Gregorio, Giovanni; Stremlau, Nicole

Article

Information interventions and social media

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Gregorio, Giovanni; Stremlau, Nicole (2021) : Information interventions and social media, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 10, Iss. 2, pp. 1-25, https://doi.org/10.14763/2021.2.1567

This Version is available at:
https://hdl.handle.net/10419/235970

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 10 **Issue** 2

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Information interventions and social media

Giovanni De Gregorio *University of Oxford* giovanni.degregorio@csls.ox.ac.uk
Nicole Stremlau *University of Oxford*

DOI: https://doi.org/10.14763/2021.2.1567

Published: 30 June 2021
Received: 11 November 2020 **Accepted:** 23 April 2021

Funding: This research is part of the ConflictNet project (The Politics and Practice of Social Media in Conflict). It has been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 716686, ConflictNET).
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: De Gregorio, G. & Stremlau, N. (2021). Information interventions and social media. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1567

Keywords: Social media, Hate speech, Content moderation, Conflicts, Responsibility to protect

Abstract: Recent conflicts, particularly in Asia and Africa, have highlighted the potential for social media to provoke or exacerbate violent conflict and mass atrocities. The role of media and propaganda in disseminating hate and violence has been a longstanding aspect of war. In some cases of violent conflict, international actors—including the United Nations (UN)—have undertaken 'information interventions', a term that came into its own in the mid-1990s in response to the ongoing conflict in the Balkans, and the use of radio in the Rwandan genocide in 1992. While information intervention has historically been applied to mass media, this article explores the relevance and applicability of this approach to online communications, and social media in particular. We unpack whether and how information intervention might apply when social media has a role in inflaming extreme violence, or genocide, by disseminating disinformation and hate speech and international actors have a responsibility to protect and halt mass atrocities.

Introduction

When speaking to a group of reporters in 2018, the chairman of the UN Independent International Fact-Finding Mission on Myanmar, Marzuki Darusman, was clear that social media had a "determining role" in suspected acts of genocide in the country, arguing "[i]t has… substantively contributed to the level of acrimony, dissention and conflict…Hate speech is certainly part of that. As far as the Myanmar situation is concerned, social media is Facebook, and Facebook is social media" (Miles, 2018, n.p.). This view was further outlined with supporting evidence in the UN Mission's report later that year, which also referred to reports from human rights observers going back until at least 2012 identifying the role of social media in provoking violence by promoting anti-Rohingya discourse along with inaccurate and inflammatory images of violence.

The role of social media and conflict in Myanmar is not an isolated case. The use of social media has been deeply intertwined with the decade-long war in Syria (O'Neil, 2013), while in the Central African Republic, online hate speech has been directly attributed to provoking mass atrocities between Christians and Muslims (Schlein, 2018). In Sri Lanka, rumours on social media are widely regarded as provoking a number of religious attacks, including the 2019 Easter Sunday church and hotel bombings (Fisher, 2019). This follows a longer legacy of the role of mass media in violent conflict, from the use of newspapers in Nazi propaganda campaigns (Herzstein, 1978) to the more recent conflicts involving radio in Rwanda and satellite television in Somalia (Allen and Stremlau, 2005; Stremlau, 2018).

In some cases of conflict involving mass media, international actors— including the United Nations and African Union—have undertaken 'information interventions', a term that came into its own in the mid-1990s in response to the ongoing conflict in the Balkans, and the use of radio in the Rwandan genocide (Metzl, 1997; Price and Thompson, 2002). Measures that fall under the broad umbrella of information intervention include the use of force to close newspapers or bomb radio transmitters, or softer interventions such as peace broadcasting (which entails supplementing existing media content with programming aimed at bridging divisions and encouraging reconciliation between conflicting parties), and conflict-sensitive journalism training.

While the harder forms of information interventions (such as the shutdowns of outlets) have been applied to mass media, in this article we focus specifically on the relevance of information interventions for online communications, and social media in particular. In doing so, we are primarily concerned with information inter-

ventions as a tool to forcibly silence certain voices or outlets, for example, the shutdown of social media sites, or even a partial or complete internet shutdown, on the part of international actors (including the United Nations (UN), the African Union (AU), or other multilateral organisations) to halt mass atrocities. If the state is the perpetrator of the violence, there might be a greater argument for external intervention, but if non-state actors are involved, the state itself might wish to respond by temporarily blocking social media or blocking the internet (although we recognise that rarely are conflicts so clear). Our focus here, however, is primarily on the potential role of international actors, rather than the state response. In other words, in cases where social media are misused to instigate violence and spread online hate that encourages genocide or mass atrocities, do international actors have a responsibility to launch an information intervention? And what exactly does an information intervention look like in the context of information spreading online?

The growing prominence of social media in disseminating disinformation, hate and inciting violence prompts urgent questions about whether—and to what extent—the doctrine of information intervention can be applied during mass atrocities when violence and hate is promoted through social media channels. As also underlined by Tufekci, social media can be drivers of hate and radicalisation (2018). Their role in contributing to the spread of hate and violence leads to questioning how to mitigate this situation. We recognise that the doctrine of information intervention needs to account for the peculiarities of the digital information landscape and the affordances of social media platforms. Unlike traditional media outlets which often are physically located in the countries where mass atrocities occur, in the absence of cooperation from a particular company, the only way to restrict such content may be to limit access to particular sites or to shut down the internet. The blunter the tool, such as an internet shutdown, the more problematic it is as it will have wider implications on society, not least in terms of expression, governance, and commerce (Marchant and Stremlau, 2020a, b).

Thus, against a worrying backdrop of inflammatory voices and online incitement to violence, we unpack what information interventions might mean where social media are involved, and to what extent it can be justified according to international law. While engaging with the notion of 'intervention', we focus on whether international actors could be given the legitimacy to intervene in a target country to shut down social media activities as part of an effort to cease violent conflict. While states have agreed that international law, including the principles of sovereignty and non-intervention, applies to states' activities in cyberspace (UNGA, 2013), at

the same time, as observed by Efrony and Shany, states rely on a 'policy of silence and ambiguity' to ensure broad margins of flexibility within the digital realm (2018, pp. 583-657).

We are aware that other forms of interventions may occur, such social media companies limiting or blocking content associated with an escalation of violence, as in the case of Facebook in Myanmar (Perrigo, 2021). But at present, social media companies have not demonstrated a consistent ability to effectively moderate content, particularly in the global south, and even in cases of genocide. Despite efforts to develop artificial intelligence capabilities to proactively identify and take down content, companies are still dependent on human monitors of which they simply do not have enough to address the scale of content being posted daily and the diverse contexts in which hate speech occurs (Barrett, 2020). As the UN Human Rights Council noted in their report on Myanmar, international law is clear about expressions of hate that *must* be prohibited (in contrast with those that *may* be prohibited and those that *should be* protected). Our concern is with those expressions that must be prohibited including incitement to commit genocide or incitement to violence. [1]

We explore this issue primarily through the lens of the United Nations. Although we recognise that other international and non-governmental organisations, including regional organisations such as the African Union, could also play a critical role, we begin with the applicability of Chapter VII of the UN Charter, which outlines the UN Security Council's powers to maintain peace, along with Chapters VI and VIII of the Charter which set out the responsibilities to protect populations from genocide, war crimes, ethnic cleansing and crimes against humanity. While we consider the potential consequences on the (digital) media environment of the target state, we underline that intervening in social media is not just a matter of legitimacy but also a measure that could significantly impact the digital information ecosystem of the target state.

In the first section of this article, we outline the doctrine of information intervention, grounding our arguments in a historical review of the debate. We consider information intervention by, firstly, analysing the principle of non-intervention under

1. Further hate speech that must be prohibited includes "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence; and all dissemination of ideas based on racial superiority or hatred, and on incitement to racial discrimination". Still concerning but where restrictions must be targeted and proportionate include speech that "presents a serious danger for others and for their enjoyment of human rights…[or] be necessary in a democratic society for the respect of the rights or reputation of others or for the protection of national security or public order".

international law; and, secondly, examining how the human rights law framework provides legal justification to prevent mass atrocities, thus, triggering the responsibility to protect authorising intervention under Chapter VII.

In the second section, we focus on information intervention within the framework of social media, precisely focusing on the unique aspects of social media relative to traditional media outlets, such as radio and television, with regard to spreading hate and escalating violent conflicts. In particular, this comparative analysis highlights the peculiarities of the social media environment as well as the politics of online content moderation on a global scale.

On this basis, we offer a new framework for information intervention in the context of online media. Specifically, we consider how legal justifications are affected by the social media environment, and what suitable measures might be adopted in this new and emerging context. Drawing on the work of Metzl, who nearly 25 years ago argued the need for the UN to establish a unit to intervene when mass media (and particularly radio) is involved in genocide, we conclude by revisiting the need for an international mechanism, including, what we refer to, as a possible Information Intervention Council, that proceduralises situations of interventions. This new institution, or mechanism, would ideally be grounded in an international system, like the AU or UN, affording it legitimacy and accountability in international law. The implications of our argument are significant. We do not want to be misconstrued as advocating for the widespread use of censorship or internet shutdowns that we have seen increasing over the years as blunt tools for addressing everything from concerns around electoral fraud, to hate speech, to the leaking of exam papers (Henley, 2018). Rather, it is our belief that empowering an international council to intervene would reduce arbitrary shutdowns because claims of hate speech and the association with offline violence would be independently scrutinised, thereby offering more legitimate options for addressing the more severe cases while exposing the instances when shutdowns have been used for other reasons (they are often seen as a tool by autocratic governments to silence voices they dislike) for what they are.

The doctrine of information intervention

Weapons, troops, tanks and aircraft are not the only instruments of harm in violent conflict, with propaganda and communication channels—such as radio and television—having long demonstrated an ability to weaponise hate against minorities or targeted groups (Larson and Whitton, 1963). Media outlets have contributed to guiding and organising entire military forces, and promoting propaganda with a

view to attracting new proselytists. In the 1960s, when the debate was focused on nuclear disarmament, Whitton and Larson noted that 'while in past years we have often heard the phrase 'the propaganda of disarmament', we should now hold forth as an urgent need 'the disarmament of propaganda'' (1963, p. 1).

Information interventions are strategic efforts to interfere in (whether disrupting, manipulating or altering) a communications environment within a community, region or state afflicted by mass atrocities, in order to prevent the dissemination of violence-inciting speech. The characteristics of such an intervention can be assessed by observing its duration, goals and degree. The intervention can take place at various stages of a conflict. For example, it may attempt to tackle (in advance) the conveyance of messages that could lead to conflict escalation. However, the use of force (e.g. the takedown of a media outlet) has rarely been employed as a preventive measure, as it would be difficult to garner support for such interference in national sovereignty based solely on unsubstantiated assumptions. Interventions at this stage are therefore likely to be softer, focusing on offering alternative voices or perspectives, or training journalists (what often falls under the broader umbrella of media development).

During cases of escalating violence, an information intervention might consist of media and conflict monitoring of the target state; peace broadcasting, which seeks to provide an outlet for non-violent voices to counter; or media shutdowns, which involves censoring the media whose messages are provoking conflict, such as the bombing of radio towers (Larson and Whitton, 1963, p. 17). In situations where conflict is winding down—or at least where this is the hope—intervention usually focuses on measures promoting what the international community or funders believe to be a democratic and sustainable media environment. Thus, in conflict situations, information intervention consists of both short- and long-term strategies aimed at stabilising the media environment within a specific country (Larson and Whitton, 1963, pp. 185-186).

In terms of the level of interference a target state is subject to, international law has a role when assessing the legality of more interventionist measures, such as media restrictions or shutdowns. Our primary focus in this article is on what Metzl referred to as the 'third step', which requires the taking down or closure of a particular outlet or platform, or an internet shutdown, in situations of severe violence (e.g. genocide). In this article, we focus on the challenges arising from the legal basis justifying this level of information intervention, especially when censoring social media in the target state by relying on internet shutdowns.

Historical evidence concerning the role of media in escalating violent conflicts might constitute legitimate grounds justifying information intervention. Intervening in situations of genocide and violence promoted by the media could also be justifiable from a moral perspective. As suggested by Metzl:

We need to explore what can be done between the impossible everything and the unacceptable nothing. The political cost of doing everything is usually prohibitive. The moral cost of doing nothing is astronomical. If we accept that we are not going to do everything possible to stem a given conflict, what can we do to have as much impact as we are willing to have? (Metzl, 2002, pp. 41-42).

However, historic or moral legitimacy is not necessarily the same as legal legitimacy. The principle of non-intervention in international law aims to protect the sovereignty of each country and is recognised as a peremptory norm (i.e., *jus cogens*). The legal rank of this principle is one of the primary challenges to media intervention and specifically the use of force to block the spread of certain information. Members of the international community cannot lawfully intervene without the authorisation of the UN Security Council, except in exceptional cases like self-defence. There is, therefore, a clash between the principle of non-intervention aimed at safeguarding national sovereignty, and the need to address speech that fuels severe conflict, including genocide.

Before addressing the challenges raised by social media in spreading violence and hate online, it is necessary to outline the legal justifications that form the basis of information intervention.

The principle of non-intervention

Interventions raise serious challenges for state sovereignty. Recent conflicts, from Syria to Iraq, have contributed to promoting the debate on the boundaries of the principle of non-intervention (Chinkin and Kaldor, 2017; Davis et al., 2015), even if outside the framework of information intervention. The same principle could be extended to international telecommunications law governing territorial sovereignty as it relates to the protection of airwaves and the flow of information (Rajadhyaksha, 2006, p. 1).

The principle of non-intervention is enshrined in the UN Charter and, similar to the ban on the use of force, is derived from and supports the idea of state sovereignty. As a legal principle, it first appeared when the League of Nations was created, stipulating mutual respect for territorial integrity and sovereignty, and non-interference in the internal affairs of other states. These dispositions have been included

in the UN Charter, (1945, Art. 2) [2] while in *Nicaragua v United States* (ICJ, 1986, p. 1), the International Court of Justice ('ICJ') considered the principle of non-intervention to be a general tenant of customary international law.

The principle of non-intervention can, however, be restricted in crucial ways relevant for digital information interventions. The prohibition established by the UN Charter is not absolute, with Chapter VII allowing the UN (through the Security Council) to intervene in domestic situations provided there is a threat to international peace and security (Frowein and Krisch, 2002, pp. 701-716). Chapter VII is the only part of the UN Charter empowering the Security Council to make binding decisions that apply to all UN members (Öberg, 2005, p. 879). More specifically, this process involves two steps (UN Charter, 1945, Art. 39). First, the Security Council must determine that a threat to or breach of peace, or an act of aggression, has occurred. Second, the Security Council must propose measures aimed at maintaining international peace and security that accord with the UN's purposes and principles (Art. 24(2)). The Security Council can decide what measures are to be employed when giving effect to its decisions (Art. 41), including the 'complete or partial interruption of economic relations and of rail, sea, air, postal, telegraphic, radio, and other means of communication, and the severance of diplomatic relations' (UN Charter, 1945, n.p.).

Should the Security Council consider the aforementioned measures inadequate (Art. 42), or they have been proved to be inadequate in maintaining or restoring international peace (Art. 42), it can order that further measures be implemented based on the use of force. Within this framework, measures such as radio jamming would be included as 'the most benign form of humanitarian intervention' (Metzl, 1997, p. 628) whereas the shutdown of a media tower through the use of military force would fall under the scope of Article 42. As a result, the latter measure could only be authorised by the UN Security Council once initial measures had failed to meet their objective of restoring peace and security in the area of intervention.

The general principle underlying such measures is that any action taken must be consistent with the 'purposes and principles of the United Nations'.However, the boundaries of this are ill-defined, and the Security Council enjoys absolute discretion in deciding what actions or events constitute a breach of peace, a threat to

2. See, also, United Nations General Assembly, Declaration on the Inadmissibility of Intervention in the Domestic Affairs of States and the Protection of their Independence and Sovereignty, GA Res. 2131/XX, 21 December 1965; United Nations General Assembly, Declaration on Principles of International Law concerning Friendly Relations and Co-operation among States in accordance with the Charter of the United Nations, GA Res. 2625 (XXV), 24 October 1970.

peace, or an act of aggression (Whittle, 2015, p. 671; King, 1996, p. 509).

Therefore, Security Council authorisation is the first step in assessing an information intervention's compliance with international law and, particularly, the principle of non-intervention. Should the Security Council order an information intervention, it would no longer violate the non-intervention norm. This is because once a state enters into an international treaty, it is bound by its terms, and within the framework of the UN Charter, (almost) all recognised states are parties to a treaty binding them to Security Council decisions regarding threats to international peace and security. As a result, under Chapter VII, they have effectively consented to the Security Council intervening in their sovereign affairs in situations where it is necessary to restore peace and security. This includes information interventions.

This legal architecture could appear controversial, as the right to territorial integrity and political independence are guaranteed by the UN Charter. Furthermore, a primary principle of international radio law is the prohibition of 'harmful interference', with media jamming, for instance, potentially falling into this category (Preamble of Radio Regulations, 2016). Nevertheless, according to Blinderman:

If international law precluded a state from voluntarily delegating fragments of its sovereignty to a multinational treaty organization, the international system could not operate. As such, courts have long recognized that a state's consent to a particular treaty covering a specific matter forecloses its ability to claim that the matter is exclusively within its domestic jurisdiction (2002, p. 111).

In contrast, when states intervene in the domestic affairs of another nation without receiving authorisation from the UN or the consent of the target state, they run the risk of violating the target state's sovereign rights. While it might be argued that a particular information intervention is based on humanitarian need and the responsibility to protect, this can be challenged. As Shen (2001, n.p.) argues:

> There is no commonly acceptable standard of what humanitarianism means and what human rights embrace under international law. In the absence of common understanding, the concepts of 'humanitarianism' and 'human rights' are bound to be abused if the international community allows humanitarian intervention, or favours individual human rights over national sovereignty. The consequences of this kind of abuse use would be too dreadful to contemplate. One of the consequences of placing human rights above state sovereignty and therefore permitting humanitarian intervention, would be that the ordinary and predictable short comings (*sic*) of third-world states would be attacked as

> human rights violations. Such domestic problems would provide excuses and opportunities for major powers to intervene and to 'dominate' weaker states.

Therefore, beyond the framework of Chapter VII, states are not authorised to intervene in the internal affairs of a sovereign nation, except in self-defence when protecting its national interests or under an invitation of the target state (Thomas, 1999). From this perspective, any intervention outside the framework of Chapter VII is a violation of state sovereignty, except for self-defence. If an intervention is based on a regime of consent, the boundaries of its authority are legitimated and determined by the conditions set by the UN or target country.

Hate speech and mass atrocities

The boundaries of the non-intervention principle raise the question of whether and when information intervention can be justified when seeking to prevent mass atrocities provoked by online hate speech and disinformation (HLEG, 2018; Wardle and Derakhshan, 2017). International law does not preclude the UN Security Council deciding what kind of speech or incitement satisfies the threshold required to trigger the Chapter VII mechanism. As a result of this discretion, changes in the global political environment—such as those that took place in Rwanda or Bosnia—allow for the translation of legal considerations into policy objectives.

Although certain communication channels can enable the spread of hate speech, the degree of danger may not be considered a threat to international peace and security. In general, while there is an international presumption in support of the free flow of ideas and information, this has been mitigated by international human rights law, where the protection of free speech is subject to certain conditions (Farrior, 2002, p 69). Though the right to freedom of expression is enshrined across international, regional and national bills of rights, it is subject to exceptions that protect other rights (e.g. dignity) or that pursue legitimate interests enshrined by the Universal Declaration of Human Rights (1948, Art. 7, 19, 29, and 30) and the International Covenant on Civil and Political Rights (ICCPR, 1966). As is well-known, the right to free speech is protected by international human rights law but its exercise is not absolute. With specific regard to hate speech, the International Convention on the Elimination of Racial Discrimination (ICERD) bans incitement to racial hatred and discrimination (Art. 4).

It is the UN Convention on the Prevention and Punishment of the Crime of Genocide that provides the most persuasive statement in support of information inter-

vention (1948). Although it does not directly address hate speech, this Convention states that 'direct and public incitement to commit genocide' is a punishable crime (Art. 3). [3] The inclusion of incitement as a punishable crime could therefore support the need 'for *preventive*, *pre-emptive* and *pro-active* measures to predict and intervene in potential mass suffering due in part to hate speech propagated by incendiary media' (Erni, 2009, p. 867). While international agreements and covenants do not provide guidelines for determining an 'information intervention threshold', the responsibility to protect ('R2P') regime offers an important point of reference (Bellamy, 2014), and addresses whether and to what extent international actors should intervene in situations where state actors fail (either voluntarily or involuntarily) to protect their population from mass atrocities or genocide.

The repeated failure to prevent genocides and atrocities after the Second World War eventually led to a rethinking of the notion of state sovereignty, and eventually, the Responsibility to Protect (also known as R2P). In the aftermath of the violence in Rwanda and the NATO intervention in Kosovo without the authorisation of the UN Security Council (an intervention that has been described as 'illegal but legitimate' (Independent International Commission on Kosovo, 2002)), the International Commission on Intervention and State Sovereignty (ICISS) solidified the 'responsibility to protect' in 2001. The World Summit and the UN institutions swiftly followed suit in making use of the term (UNGA Resolution 60, 2005). The ICISS report made clear that with sovereignty comes responsibilities (Glanville, 2013), among them the responsibility of a state, its population, and also the international community to answer violations of human rights, and particularly mass atrocities.

Crucially, the World Summit clarified that the R2P principle should be implemented within the framework of the UN Charter. As a result, R2P does not permit a state to use force against another state without authorisation by the UN Security Council. Concern has been expressed that unilateral humanitarian intervention is simply another way for some countries to exert their political and technological dominance over less powerful states (Shen, 2001, p. 1). The use of jamming technology, for example, raises serious sovereignty issues, particularly for developing countries, with information intervention measures easier to implement against small-scale actors compared to states with consolidated media outlets (Varis 1970, Metzl, 1997, p. 19).

There is a growing weariness with interventions. The hubris of the 1990s and early

3. This Article includes a) genocide; b) conspiracy to commit genocide; c) direct and public incitement to commit genocide; d) attempt to commit genocide; e) complicity in genocide.

2000s, when Metzl and others were initially writing about information interventions, has shifted. Terms such as 'democracy promotion', 'peacebuilding' and 'post-conflict reconstruction' are associated with large, expensive and mostly failed initiatives in countries such as Iraq or Afghanistan. Meanwhile, interventions in countries such as Myanmar, Libya, Syria, Somalia and Yemen have been less ambitious than those advocating for the responsibility to protect have argued for.

In this context, it is important to consider whether information interventions, particularly in reference to social media, can legally be based on R2P and/or humanitarian reasons in countries afflicted by violent conflicts. The primary argument in favour would be that sovereign powers are bound to respect the rights of their communities and the limitations placed on their powers. For instance, the right to life and security should be protected against threats such as torture or genocide. Since not only the principle of non-intervention but also respect of human rights constitute *jus cogens,* sovereign nations cannot hide behind the former when violating the latter.

Outside the framework of Chapter VII, the debate shifts to other exceptional grounds that might justify intervention in the media environment of a target state, including protecting the intervening state's national interest in terms of security, or in cases of gross human rights violations—such as genocide—a humanitarian intervention (Holzgrefe et al., 2003). Nevertheless, for the UN to intervene in an information space, a lack of UN authorisation constitutes the most relevant challenge; ultimately it is a political decision, and getting all security council members to agree, in the current political climate, would be difficult.

A new framework for digital information interventions?

The period of mid-1990 to early 2000 saw information interventions being intensely debated. Since then, media environments have changed significantly, particularly with regards to the emergence of social media. The doctrine of information intervention, which has traditionally dealt with mass media outlets based in the target state, has not kept pace with these changes.

While evidence of social media's ability to disseminate messages of hate and violence worldwide is compelling, there are particularities that make the direct translation of information interventions from mass media to new media challenging. Unlike traditional media, such as radio and television, which are usually subject to extensive regulation, this has not happened to the same extent with social media. The near-infinite extent of the digital environment makes monitoring its bound-

aries complex, and this difficulty extends to tackling online hate and disinformation that could lead to offline violence and mass atrocities. While states may be considered the primary legitimate authorities when it comes to implementing and enforcing binding norms, this idea of exclusive control is challenged at the international level where states cannot exercise their sovereign powers externally. In the absence of cooperation, we have seen how, especially in countries in Africa and Asia, governments have reacted to the spread of online hate by criminalising speech or shutting down the internet (De Gregorio & Stremlau, 2020; Clark et al., 2017).

The shift from traditional media outlets to social media also reveals further challenges. The business model of social media is not based on the creation of content but on their organisation and the accumulation of data based on which social media provide tailored profiling services attracting advertising revenues. As such, their primary goal is not to protect human rights through providing platforms for free speech, but to profit from users' data which are the primary source attracting advertising revenues. The immunity or exemption of liability for hosting third-party content makes this system particularly profitable. As service providers, social media are usually exempted from responsibility for the organisation and hosting of online content. In order to manage their online spaces and profile users, social media companies use automated technologies to organise content and enforce community rules (Gillespie, 2018). The increasing involvement of online platforms in organising content and user data through artificial intelligence has reflected a shift in their role toward becoming more active curators and content providers. Social media companies largely 'govern' the digital spaces where information flows (Bloch-Wehba, 2019; Klonick, 2018), and this does not change even in situations of conflict or violence where these actors can determine how to moderate hate and disinformation according to their ethical, business and legal framework.

This framework helps explain the need for a new approach within the information intervention doctrine as we shift from considering traditional media outlets to social media. As already underlined, Chapter VII can be used to authorise international intervention in the media environment of a target state without violating the principle of non-intervention. At first glance, this would seem to provide for the doctrine of information intervention being applied to social media promoting mass atrocities. Nonetheless, any information intervention measure must take into consideration the network architecture and modalities through which it is possible to limit dissemination of online hate and violence with specific regard to internet shutdowns.

The first of the following two subsections explores the challenges of extending information interventions in a social media context, particularly looking at content moderation, and the potential for establishing an information intervention council within the UN framework.

Intervention and social media

In cases where social media are involved in the escalation of violent conflicts, the UN Security Council could, in theory, authorise an intervention under Chapter VII due to a breach of international peace and security. Therefore, the international community would be legitimised to shut down the internet or limit access to social media as part of its response to addressing mass atrocities. However, even if such authorisation was forthcoming (which would undoubtedly be a challenge), such an intervention would require due care.

Unlike content-producing media outlets that are usually aware of the peculiarities of the local media environment, social media host third-party content which, due to the scale of content moderation required, is not subject to the same degree of granular decision-making. Moreover, social media companies generally do not have a substantial (or any) presence in the country involved in violent conflict, nor do they always have nationals familiar with local context and the language of online content. Information intervention in this field could lead to a more positive framework of content moderation, with greater safeguards and care applied by social media actors to avoid interventions by the international community.

However, any such information intervention runs the risk of social media actors choosing not to operate in conflict-affected countries. This would result in collateral censorship (Balkin, 1999, p. 2295; Wu, 2011, p. 293) that could involve not just the deletion of content, but the wholesale removal of specific social media spaces. Unlike traditional media outlets, which operate within a specific region and play an important role in providing information to those in that area, the presence of a social media platform is purely down to business opportunities. Therefore, social media would potentially be incentivised to cease operating in regions where information interventions might be enacted, provoking financial and reputational losses. In particular, international recognition of social media's involvement in escalating violent conflict could lead to social media companies declining to provide services to countries afflicted by such conflicts. Effectively, this could mean the creation of a 'social media vacuum' in some areas of the world.

The consequences of such a situation could be serious, especially in countries

where social media is the most popular way people experience the internet. This is the case in many countries in Africa. Social media are used every day by billions of users and, particularly in more closed regimes, are often a valuable source for connecting with others and accessing international information. Information intervention as put forward by Metzl focuses on 'democratic objectives', such as peacekeeping in the short run, and the building of civil and democratic media space in the long run. However, information intervention could potentially have more authoritarian, or protectorate, implications, such as imposing external sovereign powers over a target state's media. Here, the line between information intervention and censorship can become blurred, with the real test being whether or not the measures address the responsibility to protect (Thompson, 2002, p. 56). [4]

Unlike peace broadcasting, radio jamming, or the seizure of broadcasting towers, the international community could not proportionately fight the spread of online hate speech or disinformation without the cooperation of social media companies. Even while it is possible to rely on access providers to restrict access to this content, only companies can granularly intervene in the architecture of their digital spaces, including their proprietary algorithms (De Gregorio, 2018, p. 65). Should such companies decline to cooperate, or fail to devote significant resources and attention to addressing concerns of hate speech and disinformation, particularly in Africa and Asia, limiting access to the internet has become a primary tool for governments, either by shutting it down, slowing it down, or discriminating internet traffic relying on access providers. Without the direct cooperation of social media companies, information interventions may face difficulties adopting the scale of the approach proposed by Metzl based on monitoring, peace broadcasting and intervention. This could impede attempts to tackle the spread of hate and violence, as international actors may wait until events are deemed sufficiently serious before shutting down social media or the internet in the target state.

All this suggests that the cooperation of social media companies is an important, but not essential, component to addressing online content promoting hate and violence. Related to this is the risk of collateral censorship. Should social media be subject to pervasive information intervention measures, it is possible that companies—in seeking to evade responsibility and avoid interference from the international community—will decrease the degree of tolerance for hosted content and/or implement blanket content moderation technologies that rapidly (but less accu-

4. Metzl's approach to information has been criticised as representing 'a fashionable means of enhancing United States predominance within the international system, using information technology'.

rately) detect hate speech and violent content. The cooperation of social media companies in removing hate speech content from target states would require them to invest additional financial and human resources, especially for states requiring moderation in different languages where specific language policies are less compatible with blanket corporate content policies.

The lack of direct engagement and efforts to avoid offline harms on the part of social media companies derives not only from the fact that they are often exempt from secondary liability with regard to the content they host (Floridi and Taddeo, 2017; Dinwoodie, 2017), but also from the lack of direct obligation to respect human rights. Within the framework of international law, state actors are the only entities permitted to become parties to (and therefore subject to the obligations of) human rights treaties (Reinisch, 2005, p. 37), whereas social media companies—in the absence of any constraining legal instruments adopted by state actors—are unfettered by the need to protect human rights (Carillo-Santarelli, 2017; Clapham, 2006). Even if we can identify responsibilities of online platforms according to the Guiding Principles on Business and Human Rights (2011) and the Rabat Plan of Action (2013), these instruments do not introduce binding obligations for online platforms but require states to intervene to protect human rights.

While this paradigm aims to protect individual liberties, it also carries serious risks when private actors start to exercise new forms of power outside the boundaries of regulation (Knox, 2008, p. 1). In the past, it was thought the private sphere must be protected from the state—rather than from private actors—through the recognition of rights and liberties. Global dynamics and, especially, digital technologies have led private actors to gather power in new and significant areas (De Gregorio, 2019). In the era of globalisation, this concentration of power in the hands of transnational private actors raised primary issues for the protection of human rights (Teubner, 2006). There is increasing pressure on private actors to comply with international human rights law when moderating online content (Kaye, 2019) particularly given that social media exercise regulatory functions in the digital environment (Report of the Special Rapporteur to the Human Rights Council on online content regulation, A/HRC/38/35, 2018). [5] Doing this would allow platforms to apply a universal reference in their content moderation activities.

It is possible that regional developments in international criminal law in Africa could, in the future, fill this gap. The Malabo Protocol (2014), for example, aims to

5. See, also, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/73/348 (2018); Guiding Principles on Business and Human Rights (2011).

add an international criminal law section within the African Court of Justice and Human Rights. This would allow for the prosecution of crimes against humanity and genocide, including hate speech. Based on the precedents set in Nuremberg and Rwanda, an extension to the prosecution of crimes against humanity and genocide could open towards more responsibility of online platforms. In the past, both the Nuremberg trials and the UN International Criminal Tribunal for Rwanda convicted media content providers and executives (Lafraniere, 2003). Nonetheless, even if the Malabo Protocol would enter into force, it is unlikely that the failure of social media companies to tackle hate speech could qualify as an offense (Irving, 2019), as social media are not content providers, but organise content published by users. However, making social media companies liable could encourage the overly ambitious censorship of content to escape responsibility. Therefore, it is unlikely that this approach would make social media more accountable for the spreading of online hate and violence.

Within this existing framework, it is necessary to focus on an alternative paradigm of information intervention in cases where social media are involved which looks at cooperation with social media as the first, and primary, step and shutdowns as the exception.

Establishing an information intervention council

The doctrine of information intervention in its traditional form sits awkwardly within the present social media environment. Simply because policy has been slow to adapt to the online environment should not excuse a lack of intervention preventing the spread of online hate and violence, especially when such content is correlated with offline violence as significant as a genocide.

In the multi-stakeholder environment of internet governance, public actors share responsibility for defining the international legal and political framework within which they operate. Therefore, a first step toward defining a doctrine of digital information intervention would involve establishing an appropriate international body within the framework of an international organisation such as the UN or AU. In the case of the UN, such a body would be responsible for addressing how international law deals with decentralised public actors operating on a global scale, as well as potentially working in collaboration with the Special Advisor for the Responsibility to Protect or the Special Adviser for the Prevention of Genocide (a role created in 2004).

The body would be responsible for conducting research, establishing guidelines

for media intervention in violent conflicts, and verifying the role played by the media environment in a potential target state. This would support the development of guidelines to nudge the private sector to comply with specific standards. There are, for example, 'due-diligence guidelines' promoting a specific code of conduct for companies operating in the import, processing and sale areas of the minerals extracted in places such as the Democratic Republic of Congo to mitigate the risk of an extension of the conflict in the Eastern part of the country (Security Council, Resolution 1952, 2010). In addition, the UN Security Council has also promoted new public-private partnerships to address global challenges like terrorism, referring especially to the role of social media (Resolution 2354, 2017). This is, however, just a first step since these measures are left to the discretion of private actors, thus raising questions about effectiveness and enforcement.

Given the aim of the body would not be to solve disputes, or interpret international law, it should not be structured like an international tribunal, but rather take the form of a dynamic council ('Information Intervention Council' or 'IIC') hosting members committed to addressing specific situations. In addition to members representing the international organisation (such as permanent members of the UN Security Council), temporary members should include representatives of social media companies operating in conflict zones, members of a target state's government, scholars and experts in the media and responsibility to protect fields, as well as members of civil society organisations. This mix of membership would provide the opportunity—from a short-term perspective—of addressing challenging situations in a comprehensive way, allowing members to come up with concrete solutions, rather than simply making declarations about the behaviour of particular states or social media. The presence of permanent members would guarantee continuity and, in ensuring that standards and guidelines for targeting state and social media involved in violent conflict zones are adhered to, strengthen the body's legitimacy.

The IIC would also contribute to 'proceduralising' how violence and hate speech within the social media environment are addressed, clarifying how potential target states and social media should behave, in particular defining the conditions whereby these actors should notify and/or report to the IIC regarding the dissemination of online hate and violent content. Such a system would allow granular information to be gathered, thereby informing the proportionality of any measures required to address specific concerns within a target state. The participation of social media companies would also make possible a more proportionate approach to media interventions and shutdowns, including—depending on the capacity and capabilities of the company involved—addressing content deemed objectionable by

the Council. This process would increase the transparency and accountability of social media companies involved in moderating online content, with the IIC monitoring measures implemented by social media companies aimed at tackling hate speech in target countries. This process would also help to avert reliance on internet shutdowns as a blunt and general measure.

A key challenge in establishing such an international body would be how to encourage the various stakeholders to participate. There are several reasons why various communities concerned might choose to participate in an IIC. States afflicted by violent conflict may view participation as an opportunity to be heard at the international level, and also to monitor external interference in their media environment. States may also see an advantage—particularly in the context of losing sovereignty over their digital media environment—in being able to draw on a broader international framework when addressing online hate speech and violent content. And freedom of expression advocates, including civil society groups, may appreciate the safeguards reducing the justifications for arbitrary shutdowns or censorship. Given concerns that online hate and violence might undermine social media companies' public standing and business, participation in the IIC would provide them with an opportunity to demonstrate their respect for human rights and peace on a global scale. This could solve the issue of enforcement since social media could be incentivised to participate and address the spread of online hate.

Developing an international framework would also help mitigate more extreme information intervention measures—such as internet shutdowns—which are often implemented in an *ad hoc* way and rarely through formal policy or legal channels. Limiting internet shutdowns is particularly relevant due to the effects they produce on population, including on opportunities for expression. The collaborative collection of relevant information by the IIC would help inform whether, and to what extent, intervention in the media environment of a target state is required. In addition, it would provide target states—which may lack remedies—with an alternative system and would also reduce the risk of collateral censorship as well as the use of internet shutdowns. Finally, the IIC would facilitate social media companies' participation in how implemented practices are defined. As a result, even without solving all the issues around information intervention, this bottom-up approach would support greater shared responsibilities between all stakeholders.

Conclusion

Social media have demonstrated their ability to influence speech transnationally and it is clear that the internet (along with other technologies) can have a role in

both enhancing and challenging freedoms and rights. Within this framework, digital information interventions can have a crucial role against the spread of genocide and mass atrocities. While mass media, including TV or radio, have long been recognised as a key actor in the escalation of violent conflicts, the scale of dissemination and the degree of accountability of digital actors involved is different. Although the doctrine of information intervention initially evolved to address concerns around the role of mass media in conflict, it can provide inspiration for adjusting legal frameworks, and core foundational tenets such as the Responsibility to Protect, to address the risks coming from the spread of hate speech and disinformation to social media channels. Nevertheless, the peculiarities of social media require a different approach, and one that includes the responsibilities of social media companies and has at its core, accountable content moderation. Private companies like social media can be both tools of intervention and barriers to intervention. Therefore, IIC could have a crucial role in increasing the degree of proceduralisation of information intervention and avoiding disproportionate interference with states' sovereignty and human rights. There are some limits regarding the role of IIC with regard to participation of stakeholders, the complexity in dealing with escalation, and the effectiveness of its guidelines. However, the establishment of such a system, within regional or international bodies, would increase global awareness while providing a framework to address the spread of online hate and disinformation escalating offline harms including genocide and ethnic cleansing.

References

Allen, T., & Stremlau, N. (2005). *Media policy, peace and state reconstruction* (Discussion Paper No. 8). Crisis States Research Centre, London School of Economics and Political Science. http://eprints.lse.ac.uk/28347/

Article 19. (1996). *Broadcasting genocide Censorship, propaganda and state-sponsored violence in Rwanda, 1990-1994*.

Article 19. (2019). *The Social Media Councils: Consultation Paper*. https://www.article19.org/wp-content/uploads/2019/06/A19-SMC-Consultation-paper-2019-v05.pdf

Balkin, J. M. (1999). Free speech and hostile environments. *Columbia Law Review*, *99*(8), 2295–2320. https://doi.org/10.2307/1123612

Barrett, P. M. (2020). *Who Moderates the Social Media Giants? A Call to End Outsourcing NYU STERN* [Report]. NYU Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/tech-content-moderation-june-2020

Barrie, S. (2019, December 19). Mass Atrocities in the Age of Facebook—Towards a Human Rights-

Based Approach to Platform Responsibility [Blog post]. *OpinioJuris*. http://opiniojuris.org/2019/12/16/mass-atrocities-in-the-age-of-facebook-towards-a-human-rights-based-approach-to-platform-responsibility-part-one/

Bellamy, A. J. (2014). *The Responsibility to Protect: A Defence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198704119.001.0001

Blinderman, E. (2002). International Law and Information Intervention. In M. Price & M. Thompson (Eds.), *Forging Peace: Intervention, Human Rights and the Management of Media Space* (pp. 104–138). Edinburgh University Press. https://www.jstor.org/stable/10.3366/j.ctvxcrszn.7

Bloch-Wehba, H. (2019). Global platform governance: Private power in the shadow of the state. *SMU Law Review*, *72*(1), 27–80. https://scholar.smu.edu/smulr/vol72/iss1/9/

Carillo-Santarelli, N. (2017). *Direct International Human Rights Obligations of non-State Actors: A Legal and Ethical Necessity*. Wolf Legal Publishers.

Chinkin, C., & Kaldor, M. (2017). *International law and new wars*. Cambridge University Press.

Clapham, A. (2006). *Human Rights Obligations of Non-State Actors*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199288465.001.0001

Clark, J., Faris, R., Morrison-Westphal, R., Noman, H., Tilton, C., & Zittrain, J. (2017). *The Shifting Landscape of Global Internet Censorship* [Research Publication]. Berkman Klein Center for Internet & Society Research Publication. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33084425

Davis, M. C. (2015). *International Intervention in the Post-Cold War World*. Routledge. https://doi.org/10.4324/9781315498171

De Gregorio, G. (2018). From constitutional freedoms to the power of the platforms: Protecting fundamental rights online in the algorithmic society. *European Journal of Legal Studies*, *11*(2), 65–103. http://ejls.eui.eu/wp-content/uploads/sites/32/2019/05/4-EJLS-112-De-Gregorio.pdf

De Gregorio, G., & Stremlau, N. (2020). Internet shutdowns and the limits of law. *International Journal of Communication*, *14*, 4224–4243. https://ijoc.org/index.php/ijoc/article/view/13752

Dinwoodie, G. B. (Ed.). (2017). *Secondary liability of internet service providers*. Springer International Publishing. https://doi.org/10.1007/978-3-319-55030-5

Efrony, D., & Shany, Y. (2018). A rule book on the shelf? Tallinn manual 2.0 on cyberoperations and subsequent state practice. *American Journal of International Law*, *112*(4), 583–657. https://doi.org/10.1017/ajil.2018.86

Erni, J. N. (2009). War, 'incendiary media' and international human rights law. *Media, Culture & Society*, *31*(6), 867. https://doi.org/10.1177/0163443709343792

Facebook. (2019). *Draft Charter: An Oversight Board for Content Decision*. Facebook Newroom US. https://fbnewsroomus.files.wordpress.com/2019/01/draft-charter-oversight-board-for-content-decisions-1.pdf

Farrior, S. (2002). Hate Propaganda and International Human Rights Law. In M. Price & M. Thompson (Eds.), *Forging Peace: Intervention, Human Rights and the Management of Media Space* (pp. 69–103). Edinburgh University Press. https://www.jstor.org/stable/10.3366/j.ctvxcrszn.6

Fisher, M. (2019, April 21). Sri Lanka Blocks Social Media, Fearing More Violence. *The New York Times*. https://www.nytimes.com/2019/04/21/world/asia/sri-lanka-social-media.html

Floridi, L., & Tadeo, M. (Eds.). (2017). *The Responsibilities of Online Service Providers*. Springer. https://doi.org/10.1007/978-3-319-47852-4

Frowein, J. A., & Krisch, N. (2002). Introduction to chapter VII. In *The Charter of the United Nations* (pp. 701–716). Oxford University Press.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Glanville, L. (2013). *Sovereignty and the responsibility to protect: A new history*. University of Chicago Press. https://doi.org/ 10.7208/chicago/9780226077086.001.0001

Guiding Principles on Business and Human Rights. (2011).

Henley, J. (2018, June 22). Algeria blocks internet to prevent students cheating during exams. *The Guardian*. https://www.theguardian.com/world/2018/jun/21/algeria-shuts-internet-prevent-cheating-school-exams.

Herzstein, R. E. (1978). *The war that Hitler won: The most infamous propaganda campaign in history*. Putnam Publishing Group.

Holzgrefe, J. L., Keohane, R. O., & Tesón, F. R. (2003). *Humanitarian Intervention: Ethical, Legal and Political Dilemmas*. Cambridge University Press.

Independent International Commission Kosovo. (2002). *Kosovo Report: International Responses, Lessons Learned*. Oxford University Press. https://doi.org/10.1093/0199243093.001.0001

International Commission on Intervention and State Sovereignty. (2001). *The Responsibility to Protect: Report of the International Commission on Intervention and State Sovereignty*. International Development Research Centre.

Irving, E. (2019). Suppressing Atrocity Speech on Social Media. *AJIL Unbound*, *113*, 256–261. https://doi.org/10.1017/aju.2019.46

Kaye, K. (2019). *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global.

King, F. P. (1996). Sensible Scrutiny: The Yugoslavia Tribunal's Development of Limits on the Security Council's Powers Under Chapter VII of the Charter. *Emory International Law Review*, *10*, 509.

Klonick, K. (2018). The New governors: The People, rules, and processes governing online speech. *Harvard Law Review*, *131*, 1598–1670. https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

Knox, J. H. (2008). Horizontal Human Rights Law. *American Journal of International Law*, *102*(1), 1–47. https://doi.org/10.1017/S0002930000039828

Lafraniere, S. (2003). Court Finds Rwanda Media Executives Guilty of Genocide. *New York Times*. https://www.nytimes.com/2003/12/03/international/africa/court-finds-rwanda-media-executives-guilty-of-genocide.html,

Larson, A., & Whitton, B. (1963). *Propaganda towards Disarmament in the War of Words*. World Rule of Law Center, Duke University; Oceana Publications.

Manila Principles on Intermediary Liability and the DCPR Best Practices on Platforms' Implementation on the Right to Effective Remedy. (2017). https://www.intgovforum.org/multilingual/index.php?q=filedepot_download/4905/1550

Marchant, E., & Stremlau, N. (2020a). A Spectrum of Shutdowns: Reframing Shutdowns from Africa. *International Journal of Communication*, *14*, 4327–4342. https://ijoc.org/index.php/ijoc/article/view/15070

Marchant, E., & Stremlau, N. (2020b). The Changing Landscape of Internet Shutdowns in Africa – Introduction. *International Journal of Communication*, *14*, 4216–4223. https://ijoc.org/index.php/ijoc/article/view/11490

Metzl, J. F. (1997a). Information intervention: When switching channels isn't enough. *Foreign Affairs*, 15–20.

Metzl, J. F. (1997b). Rwandan genocide and the international law of radio jamming. *American Journal of International Law*, *91*(4), 628–651. https://doi.org/10.2307/2998097

Miles, T. (2018, March 12). UN investigators cite Facebook role in Myanmar crisis. *Reuters*. https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN

Nicaragua v United States, (International Court of Justice 1986).

Öberg, M. D. (2005). The legal effects of resolutions of the UN Security Council and general assembly in the jurisprudence of the ICJ. *European Journal of International Law*, *16*(5), 879–906. https://doi.org/10.1093/ejil/chi151

O'Neil, P. (2013, September 18). *Why the Syrian uprising is the first social media war*. http://www.dailydot.com/politics/syria-civil-social-media-war-youtube/

Patrikarakos, D. (2017). *War in 140 characters: How social media is reshaping conflict in the twenty-first century*. Hachette UK.

Perrigo, B. (2019, October 23). Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch. *Time*. https://time.com/5739688/facebook-hate-speech-languages/.

Preamble of Radio Regulations. (2016).

Price, M. E., & Thompson, M. (Eds.). (2002). *Forging peace: Intervention, human rights, and the management of media space*. Edinburgh University Press. https://www.jstor.org/stable/10.3366/j.ctvxcrszn

Protocol on Amendments to the Protocol on the Statute of the African Court of Justice and Human Rights. (2014).

Rabat Action Plan. (2013).

Rajadhyaksha, M. (2006). Genocide on the Airwaves: An Analysis of the International Law Concerning Radio Jamming. *Journal of Hate Studies*, *5*(1). https://doi.org/10.33972/jhs.43

Reinisch, A. (2005). The Changing International Legal Framework for Dealing with Non-State Actors. In P. Alston (Ed.), *Non-State Actors and Human Rights*. Oxford University Press.

Santa Clara Principles on Transparency and Accountability in Content Moderation. (2018).

Schlein, L. (2018, June 2). Hate Speech on Social Media Inflaming Divisions in CAR. *Voice of America*. https://www.voanews.com/africa/hate-speech-social-media-inflaming-divisions-car.

Shen, J. (2001). The non-intervention principle and humanitarian interventions under international law. *International Legal Theory*, *7*.

Stecklow, S. (2018, August 15). Why Facebook is losing the war on hate speech in Myanmar. *Reuters*. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

Stremlau, N. (2018). *Media, Conflict, and the State in Africa*. Cambridge University Press. https://doi.org/10.1017/9781108551199

Teubner, G. (2006). The Anonymous Matrix: Human Rights Violations by 'Private' Transnational Actors. *The Modern Law Review*, *69*(3), 327–346. https://doi.org/10.1111/j.1468-2230.2005.00587.x

Thomas, G. (1999). NATO and International law. *ON LINE Opinion*. https://www.onlineopinion.com.au/view.asp?article=1647&page=0

Tufekci, Z. (2018). YouTube, the great radicalizer. *The New York Times*. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

United Nations. (1945). *Charter of the United Nations*.

United Nations Convention on the Prevention and Punishment of the Crime of Genocide. (1948).

United Nations General Assembly. (1965). *Declaration on the Inadmissibility of Intervention in the Domestic Affairs of States and the Protection of their Independence and Sovereignty, GA Res. 2131/XX*.

United Nations General Assembly. (1970). *Declaration on Principles of International Law concerning Friendly Relations and Co-operation among States in accordance with the Charter of the United Nations (GA Res. 2625 (XXV))*.

United Nations General Assembly. (2005). *Resolution 60, 2005 World Summit Outcome (A/RES/60/1)*.

United Nations General Assembly. (2009). *The responsibility to protect (A/RES/63/308)*.

United Nations General Assembly. (2013). Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of. *International Security*, *68*(9824). https://undocs.org/A/68/98.

United Nations Guiding Principles on Business and Human Rights. (2011).

United Nations Human Rights Council. (2018). *Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar*. https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_CRP.2.pdf

United Nations Human Rights Council, Report of the Special Rapporteur on minority issues, Fernand de Varennes. (2021). https://undocs.org/A/HRC/46/57

United Nations International Covenant on Civil and Political Rights. (1966).

United Nations Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. (2018).

United Nations Report of the Special Rapporteur to the Human Rights Council on online content regulation. (2018).

United Nations Security Council, Resolution 1952. (2010).

United Nations Security Council, Resolution 2354. (2017).

United Nations Universal Declaration of Human Rights. (1948).

Uyheng, J., & Carley, K. M. (2020). Bots and online hate during the COVID-19 pandemic: Case studies in the United States and the Philippines. *Journal of Computational Social Science*, *3*, 445–468. https://doi.org/10.1007/s42001-020-00087-4

Varis, T. (1970). The Control of Information by Jamming Radio Broadcasts. *Cooperation and Conflict*, *5*(3), 168–184. https://doi.org/10.1177/001083677000500303

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (Report DGI(2017)09). Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Whittle, D. (2015). The Limits of Legality and the United Nations Security Council: Applying the Extra-Legal Measures Model to Chapter VII Action. *European Journal of International Law*, *26*(3), 671. https://doi.org/10.1093/ejil/chv042

Wu, F. T. (2011). Collateral Censorship and the Limits of Intermediary Immunity. *Notre Dame Law Review*, *87*, 293.

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya