dilemmas in Hong Kong, India, Myanmar, Thailand, Turkey, Vietnam, and elsewhere about whether to restrict the content of some users to preserve platform access for everyone—or whether to challenge government demands for content restrictions, but place the availability or functionality of their platform at risk. Other potential adverse and severe human rights impacts, such as employee safety, are also at stake. A human rights-based approach to content governance requires conscious attention to potentially conflicting rights, deliberate decision-making when difficult tradeoffs are involved or where different courses of action could reasonably be taken, and efforts to collaborate with others to increase company leverage to mitigate potential adverse human rights impacts.[3] Referring to content policies based on international human rights standards and instruments can help when faced with conflicting requirements.

[3] See Principle 19 of the UN Guiding Principles on Business and Human Rights

- **Policy implementation should be prioritized based on severity.** Principle 24 of the UNGPs states that where it is necessary to prioritize actions to address adverse human rights impacts, companies should first seek to prevent and mitigate those that are most severe. The notion of prioritization takes on a special significance in the field of content governance given the sheer volume of content and the impact even a small error rate can have—an error rate of one percent on a million posts a day is 10,000 errors. Prioritizing implementation efforts (including access to remedy) on content most likely to be linked to severe offline harm does not just seem reasonable; it seems essential. This implies that policy enforcement—both what content to limit and what user appeals to pay attention to—needs to be truly global and not limited to those countries receiving the most media attention or political pressure.

- **Policy implementation in conflict-affected areas should receive enhanced and heightened attention.** Guiding Principle 12 of the UNGPs states that the scope of corporate responsibility to respect human rights may be broader in conflict-affected contexts and that businesses should respect the standards of international humanitarian law. Policy implementation in conflict-affected areas should therefore assume that the likelihood and severity of online-to-offline harm is greater than in other areas and that the threshold for hate speech, incitements to violence, and other conflict drivers may be comparatively lower. Guiding Principle 23 of the UNGPs states that operations or business relationships may increase the risk of businesses being complicit in gross human rights abuses committed by other actors and that this necessitates extra care through "heightened" or "enhanced" due diligence. Companies would be wise to integrate conflict sensitivity analysis with human rights due diligence in these settings.

- **Appeals mechanisms should be established and meet minimum effectiveness criteria.** Principle 29 of the UNGPs states that companies should establish or participate in effective operational-level grievance mechanisms for those who may be adversely impacted, while Principle 31 of the UNGPs provides effectiveness criteria for these operational-level grievance mechanisms, such as legitimacy, accessibility, predictability, equitability, and transparency. In the social media context, operational-level grievance mechanisms generally take the form of appeals mechanisms for when users feel that content has been wrongly removed or left in place. We believe that social media companies should systematically assess their content appeals mechanisms against these effectiveness criteria, as BSR did in our human rights assessment of the Facebook Oversight Board. In the context of social media, an assessment against the notice and appeal expectations of the Santa Clara Principles will also be essential.

- **Effective remedy should be provided when mistakes are made.** Principle 22 of the UNGPs states that where companies have caused or contributed to adverse impacts, they should provide for or cooperate in their remediation.[4] Effective remedy restores the victim as much as possible to their position prior to when the harm occurred and can be provided in five different pathways—satisfaction, restitution, guarantee of non-repetition, rehabilitation, and compensation. A human rights-based approach to content implies understanding what these five pathways mean in a social media context, such as:

    a. **Satisfaction** might include an apology, an acknowledgement of the harm, and a public disclosure of the error.
    b. **Restitution** will likely include restoring or removing content.

[4] Defining how "cause," "contribute," and "directly linked" should be interpreted in the context of social media is beyond the scope of this paper.

c. **Guarantees of non-repetition** might include changes to platform policy, issuing new guidance to content moderators, or removing repeat offenders from the platform.
d. **Rehabilitation** might include providing psychological support or social services that help restore the victim to their prior condition, such as paying into funds that seek to support the rehabilitation of victims.
e. **Compensation,** such as money or other benefits, where damage can be economically assessed.

The first three (satisfaction, restitution, and non-repetition) are increasingly common in the social media industry, where users have expanding options to appeal content decisions and with companies increasingly using appeals and reporting channels to refine their content policy implementation strategies.

By contrast, the latter two (rehabilitation, compensation) would be breaking new ground and would require further exploration. Our instinct is that these two pathways only merit consideration in rare cases where impacts such as severe psychological harm, physical security, and bodily integrity have been demonstrated and where it is clear that the social media company has caused or contributed to the harm. A large number of frivolous or distracting cases in search of compensation would take attention away from more severe human rights impacts, and present undue risks to freedom of expression.

PART THREE: PRODUCT DEVELOPMENT

The features, services, and functionalities of social media platforms are constantly evolving, and it is important that human rights are integrated into this process. This meets Principle 17 of the UNGPs that companies should undertake ongoing human rights due diligence, initiated as early as possible in the development of a new activity.

- **Companies should undertake human rights assessments of new or evolving features, services, and functionalities.** Principle 18 of the UNGPs states that, as an initial step in human rights due diligence, companies should identify and assess any actual or potential adverse human rights impacts with which they may be involved and that these assessments should take place prior to any major decisions or changes to operations. In the context of social media, this implies undertaking human rights assessments prior to the introduction of new features, services, and functionalities, as well as other changes that might impact human rights, such as permissions, data visibility, or automated decision making.[5] Since content decisions can impact all human rights—not just freedom of expression—assessments should include all internationally recognized human rights as a reference point. While not focused on social media, BSR's human rights assessment of Google's celebrity recognition tool provides an indication of what form these assessments may take. It is noteworthy that the Data Protection Impact Assessment requirement of the EU's General Data Protection Regulation includes a review against all rights and freedoms contained in the EU Charter of Fundamental Rights—not just privacy—opening up opportunities for synergy with human rights assessments.

- **Companies should assume that new or evolving features, services, and functionalities may have different, unintended, or more severe consequences in higher risk markets.** The impact of social media platforms varies significantly according to context, and companies should be deliberate in identifying and addressing how products may adversely impact human rights in different countries. For example, tools that automate the detection of content policy violations may be less effective, and easier to intentionally misuse, in

[5] As referenced elsewhere in this paper, prioritization—such as focusing on the most material or impactful changes—will be important given the sheer volume of new or evolving features, services, and functionalities.

markets where the capacity of natural language processing is limited; similarly, certain features (e.g., news content) may take on special significance where mainstream media is restricted, censored, or state controlled.

- **Informed consent should be integrated into product choices.** The notion of informed consent is defined by both participation (i.e., the ability to participate in decisions) and empowerment (i.e., the ability to understand both risks and rights when consenting), and it is an important principle in a human rights-based approach. To obtain consent, information upon which a decision is made should be accurate and in a form that is accessible and understandable, including in a language that vulnerable groups such as children, indigenous peoples, and persons with disabilities will fully understand.

PART FOUR: TRACKING AND TRANSPARENCY

Human rights performance improvement and public accountability will be facilitated by measures, benchmarks, and public communications against which outcomes, progress, and effectiveness can be reviewed and judged.

- **Quantitative and qualitative indicators of effectiveness should be established.** Principle 20 of the UNGPs states that to verify whether adverse human rights impacts are being addressed, business enterprises should track the effectiveness of their response. As described in the point below, companies are already publishing significant volumes of data about content removals and appeals, as well as the prevalence of violating content—however, there are opportunities to broaden this to include other management and performance measures, such as employee well-being for content moderators, error rates, and timeliness. Identifying true indicators of effectiveness (i.e., addressing adverse human rights impacts) rather than simply process-based indicators will be important.

- **Companies should publish annual public reports that provide sufficient quantitative and qualitative information to evaluate the adequacy of their approach.** Principle 21 of the UNGPs states that to account for how they address their human rights impacts, companies should be prepared to communicate externally, particularly when concerns are raised by or on behalf of affected stakeholders. In the context of content moderation, this can include disclosing data about the type, volume, and speed of content removal, building a repository of cases to increase understanding of how content policies are enforced, and a narrative describing challenges and future plans. Importantly, the new gender framework for the UNGPs emphasizes the value of gender-disaggregated data as a means of informing more gender-responsive and gender-transformative approaches. Segmentation by other relevant boundaries—such as region, language, or product type (e.g., video, text, image, messaging, app)—may also be useful for report readers.[6]

- **Companies should be transparent about the rationale for important content decisions.** Twitter's blog setting out its rationale for the permanent suspension of @realDonaldTrump provides good direction for what this might look like and an example for others to emulate, and it would be even more beneficial for a succinct assessment of human rights impacts and considerations to be included in these communications. In the Twitter case, international human rights principles on whether democratically elected officials should be treated differently online can be interpreted differently—some make the case that elected officials should be afforded more freedom (because hearing directly from politicians is needed to hold them accountable), while others make the case that elected officials should be afforded less freedom (because the Rabat Plan of

[6] Segmentation of data is a long-term challenge that may require significant reengineering of existing processes. For this reason, it will be important to be clear on what data will be decision-useful for report readers and to consider potential adverse or unintended consequences (such as arming adversaries with increased insight). This is a priority worthy of deeper exploration.

Action six-part threshold test for incitement to hatred includes the "status of the speaker" and standing in the context of the audience to whom the speech is directed). An explanation of company thinking in these and other cases can form a kind of "case law" that makes clearer to the public, researchers, and policy makers how international human rights standards are being applied in practice.[7] Facebook's recent decision to refer its suspension of Donald Trump to its Oversight Board, which in turn emphasized the importance of international human rights standards in its decision making, looks set to establish a new precedent for transparency when the Oversight Board's rationale for its case decision is published later this year.

THROUGHOUT CONTENT GOVERNANCE: VULNERABLE GROUPS

All human beings are born free and equal in dignity and rights. However, while the UNGPs should be implemented in a nondiscriminatory manner, they emphasize that companies should pay particular attention to the rights, needs, and challenges of individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized.

- **Stakeholder-inclusive approaches should be taken throughout all elements of content governance.** A human rights-based approach implies placing the interests of those whose rights are affected at the center, and for this reason, it is essential that platform policies are developed through meaningful consultation. In the social media industry, user "personas" are often created to inform policy and product development, and from a human rights point of view, it is important that these personas are drawn from a range of different vulnerable groups.

- **Companies should pay special attention to human rights defenders.** Human rights defenders are people who, individually or with others, engage in activities and advocacy that contribute to the protection of human rights. By providing a space for expression, association, and organization, online platforms play an especially important role in the work of human rights defenders—but online platforms are also associated with risks, such as when governments often use social media content to threaten, intimidate, and persecute human rights defenders. Being alert to the specific needs of human rights defenders (such as digital security), advocating for civic space, and being responsive when threats happen is an essential part of a human rights-based approach to content governance—and indeed will likely surface risks of relevance to the wider user population, too.

- **Companies should take a structured approach to vulnerability.** The need to pay special attention to vulnerable groups exists throughout a human rights-based approach, from policy formation right through to access to remedy. To achieve this effectively, we believe it is important to recognize that vulnerability is contextual and that someone considered powerful in one context may be vulnerable in a different context. Social media companies often highlight that content should be reviewed considering its context—for example, hate speech in one context may be harmless satire in another context—and the concept of vulnerability has similar characteristics. Rather than simply list vulnerable groups (e.g., women, children, LGBTI users), companies can consider the four dimensions of vulnerability when making content decisions:

    a. **Formal discrimination**—laws or policies that favor one group over another.
    b. **Societal discrimination**—cultural or social practices that marginalize some and favor others.
    c. **Practical discrimination**—marginalization due to life circumstances, such as poverty.

[7] The concept of "case law" in the context of content governance is introduced in the Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression to the UN Human Rights Council in April 2018

d. **Hidden groups**—people who might need to remain hidden and consequently may not speak up for their rights, such as undocumented migrants.

Deploying these four dimensions of vulnerability will assist social media companies with the difficult challenge of adapting and tailoring content policy implementation across a wide variety of jurisdictions, cultures, and contexts. For example, in conflict-affected areas, social media companies can use these four dimensions to identify which communities may be at heightened risk during conflict and prioritize efforts accordingly.

CONCLUSION

Over the past year, a debate has arisen around the size of technology companies and whether they should be broken up. BSR does not take a position on this debate, but we note that the need for content governance at scale represents one "defense" of large social media companies—that all social media companies would face content governance challenges on a massive scale, but only large companies would have the capability to implement content governance effectively. This perspective has generated counterproposals, such as requiring larger companies to provide "content moderation as a service" to smaller companies, who would then be able to make use of the sizeable content moderation infrastructure of their larger competitors.

Regardless of how this debate unfolds over the coming years, we believe that human rights-based approaches should be non-negotiable, whether undertaken inside companies or outsourced "as a service." The UNGPs clearly state that companies have a responsibility to address the adverse human rights impacts with which they are involved, including the responsibility to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products, or services. User-generated content clearly has a connection to adverse human rights impacts; therefore, a human rights-based approach to content governance is essential to meet the responsibility that companies have to address this connection.

However, while the elements we present here constitute the foundations of a human rights-based approach to content governance, they are no more than foundations. Significant work is needed to build out the details of each element as well as the important relationship between action taken by companies and regulatory requirements mandated by governments.

Notably, this paper has not touched upon the human rights impacts of content moderation on those actually doing the work, which can be considerable. We have not explored the role of multi-company and multi-stakeholder initiatives, such as the Global Internet Forum to Counter Terrorism and the Global Network Initiative, and we have not addressed public policy reforms, such as the UK Online Harms Bill, changes to Section 230 in the US, or the proposed EU Digital Services Act, that may support, hinder, or event require a human rights-based approach. We have not attempted to define how the UNGPs concepts of "cause," "contribute," and "directly linked" should be interpreted in the context of user-generated content, and we have not identified specific metrics companies can use to track the effectiveness of their approach. We believe more work is needed to consider the roles and responsibilities of companies in different layers of the "tech stack."

As these limitations indicate, the content governance debate has some distance still to travel, and there are many elements in need of deeper exploration. We look forward to playing our part.

ABOUT THIS PAPER

This paper was written by Dunstan Allison-Hope, Lindsey Andersen, and Joanna Lovatt. BSR wishes to thank all those who commented on drafts and have contributed to our thinking.

BSR publishes occasional papers as a contribution to the understanding of the role of business in society and the trends related to responsible business practices. BSR maintains a policy of not acting as a representative of its membership, nor does it endorse specific policies or standards. The views expressed in this publication are those of its authors and do not necessarily reflect those of BSR members.

ABOUT BSR

BSR is a global nonprofit organization that works with its network of more than 250 member companies and other partners to build a just and sustainable world. From its offices in Asia, Europe, and North America, BSR develops sustainable business strategies and solutions through consulting, research, and cross-sector collaboration. Visit www.bsr.org for more information about BSR's 25 years of leadership in sustainability.

SUGGESTED CITATION

BSR, 2021. "A Human Rights-Based Approach to Content Governance."

ANNEX 1: A SUMMARY OF HOW THE UNGPS CAN BE APPLIED TO THE CONTENT GOVERNANCE CHALLENGE

UNGP Principle	Application to Content Governance
Principle 16: Policy Commitment As the basis for embedding their responsibility to respect human rights, business enterprises should express their commitment to meet this responsibility through a statement of policy, and work towards policy coherence in their wider activities.	• Content policies should be founded upon and reference international human rights standards and instruments. • Content policies should refer to all internationally recognized human rights. • Content policies should be informed by the perspectives of potentially affected stakeholders and relevant experts.
Principle 17: Human Rights Due Diligence To identify, prevent, mitigate, and account for how they address their adverse human rights impacts, companies should carry out ongoing human rights due diligence that is relevant for the nature and context of business activities.	• Content decisions should be made in the context of an overall decision-making framework developed over time, not "on the go." • The implementation of content policies should benefit from ongoing dialogue and consultation with affected stakeholders and experts with informed insight into the relevant decisions and context.
Principle 18: Assessment Companies should identify and assess any actual or potential adverse human rights impacts by drawing upon internal and/or independent external human rights expertise and involving meaningful consultation with potentially affected groups and other relevant stakeholders.	• Companies should undertake human rights assessments of new and evolving features, services, and functionalities. • Content policy implementation should be informed by engagement with affected stakeholders and experts that understand the relevant context.
Principle 19: Appropriate Action To prevent and mitigate adverse human rights impacts, companies should integrate the findings from their impact assessments across relevant internal functions and processes, and take appropriate action.	• Companies should pay attention to the cumulative impact of multiple individual content decisions. • Companies should consider how their policy enforcement actions on a specific topic (such as expressions of hatred) exist in and are connected to a broader system (such as a history of structural racism).
Principle 20: Tracking To verify whether adverse human rights impacts are being addressed, companies should track the effectiveness of their response using (a) qualitative and quantitative indicators and (b)	• Quantitative and qualitative indicators of content policy effectiveness should be established. • Companies should review the impact of content decisions previously taken—

feedback from both internal and external sources, including affected stakeholders.	especially difficult and complex ones—and integrate lessons learned into future plans.
Principle 21: Communications To account for how they address their human rights impacts, companies should communicate externally in a form and frequency that reflects human rights impacts, is accessible to intended audiences, and does not pose risks to stakeholders.	• Companies should publish annual public reports about their content moderation governance, strategy, risks, and metrics. • Companies should be transparent about the rationale for important content decisions and should create a compilation of examples that in combination forms a kind of "case law."
Principle 22: Remedy Where companies have caused or contributed to adverse impacts, they should provide for or cooperate in their remediation through legitimate processes.	• Company reporting and user appeals mechanisms should meet minimum effectiveness criteria based on Principle 31 of the UNGPs (legitimate, accessible, predictable, equitable, transparent, rights respecting, source of continuous learning). • Companies should provide remedy by satisfaction (e.g., an apology and explanation), restitution (e.g., by removing or restoring content), and guarantee of non-repetition (e.g., by revising policy, training staff, or removing repeat offenders).
Principle 23: Context In all contexts, business enterprises should: comply with all applicable laws and respect internationally recognized human rights; seek ways to honor the principles of internationally recognized human rights when faced with conflicting requirements; and treat the risk of causing or contributing to gross human rights abuses as a legal compliance issue wherever they operate.	• Companies should engage with stakeholder and experts when faced with conflicts between content policy and local law—especially when addressing the dilemma of whether to restrict some content for some users to preserve the availability of an overall service for everyone. • Companies should build and maintain relationships with affected stakeholders and experts in all relevant countries and regions, but especially those where risks of adverse human rights impacts—including violations of international humanitarian law—is more likely.
Principle 24: Prioritization Where it is necessary to prioritize actions to address adverse human rights impacts, companies should first seek to prevent and mitigate those that are most severe or where delayed response would make them irremediable.	• Companies should prioritize implementation of content policy in countries, regions, and contexts where the human rights risks are most severe—not just "in the local backyard."

ANNEX 2: FRAMEWORK SUMMARY

Content Policy	Content Policy Implementation	Product Development	Tracking and Transparency
Statements about what content is and is not allowed, as well as about the visibility and distribution of content.	How content decisions are executed in practice.	How new features, services, and functionalities are introduced and evolve.	How the outcomes and effectiveness of a human rights-based approach is measured and communicated.
Founded upon human rights standards and instruments. Encompass all human rights. Informed by stakeholder engagement. Distinguish between organic and paid speech.	Informed by affected stakeholders and experts in the relevant context. Based on international human rights principles (necessary, proportionate, legitimate, non-discrimination). Special attention to cumulative impacts of individual content decisions. Honor the principles of internationally recognized human rights when faced with conflicting laws. Prioritize based on severity. Heightened attention to conflict-affected areas. Appeals mechanisms meet UNGP effectiveness criteria. Effective remedy when mistakes are made.	Human rights assessments of new or evolving features, services, and functionalities. Assume new or evolving features, services, and functionalities may have different, unintended, or more severe consequences in higher-risk markets. Informed consent integrated into product choices.	Quantitative and qualitative indicators of effectiveness. Annual public reports providing sufficient quantitative and qualitative information to evaluate the adequacy of approach. Transparency about the rationale for important content decisions.
Special attention to human rights defenders. Structured approach to vulnerability.			