VERITÉ RESEARCH
Strategic Analysis for Asia

BETTER MODERATION OF HATE SPEECH ON SOCIAL MEDIA

A SRI LANKAN CASE STUDY FOR REPUTATIONAL-COST APPROACHES

STRATEGY BRIEF

July 2021

BETTER MODERATION OF HATE SPEECH ON SOCIAL MEDIA

A SRI LANKAN CASE STUDY FOR REPUTATIONAL-COST APPROACHES

The Argument for a Reputational-Cost Approach – Drawing on a Sri Lankan Case Study

July 2021

The Legal Team of Verité Research compiled this study. The principal authors are Uween Jayasinha and Nishan de Mel. Research support was provided by Sierra Amarasiri and Shamana Amjah. The authors wish to thank Gehan Gunatilleke for his continuous contributions in brainstorming and review and Nishana Weerasooriya for research management. Stefan Theil, Michael Wiener shared valuable observations and feedback on an early version of this paper.

Verité Research aims to be a leader in the provision of information and analysis for negotiations and policy making in Asia, while also promoting dialogue and education for social development in the region. The firm contributes actively to research and dialogue in the areas of economics, sociology, politics, law, and media, and provides services in data collection, information verification, strategy development and decision analysis.

Email comments to: publications@veriteresearch.org

ISSN: 2806-5360

Other research briefs can be downloaded at: https://www.veriteresearch.org/publication_type/research-briefs/

Contents

List of Tables and Boxes

1
Overview

The spread of disinformation and hate speech on social media is a recurring issue of concern in the global digital space. From ethnic riots in South Asia to the recent mob invasion of the United States Capitol, acts of offline violence have been disturbingly linked to disinformation and hate speech on social media. In Sri Lanka, inflammatory, discriminatory, and deceptive content regularly forms part of political discourse on social media platforms. Such content has, on occasion, advocated and incited hatred and may have even been catalytic in triggering widespread communal violence. Although mainstream media has a wider reach in Sri Lanka, social media has been growing in influence with regard to socio-political discourse. Given the increasing reach and influence of social media, it is prudent to think ahead about durable policies that can address issues of hate speech and disinformation on social media.

1.1 The two-fold problem of content moderation

At present, the social media space is predominantly regulated by internal mechanisms involving voluntary self-enforcement of platform community standards. However, social media service providers have faced significant criticism with regard to the quality and adequacy in enforcing compliance with these self-made community standards. Many governments have attempted to address the possible weaknesses in self-regulation by enacting laws that require service providers to moderate content in a manner prescribed by the government. Such laws have in turn attracted significant criticism for advancing inappropriate censorship and restricting free speech.

In Sri Lanka as well, there have been many instances where social media service providers were ineffective in moderating hate speech and disinformation, particularly leading up to the ethno-religious riots in Ginthota (in 2017) and Ampara (in 2018).[1] Although Sri Lanka does not currently have laws that impose legal obligations on service providers, Sri Lanka has laws that prohibit hate speech, such as the International Covenant on Civil and Political Rights Act and the Prevention of Terrorism Act. However, these laws have been abused by successive governments to silence free speech and political dissent, as is evident by, for instance, the recent arrests of lawyer Hejaaz Hizbullah and author Shakthika Sathkumara.[2]

Therefore, in Sri Lanka, as in many other jurisdictions, the two-fold problem of social media content moderation is: (1) the inadequate efficacy of voluntary content moderation by social media companies; and (2) the overreach and abuse of power by government, when it forays into the domain of regulating content moderation. The question of finding a path between these two problems continues to bedevil the discussion on social media content moderation.

1.2 The reputational-cost solution

A trend that has been observed in the global social media landscape is that user activism, through actions such as public condemnation and boycotting, has pushed service providers to actively review their policies. Societal engagement of this nature can significantly affect a service provider's reputation. A reputational-cost has the potential to translate into a loss of societal "good-will", and results in risks to revenue growth of social-media platforms. Thus, service providers have a significant motivation to respond to the risk of reputational-cost brought about by societal engagement on shortfalls in content moderation, more so than if the task is left to voluntary due-diligence.

This strategy brief explores the regulatory framework of Sri Lanka's social media space and presents an approach to enhance the quality of content moderation through societal initiatives that leverage the potential of reputational-cost to result in better content moderation by social media companies. It is based on the idea that accountability is important for improving content moderation by social media companies, and that risks created by horizontal accountability structures that sanctioned more power to government can be reduced if there were better means of generating vertical accountability to society. Such vertical accountability can be generated through more direct social engagement initiatives.

The research on which this strategy brief is based draws significantly from secondary sources of information that are publicly accessible. This strategy brief is divided into four sections: i) an explanation of the concepts of hate speech and disinformation; ii) a discussion on existing models of content moderation; iii) a case study, based on Sri Lanka's past experiences on the spread of hate speech and disinformation on social media, which demonstrates why existing models of content moderation are inadequate; and iv) a discussion on how reputational-cost approaches can be effective, and the role civil society plays in such approaches.

The development of this strategy brief is centred primarily around the case study of the Sri Lankan experience on Facebook, YouTube and Twitter between the period from 2014 to 2020. As such, reference to the term 'social media' in this strategy brief is limited to Facebook, YouTube and Twitter, as these three platforms feature the highest usage and engagement in Sri Lanka.[3]

2
Why Hate Speech and Disinformation?

A wide range of behaviours ranging from misinformation to sexual predation fall within the larger category of, 'problematic-behaviours' on social media. These problematic-behaviours have been recognised along many dimensions and types of responses to social media. This brief focuses on two classes of problematic-behaviours; (1) hate speech; and (2) disinformation. Both of these problems are often linked to one another.

Hate speech, as a concept, has no universally accepted definition. Generally, speech that is merely offensive or meant to humiliate does not constitute hate speech. Nevertheless, three forms of speech can be identified as speech that falls within the scope of 'hate speech' (in increasing levels of concerns): (1) speech that conveys hatred to a person or group based on an identity (such as calling a religious community disgusting); (2) speech that not only conveys hatred but also persuades others to act in a harmful manner (such as encouraging others to refrain from selling goods to the LGBTQI + community); and (3) speech that incites violence (such as calling on people to attack an ethnic group).[4]

Disinformation, in its simplest form, is harmful information that is: (1) false; and (2) disseminated to mislead or deceive.[5] Disinformation, therefore, must be distinguished from false expressions that are satirical or merely meant to humour, or inadvertent mistakes in communication.

The uncontrolled spread of hate speech and disinformation, in tandem, can fuel communal tensions and intolerance, incite violence, foster distrust in public institutions, and undermine democratic and public processes.[6] Therefore, this brief will focus on the harms brought about by hate speech and disinformation, not just through online interactions, but also through the resulting polarisation of communities and violence that can occur offline.

For the purpose of this strategy brief, every reference to 'hate speech' and 'disinformation' shall be within the parameters discussed above.

3
Current Approach to Content Moderation: Service Provider-Driven & State-Driven Mechanisms

Before delving into the different approaches of content moderation, it is pertinent to discuss the nature of social media service providers and why they undertake content moderation. Social media service providers identify themselves as 'platforms'; that is, a digital medium that allows third-parties (users) to publish content on.[7] As more and more users begin to use a social media platform and have increased interactions, content that is illegal or harmful may also be published on the platform, which could lead to offline-harm. Thus, social media platforms are compelled to undertake a degree of content moderation, even though it involves regulating speech and expression. In these circumstances, current discourse is not *whether* platforms should be held accountable for failures in content moderation, but rather *when* (how and to what extent) platforms should be held accountable.[8]

Social media platforms have voluntarily sought to moderate content with a view to reducing potential harm, and governments and society have an interest in holding social media service providers accountable for effective content moderation in that regard.

This section discusses the three types of approaches that are prevalent in content moderation: 1) self-regulation (internal accountability); 2) government/statutory regulation (horizontal accountability); and 3) societal engagement (vertical accountability).

3.1 Self-regulation (internal accountability)

Self-regulation is a voluntary and internal mechanism adopted by social media service providers to moderate the content posted on their platforms and to safeguard 'their interests in a way that it aligns with the public's interest'.[9] Service providers generally seek to moderate content by: (a) creating community standards; and (b) enforcing the community standards to curb problematic-behaviours.

a. Community Standards

A key instrument implemented by social media service providers to self-regulate are 'Community Standards' that all users are required to follow. Such 'community standards' outline the nature and type of content that is permitted or prohibited on social media platforms.

At present, the community standards of all major social media platforms contain guidelines on monitoring and removing hate speech and disinformation. See Table 1: Facebook, YouTube and Twitter Community Standards and its Enforcement (Community Standards as at 18 March 2021)

Table 1: Facebook, YouTube and Twitter Community Standards and its Enforcement (Community standards as at 18 March 2021)

Platform	Community standard on hate speech and disinformation	Enforcement of community standard
Facebook	▪ Prohibits 'violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation' in relation to 'protected characteristics' of people (race, ethnicity, nationality, religion, caste, gender, and sexual orientation) and vulnerable groups (immigrants, refugees, and asylum seekers).[10] ▪ No clear parameters of what constitutes disinformation (referred to as 'false news' in the community standard).[11]	▪ Hate speech is removed unless it is shared with condemnation or to raise awareness.[12] ▪ There is no clear distinction between misinformation and disinformation and false news is not removed but is brought to the bottom of the user's news feed.[13]
YouTube	▪ Promoting violence or hatred against individuals or groups based on age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin and veteran status.[14] ▪ Information that is 'deliberately trying to deceive or mislead people'.[15]	▪ The enforcement against hate speech and misinformation ranges from a preliminary warning to restricting the content that can be uploaded. Access to other features on YouTube may be restricted or the user's channel may be removed.[16]
Twitter	▪ Prohibits violence against, threatening, or harassing other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.[17] ▪ Prohibits any activity that involves disrupting civic processes (like elections) and sharing synthetic and manipulated media that is likely to cause harm.[18] Which in other words mean disinformation is prohibited.	▪ *Tweet level enforcement* Includes limiting tweet visibility, requiring tweet removal or hiding a tweet while awaiting its removal.[19] ▪ *Direct-Message level enforcement* Includes stopping conversations between a reported violator and the reporter's account or placing a direct message behind a notice.[20] ▪ *Account level enforcement* Includes requiring media or profile edits, placing an account in read-only mode, verifying account ownership or a permanent suspension.[21]

b. Content moderation

Content moderation is the process of monitoring the content published on social media platforms, and evaluating whether such content is harmful.[22] Service providers assess whether content that is shared on their platform complies with their community standards and attempt to remove content deemed to be in violation of those standards, ideally before it becomes 'viral and visible to others'.[23] Content moderation is usually performed through methods such as: (i) artificial intelligence; (ii) in-house content moderators; and (iii) third-party moderators.

Of the content moderation methods used by service providers, machine based "artificial intelligence" techniques (AI) has taken on increasing importance. AI techniques allow machines to quantify, process, structure, and analyse data quickly and cost-effectively on a large scale.[24] AI techniques can also generate insights on patterns, by which harmful content can be rapidly identified and removed.[25] Social media service providers also have their internal content moderation teams that review objectionable content that requires more nuanced review. As a recent tool to combat widespread disinformation, service providers also employ fact-checkers within certain regions to review information and direct the attention of users to fact-checks on disinformation.[26]

3.2 Government/ Statutory regulation (horizontal accountability)

Statutory approaches to content moderation entails imposing obligations on actors to curb hate speech and disinformation and setting out the liabilities for failing to comply with such obligations. There are two broad approaches taken by countries in this regard: (a) enacting laws that impose liability on social media users who spread hate speech and disinformation (user liability); and (b) enacting laws that impose liability on service providers who fail to monitor and remove hate speech and disinformation on their platforms (intermediary liability).

a. User liability

Most countries have laws that prohibit expressions that incite hatred and violence, or the spread of false information that can result in discrimination, or the breakdown of public order.[27] Such laws traditionally impose liability on the individuals responsible for such expressions. For instance, Sri Lanka has several laws that impose such prohibitions, such as the International Covenant on Civil and Political Rights (ICCPR) Act, No.56 of 2007, Prevention of Terrorism (Temporary Provisions) Act, No. 48 of 1979 (PTA), and the Penal Code, No. 2 of 1883, and the Computer Crimes Act, No. 24 of 2007.

Although these laws can effectively punish perpetrators of hate speech and disinformation, and act as a deterrence, they are not effective in terms of removing harmful content before they go viral and possibly inciting discrimination, hatred, or violence. This is primarily because such laws are generally not accompanied with complementary mechanisms that require social media service providers to rapidly take down harmful content. Moreover, governments do not have the requisite access to social media data or control over platforms to take rapid action to stop the spread of any specific content on social media platforms.[28]

b. Intermediary liability

Over the years, social media has transformed from platforms that merely host publications to multi-billion-dollar organisations that employ algorithms and technology to harvest personal data and provide curated and personalised content for users. The use of such algorithms and technology have enabled the phenomenon of "echo-chambers" that amplify certain types of content among a certain group of users.[29] Inflammatory content and disinformation (which are contrary to the Community Standards of social media platforms) can also be amplified by such algorithms and technology, thereby increasing the potential of such content to result in problematic-behaviours, to gain 'viral' traction, and thereby to even catalyse violence.[30]

Therefore, governments have attempted to hold service providers accountable by imposing liabilities on service providers for failing to remove harmful content or by carving-out exceptions to immunities granted to platforms, for liabilities arising from third party publishers. Presently, only a few countries have enacted laws that cast direct obligations and liabilities on service providers in respect of monitoring and removing harmful content. For the purposes of this brief, the approaches taken by the United States, India, and Germany on intermediary liability are set out as a means of focusing the analysis.

i. **United States:** Communications Decency Act of 1996 (CDA)

In the United States, regulating social media services concerns the application of section 230 of the CDA. Section 230 makes a distinction between a service provider of a platform and the users of such a platform.[31] Section 230 grants immunity to a service provider in respect of harm arising from content published by users. In essence, this legal framework is premised on the idea that service providers of platforms (such as social media websites, message boards, and hosting services) are protected from legal claims made in respect of harm arising out of content published by a third-party.

However, this immunity is not without exception. For instance, a service provider is precluded from claiming immunity under section 230 where the platforms are used to commit illegal acts under federal criminal law (such as human trafficking) by users, or where the service provider has 'materially contributed' to unlawful activity or to generating the harmful content.[32]

Section 230 protects the freedom of expression of social media users while also recognising that service providers should be free to regulate their platforms as they see fit.[33] Section 230 places the onus for moderating online speech exclusively in the hands of social media service providers, which does not guarantee fair and uniform content moderation protocols on larger social media platforms. The challenges that are posed by this dichotomy in section 230 are exemplified by the recent controversies concerning Facebook and U.S. politics, as the social media giant has been repeatedly called out for inconsistent content moderation leading to the spread of disinformation and the undermining of U.S. institutions and democratic processes.[34]

ii. **India:** Information Technology Act of 2000 (ITA)

India espouses the 'safe harbour regime' of regulating social media, through the ITA. Similar to the U.S. approach, the ITA grants immunity to social media service providers in relation to unlawful content by third-party users. However, the ITA differs from the U.S. approach as immunity is afforded only if the social media service provider can prove that it has adhered to due diligence requirements that are prescribed in the Information Technology (Intermediary Guidelines) Rules 2011 (IT Rules).[35]

One of the key requirements of the IT Rules is disclosing the internal rules and regulations regarding using and accessing a social media platform. These rules must clearly warn users against publishing content that is 'prohibited' by the IT Rules, which includes content that is 'grossly harmful, obscene, pornographic, libellous, hateful, unlawful'.[36] The IT Rules also require service providers to cooperate with authorised government agencies, to report cybersecurity incidents to the Indian Computer Emergency Response Team, and to appoint a designated Grievances Officer. Service providers that fail to observe these due diligence requirements or conspire, aid, abet or induce unlawful conduct will be disqualified from the safe harbour exemption.[37]

However, section 69 of the ITA vests any officer of the government with the power to order any service provider to 'block for access by the public… any information generated, transmitted, received stored or hosted in any computer resource.' Failure to comply with such an order is an offense punishable by imprisonment and fine.

The ITA framework has received criticism due to the ambiguity of what constitutes 'prohibited' content and the power it vests in the government to order the take down of content.[38] For instance, in 2019 the Committee to Protect Journalists expressed concerns that the

Indian government was using the ITA to suspend Twitter accounts that were disseminating information regarding the disputes in the Jammu and Kashmir regions. In the Indian Supreme Court's judgment *Shreya Singhal v. Union of India* (2015), the court emphasised the obligation cast on service providers under the ITA to remove within 36 hours content deemed to be unlawful by order of the court or a government directive, thereby entrenching the government's broad power to censor content published online through the ITA.[39]

iii. Germany: Network Enforcement Act (NetzDG)

The NetzDG requires that when a social media service provider receives a complaint regarding unlawful content, it must remove such content within prescribed timeframes, i.e., within 24 hours for 'manifestly' unlawful content, and within seven days for content that is not manifestly unlawful. The NetzDG further requires service providers to regularly disclose information on complaint mechanisms and other efforts taken to 'eliminate criminally punishable activity' on their platforms.[40]

Under the NetzDG, content is considered unlawful if it violates any of 22 specified provisions of the German Criminal Code. These 22 provisions correspond to several offenses, including incitement to hatred, and criminal defamation (which is no longer an offense in many countries, due to its implicit chilling effect on critics and commentators). Service providers are vested with the responsibility for interpreting and determining if questionable content violates the provisions of the German Criminal Code. Failure by service providers to adequately monitor and handle complaints or act within the prescribed timeframes can result in steep fines and sanctions.[41]

The United Nations Special Rapporteur on Freedom of Opinion and Expression is among those who have criticised NetzDG. The Rapporteur criticised the NetzDG framework, noting that content could be unilaterally ordered to be taken down without any judicial oversight and, as such, was 'incompatible with Article 19 of the ICCPR'.[42] Human Rights Watch has also noted that with the short timeframes to review and take down unlawful content and the risk of steep fines, social media companies have limited incentive to uphold a user's right to freedom of expression.[43] Further, the NetzDG does not require service providers to accept appeals in respect of removed content. Thus, Facebook, YouTube and Twitter have, declined to set up appeal mechanisms for users who are aggrieved by a unilateral take down of their content.[44]

It must be noted that attempts by countries other than the U.S. to impose intermediary liability have encountered the challenge of extraterritorial application. For instance, representatives of Facebook explicitly defied summons for hearings issued by authorities in the United Kingdom and Canada.[45] Even in India, a summon issued by the Peace & Harmony Committee of the Delhi Legislative Assembly on Ajit Mohan, Vice President of Facebook India, in September 2020 was ignored. The summon was in relation to the failures of Facebook to curb hateful content that facilitated riots that broke out in Delhi in early 2020. Ajit Mohan has presently challenged the legality of such a summon in the Indian Supreme Court, on the basis that such a summon by a federal legislative assembly was a violation of his fundamental rights.[46]

Thus, laws that impose intermediary liability may be toothless without cooperation by service providers that dominate the social media space.

3.3 Societal regulation (vertical accountability)

Arguably, social acceptance is the most important factor for the success of social media platforms. The public perception of a social media service provider proportionately influences the success of that service provider. Service providers require positive public perceptions of their platform to attract and retain users. Retaining a high number of users and engagement allows the service provider to accumulate larger volumes of user data, and explore wider revenue generation methods, such as personalised advertising and selling licenses to access and re-use user data.[47] Therefore, a loss of reputation that results in reducing users or reducing engagement can significantly disrupt the operations of a service provider. See Box 1 for notable recent societal movements against social media service providers.

BOX 1: NOTABLE SOCIETAL MOVEMENTS THAT CAUSED REPUTATIONAL-COSTS ON SOCIAL MEDIA SERVICE PROVIDERS.

In April 2021, premier football clubs and major players in the United Kingdom held a four-day boycott of social media platforms, to show their dissatisfaction with the failings of social media companies to curb racist abuse faced by players. Other regional football organisations also joined the boycott. Both Facebook and Twitter issued statements in response, pledging to improve their content moderation policies.[48]

In July 2020, the 'Stop Hate for Profit' movement commenced against Facebook in the face of mounting evidence that lapses in moderating hate speech were undermining civil rights and racial justice. The 'Stop Hate for Profit' movement resulted in at least 1,200 companies boycotting Facebook and suspending advertisements on the platform.[49] This prompted Facebook to immediately reassess its content moderation policies.

The #LogOutFacebook campaign was an initiative commenced by the National Association for the Advancement of Coloured People (NCCAP) in order to demonstrate opposition against Facebook's 'history of data hacks which unfairly target its users of colour'. As many other organisations and users joined the campaign, Facebook conducted civil rights audits and pledged 'to do more'.[50]

4
Limits of Self-Regulation and Risks of Government Regulation: Sri Lankan Case Study

The argument presented at the outset of this strategy brief states that, it is inadequate to depend on self-regulation by social media platforms in terms moderating content that may proliferate hate speech and disinformation. Furthermore, the sanctioning of government regulation for the moderation of such content can be dangerous. Sections 3 on the Current Approach to Content Moderation: Service Provider-Driven and State-Driven Mechanisms has set out the prevalent global structures in giving effect to these two accountability models. This section unpacks the problems inherent in these two accountability models using Sri Lanka as a case study.

The specific aspects of Sri Lanka that give rise to concern with regard to these two models are, arguably, more the norm than the exception among the nations of the world. The specific experiences of the United States (U.S.) and Nations in the European Union (EU) are likely to be more the exception, than the rule, in terms of how these two accountability models are likely to work out in practice.

Therefore, in reflecting on a globalised approach to better moderate the content on social media – where the powerful actors tend to be more influenced by U.S. and EU experiences - this case study of Sri Lanka could be instructive.

4.1 Content moderation design: The relevance of the socio-political landscape

By serving as a platform for expression and interaction, social media closely mirrors the socio-economic and political sentiments and impulses of a community. While social structures, as they play out in practice, influence the necessity and scope for content moderation, political structures are also important to understand how sanctioning government to engage and enforce content moderation can play out in practice.

These two practical aspects impact the two content moderation models that have been discussed: self-regulation and government-regulation. The Sri Lankan case study points to the limits of self-regulation and the risks of government regulation, by setting out the context and experience in the working out of these two models in Sri Lanka.

The Sri Lankan case highlights two features in particular that can influence the success of the above content moderation models: (i) social polarisation; and (ii) political freedom. The practical relevance of these features in Sri Lanka are explained in turn.

i. Social Polarisation

Sri Lanka, like many other countries in the world, suffers from entrenched tendencies towards social polarisation. In Sri Lanka, the polarisation takes both ethnic and religious dimensions. Ethnic tensions between Sinhala and Tamil communities, for instance, have been a defining feature of the past 7 decades of Sri Lanka's history – underscored by 3 decades of violent conflict that was brought to an end in 2009, without a political resolution of the conflict. In addition to the ethnic based Sinhala-Tamil polarisation, there is also overt religion based polarisation such as, between Buddhists and Christians, as well as Buddhists and Muslims.

Therefore, ethnic and religious polarisation discourses have long dominated the political process in Sri Lanka. Several of Sri Lanka's prominent political parties and groups are founded on or are perceived as exclusively serving the interests of an ethno-religious group or a specific demographic of the population.

Studies on the interplay between social polarisation and social media indicate that social media has the potential to exacerbate prevailing social issues.[51] This is true of Sri Lanka, which has experienced instances where inflammatory content on social media has resulted in widespread ethno-religious violence. For instance, racially, ethnically, and religiously inflammatory content that exacerbates prevailing tensions is frequently propagated on social media. In several instances, social media was weaponised to propagate hate speech and disinformation to the extent that it resulted in widespread violence:

a. In 2014, inflammatory speeches against the Muslim community by the Buddhist militant group *Bodu Bala Sena* at a rally sparked wide-spread communal violence in the towns of Aluthgama, Beruwala, and Dharga.[52] Although the government imposed a curfew and media-blackout in these areas, social media was used to spread disinformation and inflammatory content to aggravate the situation.[53]

b. In 2017, a dispute between a Sinhalese and Muslim group in Gintota escalated into communal violence in the area, with several houses owned by Muslims and a mosque being attacked by a mob. Reports claim that violence escalated due to disinformation and hate speech propagated on social media.[54]

c. In 2018, a video allegedly containing footage of a Muslim restaurant owner in Ampara being confronted for serving food that contained 'sterilisation pills' was widely circulated on social media. Over the next few days, disinformation and hate speech was propagated on social media platforms to create a false impression that there was a Muslim conspiracy to sterilise the Sinhala population, leading to violence in Ampara.[55] In the following weeks, a separate incident triggered further anti-Muslim violence and riots in Digana. These riots were coordinated using social media platforms and resulted in widespread destruction of private property belonging to Muslims.[56]

d. In 2018, the leader of the extremist group National Thowheed Jamath (NTJ) shared personalised content on social media that called for the use of explosives to target non-Muslims in Sri Lanka.[57] In April 2019, the NTJ would go on to perpetrate the Easter Sunday terrorist attack, which resulted in the death of over 250 people.

e. Following the Easter Sunday attacks, anti-Muslim sentiments escalated, and several rumour-based news reports of further attacks and conspiracies of 'Muslim expansionism' began circulating on social media. In May 2019, these incidents culminated in several acts of violence and riots in areas such as Chilaw, Minuwangoda, Kurunegala and Kandy.[58]

Another key instance of polarisation can be seen through the instrumentalisation of elections. Politicians regularly amplify prevailing social and economic tensions to bolster political support or discredit the opposition. In the emerging media context, social media has become a significant means of political communication, which in turn is amplifying these existing polarisations within society. The following are some manifest examples of the relationship between social media engagement and the political amplification of polarisations in Sri Lanka:

f. The 2019 Presidential Election and 2020 General Election witnessed several ethno-nationalistic campaigns. Independent social media monitors have observed that these elections were highly polarised and were chequered with anti-Muslim propaganda, attacks against politicians from minority ethnicities, attempts to tarnish the Election Commission, and gender-based harassment of candidates.[59]

g. Several prevailing social and ethnic tensions were exacerbated by the COVID-19 pandemic. In particular, the policy of the government to cremate the remains of any person who died as a result of COVID-19 disregarded religious beliefs that prohibit cremation. This policy was implemented despite religious beliefs against forced cremation, which heightened ethnic tensions. The outcry against this government policy was countered on social media with hate speech and racial attacks against the opponents of the policy.[60]

The Sri Lankan case study corroborates the concern that a context of such social polarisation, and the political instrumentalisation of social polarisation, creates fertile ground for online hate speech and disinformation on social media platforms to translate into off-line harm. When an increasing number of users reflect the problematic-behaviours of engaging with and sharing hate-speech and disinformation, it can overwhelm the ability of others to report and provide moderating responses; and the normal resource allocation by platforms on content moderation can prove to be inadequate.

ii. Political Suppression of the Freedom of Speech & Expression

Sri Lanka, like many other countries in the world, suffers from a tendency towards the over-reach of political power to supress the freedoms of expression – particularly when it relates to political dissent. Overtime, as in other jurisdictions, these tendencies have advanced to encompass online media as well. However, a particularly unique feature in Sri Lanka is that, increasingly, it is the laws that have been promulgated to constrain hate speech and disinformation that are being used as the instruments for suppressing freedom of expression.

In the past, the government had caused the blocking of access to several alternate news pages and citizen journalism websites that were critical of government policy.[61] For instance, the LankaNews website was blocked under the direction of the Telecommunication Regulatory Commission of Sri Lanka (TRCSL) in 2017.[62] More recently, activists, journalists, bloggers and ordinary citizens have been apprehended and harassed by law enforcement for expressing dissent or being critical of government actors on social media.[63] For instance the arrest of author Shakthika Sathkumara under the ICCPR Act and arrest of poet Ahnaf Jazeem under the Prevention of Terrorism Act are two such examples, where provisions of the law relating to hate speech have been used as instruments of suppressing the freedom of expression.[64] Such abuse and misapplication of the existing legal framework concerning hate speech laws continues to be a prevalent phenomenon.

4.2 The Sri Lankan experience of content moderation

Sri Lanka does not have laws that impose direct obligations on social media service providers. Social media regulation in Sri Lanka is primarily driven by two approaches: (1) self-regulation by service providers (internal accountability), and (2) legislative regulation on user liability (horizontal accountability).

However, on several occasions, communal violence has been attributed to hate speech on social media and freedom of speech has been restricted by misapplication of anti-hate speech laws. This section will discuss the Sri Lankan experience in respect of the current approaches to content moderation and demonstrate why these approaches have failed to serve the public interest in Sri Lanka.

4.2.1 Self-regulation: Insufficient and undermined by systemic pitfalls

The Sri Lankan digital space is miniscule in comparison to other countries. As such, there is a dearth of data on the self-regulation mechanisms employed by service providers to address issues of hate speech and disinformation in Sri Lanka. This lack of data is exacerbated by low levels of transparency by service providers in respect of the inner workings of content moderation on their platforms.[65]

In response to the 2018 communal riots in Sri Lanka, Facebook commissioned a study on the Human Rights Impact Assessment (HRIA) in relation to the Facebook platform. The study confirmed many of the issues that had already been highlighted by local experts and organisations.[66] While this HRIA was specific to the Sri Lankan context on Facebook, the findings therein can be applicable to other contexts and other social media service providers as well. The main issues identified include:

a. Language processing constraints:

A large volume of content generated in Sri Lanka is in Sinhala and Tamil. It was observed that these languages 'lack the digital lexicon required for computational analysis'.[67] Thus, service providers are inherently constrained in analysing content generated in Sinhala and Tamil, particularly when such content has historic and cultural innuendo. The unregulated spread of hate speech and disinformation that served as the precursor to the 2018 anti-Muslim riots has been attributed to the lack of language and cultural expertise by Facebook.[68] In a separate report, it has been highlighted that the AI employed by service providers lacks the capacity to detect local language texts on images and videos with overlaid text.[69] This issue further restricts effective monitoring of harmful content in local languages, particularly in light of the prevailing 'meme' culture and increase in visual-based content.

b. Engagement-driven motivations:

Service providers rely on increased engagement by users to make their platforms profitable. It was revealed that Facebook had employed digital algorithms that drive more engagement on its platform 'regardless of the veracity or intention of the content' that the user consumes.[70] The use of such tools to drive engagement can directly undermine the curbing of hate speech and disinformation.

c. Lack of due diligence:

As noted earlier, service providers have formulated and implemented community standards to set out the boundaries of acceptable conduct. Thus, users are expected to be guided by these community standards in their social media interactions. However, it was reported that these guidelines and processes have poor accessibility on their respective social media platforms, as they are not freely accessible in local languages.[71] It was recently reported that officials within social media companies had selectively applied community standards to members of different political groups in India.[72] Thus, questions in respect of the uniform application of community standards have also arisen.

d. No specified take-down times:

In both the local and global context, service providers across the industry do not undertake to resolve complaints pertaining to hate speech and disinformation within a specific period and have frequently failed to remove such content.[73] Online content can be rapidly circulated and duplicated, and delays in taking down such content can result in dire consequences. The Facebook HRIA specifically revealed how Facebook had been 'largely unresponsive' to numerous complaints against hate speech made by CSOs in Sri Lanka.[74] For instance, Hashtag Generation[75] has specially called on social media service providers to 'declare a maximum response time and an average response time for content reported as potential violations of their respective community standards/ guidelines'.[76] Even on Twitter, there have been reports of anti-Semitic content published by high profile users, being online for several hours and circulated widely before being taken down.[77]

e. Lack of transparency:

Many aspects of a service provider's self-regulation measures remain undisclosed. For instance, many service providers have not disclosed the procedures or standards adopted to evaluate whether content should be taken down.[78] Moreover, the reasons for not refusing to take down content are not always disclosed to the complainant or the creator of the content in question.[79] These issues are compounded by misleading reporting by social media platforms. For example, Facebook's reports in 2020 failed to highlight that approximately one million posts containing hate speech went undetected and made no mention of the number of hours these undetected posts remained online.[80] Thus, there are serious gaps in the transparency of self-regulation mechanisms, that call into question their impartiality and effectiveness.

The exposure of these failures in self-regulation resulted in a serious backlash against service providers. For instance, Facebook received significant backlash after it was revealed that gaps in its self-regulation processes served as a precursor to the escalation of communal tensions and subsequent violence in Sri Lanka and other parts of Asia.[81] This prompted service providers to overhaul their content moderation protocols globally,[82] which has reportedly increased their capacity to remove hate speech and harmful content.[83]

Although these measures are commendable, observers have pointed out that hate speech and disinformation continue to be prevalent in the Sri Lankan social media space.[84] It was reported that the present trends in hate speech and disinformation are similar to the trends that have been reported previously.[85] This continuing trend of hate speech and disinformation, despite the implementation of reforms by service providers, suggests that service providers are not as yet sufficiently incentivised to develop effective solutions to local context based challenges that exist when moderating content in non-mainstream languages that relate to countries with smaller populations.

4.2.2 Laws on user liability: Selective and perverse application

Sri Lanka does not have laws that contemplate intermediary liability or impose obligations on service providers to drive transparency. However, Sri Lanka has several laws that prohibit the publishing and propagation of hate speech and disinformation in the media space. These laws do not specifically deal with content moderation on social media, yet their scope is wide enough to extend to such content. Nevertheless, there is a history of subjecting these laws to selective and perverse application.

Such a history of selective and perverse application is most evident in the case of implementing the ICCPR Act, which prohibits publications that incite discrimination, hostility or violence.[86]

In the aftermath of anti-Muslim violence and other racially charged incidents that were sparked by hate speech and disinformation on social media, it was expected that laws such as the ICCPR Act would be invoked to hold the perpetrators to account. Although a few selective arrests were made in this regard, the most notable suspects in inciting violence were neither charged nor arrested under the ICCPR Act.[87]

At the same time, the ICCPR Act was invoked perversely to arrest and subject targeted persons to prolonged detention for expressions that allegedly offended the Sinhala-Buddhist community, Buddhist clergy, or Buddhism itself.[88] For example, the arrest of author Shakthika Sathkumara for posting a fictional short story insinuating sexual abuse by a Buddhist monk, and for discussing the sexuality of the Buddha; the arrest of Ramzy Razeek for sharing content on an 'ideological jihad'; and the threats of legal action against journalist Kusal Perera for publishing a column against anti-Muslim violence, stand out as clear examples of how the perverse application of the ICCPR Act has had serious implications on the freedom of expression.[89]

Other laws, such as the Penal Code, the Computer Crimes Act, the Press Council Law and the PTA, seek to prohibit the spread of misinformation and publications that can potentially harm public order, communal, religious or racial feelings. However, these laws have also been perversely misapplied in several instances. In 2017, the Supreme Court held that the Penal Code was misapplied against a British tourist who was arrested on arrival in Sri Lanka on the allegation that she had intended to wound religious feelings by having a tattoo of Lord Buddha.[90] More recently, at least two individuals who had been critical of government response to the COVID-19 pandemic on social media were arrested under the Penal Code and the Computer Crimes Act for allegedly spreading false information.[91] Other examples include, the arrests and charging of lawyer Hejaaz Hizbulla and researcher Dilshan Mohammed under the PTA for alledgedly spreading of hate speech, and the arrest and imprisonment of editor J.S. Thissanayagam in 2008 for allegedly inciting communal disharmony by accusing the armed forces of committing war crimes.[92] These incidents have all been severely criticised for their chilling effect on free speech. Emergency regulations issued by various presidents under the Public Security Ordinance, such as the regulations issued by President Maithripala Sirisena following the Easter Sunday Attacks[93], have also received criticism as being excessive in restricting individual freedoms.[94]

Laws setting up regulatory frameworks have also come into question in the recent past. The Telecommunications Regulatory Commission of Sri Lanka (TRCSL), established by the Sri Lanka Telecommunications Act, was billed as an independent regulator with broad powers in respect of the media space.[95] However, in 2011 and again in 2017, it was revealed that the TRCSL had directed all internet service providers to block access to alternate news websites that were critical of the regime at the time.[96] The fact that several of these websites were blocked due to a directive issued by the president of Sri Lanka cast serious aspersions on the independence of the TRCSL.[97] As such, the politicisation of the TRCSL has been noted to undermine the independence of the Commission.[98]

The excesses in government regulation and restriction of free speech are exemplified by the circumstances that led to the adoption of the 2018 Colombo Declaration on Media Freedom and Social Responsibility by key media stakeholders in Sri Lanka. This Declaration stressed on the arbitrariness of restrictions imposed on the media and the increases in threats and intimidation of journalists and called for reforms to strengthen free speech and accessibility to information platforms.[99]

4.3 Overall observations on content moderation from the Sri Lankan case study

The Sri Lankan case highlights two features in particular that can influence the success of the above content moderation models: (i) social polarisation; and (ii) political freedom. In light of the above analysis, there are four key observations that might be generalised regarding online content moderation and social media regulation based on the Sri Lankan case study.The Sri Lankan case also demonstrates that the reputational-cost that arose for social media platforms, from the manner in which events in Sri Lanka received global attention, led to significant improvements in content moderation.

Limits of the self-regulation model especially in the context of social polarisation.

i. Self-regulation frameworks have inherent gaps and weaknesses – fuelled by limitations in regulating content in non-mainstream languages, contrary incentives of social media providers, and the lack of means by small nations to drive accountability of global social media platforms for proper self-regulation – that limit the capacity for adequate monitoring and removal of hate speech and disinformation by social media service providers.

ii. The context of social polarisation undermines even the limited efficacy of the self-regulation model – because polarisation can result in the preponderance of user responses working to promote hate speech and disinformation in social media content, rather than assisting in its proper regulation.

Risks of the government regulation model, especially in the context of weak institutions.

iii. When social polarisation exists in the context of institutions that are not adequately independent from political influence, hate speech and misinformation can be politically instrumentalised in election campaigns. This can occur through the legal and regulatory powers of government being misapplied to take down content that is unfavourable to government and silence opposition voices and supporters, while simultaneously enabling content and users supporting the government to drive political campaigns through hate speech and misinformation.

iv. The context of weak political freedoms increases the risk of laws that are meant to protect against hate speech being selectively and perversely applied by governments to restrict free speech and criticism of the government, instead of regulating harmful content.

Hate speech and disinformation that can instigate communal and political tensions continue to be prevalent within the social media space. In the Sri Lankan case, the failures in content moderation, in the context of engagement-driving algorithms, were found to have contributed to instigating communal violence, such as the 2017 and 2018 anti-Muslim riots. The events in Sri Lanka led to a global discussion and concern with regard to the inadequacy of content moderation by the relevant social media platforms. The global coverage and discussion on the failure of content moderation in Sri Lanka resulted in significant reputation damage to the Facebook platform, which in turn also resulted in a series of proactive measures that were taken by the platform to improve the effectiveness of content moderation in relation to Sri Lanka specifically as well as more generally in the algorithmic curation of platform content.

5
The Solution of a Reputational-Cost Approach

In the midst of the growing consensus that platforms should be held accountable for failures in content moderation, the focal question is how such accountability can be instrumentalised. This strategy brief draws on the case study of Sri Lanka to suggest that the appropriateness of the methods used to drive accountability on content moderation is sensitive to the country context. Both self-regulation (by social media service providers) and government-regulation may be effective in certain social and political contexts. However, these approaches have proven to be (and are likely to remain) ineffective or constitute potentially detrimental approaches in contexts such as set out by the case study on Sri Lanka.

The Sri Lankan case study is not likely to be reflective of the dynamics of content regulation by the United States (U.S.) and the European Union (EU), which have exceptional institutions and leverage. However, the Sri Lankan case study might be typical of a majority of the nations in the world in terms of how the accountability models of self-regulation and government-regulation are likely to work out in practice. Therefore, in adopting a globalised approach to better moderate hate speech and disinformation in social media, the strategic insight that is drawn from the Sri Lankan case study is that adopting a reputational-cost approach could be an important and effective part of the solution to the existing challenges in content moderation.

5.1 The strategic case for a reputational-cost approach

Social media service providers are continually deploying better technology and algorithms that are designed to enhance the curation of content digested by users on social media platforms.[100] By deploying such technology and algorithms, social medial service providers drive greater interaction and engagement. Thus, social media is becoming increasingly central to personal and business relations. However, as individual and organisational engagement on social media increases, the social reputation of social media platforms becomes even more important for service providers in terms of retaining engagement and generating revenue.

Relevance of a reputational-cost approach

In designing approaches to instrumentalise accountability for better content moderation, one method adopted by governments has been to hold social media service providers to higher standards by enacting intermediary liability laws that penalise service providers for lapses in content moderation. In the best cases, these laws are controversial due to the risk of excessively curtailing free speech. However, in contexts that would be akin to the Sri Lankan case study, intermediary liability laws can be far more problematic for two key reasons. First, such laws can also provide undue leverage to the state to influence social media content moderation in line with state interests that are related to supressing democratic political criticism. Second, global social media service providers can resist conforming to the requirements of such laws in smaller countries where the local users constitute only a very small market share for the platform.

To overcome all these challenges, it is necessary to have a method of accountability that is on the one hand not vulnerable to abuse by government, and yet adequately compelling to motivate better content moderation by social media platforms. The relevance of a reputational-cost approach arises from the fact that it can be designed to achieve both of the above outcomes.

The final observations of the Sri Lankan case study in section 4.3 suggests that it was the wider reputational-cost that was most instrumental in improving the response of the Facebook social media platform. The reputational-cost was generated from the local and international discussion of offline harm in Sri Lanka arising from online behaviour that was poorly moderated. An important extension of that observation from the Sri Lankan case study would be to make reputational-cost part of an institutionalised strategy of driving better content moderation.

The reputational-cost strategy

Consumer responses and civil society activism that results in losses to reputation serve as an impetus for reform and improvement in the modern marketplace.

The strategy of promoting a reputational-cost approach is one where the social media platforms are subject to an increasing risk of losing commercial "good-will" on a global scale, due to content moderation failures at a local level. In commercial terms, "good will" refers to the monetary value of a business that is attributed to the intangible asset of a positive social reputation, including expectation of future performance. The majority of the stock value of social media platforms is derived from such "good-will". This is evidenced in especially high price-to-book value ratios of social media platform stocks. [101]

The strategy of a reputational-cost approach, therefore, is to promote an architecture of societal response that increases the risk of generating negative feedback to the commercial "good-will" – reflected eventually in the price-to-book value ratio – of social media platforms when the platform fails to exercise adequate due diligence in responsible management of the platform, including effective content moderation.

Box 2 sets out examples of how reputational-cost had a significant impact on stock-market valuations of two major social media platforms in 2018. Due to the high component of "good-will" in stock market pricing of shares, the reputational-cost approach generates strong incentives for better outcomes from platform management and content moderation. Because decision boards of organisations are highly sensitive to pricing of the company stock, the reputational-cost approach also locates those incentives for better content moderation at the highest level of decision making. The numbers set out in Box 2 shows that these reputational-costs could represent a much larger financial incentive relative to the costs that would be incurred by having to face legal costs or regulatory fines.

Even in respect of the Sri Lankan context, the widespread public outcry against the role played by social media in the ethnic riots in Sri Lanka in 2018 presents an example of the effectiveness of reputational-cost approaches. Facebook received widespread criticism for the failures in content moderation that permitted the platform to be used in instigating the ethnic riots in 2018. The events and circumstances surrounding the ethnic riots were reported in internationally reputed media agencies such as the New York Times, The Guardian, Bloomberg, and Al Jazeera. The social backlash prompted Facebook to commission several studies and commit greater resources to numerous initiatives designed to improve due diligence and content moderation in relation to Sri Lanka. This example from Sri Lanka is important, because it establishes a feature of the strategy: that the reputational-cost approach can succeed even in the context of small social media markets such as Sri Lanka, through leveraging the risk to commercial "good-will" on a global scale.

BOX 2: EXAMPLE OF REPUTATIONAL-COST RESULTING IN LOSS OF GOOD-WILL AND STOCK-VALUE OF SOCIAL MEDIA COMPANIES, DESPITE RISING VALUE OF NET-ASSETS.

2018 was the year in which the scandals of Cambridge Analytica and Russian Meddling in U.S. elections through social media received wide publicity.

In that year, from the second quarter of the year (30 June) to the end of the year (31 December) 2018, Facebook and Twitter both had drastic reductions in price-to-book value and stock market valuations, even while their book value increased.

Facebook price-to-book value ratio declined from 7.08 to 4.45. This represented a reduction in the stock market value of Facebook by over USD 220 billion (over 35% decline), even while the book value of the company grew by almost 8% — from USD 90 to 97 billion, during that time.

Likewise, the price-to-book value ratio of Twitter declined from 5.93 to 3.35. This represented a reduction in the stock market value of Twitter by over USD 12 billion (over 30% decline), even while the book value of the company grew by almost 15% — from 8.86 to 10.16 billion USD, during that time.[102]

While not all these negative financial impacts can be attributed to the reputational-cost consequences, there has been adequate analysis to suggest that the reputational impact from these scandals were a significant part of the financial impact.[103]

5.2 The role of civil society in a reputational-cost approach: Building the societal architecture

BOX 3: GLOBAL LABOUR STANDARDS WERE ENTRENCHED THROUGH BUILDING A SOCIETAL ARCHITECTURE FOR CREATING REPUTATIONAL-COSTS.

The global entrenchment of labour standards in the apparel and footwear sectors in the past 25 years provides an excellent example of how the civil society based societal architecture for creating reputational-costs was pivotal in entrenching better labour standards in the global supply chain.

The creation of reputational-costs which led to consumers rejecting apparel and footwear of popular brands that relied on "sweat-shop" labour in their supply chain, gained traction in the late 1990s. It began with university student movements in the United States being concerned about the labour conditions under which the clothing with university insignia were being produced in poor countries.

The United States Students Against Sweat-shops (USAS) was a university based student organisation that spawned at least a couple of hundred chapters in universities across the United States. As students from these formative movements graduated, their engagements grew into national and international civils society initiatives and partnerships. The formation of the Fair Labour Association (FLA) in the United States, in partnership with the industry, was also a consequence of these maturing initiatives.

These growing and maturing civil society movements then also proceeded to build global alliances of civil society movements that worked together to highlight concerns of "sweat-shop" labour in the global supply chain for apparel and footwear.

The above insights are drawn from an Oxford University Thesis in 2004, that is an instructive study of how civil society movements can emerge and link together to construct a global architecture of societal response that improves corporate behaviour within complex supply chains (in this case the conditions and treatment of industrial labour).[104]

For the present analysis, it is notable that this societal architecture in relation to labour standards in the apparel and footwear industry was built on leveraging reputational-costs, and that the global entrenchment of better labour standards continues to succeed on the basis of reputational-costs rather than the threat of legal sanctions.

The success of a reputational-cost approach depends on an architecture of societal response that has an appropriate impact on the commercial "good-will" of social media platforms. Building such an architecture would require an organised approach to: (a) collect information and generate credible analysis on the performance of social media platforms, and (b) communicate and build awareness among users and stakeholders. Various stakeholders can contribute to this process. Civil-society groups, especially, will have a critical role to play in building such an architecture of societal response.

Historically, civil-society activism, such as calling for product boycotts, and petitioning domestic and international forums, have proven to be effective methods used to create "good-will" costs on organisations that have engaged in unethical practices. Through such activism organisations have faced major social backlash, resulting in losses of reputation and shifts in consumer loyalty and the market position relative to their competitors. Such social pushback, in addition to impacting sales and revenue also impacts the stock market valuation of the organisation that is subject to social scrutiny and motivates its management to address the practices that are being protested.

Box 3 sets out the insight that it is the architecture of societal response created by civil society organisations, rather than government regulation or internal corporate values that has driven and entrenched this improvement of labour standards practices in the apparel and footwear manufacture and supply industry.

The strategy of a reputational-cost approach envisages the replication of such an architecture of societal response with regard to social media service providers. Activism against poor content moderation can cause major losses of reputation for service providers. As a result, targeted social media platforms can experience a loss of users and engagement, which can translate into losses in advertisements, revenue, and commercial "good-will" that impacts the all-important stock market valuations (See Box 2). These losses strike at the heart of the business model of modern social media platforms and are likely to prompt overhauls of systems that moderate content and develop user trust. As demonstrated in Box 1 above, reputational harms caused to service providers by civil-society activism, such as boycotts, have moved service providers to reinvigorate their efforts to ramp up content moderation against rampant racism, discrimination, and hate in the different social contexts.

Presently, the formal assessments of social media service providers have been, to a great extent, dominated by the service providers themselves. Service providers have commissioned impact assessments on their own, deflected independent scrutiny, and resisted attending government hearings. Thereby, they have also limited public access to content moderation data that could entail greater visibility and accountability on platform responses and responsibility.[105] In short, in the present context, reputational-costs are managed by an evaluative architecture created by social media platforms themselves.

This capture of the assessment process, through which public accountability is generated, was notably, also an early-stage feature of the global apparel and footwear industry, which was discussed in Box 3. Prior to the creation of the Fair Labour Association (FLA), similar to the initiatives of social media service providers at present, there were several industry-led initiatives to demonstrate commitments to improve labour standards. The most prominent example was perhaps the Apparel Industry Partnership (AIP) which went on to develop codes of conduct with government for the apparel and footwear companies. But these were subject to criticism by civil society organisations as being inadequate.[106] Eventually these initiatives collapsed into industry support for the more independent FLA, which has enlisted wide civil-society support despite being subject to some concern and criticism about its lack of independence, since it is substantially dependent on industry contributions for its funding.

With regard to social media platforms, there is continuing opacity with regard to the data by which self-regulation behaviour can be evaluated. A greater level of openness with regard to the data is required for the public accountability of platforms to be effective, and driven by an independent architecture of societal response. CSOs have an important role to play in engaging with social media platforms and enlisting commitments to such openness in the first instance. Additionally, when the data is more accessible, CSOs also have a critical role in: (a) generating credible analysis with regard to effective content moderation; and (b) communicating and building awareness among users and stakeholders with regard to performance of content moderation, and especially the failures in content moderation that can lead to offline-harm. The effectiveness of this vertical accountability model would therefore likely arise as a gradual process, through the progress of CSOs in building an adequate societal response architecture.

The success of this model would also grow with the development of a global network of CSOs to take local issues and have them discussed at a global level. The global networking of CSOs can generate much greater efficacy for such a vertical accountability model in content moderation of the social media space. Apart from the example of pushback against sweat-shop labour, another case in point is the proliferation of civil society actors and organisations that have been holding governments to account on issues of governance and human rights, and the international networking of such initiatives.

However, unlike the global proliferation of civil society activism to hold governments and manufacturing industry accountable, the civil society activism to hold social media platforms accountable are yet at a very nascent stage of development. In Sri Lanka, there has been, for instance, an early stage move to address issues in content moderation, with an initiative known as the Colombo Social Media Declaration. Through the Colombo Social Media Declaration, CSOs have undertaken to 'minimise/eventually eradicate' the generation and spread of, among others, discrimination, harassment, and disinformation on social media platforms.[107] This is an example and an indication that CSO initiatives in this regard are yet at a very early stage. For the reputational-cost approach to become effective, such initiatives will need to increase in scope and reach and become part of a global eco-system that builds the necessary social-response architecture.

5.3. Critical aspects of building the societal architecture for a reputational-cost approach

This section sets out some of the critical aspects of CSOs succeeding in building a social-response architecture. As mentioned before, the creation of critical user interest that can drive reputational-costs to service providers requires user and stakeholder awareness based on adequate and credible analysis. The Sri Lankan case experience highlights three critical aspects that are pivotal in generating organic user interest that can enhance effective accountability through a reputational-cost approach: (1) identifying (discovering and highlighting); (2) quantifying (analytical reporting and ranking); and (3) communicating (building global awareness and networking).

While circumstances that impose reputational-costs on social media service providers do occur organically and in response to escalated situations, such as, in the aftermath of the 2018 ethnic riots in Sri Lanka, the strengthening of reputational-cost approaches as a form of vertical accountability requires an organised and structured exploration of these three aspects.

1. Identifying (discovering and highlighting)

This refers to the critical aspect of identifying localised issues – including unrecognised issues arising from local language nuances – that prevent the current content moderation by social media companies from being adequately effective. For this purpose, civil society actors can build public facing platforms that are engaged in identifying (that is, discovering and highlighting) localised issues.

For example, *Ethics-Eye* is a public facing platform, developed by Verité Research, which flags unethical reporting in Sri Lanka's mainstream media. *Ethics-Eye* engages in discovering and highlighting unethical reporting in mainstream media by publishing the violation and calling out the newspaper that carried it. Through this process, of identifying and highlighting unethical practices, *Ethics-Eye* prompts journalists, editors, and media owners to be more mindful in adhering to standards of ethical reporting.

The first critical aspect, therefore, is the engagement of civil society actors in the process of identification (discovery and highlighting) localised failures in content moderation. When such failures are highlighted in a public manner, social media companies are prompted early to take necessary action to improve their content moderation function before the failures escalate into graver consequences.

While the institution and proliferation of such mechanisms for identification can occur organically, a more organised, collaborative, and cooperative approach between relevant CSOs could be useful for two reasons. First, identifying localised issues is resource and time intensive, given the large amount of content posted online, and therefore a structured division of labour between CSOs can enhance the outcome of the combined efforts. Second, a division of labour can enable specialisation among CSOs in the identification of localised issues, allowing for better identification of such issues and their implications. For instance, local language based issues can be complex and require specialised skills and investments to identify adequately.

The 'No Hate Speech Movement' project initiated by the Council of Europe is an innovative example of driving better content moderation through civic engagement. The 'No Hate Speech Movement' is a civil society platform that encourage users to report harmful content, such as hate speech and cyber bullying, and supports such users in creating counter narratives to such harmful content.

2. Quantifying (analytical reporting and ranking)

This refers to the critical aspect of developing quantifiable metrics and rankings of the performance of social media companies, over time, and in relation to each other, in terms of better moderation of content. The importance of quantification is that it allows for trackable metrics against which the societal concern and response can become more objectively focused when driving accountability for better moderation of content.

The EU Code of Conduct on Countering Hate Speech Online is an apt example, as it evaluates social media companies that have pledged to 'review posts flagged by users and take down those that violate EU standards.'[108] The evaluation highlights how effectively social media platforms responded to prevent and counter the spread of illegal hate speech online.

In Sri Lanka as well, initiatives to quantify and rank have proven to be effective in eliciting a positive response. The most successful posts on the *Ethics-Eye* platform, mentioned previously, tend to be those that quantify and compare ethical violations of competing media organisations in the reporting of specific issues or events. This quantification approach is also used in Sri Lanka to drive greater public accountability of parliament, through the *Manthri.lk* platform.[109] This platform ranks Members of Parliament (MPs) in Sri Lanka based on their quantifiable contribution at parliamentary sittings. It has over time become a significant basis for public evaluation of MPs, and thereby has also incentivised MPs to improve their attendance and contributions in parliament.

Globally, the Ranking Digital Rights Index evaluates digital companies and ranks them based on their disclosed policies and practices affecting people's rights to freedom of expression and privacy.[110]

The second critical aspect, therefore, is the development of a structured network of independent analysis, quantification and ranking, of content moderation and related functions of social media platforms. Such quantification is likely to focus the interest and awareness of users as well as corporate stakeholders, which then becomes a part of the societal response architecture that drives reputational-cost incentives for the better moderation of content.

3. Communicating (building global awareness and networking)

This refers to the critical aspect of making public awareness on the failure of social media content moderation a matter of global concern, and the networking of CSOs in order to do so.

Vertical accountability arises from public concern, driven by information, interest, and engagement. For actors with a global footprint effective public concern, to build adequate reputational-cost incentives, may also need to be on a global scale.

In Sri Lanka, the public nature of *Ethics-Eye* and *Manthri.lk* platforms mentioned above drives media organisation and parliamentarians, respectively to take note of potential reputational-cost and improve behaviour. However, with regard to social media platforms, public concern being limited to the discourse in Sri Lanka proves to be inadequate. For global social media platforms, negative reputational-costs are less likely to arise from a discourse that is not globalised.

The anti-Muslim riots which took place in Sri Lanka were featured in the global media, such as, BuzzFeed, The Guardian, The New York Times, Eurasia etc. and further highlighted that Facebook played a role in the escalation of the riots.[111] That is, global publications and media stations shed light on the issue and created global awareness. As a result, Facebook commissioned an independent human rights impact assessment that identified that Facebook was not doing enough in Sri Lanka. Consequently, Facebook was seen doing more and getting more involved in solving localised issues of its platform in Sri Lanka. It is this globalised discourse that impacted the reputation of the platform and led to the company taking a strong interest in improving content moderation in Sri Lanka.

There are various other instances in which public concern in particular societies had a major impact on users boycotting social media platforms. Such backlash and boycotts resulted in creating answerability, where platforms had to explain themselves to society on a global scale (due to the coverage the issues received on globalised media) and make changes in order to contain the reputational-cost. The improved responses to localised issues derives mainly from recognising society's discontent, which is affecting the credibility and the traction of the social media platform itself on a global scale. This is the result, not of external legal accountability to governments, nor internal accountability within the organisation, but the working out of vertical accountability to society on a global scale.

The third critical aspect of communicating, therefore, is about building awareness on local issues on a global scale. This would typically require the coordination of a global network of CSOs and the engagement of globalised media. It starts with identifying and quantifying localised issues as described, and then working to build user awareness and interest on a global scale.

***

The Sri Lankan social media space serves as a pertinent case study for recognising the wisdom and centrality of a reputational-cost approach as a means of improving the outcome of content moderation by social media service providers. The case for a reputational-cost approach derives from the recognition that self-regulation based on internal accountability of service providers has proven to be ineffective, and defaulting to a strategy based on accountability to local governments can be dangerous. The Sri Lankan case study sets out the recognition of the weakness of these two methods and this case study is likely to be more typical of most nations in the world outside of the exceptional contexts of the United States and the European Union.

The alternative of a reputational-cost approach which is proposed in this strategic brief draws from the Sri Lankan case study, as well as from numerous international experiences both within the social media space as well as other areas of economics and commerce. This analysis identifies the importance of the role of CSOs in building and taking forward a societal response architecture that creates adequate vertical accountability for the reputational-cost approach to gain traction and sets out some of the critical elements that are yet in need of development in order for a reputational-cost approach to become institutionalised and effective on a global scale.

End Notes

1. Article One, 'Assessing the Human Rights Impact of the Facebook Platform in Sri Lanka' (2018), at https://about.fb.com/wp-content/uploads/2020/05/Sri-Lanka-HRIA-Executive-Summary-v82.pdf [last accessed 14 June 2021].
2. 'Arrest of writer Sathkumara sparks debate on freedom of expression', Daily Mirror, 12 April 2019 , at http://www.dailymirror.lk/news-features/Arrest-of-writer-Sathkumara-sparks-debate-on--freedom-of-expression/131-165392 [last accessed 17 April 2020]. Website of the Times of India, 'Global rights groups demand release of lawyer held under Lanka's anti-terrorism law', 28 July 2021, at https://timesofindia.indiatimes.com/world/south-asia/global-rights-groups-demand-release-of-lawyer-held-under-lankas-anti-terrorism-law/articleshow/84823564.cms [last accessed on 28 July 2021].
3. Social Media Stats Sri Lanka', Statcounter GlobalStats, January 2021, at https://gs.statcounter.com/social-media-stats/all/sri-lanka [last accessed 28 February 2021].
4. Dr. Gehan Gunatilleke , 'Countering harmful speech: Why trust the State?', (3 January 2021), at http://www.themorning.lk/countering-harmful-speech-why-trust-the-state/ [last accessed 28 February 2021]. The prohibition of hate speech has also received international recognition in article 20(2) of the ICCPR and Article Article 4(a) of the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) (21 December 1965) 660 UNTS 195 (entered into force 4 January 1969), which obliges states parties to 'declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin, and also the provision of any assistance to racist activities, including the financing thereof'
5. Claire Wardle & Hoseein Derakhshan, Information Disorder: Toward an interdisciplinary framework for research and policy making (Council of Europe Report 2017), p. 5; Christina Nemr & William Gangware, Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age (Park Advisors March 2019), p. 4.
6. International Foundation for Electoral Strategy, Disinformation Campaigns and Hate Speech: Exploring the Relationship and Programming Interventions (April 2019), at https://www.ifes.org/sites/default/files/2019_ifes_disinformation_campaigns_and_hate_speech_briefing_paper.pdf, p. 4-8 [last accessed 19 July 2021].
7. Jeff Jarvis, 'Platforms are not publishers', The Atlantic, 10 August 2018, at https://www.theatlantic.com/ideas/archive/2018/08/the-messy-democratizing-beauty-of-the-internet/567194/ [last accessed 19 July 2021].
8. United Nations Human Rights Council, Statement by United Nations High Commissioner for Human Rights, Michelle Bachelet at the 13th Session of the Forum on Minority Issues: Hate speech, social media and minorities, 19 – 20 November 2020, at https://www.ohchr.org/EN/HRBodies/HRC/Pages/NewsDetail.aspx?NewsID=26519&LangID=E [last accessed 19 July 2021]; See Hebe Campbell, 'Social media firms 'must take accountability' over online racist abuse', Euronews, 11 February 2021, at https://www.euronews.com/2021/02/08/social-media-firms-must-take-accountability-over-online-racist-abuse [last accessed 19 July 2021]; Paul Waters, 'The growing movement for platform accountability', Democracy Fund, 8 March 2021, at https://democracyfund.org/idea/the-growing-movement-for-platform-accountability/ [last accessed 19 July 2021].
9. M. Todd Henderson, 'Why self-regulation of social media could work — the financial services model', The Hill, 29 July 2019, at https://thehill.com/opinion/technology/455144-why-self-regulation-of-social-media-could-work-the-financial-services?rl=1 [last accessed 2 February 2021]; See Article 19, Self-regulation and 'hate speech' on social media platforms (2018), at https://www.article19.org/wp-content/uploads/2018/03/Self-regulation-and-%E2%80%98hate-speech%E2%80%99-on-social-media-platforms_March2018.pdf [last accessed 27 March 2020] p.9.
10. Website of Facebook, Community Standards – Hate Speech, at https://www.facebook.com/communitystandards/hate_speech [last accessed 10 February 2021].
11. Website of Facebook, Community Standards – Integrity and Authenticity, at https://www.facebook.com/communitystandards/integrity_authenticity [last accessed 18 March 2021].
12. Website of Facebook, Community Standards – Hate Speech, at https://www.facebook.com/communitystandards/hate_speech [last accessed 10 February 2021].
13. Website of Facebook, Community Standards – Integrity and Authenticity, at https://www.facebook.com/communitystandards/integrity_authenticity [last accessed 18 March 2021].
14. Website of YouTube, YouTube Community Guidelines, Hate Speech Policy, at https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436 [last accessed 16 November 2020]. The site offers examples of what would amount to 'hate speech'.
15. Website of YouTube, YouTube Community Guidelines, Overview, at https://www.youtube.com/howyoutubeworks/policies/community-guidelines/ [last accessed 18 March 2021]. See also: Website of YouTube, How does YouTube combat misinformation?, at https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/#misinfo-versus-disinfo [last accessed 18 March 2021].
16. Website of YouTube, YouTube Community Guidelines, Hate Speech Policy, at https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436 [last accessed 16 November 2020]. The site offers examples of what would amount to 'hate speech'.
17. Website of Twitter, Rules and Policies, at https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [last accessed 28 March 2020].
18. ibid.
19. Website of Twitter, General guidelines and policies, Our range of enforcement options, at https://help.twitter.com/en/rules-and-policies/enforcement-options [last accessed 26 January 2021].
20. ibid.
21. ibid.
22. 'What is Social Media Content Moderation and how Moderation Companies use Various Techniques to Moderate Contents?', 11 May 2020, para 2, at https://medium.com/cogitotech/what-is-social-media-content-modera-

tion-and-how-moderation-companies-use-various-techniques-to-a0e38bb81162 [last accessed 12 February 2021].

23. ‘What is Social Media Content Moderation and how Moderation Companies use Various Techniques to Moderate Contents?’, 11 May 2020, para 3, at https://medium.com/cogitotech/what-is-social-media-content-moderation-and-how-moderation-companies-use-various-techniques-to-a0e38bb81162 [last accessed 12 February 2021].

24. ‘Artificial Intelligence: How Facebook uses Artificial Intelligence’, 5 June 2019, at https://kambria.io/blog/how-facebook-uses-artificial-intelligence/#:~:text=In%20addition%20to%20the%20already,and%20shuts%20them%20down%20instantly [last accessed 12 February 2021].

25. ibid.

26. Website of Groundviews, ‘On Facebook’s New Misinformation Policy for Sri Lanka’, 20 July 2018, at https://groundviews.org/2018/07/20/on-facebooks-new-misinformation-policy-for-sri-lanka/; Website of Time, ‘Facebook Says It’s Removing More Hate Speech Than Ever Before. But There’s a Catch’, 26 November 2019, at https://time.com/5739688/facebook-hate-speech-languages/ [last accessed 18 March 2021].

27. Website of the United Nations Human Rights Office of the High Commissioner, ‘Study on the prohibition of incitement to national, racial or religious hatred: Lessons from the Asia Pacific Region’, Vitit Muntarbhorn, at https://www.ohchr.org/Documents/Issues/Expression/ICCPR/Bangkok/StudyBangkok_en.pdf [last accessed 18 March 2021]; See The website of The Future of Free Speech, ‘Global Handbook on Hate Speech Laws’, 20 November 2020, at https://futurefreespeech.com/global-handbook-on-hate-speech-laws/ [last accessed 19 July 2021].

28. However, most platforms do have mechanisms where governments can request service providers to take down content during escalations or emergencies.

29. Yaël Eisenstat, ‘How to hold social media accountable for undermining democracy’, Harvard Business Review, 11 January 2021, at https://hbr.org/2021/01/how-to-hold-social-media-accountable-for-undermining-democracy [last accessed 19 July 2021].

30. ibid.

31. 47 U.S. Code § 230 - Protection for private blocking and screening of offensive material, at https://www.law.cornell.edu/uscode/text/47/230 [last accessed 20 January 2021].

32. United States Court of Appeals, Nos. 04-56916, 04-57173 (2008), at https://caselaw.findlaw.com/us-9th-circuit/1493375.html [last accessed 21 January 2021].

33. CDA 230 The Most Important Law Protecting Internet Speech, Electronic Frontier Foundation, at https://www.eff.org/issues/cda230 [last accessed 20 January 2021].

34. ‘Social Media Liability Law Is Likely to be Reviewed Under Biden’, Washington Post, 18 January 2021, at https://www.washingtonpost.com/politics/2021/01/18/biden-section-230/ [last accessed 20 January 2021]. See also Joe Biden, The New York Times, 17 January 2020, at https://www.nytimes.com/interactive/2020/01/17/opinion/joe-biden-nytimes-interview.html?smid=nytcore-ios-share [last accessed 20 January 2021].

35. Vikram Jeet Singh and Prashant Mara, ‘India: Liable vs. Accountable: How Criminal Use Of Online Platforms And Social Media Poses Challenges To Intermediary Protection In India’, mondaq, 5 May 2020, at https://www.mondaq.com/india/social-media/928106/liable-vs-accountable-how-criminal-use-of-online-platforms-and-social-media-poses-challenges-to-intermediary-protection-in-india [last accessed 10 February 2021].

36. Website of Ministry of Electronics & Information Technology – Government of India, “Information Technology (Intermediaries guidelines) Rules, 2011”, at https://www.wipo.int/edocs/lexdocs/laws/en/in/in099en [last accessed 19 January 2021].

37. Christian Louboutin SAS v. Nakul Bajaj & Ors, [2018(76) PTC 508(Del)].

38. Official website of Scroll.in, ‘Internet Laws: How India is using its Information Technology Act to arbitrarily take down online content, Torsha Sakar and Gurshabad Grover, 15 February 2020, at https://cis-india.org/internet-governance/blog/content-takedown-and-users-rights-1 ; https://scroll.in/article/953146/how-india-is-using-its-information-technology-act-to-arbitrarily-take-down-online-content [last accessed 4 March 2021].

39. Shreya Singhal v. Union of India (2015), Writ Petition (Criminal) NO.167 OF 2012, at https://indiankanoon.org/doc/110813550/

40. Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act), at https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=publicationFile&v=2 [last accessed 21 January 2021].

41. Germany’s Network Enforcement Act and its impact on social networks, at https://www.taylorwessing.com/download/article-germany-nfa-impact-social.html [last accessed 21 January 2021].

42. Official website of Office of the United Nations High Commissioner for Human Rights, ‘Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression’, 1 June 2017, at https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf [last accessed 4 March 2021].

43. Germany: Flawed Social Media Law, at https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law [last accessed 21 January 2021].

44. ibid.

45. A. Panetta and N. Scola, Facebook’s Zuckerberg faces summons after snubbing Canada, Politico, 28 May 2019, at https://www.politico.com/story/2019/05/28/canada-facebook-mark-zuckerberg-1475782 [last accessed on 01 March 2021]; Varun Thomas Mathew, The arrogance of being Facebook, a serious tragedy for the rule of law, The Hill, 08 October 2020, at https://thewire.in/law/facebook-delhi-assembly-summons-rule-of-law-riots [last accessed on 01 March 2021];

46. Website of Scroll.in, Social Media Politics, ‘Delhi Assembly panel asks Facebook India’s vice president to appear before it on Wednesday’, 21 September 2020, at https://scroll.in/latest/973664/delhi-assembly-panel-asks-facebook-indias-vice-president-to-appear-before-it-on-wednesday [last accessed 4 May 2021].

47. Website of Investopedia, Company Profiles, ‘How Facebook, Twitter, Social Media Make Money From You’, 7 February, at https://www.investopedia.com/stock-analysis/032114/how-facebook-twitter-social-media-make-money-you-twtr-lnkd-fb-goog.aspx#:~:text=The%20primary%20way%20social%20media,before%20social%20media%20companies%20existed [last accessed 4 May 2021]; How does Twitter make money? [2020 Revenue Facts], Alejandro Rioja, 28 January 2020, at https://alejandrorioja.com/how-does-twitter-make-money/ [last accessed 4 May 2021].

48. The Athletic Staff, ‘UEFA joins English football’s social media boycott’, 29 April 2021, at https://theathletic.com/news/football-uefa-social-media-blackout/k7VfuGDvNis6 [last accessed 19 July 2021]; See Sean Ingle, ‘Sports bodies to boycott social media for bank holiday weekend over abuse’,

The Guardian, 29 April 2021, at https://www.theguardian.com/sport/2021/apr/29/major-sports-bodies-84-hour-social-media-boycott-over-online-abuse-facebook-twitter [last accessed 19 July 2021].

49. Website of Stop Hate for Profit, 'One Year After Stop Hate for Profit: Progress Update from the Coalition' (30 July 2020), at https://www.stophateforprofit.org/ [last accessed 30 June 2021].
50. Aimee Picchi, '#LogOutFacebook: The social media giant faces protests and canceled accounts', CBS News, 18 December 2018, at https://www.cbsnews.com/news/logoutfacebook-the-social-media-giant-faces-naacp-protests-and-canceled-accounts/ [last accessed 19 July 2021].
51. Gilat Levy and Ronny Razin, 'Social Media and Political Polarization', LSE Public Policy Review, 20 July 2020, at https://ppr.lse.ac.uk/articles/10.31389/lseppr.5/ [last accessed 19 July 2021]; See Erica Bell, 'How Social Media Algorithms Are Increasing Political Polarisation', Young Australians in International Affairs, 1 March 2021, at https://www.youngausint.org.au/post/how-social-media-algorithms-are-increasing-political-polarisation [last accessed 19 July 2021]; Dr Sander van der Linden, 'Social networks are built to turn us against each other. Can we fix them?', Science Focus, 12 May 2021, at https://www.sciencefocus.com/future-technology/social-networks-are-built-to-turn-us-against-each-other-can-we-fix-them/ [last accessed 19 July 2021].
52. Amnesty International, Sri Lanka: Act now to prevent further bloodshed in anti-Muslim violence, 17 June 2014, at https://www.amnesty.org/en/latest/news/2014/06/sri-lanka-act-now-prevent-further-bloodshed-anti-muslim-violence/ [last accessed on 08 February 2021].
53. Sanjana Hattotuwa, supra note 68.
54. Global Voices, Social Media Rumours Escalate Buddhist-Muslim Violence in Sri Lanka, 23 November 2017, https://globalvoices.org/2017/11/23/social-media-rumors-escalate-buddhist-muslim-violence-in-sri-lanka/ [last accessed on 08 February 2021].
55. AlJazeera, In Sri Lanka, hate speech and impunity fuel anti-Mulsim violence, 13 March 2018, at https://www.aljazeera.com/news/2018/3/13/in-sri-lanka-hate-speech-and-impunity-fuel-anti-muslim-violence [last accessed on 08 February 2021].
56. Reutuers, Police, politicians accused of joining Sri Lanka's anti Muslim riots, 24 March 2018, at https://www.reuters.com/article/us-sri-lanka-clashes-insight/police-politicians-accused-of-joining-sri-lankas-anti-muslim-riots-idUSKBN1H102Q [last accessed on 08 February 2021].
57. The Wall Street Journal, Sri Lankan Islamist called for violence on Facebook before Easter Attacks, 30 April 2019, at https://www.wsj.com/articles/sri-lankan-islamist-called-for-violence-on-facebook-before-easter-attacks-11556650954 [last accessed on 08 February 2021].
58. Aljazeera, Sri Lanka orders nationwide curfew amid anti-Muslim riots, 14 May 2019, at https://www.aljazeera.com/news/2019/5/14/sri-lanka-orders-nationwide-curfew-amid-anti-muslim-riots [last accessed 22 June 2021]; also see Michael Safi, 'Sri Lanka imposes curfew after mobs target mosques', The Guardian, 13 May 2019, at https://www.theguardian.com/world/2019/may/13/sri-lanka-imposes-curfew-after-mobs-target-mosques [last accessed 22 June 2021].
59. Hashtag Generation, Sri Lanka: Social Media and Electoral Integrity: Findings from Sri Lanka's 2020 Parliamentary Elections; Hashtag Generation: Findings from Sri Lanka's 2019 Presidential Elections
60. Sanjana's tweets, at https://twitter.com/sanjanah/status/1346669769601126401
61. Groundviews, RTI Reveals Lanka E News Blocked on Order from President's Office; also see Center for Policy Alternatives, 'The Internet as a Medium for Free Expression: A Sri Lankan Legal Perspective', J C Weliamuna, (2013), p. 30 – 36, at https://www.cpalanka.org/the-internet-as-a-medium-for-free-expression-a-sri-lankan-legal-perspective/ [last accessed 26 March 2021].
62. Verité Research, Regulating Social Media in Sri Lanka: An Analysis of the Legal and Non-Legal Regulatory Frameworks in the Context of Hate Speech and Disinformation, December 2020 (revised March 2021), Forthcoming.
63. Website of the Human Rights Watch, 'Sri Lanka: Increasing Suppression of Dissent', 8 August 2020, at https://www.hrw.org/news/2020/08/08/sri-lanka-increasing-suppression-dissent [last accessed 3 March 2021]; Website of Amnesty, 'Sri Lanka: Government suffocating dissent and obstructing justice for historic crimes says Amnesty report', 17 February 2021, at https://www.amnesty.org/en/latest/news/2021/02/sri-lanka-government-suffocating-dissent-and-obstructing-justice-for-historic-crimes-says-amnesty-report/ [last accessed 3 March 2021].
64. Verité Research, Regulating Social Media in Sri Lanka: An Analysis of the Legal and Non-Legal Regulatory Frameworks in the Context of Hate Speech and Disinformation, December 2020 (revised March 2021), Forthcoming.
65. United Nations Human Rights Council, Report of the Special Rapporteur on the promotion and protection of the right freedom of opinion and expression, supra note 1, p. 14; Amnesty International, #ToxicTwitter – The Reporting Process (Chapter 4), at https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-4/#-topanchor [last accessed on 10 February 2021]; ARTICLE 19, Regulating social media: We need a new model that protects freedom of expression, at https://www.article19.org/resources/regulating-social-media-need-new-model-protects-free-expression/ [last accessed on 10 February 2021].
66. Article One, Assessing the human rights impact of the Facebook platform in Sri Lanka, 2018, ['Facebook HRIA'] at https://about.fb.com/wp-content/uploads/2020/05/Sri-Lanka-HRIA-Executive-Summary-v82.pdf [last accessed on 10 February 2021].
67. Yudanjana Wijeratne, The Control of Hate Speech on Social Media: Lessons from Sri Lanka Wijeratne, Yudhanjaya, The Control of Hate Speech on Social Media: Lessons from Sri Lanka (October 30, 2018). CPR South, 2018, Policy Brief, at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3275106 [last accessed on 10 February 2021].
68. Facebook HRIA, p. 6.
69. Mark Sullivan, 'Facebook's AI for Detecting Hate Speech is Facing its Biggest Challenge yet', 14 August 2020, at https://www.fastcompany.com/90539275/facebooks-ai-for-detecting-hate-speech-is-facing-its-biggest-challenge-yet [last accessed 12 February 2021].
70. Ibid. See also, U.S. Congressional Research Service, Social Media: Misinformation and content moderation issues for Congress, 27 January 2021, at https://crsreports.congress.gov/product/pdf/R/R46662 [last accessed on 10 February 2021], p. 8-9.
71. Reuters, Facebook's flood of languages leave it struggling to monitor content, 23 April 2019, at https://www.reuters.com/article/us-facebook-languages-insight/facebooks-flood-of-languages-leaves-it-struggling-to-monitor-content-idUSKCN1RZ0DW [last accessed on 10 February 2021]; Forbes, Delving deeper into Facebook's murky community standards, 08 June 2020, at https://www.forbes.com/sites/petersuciu/2020/06/08/delving-deeper-into-facebooks-murky-community-standards/?sh=5f4779521314 [last accessed on 10 February 2021]; ProPublica, Civil rights

groups have been warming Facebook about hate speech in secret groups for years, 02 July 2019, at https://www.propublica.org/article/civil-rights-groups-have-been-warning-facebook-about-hate-speech-in-secret-groups-for-years [last accessed on 10 February 2021].

72. Does FB's Selective Hate Speech Rules Threaten India's Democracy?', Shorbori Purkayastha, 18 August 2020, at https://www.thequint.com/podcast/fbs-selective-hate-speech-rules-threatens-indias-democracy#read-more [last accessed 26 January 2021].

73. See generally, C.R. Carlson and H. Rousselle, Invesigating Facebook's hate speech removal process, First Monday, Volume 25, Number 2, 03 February 2020, at https://firstmonday.org/ojs/index.php/fm/article/download/10288/8327; Global Centre for the Responsibility to Protect, Hate Speech and Social Media: Preventing Atrocities and Protecting Human Rights Online, 16 February 2020, at https://www.globalr2p.org/wp-content/uploads/2020/02/2020-February-Adams-Doha-Speech.pdf [last accessed on 10 February 2021].

74. Facebook HRIA, p. 6.

75. Hashtag Generation is a 'A youth-led movement advocating for full and effective participation of young people in policy making, implementation and evaluation at local, national, regional & international levels.' See https://hashtaggeneration.org/about/ [last accessed 7 July 2021].

76. Hashtag Generation, Sri Lanka: Social Media and Electoral Integrity – Findings from Sri Lanka's 2020 Parliamentary Election, at https://hashtaggeneration.org/publications/ [last accessed on 10 February 2021], p. 35.

77. 'Lammy Says Twitter Must Get Much Faster At Removing Hate After Racist Abuse', Faith Matters, 3 August 2020, at https://www.faith-matters.org/lammy-says-twitter-must-get-much-faster-at-removing-hate-after-racist-abuse/ [last accessed 11 February 2021]; Official Website of BBC, 'Wiley: Priti Patel probes Twitter and Instagram delay in removing 'appalling' posts', 26 July 2020, at https://www.bbc.com/news/uk-53544902 [last accessed 11 February 2021].

78. United Nations Human Rights Council, Report of the Special Rapporteur on the promotion and protection of the right freedom of opinion and expression, supra note 1, p. 14; Amnesty International, #ToxicTwitter – The Reporting Process (Chapter 4), supra note 2.

79. ibid.

80. Facebook's Technocratic Reports Hides its Failures on Abuse', Chirs Gillard, 27 August 2020, at https://onezero.medium.com/facebook-is-hiding-its-failure-to-keep-abuse-off-its-platform-behind-technocratic-reports-682d871ef1ca [last accessed 10 February 2021].

81. Bloomberg, Facebook Apologises for role in Sri Lankan violence, 12 May 2020, at https://www.bloomberg.com/news/articles/2020-05-12/facebook-apologizes-for-role-in-sri-lankan-violence [last accessed on 10 February 2021]; NikkeiAsia, Facebook must confront its Asian shortcomings, 23 March 2018, at https://asia.nikkei.com/Opinion/Facebook-must-confront-its-Asian-shortcomings2 [last accessed on 10 February 2021].

82. Facebook HRIA, p. 6.

83. Website of CNBC, 'Facebook claims A.I. now detects 94.7% of the hate speech that gets removed from its platform', Sam Shead, 19 November 2020, at https://www.cnbc.com/2020/11/19/facebook-says-ai-detects-94point7percent-of-hate-speech-removed-from-platform.html [last accessed 12 February 2021]; Mark Sullivan, 'Facebook's AI for Detecting Hate Speech is Facing its Biggest Challenge yet', 14 August 2020, at https://www.fastcompany.com/90539275/facebooks-ai-for-detecting-hate-speech-is-facing-its-biggest-challenge-yet [last accessed 12 February 2021].

84. Hashtag Generation, Sri Lanka: Social Media and Electoral Integrity – Findings from Sri Lanka's 2020 Parliamentary Election, supra note 10, p. 34-35; Hashtag Generation, Findings from the Social Media monitoring exercise during the 2019 Sri Lankan Presidential Election, at https://hashtaggeneration.org/publications/ [last accessed on 10 February 2021], p. 19; Minor Matters, Hate Speech in Sri Lanka during the Pandemic, 2020, at https://www.minormatters.org/storage/app/uploads/public/5fc/76b/014/5fc76b014d43f554793096.pdf [last accessed on 10 February 2021], p. 21.

85. Minor Matters, Hate Speech in Sri Lanka during the Pandemic, supra note 15, p. 21.

86. International Covenant on Civil and Political Rights Act No. 56 of 2007, at https://citizenslanka.org/wp-content/uploads/2015/12/International-Covenant-on-Civil-Political-Rights-ICCPR-Act-No-56-of-2007E.pdf; Gehan Gunatilleke, 'Broken shield and weapon of choice', at https://www.veriteresearch.org/2019/06/24/iccpr-act-sri-lanka/ [last accessed 10 February 2021].

87. ColomboPage, 13 arrested for riots in Minuwangoda to be produced in court today, 14 May 2019, at http://www.colombopage.com/archive_19A/May14_1557817003CH.php [last accessed on 10 February 2021]; The Diplomat, The Problem with Sri Lanka's New 'False News' Law, 07 August 2019, at https://thediplomat.com/2019/08/the-problem-with-sri-lankas-new-false-news-law/ [last accessed on 10 February 2021]; See Dr. Gehan Gunatilleke, 'Countering Harmful Speech: Why Trust the State?', The Morning, 3 January 2021, at https://www.themorning.lk/countering-harmful-speech-why-trust-the-state/ [last accessed 19 July 2021]; Dr. Gehan Gunatilleke, 'The Constitutional Practice of Ethno-Religious Violence in Sri Lanka', Research Gate, September 2018, at https://www.researchgate.net/publication/327702779_The_Constitutional_Practice_of_Ethno-Religious_Violence_in_Sri_Lanka [last accessed 19 July 2021].

88. United States Department of State - Office of International Religious Freedom, International Religious Freedom Report for 2019: Sri Lanka 2019 International Religious Freedom Report, at https://lk.usembassy.gov/wp-content/uploads/sites/149/SRI-LANKA-2019-INTERNATIONAL-RELIGIOUS-FREEDOM-REPORT.pdf, p. 8. [last accessed 10 June 2021].

89. Aanya Wipulasena, Abuse of ICCPR Act has 'chilling effect' on fundamental freedoms, Sunday Observer, 16 June 2019, at http://www.sundayobserver.lk/2019/06/19/news-features/abuse-iccpract-has-%E2%80%98chilling-effect%E2%80%99-fundamentalfreedoms [last accessed on 10 February 2021]; Civicus, Abuse of ICCPR Act and Judicial System to Stifle Freedom of Expression in Sri Lanka, 05 July 2019, at https://monitor.civicus.org/updates/2019/07/05/iccpr-act-and-judicial-system-being-misused-stifle-freedom-expression-sri-lanka/ [last accessed on 10 February 2021]; Laksara, Ramzy Razeek arrested under ICCPR Act granted bail after 161 days, 17 September 2020, at https://lankasara.com/en/news/ramzyrazeek-arrested-under-iccpr-granted-bail-after-161-days/ [last accessed on 10 February 2021].

90. S S Selvanayagam, SC awards compensation to British tourist with Buddha tattoo, Daily Mirror, 15 November 2017, at http://www.dailymirror.lk/article/SC-awards-compensation-to-British-tourist-withBuddha-tattoo-140437.html [last accessed on 10 February 2021].

91. World Socialist Web Site, Sri Lanka government intensifies crackdown on social media, 09 April 2020, at https://www.wsws.org/en/articles/2020/04/09/medi-a09.html [last accessed on 10 February 2021].

92. Civicus, Abuse of ICCPR Act and Judicial System to Stifle Freedom of Expression in Sri Lanka, supra note 15; Amnesty International, Sri Lanka jails journalist for 20 years for exercising his right to freedom of expression, 01 February 2009,

at https://www.amnesty.org/en/latest/news/2009/09/sri-lanka-condena-periodista-20-anos-prision-ejercer-libertad-expresion-20090901/ [last accessed on 10 February 2021].

93. Extraordinary Gazette, No. 2120/5 issued on 22 April 2019, at http://www.documents.gov.lk/files/egz/2019/4/2120-05_E.pdf [last accessed on 7 May 2021].

94. International Commission of Jurists, Sri Lanka: Briefing Paper – Emergency Laws and International Standards, March 2009, at https://www.icj.org/wp-content/uploads/2012/05/SriLanka-emergencylaws-advocacy-2009.pdf [last accessed on 10 February 2021], pp. 23-27; Groundviews, Explaining Sri Lanka's New Emergency Regulations on 'Publication', 25 April 2019, at https://groundviews.org/2019/04/25/explaining-sri-lankas-new-emergency-regulations-on-publication/ [last accessed on 10 February 2021].

95. Sri Lanka Telecommunications Act No. 22 of 1991, as amended by Act No. 27 of 1996, at http://www.trc.gov.lk/2014-06-09-09-55-30/2014-07-15-04-54-15/2014-05-13-12-24-49.html [last accessed on 10 February 2021].

96. Groundviews, RTI reveals Lanka E News blocked on Order from President's Office, 04 November 2018, at https://groundviews.org/2018/04/11/lanka-e-news-blocked-on-order-from-presidents-office-rti-reveals/ [last accessed on 10 February 2021].

97. ibid.

98. Verité Research, Regulating Social Media in Sri Lanka: An Analysis of the Legal and Non-Legal Regulatory Frameworks in the Context of Hate Speech and Disinformation, December 2020 (revised March 2021), Forthcoming.

99. Colombo Declaration on Media Freedom and Social Responsibility, 30 September 2018, at http://www.pccsl.lk/index.php/colombo-declaration-on-media-freedom-and-social-responsibility-2/ [last accessed on 04 March 2021].

100. Yaël Eisenstat, 'How to hold social media accountable for undermining democracy', *Harvard Business Review*, 11 January 2021, at https://hbr.org/2021/01/how-to-hold-social-media-accountable-for-undermining-democracy [last accessed 19 July 2021].

101. Since the book value of a company – the redeemable value of its net assets – is dynamic, and different from the market value of the stocks, changes to the "good-will" value of a company is best traced not by observing the changes in the stock price, but by changes in the price-to-book value ratio. This ratio automatically adjusts changes in price resulting from the inherent valuation of increasing/decreasing net assets (book value).

102. Data from https://www.macrotrends.net/

103. Website of CNBC, 'Here are the scandals and other incidents that have sent Facebook's share price tanking in 2018', Salvador Rodriguez, 20 November 2018, at https://www.cnbc.com/2018/11/20/facebooks-scandals-in-2018-effect-on-stock.html [last accessed 27 July 2021].

104. Donna Chung, "Transnational Advocates, Norm Advancement, and US Corporate Self-Regulation in China (1993-2000)", DPhil Thesis, Nuffield College, University of Oxford.

105. Naomi Nix and Nico Grant, 'Twitter CEO Takes Some Responsibility for Stop the Steal Spread' (25 March 2021), at https://www.bloomberg.com/news/articles/2021-03-25/tech-ceos-face-lawmakers-queries-on-dangers-of-disinformation; Website of The New York Times, 'Delay, Deny and Deflect: How Facebook's Leaders Fought Through Crisis' (14 November 2018) at https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html [last accessed 30 June 2021].

106. See for example the commentary at: https://www.corpwatch.org/article/clintons-new-no-sweatshop-agreement [last accessed 19 July 2021]

107. Colombo Declaration on Media Freedom and Social Responsibility, 30 September 2018, at http://www.pccsl.lk/index.php/colombo-declaration-on-media-freedom-and-social-responsibility-2/ [last accessed on 04 March 2021].

108. The Website of the European Commission, 'The EU code of conduct on countering illegal hate speech online: The robust response provided by the European Union', at https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en [last accessed 19 July 2021].

109. The platform is managed by Verité Research and is available at http://www.manthri.lk.

110. Website of Ranking Digital Rights, '2020 Ranking Digital Rights Corporate Accountability Index', at https://rankingdigitalrights.org/index2020/ [last accessed 19 July 2021].

111. Amanda Taub and Max Fisher, 'Where Countries Are Tinderboxes and Facebook Is a Match', *The New York Times*, 21 April 2018, at https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html [last accessed 19 July 2021]; Also see Pranav Dixit, 'Sri Lanka Has Blocked Most Major Social Networks After A Facebook Post Sparked Anti-Muslim Riots', *BuzzFeed* News, 13 May 2019, at https://www.buzzfeednews.com/article/pranavdixit/sri-lanka-has-blocked-most-major-social-networks-after-a [last accessed 19 July 2021]; Julia Carrie Wong, 'Sri Lankans fear violence over Facebook fake news ahead of election', *The Guardian*, 12 November 2019, at https://www.theguardian.com/world/2019/nov/11/facebook-sri-lanka-election-fake-news [last accessed 19 July 2021]; The Website of the EurAsian Times, 'After Rohingya Genocide; Facebook Now Apologizes To Sri Lanka Over Anti-Muslim Riots', 15 May 2020, at https://eurasiantimes.com/after-rohingya-genocide-facebook-now-apologizes-to-sri-lanka-over-anti-muslim-riots/ [last accessed 19 July 2021].

Design and layout by Dinuk Senapatiratne

ISSN 2806-5360

9 772806 536007

VERITÉ RESEARCH
Strategic Analysis for Asia

A | No. 5A, Police Park Place, Colombo 5
T | +94 11-2055544
E | reception@veriteresearch.org
W| www.veriteresearch.org