UNITED STATES COMMISSION *on*
INTERNATIONAL RELIGIOUS FREEDOM

FACTSHEET

December 2021

PROTECTING RELIGIOUS FREEDOM ONLINE

USCIRF's Mission

To advance international freedom of religion or belief, by independently assessing and unflinchingly confronting threats to this fundamental right.

By: Kirsten Lavery, Supervisory Policy Analyst

Introduction

As explored in USCIRF's 2020 hearing on *Combatting Online Hate Speech and Disinformation Targeting Religious Communities*, social media platforms can mobilize violence, discrimination, and hatred towards religious communities, negatively impacting their freedom of religion or belief. Governments' and social media companies' insufficient responses to online hate can result in grave human rights violations, as illustrated by Facebook's failure to address *incitement against Rohingya Muslims in Myanmar*. At the same time, social media platforms have become the primary forum for public and private expression in many contexts. Removing or censoring protected speech online can also impede human rights.

This factsheet outlines international human rights standards relevant to the regulation of speech on social media platforms, highlighting the impact on the freedom of religion or belief. States have a direct duty to protect human rights online, which requires balancing removing unlawful content with allowing protected speech. Social media companies, as private companies, are only bound by human rights standards enshrined in national legislation. However, this factsheet considers the impact of social media companies' current content moderation policies on religious freedom and explores the opportunities for those companies to incorporate international human rights standards to address the harms that may occur on their platforms.

The Scale of Online Hate Speech Targeting Religious Communities

The volume of speech, and hate speech, being shared and regulated online is astonishing. Facebook, for instance, *governs* more communication than any government and *removes* three million pieces of hate speech a month, or more than 4,000 an hour. According to the *UN Special Rapporteur on minority issues*, at least 70% of the victims of hate and incitement to violence and discrimination online are minorities, including many members of religious minority communities. The UN Special Rapporteur on freedom of religion or belief has raised concern regarding the wide dissemination of hate speech online targeting religious groups, including *anti-Muslim* and *antisemitic* hatred and conspiracy theories.

www.USCIRF.gov
@USCIRF
Media@USCIRF.gov

732 N. Capitol Street, NW, Suite #A714
Washington, DC 20401
202-523-3240

International Standards Related to Hate Speech

As online platforms have *revolutionized* the public square, the starting point for tackling online hate and incitement to discrimination or violence must be the protection of international human rights. Article 19 of the *International Covenant on Civil and Political Rights (ICCPR)* protects the freedom of expression or opinion "through any media" and "regardless of frontier." All rights are interrelated, and the freedom of religion or belief—protected under Article 18 of the ICCPR—relies heavily on the freedom of expression or opinion and vice versa. As *explained* by the UN Special Rapporteur on freedom of religion or belief, "the fates of these two rights are entwined, such that the violation of one is frequently tantamount to contrivance to undermine the other."

Although expansive, freedom of speech is not absolute and must be balanced against other important values. To assist states in this balancing act, the ICCPR and other international instruments define narrow circumstances where speech can be limited. Since the freedom of expression is fundamental to the enjoyment of all other rights, restrictions on free speech must be exceptional. Hateful speech that does not meet the high standards to be restricted under international law should be combated in other ways, such as with education, condemnation by public officials, and the promotion of pluralism.

- **Hate Speech:** There is no definition in international human rights law of the colloquial term "hate speech," but it is typically understood to mean speech that expresses prejudice against a specific group. Any legal restriction on hate speech must be narrow and meet the strict requirements of Article 19(3) of the ICCPR, which provides that any speech *limitation* must be legal, proportional, and necessary to protect the rights or reputations of others, national security, public order *(ordre public)*, public health, or public morality. In other words, restrictions can only be imposed by law when necessary to protect one of these legitimate aims. Otherwise, the freedom of expression is expansive and protects the right to express offensive, controversial, or disturbing information and ideas.
- **Imminent Incitement:** Under Article 20(2) of the ICCPR,[1] states are required to prohibit the most severe forms of hate speech, specifically, any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence (hereinafter "imminent incitement"). Imminent incitement is a particularly small subset of the most dangerous hate speech, given its high likelihood of leading to discrimination, hostility, or violence. The question of what constitutes imminent incitement is complex, as there are no clear and universally accepted definitions of its key components under international law. The terminology and standards in domestic laws seeking to address imminent incitement vary from country to country, and a state can implement this framework within its own moral and social context as long as the standards of both Article 20(2) and Article 19(3) are met.

  As explained in *the Rabat Plan of Action* and other instruments, prohibitions on imminent incitement should: (1) expressly incorporate the language of Article 20(2); (2) narrowly define key terms; (3) require a high threshold for when speech can be limited; (4) include a contextual assessment; and (5) criminalize speech only as a last resort. Further, prohibitions of imminent incitement must also meet the strict and narrow conditions established under Article 19(3).

  Maintaining a high threshold to limit speech is essential to protecting both freedom of expression and religious freedom. As the *Special Rapporteur on freedom of religion or belief* has noted, "any attempt to lower the threshold of article 20 … would not only shrink the frontiers of free expression, but also limit freedom of religion or belief itself. Such an attempt could be counterproductive and may promote an atmosphere of religious intolerance." *UN Human Rights Council Resolution 16/18* further reiterates the necessity of a high threshold of limiting speech pursuant to Article 20(2) as "the open public debate of ideas can be among the best protections against religious intolerance." Even in the context of *violent extremism* or *the denial of past incidents of severe persecution or genocide* (so called "*memory laws*"), the threshold for speech that can be prohibited as incitement to violence remains quite high.

1 The United States and a number of other countries have *reservations* to ICCPR Article 20. The United States' reservation applies to the extent that this provision interferes with the rights protected under the U.S. Constitution.

The Relationship between Hate Speech and Blasphemy

Some states have *enacted* hate speech laws that effectively operate as *blasphemy laws* prohibiting speech insulting to religion or religious beliefs, symbols, or feelings, which are not permitted under international human rights law. These laws are often formulated in vague terms and define their purpose as providing protection against hostility that might arise following speech offensive to religion. Under international human rights law, however, advocacy of religious hatred can only be prohibited when it meets the *strict standards of Article 20(2)*. These laws are also problematic because rights holders are human beings, not religions or belief systems.

USCIRF's report *Apostasy, Blasphemy, and Hate Speech Laws in Africa* examined the intersection between blasphemy and hate speech laws, finding significant overlap between the speech prohibited by these laws at times. This is particularly true when hate speech laws are formulated in vague terms, target specific content, and lack clarity as to whether the laws' purpose is to protect individuals belonging or subscribing to the belief of a group or to protect the beliefs. For these reasons, states must be particularly careful and precise in regulating hate speech to ensure that such laws comply with international human rights standards.

- **Incitement to genocide:** In recognition that genocide is often preceded and accompanied by widespread hate speech, *the Convention on the Prevention and Punishment of the Crime of Genocide (CPPCG)* requires states to prohibit incitement to genocide, along with the act of genocide itself. Direct and public incitement to commit genocide is a punishable offence pursuant to the CPPCG. The CPPCG defines genocide as committing specific acts with the intent to destroy, in whole or in part, a national, ethnic, racial, or religious group. Like imminent incitement, incitement to genocide is a narrow subcategory of the most dangerous hate speech, given its serious likelihood of sparking genocide.

The Obligation of States to Protect Human Rights on Social Media Platforms

International human rights obligations apply both *online and offline*. When regulating intermediaries, states have a dual responsibility to protect rights by ensuring legal content remains online, while also enforcing the boundaries of the freedom of expression by removing unlawful content. Under international human rights law, states are required to take measures to ensure that incitement to genocide and imminent incitement are prohibited online. In addition to the inclusion of these narrow prohibitions, any regulation of online hate speech must meet requirements of *legality, necessity and proportionality, and legitimacy*. Further, states should not create special categories for online hate speech that impose higher penalties than offline hate speech.

States cannot require that social media companies restrict expression that states themselves cannot directly prohibit. The UN Special Rapporteur on freedom of religion or belief has *expressed* concern regarding the scale of some governments' censorship and filtering, noting that "[w]hile there is a need to prevent and punish online incitement to violence, some of the current approaches, characterized by vaguely worded laws on what is proscribed and draconian intermediary penalties, are likely to be highly counterproductive, with chilling effects." Online blasphemy prohibitions and overly broad definitions of hate speech, such as those that ban incitement of "religious discord," are problematic for this reason.

Enforcement of Blasphemy Laws Online

Even though international human rights law prohibits the enforcement of blasphemy laws, some countries maintain blasphemy laws that criminalize expression, including on social media. *USCIRF's 2020 report Violating Rights: Enforcing the World's Blasphemy Laws* examined the global enforcement of blasphemy laws between 2014 to 2018. More than one-quarter of reported cases involved alleged blasphemous speech posted on social media platforms. Of the enforcement cases and incidents in which social media was implicated, Facebook was involved in nearly half of the cases. Cases were also found involving Twitter, Vkontakte, YouTube, Instagram, WhatsApp, and Telegram. Enforcement of blasphemy laws has broad implications for the freedom of expression and, consequently, the freedom of religion or belief. As noted in the report, "social media blurs the distinction between the public and private spheres, permitting the state to enforce against conduct that may never have been intended for public consumption or widespread dissemination."

In addition to criminal enforcement for online speech, some governments censor online publications to protect religious beliefs. Often, these laws require social media companies to take down alleged blasphemy posted online, in violation of the rights of users. While many social media companies lack transparency regarding their compliance with requests from governments to take down content, Facebook publishes high-level statistics regarding the content it removes based on local restrictions. According to these *transparency reports*, between July and December 2020 Facebook *restricted access* in Pakistan to 1,531 items reported by the Pakistan Telecommunication Authority as allegedly in violation of the local blasphemy law. During the same period, Facebook restricted access to additional content pursuant to government requests that it allegedly violated local blasphemy laws, including in *Bangladesh* and *Indonesia*.

Content Moderation and Religious Freedom Online

Every day social media companies make millions of *decisions* that regulate speech. Most social media companies, such as *Facebook* and *Twitter*, have policies banning certain types of hate speech, including against religious communities, in their community standards and platform policies. Other harmful types of speech, such as *Holocaust denial* and *violent extremism*, are also frequently banned. These rules are enforced through a combination of artificial intelligence (AI) and human content moderation to filter, review, flag, and remove or downgrade potentially problematic posts. Notably, other social media platforms have few if any policies for addressing hate and moderating content.

While social media companies should provide greater *transparency* to permit a full understanding of their hate speech rules and enforcement, it is clear that current content moderation processes and tools can harm the rights of religious communities. AI software allows companies to identity potentially problematic content, but it is not always effective in removing hate speech that targets religious communities. Hate speech is *nuanced and context-dependent*, *varies* across *jurisdictions* and *languages*, and often involves unknown speakers and coordinated bots, which can be hard for AI tools to detect.

Restrictions and removal measures can also *disproportionately* affect religious communities. Speech that seeks to counter hate speech narratives can be indistinguishable to an algorithm from imminent incitement or other forms of hate speech. According the UN Special Rapporteur on freedom of religion or belief, this has *led* to "hampering and potentially de-platforming targeted [religious] communities' own efforts to counter the discrimination they face." There have also been reports of AI tools on platforms such as *YouTube* confusing documentation of war crimes and other atrocities committed against religious communities with extremist content, leading to its removal. As algorithms frequently remove content before it is posted, it is impossible to know the scale of these issues.

Alongside automated tools, some social media companies also rely on users and third parties to flag hate speech. The bias of those using these notification systems can also impact religious communities, alongside the human bias inherent in algorithms. As *explained* by the UN Special Rapporteur on freedom of religion or belief, the use of these tools "might reinforce societal prejudices against minorities, exposing them to further stigmatization, discrimination and marginalization. Their use in a climate of intolerance, for example, at times, can result in the over-policing of certain faith communities and further inhibit communicative action." To better protect religious

communities, algorithms can be improved to combat these biases. Facebook, for example, recently *updated* its *algorithm* to prioritize the flagging of hate speech targeting minorities, including Muslims and Jews.

Given their limitations, AI tools are often supplemented with human content moderators. The UN Special Rapporteur on freedom of religion or belief *noted* that the hate speech policies of some social media companies have improved significantly in recent years, resulting in the removal of some of the most egregious content. However, there has been an increase in "borderline content," which requires in-depth analysis to decide whether it violates a company's hate speech policy. Tackling these borderline cases has led to increased human augmented moderation, which is a "welcomed change," although the training and decision-making processes used by moderators are still generally not transparent.

The Human Rights Obligations of Social Media Companies

Social media platforms, as private companies, do not have the same obligations to protect human rights as governments. Still, given the impact of content moderation on human rights, along with the harms caused by unchecked hate speech and excessive censorship, social media companies are still forced to grapple with the protection of rights.

The *UN Guiding Principles on Business and Human Rights* (UNGPs) provides guidance on how *technology companies* can incorporate human rights into their products and address adverse human rights impacts, including a framework for a human rights-based approach to *content moderation*. The UNGPs outline principles of due diligence, transparency, accountability, and remediation that can help mitigate human rights impacts.

While some social media companies, such as *Facebook*, have recently made efforts to incorporate human rights concerns into their products and policies, many social media companies make little to no reference to human rights. However, as *articulated* by the UN Special Rapporteur on freedom of expression or opinion, "this is a mistake, as it deprives the companies of a framework for making rights-compliant decisions and articulating their enforcement to [g]overnments and individuals, while hobbling the public's capacity to make claims using a globally understood vocabulary." An anchor in international human rights standards can *provide* a tool and common language for combating demands by government for excessive content removal pursuant to blasphemy laws or other problematic types of speech prohibitions. In many of the countries where hate speech targeting religious communities is the most *prevalent*, the government either produces such content or tacitly encourages it. International human rights principles can similarly aid social media companies in setting standards to regulate hate speech online and other harmful content in these jurisdictions where government sponsored hate speech is prevalent.

Applying Human Rights Norms to Content Moderation

A recent initiative to improve content moderation policies and transparency is Facebook's *Oversight Board* (OSB), which serves as the platform's court of final appeal on content moderation decisions and applies international human rights norms. Several of the OSB's first *decisions* addressed hate speech targeting religious groups, and its *recommendations* have included calling on Facebook to take action to avoid mistakes that silence the voices of religious minorities. While the establishment of the OSB may lead to increased respect for the rights of religious communities on the platform, the OSB has a limited mandate and will only be involved in a very small percentage of content moderation decisions. Further, while its decisions regarding take-downs are binding, the OSB's *policy recommendations* are not binding on Facebook.

Conclusion and Recommendations

It is equally important to protect the freedom of religion or belief offline and online, and the U.S. government can play an important role in ensuring the protection of religious freedom digitally. Following USCIRF's hearing on online hate speech, USCIRF *recommended* that the U.S. government take the following actions to protect the freedom of religion or belief and related rights on social media platforms:

- Highlight the abuse of social media by foreign governments that create a hostile environment to religious freedom and freedom of expression in bilateral dialogues and multilateral forums, along with including pertinent examples in the State Department's annual International Religious Freedom report;
- Engage in and promote counterspeech on U.S. government social media accounts to assist in combating disinformation and hate speech directed at religious communities; and
- Fund programs that develop and utilize early warning mechanisms in countries with widespread hate speech and misinformation directed at religious communities, to better develop tools and processes to monitor harmful speech on social media and prevent offline violence and discrimination.

The U.S. government can also play a vital role in ensuring that social media companies protect human rights and religious freedom. In May 2021, the U.S. Department of State *emphasized* "the importance of technology companies developing transparent criteria and robust safeguards to ensure the application of any terms of service is consistent with fundamental freedoms." As this must include the protection of religious freedom online, USCIRF has *recommended* that the U.S. government:

- Work with other governments to consider and define the responsibility of social media companies to abide by international human rights law on their platforms.

Professional Staff

Danielle Ashbahian
Supervisory Public Affairs Officer

Keely Bakken
Senior Policy Analyst

Dwight Bashir
Director of Outreach and Policy

Susan Bishai
Policy Analyst

Elizabeth K. Cassidy
Director of Research and Policy

Mingzhi Chen
Policy Analyst

Patrick Greenwalt
Policy Analyst

Gabrielle Hasenstab
Communications Specialist

Roy Haskins
Director of Finance and Operations

Thomas Kraemer
Director of Human Resources

Kirsten Lavery
Supervisory Policy Analyst

John Lechner
Policy Analyst

Niala Mohammad
Senior Policy Analyst

Jason Morton
Senior Policy Analyst

Mohyeldin Omer
Policy Analyst

Dylan Schexnaydre
Victims List and Outreach Specialist

Jamie Staley
Supervisory Policy Advisor

Zack Udin
Researcher

Nina Ullom
Congressional Relations and Outreach Specialist

Madeline Vellturo
Policy Analyst

Scott Weiner
Supervisory Policy Analyst

The U.S. Commission on International Religious Freedom (USCIRF) is an independent, bipartisan federal government entity established by the U.S. Congress to monitor, analyze, and report on religious freedom abroad. USCIRF makes foreign policy recommendations to the President, the Secretary of State, and Congress intended to deter religious persecution and promote freedom of religion and belief.

www.USCIRF.gov
@USCIRF
Media@USCIRF.gov
732 N. Capitol Street, NW, Suite #A714
Washington, DC 20401
202-523-3240