Facebook: Regulating Hate Speech in the Asia Pacific

By Aim Sinpeng (University of Sydney), Fiona Martin (University of Sydney), Katharine Gelber (University of Queensland), and Kirril Shields (University of Queensland).

July 5th 2021

Final Report to Facebook under the auspices of its Content Policy Research on Social Media Platforms Award

Acknowledgements

We would like to thank our Facebook research colleagues, as well as our researchers Fiona Suwana, Primitivo III Ragandang, Wathan Seezar Kyaw, Ayesha Jehangir and Venessa Paech, co-founder of Australian Community Managers.

The research grant we received from Facebook was in the form of an "unrestricted research gift." The terms of this gift specify that Facebook will not have any influence in the independent conduct of any studies or research, or in the dissemination of our findings.

Cover photo: Jon Tyson on Unsplash

Table of Contents

1. Executive summary

This study examines the regulation of hate speech on Facebook in the Asia Pacific region, and was funded through the Facebook Content Policy Research on Social Media awards. It found that:

- The language and context dependent nature of hate speech is not effectively captured by Facebook's classifiers or its global Community Standards and editorial policy. It requires local knowledge to identify, and consultation with target groups to understand the degree of harm they experience.
- Facebook's definition of hate speech does not cover all targets' experiences of hateful content.
- Facebook's definition of hate speech is more comprehensive than most legislation in the Asia Pacific.
- There is little specific legislation regarding hate speech in the Asia Pacific and any proposed hate speech legislation, such as in Myanmar or The Philippines, remains stalled at Bill stage.
- Laws that might legislate hate speech, including cybersecurity or religious tolerance laws, are sometimes employed by governments to suppress political opposition and inhibit freedom of speech.
- Facebook requires more local policy expertise in the Asia Pacific region, including market specialists with minority political, cultural and language expertise.
- Facebook's trusted partner program in the APAC region needs auditing to ensure it is comprehensive and the parameters for membership are clearly stated and publicly accessible.
- Facebook's public policy team requires more regular outreach with key minority groups to document evolving hate speech forms.
- Publicly visible hate speech comments on Facebook against LGBTQ+ groups are more commonly found in countries where the issue of LGBTQ+ is politicised at the national level.
- LGBTQ+ page administrators interviewed for our case studies had all encountered hate speech on their group's account.
- All LGBTQ+ page administrators interviewed were key actors in hate speech moderation but were volunteers, not professional community managers.
- In Myanmar, The Philippines and Indonesia, most LGBTQ+ page admins interviewed have either not read or do not understand Facebook community standards. They are self taught or receive training from third party organisations.
- Most LGBTQ+ page moderators interviewed do not engage directly with hate speakers, believing it will escalate their risk online and offline.
- The most common form of hate speech management by LGBTQ+ page admins interviewed in Myanmar, The Philippines and Indonesia is to leave the violating content alone. This means if proactive detection technology fails to take down this content, it remains on the platform.
- LGBTQ+ page admins interviewed in India and Australia commonly removed and reported hate speech.
- All page administrators said Facebook had failed to take down material that they had reported, and indicated that they felt disempowered by the process of flagging hate speech.
- The utility and integrity of the content flagging process may be affected by 'reporting fatigue', where individuals are disinclined to report violating content as a result of their perceived lack of impact on Facebook's moderation practices.

We recommend that Facebook:

- Work with protected groups to identify the commonly expressed forms of hate speech that deprive targets of powers and codify these in its reviewing policy.
- Make transparent the types and weight of evidence needed to take action on hate figures and groups, to assist law enforcement agencies and civil society groups in collating this information.
- Audit its trusted partners program in the APAC region to ensure it is comprehensive and the parameters for membership are clearly stated and publicly accessible.
- Make public a trusted partner in each country, or nominate a supranational trusted partner for the APAC region, so that individuals and organisations have a direct hate speech reporting partner for crisis reporting issues.
- Conduct an annual APAC roundtable on hate speech involving key non-governmental stakeholders from the protected groups in all countries.
- Better recognise the role of page administrators as critical gatekeepers of hate speech content, and support their improved regulatory literacy via training and education.
- Improve the regulatory literacy of all page admins by providing mandatory hate speech moderation training modules in major languages.
- Support extended training in hate speech management to APAC located page admins working for groups in protected categories.
- Make publicly transparent all content regulation procedures, in an easy to follow guide, including penalties for violations and the appeals process. The guide should be available in as many regional languages as possible and automatically recommended to all users who start a Facebook account.
- Facilitate regular consultative forums with target groups and protected category page owners to discuss best practice moderation approaches.

2. Introduction

It has been widely recognised that Facebook has an ongoing problem with the scale and scope of hate speech posted on its platform. In response to criticisms about the political, social and cultural impact of such content in Asia (Asia Centre 2020; Lee 2019; Reporters Without Borders 2018; Stecklow 2018), and regulatory moves in Europe to control the proliferation of illegal content (Goggin et al. 2017), in recent years the company has sought to improve its proactive machine-learning detection filters, expand its human moderation operations, and improve its content moderation policies and accountability measures (Facebook 2020a; Murphy 2020). It has begun to hire market specialists to bolster its capacity to respond to discrimination at a national and local level. The company has also strengthened its stakeholder engagement, involving a growing number of academics and civil society organisations to help improve their hate speech monitoring. Despite these efforts, recent civil society reports have found Facebook failing to respond to calls to address organised hate directed at ethnic, religious and gender minorities, amongst others (Aavaz 2019; Murphy 2020; Vilk et al 2021). Further, while the Christchurch Call demanded governments and tech companies work together to quash the type of terrorist content posted during the Christchurch attacks on Muslim communities (Ministry of Foreign Affairs and Trade 2019), nation-states have introduced sometimes ad-hoc and largely varied regulatory responses to harmful content that make the regulatory field more complex for the company to negotiate. By its own admission, Facebook continues to find it challenging to detect and respond to hate speech content across dynamic speech environments, multiple languages and differing social and cultural contexts (Facebook 2020b; Perrigo 2019).

In this context, our study was funded as part of the Facebook Content Policy Research on Social Media awards to examine hate speech regulation challenges in the Asia Pacific (APAC). The APAC region is the fastest growing market for Facebook and a key region for market expansion (Statista 2020). It also presents a host of challenges for Facebook's content regulation approach. It is a culturally, linguistically, and religiously diverse region with complicated political landscapes. Its people speak more than 2,300 languages, often mixing multiple languages together. Despite recent improvements, the language gap in the automated content moderation of Facebook leaves hate speech monitoring vulnerable to error (Wijeratne 2020). As of late 2019, Facebook's machine learning classifiers monitored only 40 languages for hate speech content (Perrigo 2019), although the company was researching more efficient comparative language translation models (Hao 2020). Facebook continues to hire additional human moderators for languages in 'at-risk countries', many of which are in the Asia Pacific.[1] However, it remains unclear how many languages, and particularly sub-national languages and dialects, are being covered in its automated and human content moderation processes. Further, it appears that Facebook's moderation procedures are more effective against racial and ethnic hate than against gender-based hate speech (Carlson and Rousselle 2020).

The task of combating hate speech online is now a collaborative project, which involves corporate, civil society and government actors in various internet governance arrangements. Facebook is a member of three international content governance bodies, the Global Internet Forum to Counter Terrorism, the Global Network Initiative and the Global Alliance for Responsible Media. In Europe, Facebook Inc. has also taken significant steps to work with government and civil society to rapidly identify and remove hate speech from its platforms. It is, for example, a signatory to the 2016 European Commission's *Code of Conduct on Countering Illegal Hate Speech Online* (European Commission 2016). The Commission has produced regular monitoring reports on the code's implementation, and the most recent (5th) monitoring report found that, "on average 90% of the notifications are reviewed within 24 hours and 71% of the content is removed." The Commission reported that Facebook's removals had increased from 28.3% of notified content, to 87.6% in the period 2016-2019 (European Commission 2020, 3). However, there is no comparable process in the Asia Pacific region.[2]

1 'At-risk countries' is a term from Facebook's independent human rights assessments, conducted by CSO Article One. It refers to nations such as Myanmar, Sri Lanka and Indonesia that are exposed to election related unrest and coordinated misinformation campaigns.

2 We do, however, see lobbying of national governments by the Asia Internet Coalition, which collectively represents top tech firms like Facebook, Google and Twitter, to advocate for their interests.

In our project, we seek to answer three research questions:

1. What constitutes hate speech in different Asia Pacific jurisdictions?
2. How well are Facebook's policies and procedures positioned to identify and regulate this type of content?
3. How can we understand the spread of hate speech in this region, with a view to formulating better policies to address it?

To address the first question we identified and mapped hate speech law in five case study countries in the Asia Pacific region, to understand how this problem is framed nationally, and what regulatory gaps exist that might enable hate speech to proliferate on Facebook. We also developed an ideal definition of hate speech derived from scholarly literature and compared that to Facebook's policy versions in its Community Standards and editorial procedures, to establish if the company's policy could be improved. The definition that we have used in this study is concerned not only with egregious examples, but also with everyday, banal forms of hate speech that also have a corrosive effect on people's wellbeing, identity and community.

We then explored how Facebook hate speech policies and procedures seek to moderate this harmful content, by examining corporate literature, conducting interviews with Facebook staff, and mapping the organisational response to this problem.

Finally, we examined hate speech from the targets' perspective (Bliuc et al. 2018). Given the level of discrimination and vilification experienced by lesbian, gay, bisexual, transgender, intersex and queer identifying people across Asia (Radics 2019), we collected data from the public Facebook pages of major LGBTQ+ groups in our case study countries to examine the incidence of hate speech that had escaped Facebook's automated filters.[3] We also interviewed page administrators of these groups to better understand their conception and management of hate speech, including their experience of reporting hate posts to Facebook. Along with expert online community management input from the Australian Community Managers network, these civil society interviews provide a framework for understanding the 'regulatory literacy' of those who are at the frontline of Facebook's efforts to minimise hate on its platform.[4] [5]

The project identified varying national legal approaches to regulating hate speech, most often framed in religious terms, and in provisions that are not hate speech specific. This contributes to a lack of legal clarity about what content is harmful, and what may be legally actionable. It is also notable that the term 'hate speech' itself is not directly translatable in some of the Asian languages, or has not been used in the same legislative spirit as in the English language. In Filipino, *paninirang* is popularly used to describe hate speech but it is directly translated to 'oral defamation.' In Bisaya (Cebuano), the term *panghimaraot* is used to describe hate speech but it translates to 'cursing.' In Thai, the term วาจาที่สร้างความเกลียดชัง is used to describe hate speech. The direct translation is 'words that create hate' but this is a newly constructed term, and its use and meaning are evolving.

An organisational review of Facebook's own internal structure, policy and processes for addressing hate speech suggests these arrangements are rapidly evolving, but require further development in the Asia Pacific context. We note that the processes used to identify hate speech triggers and forms, and for undertaking hate speech moderation, need further building, particularly in global south countries with a history of ethnic and gender tensions.

Our country studies found, overall, that Facebook users in a protected group such as lesbian, gay, bisexual, transgender, queer, intersex and asexual (hereafter LGBTQ+) individuals, may experience a great deal of hate speech content online.[6] Some of the material which is disempowering for targets, and experienced as hateful, does not align with the definition of hate speech used by Facebook, and is therefore left entirely unremedied. Some of the material which does align with the definition of hate speech used by Facebook is not automatically removed, and there is a belief among page moderators that Facebook is not always responsive to their flagging of hate speech content. This may generate what we call 'reporting fatigue', where slow or negative responses from Facebook reviewers reduce the likelihood of users flagging activity. This study suggests that there is a need for more mandatory page admin training in identifying and moderating hate speech, more outreach to protected groups to support their regulatory literacy, and better platform tools for reporting this type of speech. It also suggests that Facebook needs to be more consistent and comprehensive in its proactive strategies for identifying hate speech forms and trends in the APAC region.

3 We had originally requested Facebook provide us with access to a more general sample of hate speech removed from Asian region based accounts, in order to do a comparative analysis of regional hate speech characteristics, but this request was denied on privacy grounds.

4 Australian Community Managers, https://www.australiancommunitymanagers.com.au/

5 Regulatory literacy is a term borrowed from Teeni-Harari and Yadin (2019) in their discussion of media classification. Here it speaks to the need for ordinary citizens to understand speech law - hate speech, defamation and the like - in an age of platform communications, where they are regarded as authors and publishers.

6 We use the term LGBTQ+ in this report to recognise the diversity of lesbian, gay, bisexual and transgender communities, along with other queer, intersex, asexual and non-binary identifying folk. We recognise that other terms include LGBTQIA+ and LGBTQI+ but we use this shorter term for the sake of consistency.

This report concludes with recommendations designed to enhance Facebook's ability to respond to hate speech in ways that benefit target communities and online communicative freedoms, with a particular focus on the company better understanding the local specificities from which hate speech arises.

3. Literature review

Hate speech

Hate speech is regarded as a kind of speech that requires a policy response due to the harms it causes. Although the concept is heavily contested (Brown 2015, 2017), hate speech is widely understood to be a type of expression that harms in a manner that is comparable to more obvious physical injury. Lawrence, for example, argues that what he calls "assaultive racist speech" is "like receiving a slap in the face" and is experienced as a "blow" that, once struck, reduces the likelihood of dialogue and therefore engagement in free speech (1993, 68). Brison suggests that "verbal assaults" can cause "psychic wounds" that constitute ongoing, long term injury (1998, 42, 44). Tirrell describes "toxic speech" as a "threat to the well-being and even the very lives" of its targets. She describes it as "deeply derogatory" speech that, over time, accumulates in its effects like a slow acting poison (2017, 141-42).

We understand hate speech in these terms. That is to say, we understand it not as speech that merely offends someone, or hurts their feelings, but as speech that can harm immediately and over time, and that therefore, and to that extent, warrants policy and regulatory responses. Hate speech discriminates against people on the basis of their perceived membership of a group that is marginalised, in the context in which the speech is uttered. For example, Parekh defines hate speech as "directed against a specified or easily identifiable individual or … a group of individuals based on an arbitrary and normatively irrelevant feature", which "stigmatizes the target group by implicitly or explicitly ascribing to it qualities widely regarded as highly undesirable" and treating "the target group … as an undesirable presence and a legitimate object of hostility" (Parekh 2012, 40-41).

In scholarly literature, it is argued that hate speech can harm in two ways: causally and constitutively (Maitra and McGowan 2012, 6). Causal harms are those that arise as a direct consequence of hate speech being uttered. These can include the adoption of discriminatory beliefs against the target groups, the incitement of discrete acts of discrimination against members of the target group or, at the extreme, discrete acts of violence against those target group members. Constitutive harms are those that are occasioned as a result of the utterance being made. That is to say, the utterance in and of itself is regarded as harmful. Examples of constitutive harms include degrading and persecuting target group members, ranking them as inferior, subordinating them, and legitimating discrimination against them (Langton 2012, 76-80, 86-89; Maitra and McGowan 2007, 62).

The capacity of a speaker to harm, and the vulnerability of a target to be harmed (see esp. Langton 1993; McGowan 2009), depend on the context within which the speech takes place and the norms it both reflects and reproduces. Research on the efficacy of Facebook moderation indicates that the language and context dependent nature of hate speech is not effectively captured by Facebook's classifiers or its global Community Standards and editorial policy (Carlson and Rousselle 2020; Soundararajan et al 2019). It requires local knowledge to identify, and consultation to understand the degree of severity experienced by target groups.

A "systemic discrimination approach" to defining hate speech clarifies that when a person makes an utterance that reinforces and perpetuates extant systemic discrimination against a marginalised group, their speech has the capacity to oppress by virtue of it having taken place in a social context imbued with that discrimination (Gelber 2019). Taken together, this means that hate speech is a discursive act of discrimination, which operates against its targets to deny them of equal opportunity and infringes on their rights (Gelber 2019, 5-6), in much the same way as other acts of discrimination.

Hate speech online

The rapid expansion of online communications and self publishing has rendered the issue of hate speech a growing and urgent problem for regulators and platforms. With around 4 billion internet users globally (ITU 2020), a 122% rise within a decade, and increasing social media use in the Asian region (We Are Social 2021), online platforms have become vital arenas for communication, connection and freedom of expression. At the same time there has been a burgeoning incidence of hate speech online across the globe, interlinked with misinformation and extremist political material, particularly in politically turbulent countries, countries with a history of racism, religious and gender discrimination, and in association

with mass migration due to war, famine, political persecution and poverty (Brooking and Singer 2018).

The problem of how best to combat hate speech online has been at the forefront of public debate about digital content regulation. In May 2016, the European Commission and four major online platforms agreed to cooperate on a voluntary hate speech monitoring scheme (European Commission 2016). In April 2017, the United Kingdom Home Affairs Committee (HAC) concluded an inquiry into "Hate Crime: Abuse, Hate and Extremism", which reported that not enough was being done to combat hate speech online (HAC 2017). In June 2017 Germany passed a law that imposes fines of up to €50m for social media companies that fail to delete hate speech regarded as "evidently unlawful" material (McGoogan 2017). In 2019, following the Christchurch mosque attacks, Australia passed the Sharing of Abhorrent Violent Material Act which requires online service providers, content hosts and social media platforms to remove extremely violent audio-visual streaming content 'expeditiously' (Douek 2020).

At the same time, a voluminous scholarly literature has emerged in the study of specific issues associated with the harms of speech online (e.g. Barker and Jurasz 2019; Citron 2014; Foxman and Wolf 2013; Jane 2018; Leets 2001; Levmore and Nussbaum eds. 2011; Rolph 2010; Williams et al 2020) and attempts at content regulation on social media (e.g. Alkiviadou 2019; Heldt 2019; Matamoros-Fernández and Farkas 2021; Rochefort 2020; UNESCO 2016). There is a growing chorus of voices suggesting that platforms should take greater initiative in regulating harmful speech on their sites, some of which criticise their existing methods (e.g. Gillespie 2018; Murphy 2020; Noble 2018; Suzor 2019), and others that investigate the possibility of automated detection taking down such material (e.g. Del Vigan et al. 2017; Rodriguez, Argueta and Chen 2019).

In this context, it is vital to understand how well hate speech is regulated at a national level as a preventative strategy against social media violations, how well Facebook is addressing the scope of hate speech in specific national contexts, how its users experience and understand hate speech, and how it responds to their concerns. Facebook is facing the joint challenges of devising more effective policy and procedures to better manage harmful content and, in the interests of free speech, being more transparent about how this occurs. Our work addresses aspects of these challenges using a mixed methods, case study-based approach to investigating different aspects of the regulatory puzzle.

4. Methodology

Our study began with two interlinked steps: the establishment of an ideal definition of hate speech, based on the scholarly literature, and an analysis of hate speech legislation in five case study countries from across the region: India and Myanmar (south Asia), Indonesia and The Philippines (south east Asia) and Australia (Pacific).

Nation-states have been, historically, the primary regulators of public expression. However, in the social media era they have been slow to regulate internet expression, with some exceptions in authoritarian regimes such as China and Vietnam. We chose to analyse and compare regulatory responses to hate speech in what were, at the time, five democratic countries.[7] We chose them for comparison as:

1. each has a significant rate of Facebook penetration (We Are Social 2021),[8]
2. each has seen evidence of increasing hate speech content on Facebook (Freedom House 2019; Reuters 2019; SBS News 2021), and
3. they all have growing online LGBTQ+ communities, which have been subject to discrimination (Radics 2019).

Facebook is also a key internet service provider in Indonesia, The Philippines and Myanmar through its Free Basics program, where it partners with local telecommunications companies to provide cheap access, and in India through its direct investment in Reliance Jio, the country's largest telecommunications provider. In each of these countries Facebook faces unique regulatory problems, including national regulation (India, Indonesia and Australia) and coordinated violence (The Philippines, India and Myanmar). Use of the platform to persecute minorities in India and Myanmar has focused international attention on the organisation's difficulties in effectively identifying and removing hate speech (Avaaz 2019; Lee 2019; Purnell and Horowitz 2020).

Our comparative legal analysis of existing hate speech regulation sought to understand the role legislation plays in curbing the rise of hate speech in the region. The task was to determine if specific hate speech legislation existed in our case studies and, if not, what form of regulation a government might instead employ. The study took into account a nation-state's constitution and various provisions within these documents, alongside criminal and penal laws, civil laws, and proposed laws aimed at hate speech, either directly or via mechanisms that included cyber-criminality or inter-faith bills.

We then analysed how Facebook has sought to regulate hate speech, and where it could improve its approach, by undertaking a socio-technical systems analysis of its regulatory roles, policies, procedures and regulatory culture. This approach to investigating organisational change seeks to map the complexity of a system and understand the interdependent factors that contribute to its work or services, and their social application (Figure 1 next page). It has been used to inform the design of new information technology systems, to explore the role of users and the integration of new technology in social contexts (Davis et al. 2014; McEvoy and Kowalski 2019).

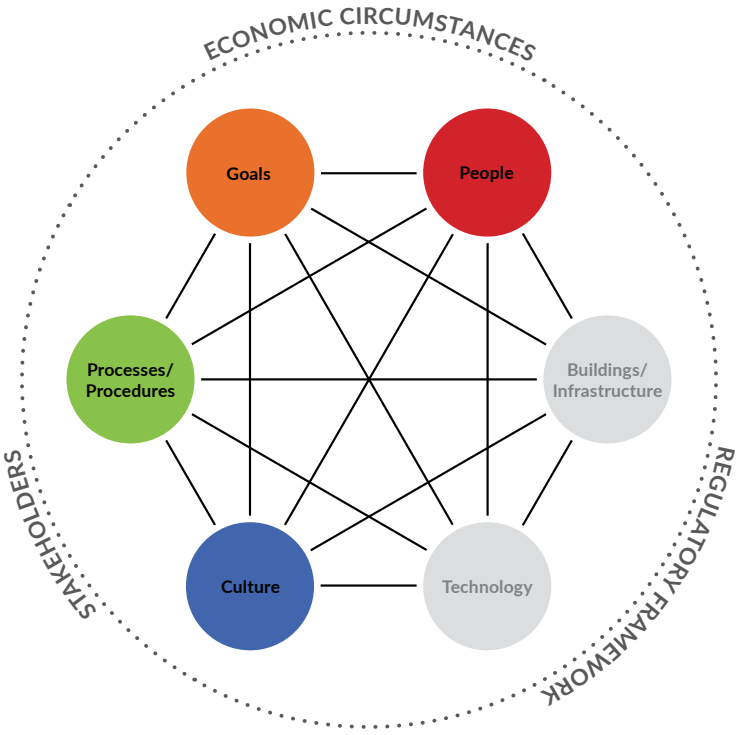
We did not investigate the infrastructure and technology aspects of the system, except where they affected regulatory roles, policies, procedures and culture. Much has already been written about social media as an infrastructure which can be used to extend the reach of hate networks and amplify hate speech (Ebner 2020; Klein 2017; Vaidyanathan 2018; Williams et al 2020). There is an emerging body of studies on how platform design and algorithmic operations support the posting and sharing of harmful content (Munn 2020; Reider, Matamoros-Fernandez and Coromina, 2018), and many studies including Facebook's own on how to design better algorithmic classifiers for harmful content (e.g. Xu et al. 2021).

Our analysis involved semi-structured interviews with Facebook policy staff in the APAC headquarters in Singapore, and information on LGBTQ+ hate speech moderation activities provided by country specialists in each of the case study countries. We also analysed relevant publicly viewable standards and policies, and conducted a literature review of critical responses to its regulatory culture, including the effectiveness of its automated and human moderation approaches.

7 Myanmar's government has since fallen to a military coup.

8 As a percentage of internet users 16-64 years of age, Facebook is used by 96.8% in The Philippines, 85.5% in Indonesia, 77.7% in Australia and 75.7% in India. 63.1% of Myanmar's total population over 13 were Facebook users.

Figure 1. Socio-technical systems analysis framework, adapted from David et al., 2014.



We did not attempt to review the operations of Facebook's contracted content review companies, as their activities have been the subject of significant critical scrutiny, public debate and legal challenge (Barrett 2020; Dwoskin, Whelan and Cabato 2019; Fick and Dave 2019). We were not given access to any internal interpretative materials provided to reviewers and could not assess the extent to which reviewer training materials, and the editorial guidelines which form part of the Community Standards, are available in local languages throughout the APAC region. This is important as it has been alleged that in the past, reviewers had used Google Translate to understand English language versions of editorial policy documents (Fisher 2018). As the identities of many of Facebook's trusted partners remain confidential, we did not attempt a comparative engagement with those that publicise their links to the company, to avoid institutional bias. Our qualitative engagement was limited to interviews with content policy and market specialist staff, although we requested access to other members of the Global Operations and the Community Integrity (Engineering and Product) teams to further explore communication and policy across the regulatory ecosystem.

Finally, we sought to explore the scale and scope of hate which had evaded Facebook's moderation processes, and also user experiences of this violating content, by analysing posts on the pages of LGBTQ+ communities and interviewing the page administrators responsible for moderating this content. Our decision to focus on hate speech against LGBTQ+ communities was driven by three main considerations.

Firstly, LGBTQ+ communities around the world suffer systematic discrimination and persecution both offline and online, making them suitable target communities for our study. Some of the case study countries have experienced a surge in the politicisation of LGBTQ+ rights, while others have not - making them strong candidates for comparative analysis. Secondly, Facebook's community standards recognise sexual orientation and gender identity as protected categories with reference to hate speech regulation, suggesting the efficacy of its policies and procedures could be tested by analysing how well it captured hate directed at these groups. Direct attacks on sexual orientation, sex, gender and gender identity constitute violations against Facebook's community standards for Tier 1, Tier 2 and Tier 3 offences.[9] Thirdly, all five case study countries have national LGBTQ+ communities hosted on public Facebook pages. As ordinary users do not require permission to post comments on these pages, they allow the potential for hate speech to be freely posted. We identified the top three most liked LGBTQ+ pages on Facebook from each case study country (n=15), online communities where we would expect to find significant levels of public interaction, and then examined these pages for evidence of unmoderated discriminatory speech.

We built a corpus of all posts made during 2019 from these public pages, and based our manual content

9 https://www.facebook.com/communitystandards/hate_speech

analysis on a randomly selected sample of approximately 10% of all comments received on these posts. The data collected were systematically anonymised prior to analysis, to protect user privacy. For the purposes of this project, we only examined comments that were text-based and not visual comments containing emojis, memes and GIFs. Replies to comments were also not analysed to maintain consistency. While content analysis is a well established methodology for analysing social media data (Sloan and Quan-Haase 2017), hate speech is a challenging subject for interpretation, so we employed manual content analysis to identify and analyse the subtleties of hateful expression. We undertook a manual sentiment analysis, dividing content into positive, negative, neutral and junk, and identified whether comments could be categorised into hate speech based on Facebook's existing Tier 1-3 categories, and the additional "deprivation of powers" category our analytical framework has proposed.

Given the capacity of page administrators to hide, delete and report hateful comments, we recognised that what we could sample publicly from these pages may not reflect the actual degree of hate speech originally posted on these pages. For that reason we also needed to explore the capacity of these regulatory actors to identify and manage violating content. We sought to conduct semi-structured interviews with at least one page administrator of each case study LGBTQ+ page, hoping to gain deeper insights by discussing our preliminary data analysis findings. Our discussions with page administrators highlighted their critical role in regulating illegal and offensive speech on Facebook. We were interested in finding out how page admins understand hate speech in their given social and cultural contexts, if they are aware of Facebook's community standards on hate speech, and how they manage hate speech content on their pages. Through these interviews we were able to construct a matrix of hate speech management actions that provide additional comparative insights into the way each group manages hate speech content. We were also able to confirm whether the amount of hate speech we had identified from the publicly available data on their pages was approximate to the actual degree of hate speech content they saw posted. This process allowed us to establish greater confidence in our data.

5. Defining hate speech

The term hate speech is used in many different ways. However, it should be used to denote a type of speech that is sufficiently harmful to be regulated, in contrast to types of insulting or offensive speech that should not. Mere offence, or having one's feelings hurt, should not be a standard for the regulation of speech either in civil or in criminal law. Hate speech, by contrast, is viewed as regulable in international human rights law, and in the domestic law of most liberal democratic states (the United States being the exception).

In order to be regulable, hate speech must harm its target to a sufficient degree to warrant regulation, consistent with other harmful conduct that governments regulate. So what is a regulable standard for harmful speech? In this project, we have utilised a definition of hate speech derived from the scholarly literature on hate speech (Gelber 2019). This work argues that, in order for speech to be capable of occasioning harm to a sufficient degree to warrant regulation, it needs to have the following characteristics:

- It needs to take place in 'public', by which we mean that it is reasonably foreseeable that other people will come into contact with the speech unintentionally. Private conversations ought not to be regulable. In this project, we presume that posts on Facebook in anything other than purely private conversations or 'direct messages' can be deemed 'public.'
- It needs to be directed at a member of a systemically marginalised group in the social and political context in which the speech occurs. This means that groups that are not systemically marginalised ought not to be able to claim the protection of hate speech laws. In most Western liberal democratic states, this would include, for example, white, i.e. Anglo Saxon or Caucasians on the ground of their race, or men on the ground of their gender. Note also that in assessing this element, it is important to consider whether the impugned speech:
  - is using terms in a critical way to condemn, or raise awareness about, discrimination and harm.
  - occurs as part of a critical practice of art, journalism, or genuine research.
- Systemic marginalisation is taken to mean pervasive, institutionalized exclusion presenting in patterns and practices that are perpetuated in, and through, ostensibly neutral institutional principles, e.g. racism, sexism, ageism and discrimination against the disabled.
- The speaker needs to have the 'authority' to carry out the speech. This authority can be obtained in three ways:
  1. It can be formally and institutionally derived, e.g. a manager or supervisor in a workplace, a police officer, or parliamentary representative, all of whom possess formal, institutional authority.
  2. It can be informally derived, e.g. when others 'like' or positively react to a hate speech comment, or when they share it, both being acts that can amplify its visibility in news feeds. It is also derived when onlookers remain silent in the face of hate speech; that is, when they accommodate hate and fail to counter its presuppositions.
  3. It can be structurally derived when the speaker enacts oppressive permissibility facts (e.g. racist or sexist comments) in speech that perpetuates systemic discrimination against the group to whom the target of hate speech is perceived by the speaker to belong. This means that sexist or racist comments that take place in a society infused with sexism or racism are granted authority by the very fact of their taking place in that context.
- The speech needs to be an act of subordination that interpolates structural inequality into the context in which the speech takes place. In doing so **it ranks targets as inferior, legitimates discriminatory behaviour against them, and deprives them of powers**. Hate speech achieves this by being an action that sets limits on what is speakable by the target group; this makes it much harder for the target group to speak back with their own counter speech, either because they are fearful to speak out, or because they do speak but their words do not count in the way they intend.

In this project, we adopt the following coding to reflect this definition (Table 1 next page).

We recommend that Facebook consider refining its Tier 3 Community Standards and its editorial policy to better recognise that hate speech ought not to be restricted to the most vituperative or egregious instances of expression.[10] Facebook's current reviewing policy seems

10 The three tiers are defined in Facebook, 2020 Community Standards, Objectionable Content: Hate Speech. https://www.facebook.com/communitystandards/hate_speech

Table 1. Mapping ideal hate speech definition against Facebook Community Standards

	Discrimination	Inferiorisation	Deprivation
Definition	Legitimating discriminatory behaviour such as denying their right to do ordinary things, excluding them, or discriminating against them	Ranking targets as inferior through dehumanisation, using terms that are connected to historical violence and discrimination	Depriving targets of powers
Facebook Community Standard equivalent	Tier 3 (partially)	Tier 1and 2	Tier 3 (partially)
Examples	Saying it's OK for targets to be arrested, denied jobs or housing or rights, or be discriminated against by the bible	Labelling targets as insects, viruses, sub-human, toxic, harmful, poisonous	Denying targets their rights to make everyday decisions, their right to have agency in their own life (so others need to decide for them), their right to exist, existentially denying the validity of their voice

to associate hate speech with vehemence of expression, when it can also be subtle and not overtly aggressive in its framing. One such example is suggesting LGBTQ+ people can simply choose to change their gender identity or sexual orientation, a view that denies their lived experience and identity and underpins the dangerous, human rights infringing practice of conversion therapy.[11]

During our research Facebook expanded its definition of hate speech, which it now publicly updates (Facebook 2021a). However, our data analysis research shows that APAC LGBTQ+ identifying individuals experience forms of hate speech that are not being removed by Facebook, some of which are culturally specific to intersectional experiences, gender communities or ethnolinguistic groups, and some which are focussed on depriving them of powers; that is, denying targets their right to make everyday decisions, their right to have agency in their own life (so others need to decide for them), their right to exist, or that existentially deny the validity of their voice.

While we acknowledge that codifying this latter type of hate speech may be challenging, Facebook should work with protected groups to isolate the expressed forms of deprivation of powers that are commonly applied by hate speakers.

Recommendations

Facebook should work with protected groups to identify the commonly expressed forms of hate speech that deprive targets of powers and codify these in its reviewing policy.

11 For background on the problems with conversion therapy see OHCHR (2020a).

6. Legal analysis and findings

Domestic and regional law

Hate speech legislation across the Asia Pacific is a mixed and varied array of rules, or proposed rules, based in some measure on constitutional law, but also reliant on penal codes and civil laws. These rules often have limited ability to regulate what each nation deems to be hate speech. To date, very little specific legislation targets hate speech explicitly, for varied reasons. Instead, some governments hoping to curb the rise of this activity draw on legislation that targets hate speech tangentially including through cybercriminal law, telecommunications law, safe spaces laws, and other measures. Legislation that specifically targets hate speech has, in countries including The Philippines and Myanmar, been proposed, yet the passing of these bills into national law has been hindered due to their turbulent politics.

Curbing hate speech via legal processes in our case study nations has been a constant struggle with each nation's constitutional right to freedom of speech. Most constitutions include some reference to this foundational principle, though it is expressed differently from nation to nation. India makes direct reference to a person's right to freedom of speech (Article 19 [1]), while Indonesia words this as freedom of expression (Article 28 [e]). The exception is Australia whose constitution only in a limited way, and impliedly, protects freedom of 'political communication.' Across all case studies, a right to a person's freedom of speech, expression and/or personal or political convictions continues to influence governments' attempts at curbing hate speech. As a result, direct hate speech laws, or those laws that act against the crime via other legal mechanisms, have been interpreted by critics, political opposition, and by those in power, to hinder freedom of speech. Hate speech laws, or the amendment of other laws with hate speech as a side consideration, therefore often stall at proposition stage.

Alongside constitutions, Asia Pacific nations need to take into account a region's cultural, religious and ethnic differences, and therefore might attempt to curb hate speech using eclectic legal measures that consider this diversity. These are acts and bills that could be employed (yet often are not) to curb hateful verbal or non-verbal acts committed against religions and religious organisations, ethnic minorities, varied cultures and peoples, and a person's gender and/or sexuality. These legal measures often underscore a nation's inability to accurately define hate speech. Included in legislation that nations think applicable to hate speech are laws that prohibit unwanted or undesirable public interactions (including wolf whistling), nuisance speech that might annoy persons, and speech that humiliates people due to social standing, disability, place of birth, and belief system.

Among this array of laws are criminal laws that fine or jail persons guilty of promoting animosity towards religions, race, groups of people, and persons in positions of power. These laws largely apply to face-to-face interaction (and not the online world), though some Acts have been either amended - or there are proposals for their amendment - to take into account specific hate speech wording, as noted in some of India's penal codes. A different set of laws across our case study nations addresses online or virtual hate speech crimes via cyber-criminality or online space legislation, and these work to limit the publication of disturbing content on the internet that terrorizes or intimidates an online user emotionally and psychologically.

Attempts to curb online hate speech in the Asia Pacific are built into telecommunication, electronic information and cyber-criminality laws, in some instances to directly address hate speech, while in others hate speech comes under their broader remit. For example, The Philippines 2021 Cybercrimes Act does not mention hate speech specifically but is a safeguard of online information. Similarly, The Philippines 2018 Safe Spaces Act prohibits terrorizing and intimidating persons based on gender and sexuality. In Indonesia, the 2008 Electronic Information and Transactions Law was amended in 2016 to allow the government to terminate a person's access to "electronic information or documents with content that violates applicable laws and regulations, such as immoral content, hate speech, insult or defamation" (Molina et al. 2019). These legal measures are not concerned directly with criminalising hate speech insomuch as hate speech is just one form of undesirable online content, but they see hate speech as a possible outcome of telecommunication and social media platforms that might warrant investigation akin to insulting or defaming a person.

Similarly, though from a different legal perspective, in 2017 Myanmar introduced a third draft "Interfaith Harmonious Coexistence" bill that includes hate speech

definitions in its preamble, and the proposed legislation would have criminalised acts of hate speech aimed at religions and faiths (Article 19 2017). Myanmar's bill defined hate speech and proposed the crime to be: "Utterance of hate-speech, reiteration of hate-speech and spreading it out, publicity for hate-speech through information communication technology for the purpose of creating dissent and conflict amongst diverse religious followers and ethnic groups, are strictly prohibited" (Chpt. 6 [10]). While Myanmar's interfaith harmonious bills are proposed laws that define and incorporate hate speech, they can also confuse hate speech with blasphemy.

On hate speech directly, there has been very little legislation in the Asia Pacific region. The Philippines proposed a hate speech law in 2019 (House bill no. 3672), but this has been constantly sidelined for a variety of political reasons. Myanmar proposed the Hate Speech Prevention bill in 2017, and this is yet to become law. Contrastingly, in the case of The Philippines, Myanmar and Indonesia, laws are being designed and enacted to address criminality deemed 'terrorism', including a terrorist organisation's supposed use of hate speech. Yet these legal measures are employed by governments to bolster their own positions; for example, wrongly earmarking anti-government sentiment as hate speech. In so doing, these laws limit freedom of speech while concurrently hindering broad public opinion and stopping criticism of the government (Beltran 2020).

In Australia, a variety of civil and criminal law has been implemented since the late 1980s at a Commonwealth and a state/Territory level. The wording of legislation differs between jurisdictions, and often state/Territory legislation appears a more effective legal measure against hate speech than Commonwealth law, which is only civil. In other case study nations, this legislation is also sometimes tiered at national, regional, and even a city's jurisdiction, though in this study we focussed on national legislation rather than region and city. Unlike Australia in which states remain somewhat autonomous from the Commonwealth, in many Asia Pacific nations including those we study, legal mechanisms at all levels conjoin with national legislation and the ruling party.

Lastly, across the region the only nation to include the term 'incite' or 'incitement' in their laws in relation to acts of discrimination is state/Territory legislation in Australia. Other nations have opted for softer language including 'the promotion of', 'to offend', 'to insult' or 'the dissemination of information.' This complicates the legislation of hate speech in two ways: 1) this form of activity, even at the most egregious end of the spectrum, is signalled as having no serious outcome aside from offence or insult, or the wrongful passing-on of information; 2) repercussions of hate speech such as violence and atrocity criminality are possibly deemed hyperbolic. Hate speech and the incitement of peoples to hatred are therefore semantically coded in these laws as the equivalence of annoyance, teasing or belittling, and carrying out a hate speech act is viewed as just another example of less insidious rhetorical or visual forms of segregation deemed only hurtful or offensive (please see Table 2).

International human rights law

All five case study countries are signatories to, or have ratified a number of international laws pertaining to human rights. An example is the often cited International Convention for the Elimination of Racial Discrimination (ICERD) that The Philippines (1967), India (1968), Australia (1975), and Indonesia (1999) have ratified. Myanmar has neither ratified, nor is a signatory to this convention. Similarly, all five case study nations have ratified the Convention on the Elimination of All Forms of Discrimination against Women (1979) that states that all signatories will "take all appropriate measures to eliminate discrimination against women by any person, organisation or enterprise" (Art. 2 e). These are two examples of international law that might be used by governments to help prevent the spread of hate speech in the region.

Yet forms of hate speech that have been deemed illegal by the international community are speech acts that incite violence. In so being, international law that directly hopes to hinder incitement, such as The Convention on the Prevention and Punishment of the Crime of Genocide (1948), has a better chance of addressing illegal forms of hate speech than select sections taken from other international conventions and charters. In the words of human rights lawyer, Geoffrey Robertson, " . . attempts to stop radical preachers by inventing new hate-speech crimes are doomed to failure. It is an offence to incite violence or incite others to commit murder or to bring about civilian deaths recklessly . . ." (2012, 669). It could be proposed, from an international perspective, that countering hate speech via international human rights law falls under the remit of prohibiting acts of incitement.

Table 2. Case Study Countries' Constitutions, Laws, Bills and Acts

Country	Constitutional protection for free speech?	Other relevant constitutional provisions?	Criminal Law	Civil Law	Proposed new laws?
The Philippines	Bill of Rights (s4)	Human dignity and human rights (s11 Promotes social justice (s10) State has responsibility for full development and communication	Penal Code (1930): Offense against persons (Art 4) Offend any race or religion (Art 201(2)(3))	Civil Code (1950): Article 26(4) Vexing or humiliating … on account of religious beliefs, lowly station in life, place of birth, physical defect Article 694(2) Nuisance that annoys or offends the senses	House Bill No. 6963 proposed 2018: Hate Speech Act: ethnicity, race, religion. Defined as discriminate against and actively incites hostility or foments violence
Indonesia	Freedom of expression (Art 28) Freedom to express views in accordance with conscience (Art 28E)	Article 28I: Protection from discrimination on any ground Article 28J: Respect human rights of others	Articles 154, 155 (1) Expression of hostility, hatred, contempt of government Article 156: Expression of hostility, hatred or contempt against one or more groups of the population Article 157: Dissemination of hostility, hatred or contempt against or among groups of the population	Law No. 40 of 2008: Public expressions of hostility or hatred due to racial and ethnic differences (Arts 4(b), 16). Law No. 39 of 1999: discrimination defined to include degradation Law No. 11 of 2008: deliberate dissemination of information with intention of inflicting hatred or dissession on individuals or groups based on ethnicity, religion, race (Art 28(2))	
Myanmar	(2008): 354 (a) Freedom to express and publish freely their convictions and opinions (d) to develop their language, literature, culture they cherish, religion they profess, and customs without prejudice to the relations between one national race and another or among national races and to other faiths.	Article 364 The abuse of religion for political purposes is forbidden. Moreover, any act which is intended or is likely to promote feelings of hatred, enmity or discord between racial or religious communities or sects is contrary to this Constitution. A law may be promulgated to punish such activity. Article 365 Every citizen shall, in accord with the law, have the right to freely develop literature, culture, arts, customs and traditions they cherish. In the process, they shall avoid any act detrimental to national solidarity.	Penal Code Article 153(a): Whoever by words, either spoken or written, or by signs, or by visible representations, or otherwise, promotes or attempts to promote feelings of enmity or hatred between different classes of [persons resident in the Union] 'shall be punished with imprisonment which may extend to two years, or with fine, or with both. Explanation.-- It does not amount to an offence within the meaning of this section to point out, without malicious intention and with an honest view to their removal, matters which are producing, or have a tendency to produce, feelings of enmity or hatred between different classes of [persons resident in the Union]'	(2013) Telecommunications Law 66(d): Against 'extorting, coercing, restraining wrongfully, defaming, disturbing, causing undue influence or threatening any person using a telecommunications network'	(2017) Hate Speech Prevention Bill. According to leaked document, hate speech described as 'as any language or action that spreads disunity, discrimination and hatred in matters of religion and race or causes racial disputes and conflict.' (2017) Interfaith Harmonious Coexistence Bill (3rd version). Chapter 1 (2). Hate speech defined as 'Hate speech denotes any bodily or verbal action by any manner or by a certain language which can create conflict among diverse religious followers and ethnic groups.' Chapter 6 (10). Prohibited is: 'Utterance of hate-speech, reiteration of hate-speech and spreading it out, publicity for hate-speech through information communication technology for the purpose of creating dissent and conflict amongst diverse religious followers and ethnic groups, are strictly prohibited.'

Country	Constitutional protection for free speech?	Other relevant constitutional provisions?	Criminal Law	Civil Law	Proposed new laws?
Australia	Doctrine of implied freedom of political communication acts as constraint on government, but very weak in practice.	N/A	Commonwealth: urging violence against groups and individuals based on race, religion, nationality, national or ethnic origin, political opinion (Criminal Code ss 80.2A(2), 80.2B(2)). Advocacy of genocide (Criminal Code s80.2D). NSW: public threat or incitement of violence on grounds of race, religion, sexual orientation, gender identity, intersex or HIV/AIDS status (Crimes Act 1900, s93Z). Queensland: serious racial, religious, sexuality or gender identity vilification (Anti-Discrimination Act 1991, s131A). Victoria: intentionally incite hatred and threaten, or incite others to threaten, on ground of race or religion (Racial and Religious Tolerance Act 2011, ss24, 25) WA: Conduct intended or likely to racially harass; incite racial animosity;, threaten seriously and substantially abuse or severely ridicule (Criminal Code ss 77-80D) South Australia: incite hatred, serious contempt, severe ridicule by threatening physical harm, or inciting others to threaten physical harm (Racial Vilification Act 1996)	Commonwealth: offend, insult, humiliate or intimidate on ground of race (Race Discrimination Act s18C). NSW: incite hatred, serious contempt, severe ridicule on ground of race, transgender, sexuality, HIV/AIDS status (Anti-Discrimination Act 1977, s20C, 38S, 49ZT, 49ZXA,) Queensland: incite hatred, serious contempt, severe ridicule on ground of race, religion, sexuality, gender identity (Anti-Discrimination Act 1991, s124A). Victoria: incite hatred, serious contempt, revulsion, severe ridicule (Racial and Religious Tolerance Act 2011, ss7, 8) ACT: incite hatred, revulsion, serious contempt or severe ridicule on disability, gender identity, HIV/AIDS status, intersex, race, religion, sexuality (Discrimination Act 1991, s67A) Tasmania: incite hatred, serious contempt, severe ridicule on ground of race, disability, sexuality, religion, gender identity (Anti-Discrimination Act 1998, s19) South Australia: incite hatred, serious contempt, or severe ridicule on ground of race (Civil Liability Act 1936, s73)	Attempts in 2014 and in 2016 to repeal Commonwealth law.

Country	Constitutional protection for free speech?	Other relevant constitutional provisions?	Criminal Law	Civil Law	Proposed new laws?
India	Article 19. (1) All citizens shall have the right— (a) to freedom of speech and expression.		153A. 'Promoting enmity between different groups on ground of religion, race, place of birth, residence, language, etc., and doing acts prejudicial to maintenance of harmony. (1) Whoever- (a) by words, either spoken or written, or by signs or by visible representations or otherwise, promotes or attempts to promote, on grounds of religion, race, place of birth, residence, language, caste or community or any other ground whatsoever, disharmony or feelings of enmity, hatred or ill-will between different religious, racials, language or regional groups or castes or communities …' 295A Penalises 'deliberate and malicious acts, intended to outrage religious feelings of any class by insulting its religion or religious beliefs.' 298 Penalises 'uttering, words, etc., with deliberate intent to wound the religious feelings of any person' 505 (2) Statements creating or promoting enmity, hatred or illwill between classes. Whoever makes, publishes or circulates any statement or report containing rumour or alarming news with intent to create or promote, or which is likely to create or promote, on grounds of religion, race, place of birth, residence, language, caste or community or any other ground whatsoever, feelings of enmity, hatred or ill-will between different religious, racial, language or regional groups or castes or communities …	The Representation of the People Act (1951) 123. Corrupt practices.—The following shall be deemed to be corrupt practices for the purposes of this Act. (3A) The promotion of, or attempt to promote, feelings of enmity or hatred between different classes of the citizens of India on grounds of religion, race, caste, community, or language, by a candidate or his agent or any other person with the consent of a candidate or his election agent for the furtherance of the prospects of the election of that candidate or for prejudicially affecting the election of any candidate.] 125. Promoting enmity between classes in connection with election.—Any person who in connection with an election under this Act promotes or attempts to promote on grounds of religion, race, caste, community or language, feelings of enmity or hatred, between different classes of the citizens of India. Protection of Civil Rights Act (1955) Section 7 penalises incitement to, and encouragement of untouchability through words, either spoken or written, or by signs or by visible representations or otherwise … Religious Institutions (Prevention of Misuse) Act (1988) Section 3(g) prohibits religious institution or its manager to allow the use of any premises belonging to, or under the control of, the institution for promoting or attempting to promote disharmony, feelings of enmity, hatred, ill-will between different religious, racial, language or regional groups or castes or communities.	Suggested amendments to the Penal Code and the Code of Criminal Conduct to curtail hate speech, as outlined in 'Law Commission of India. Report No. 267. Hate Speech. March 2017) Information Technology Act (2000) Section 66a: criminalises online hate speech, but Supreme Court ordered this unconstitutional as it restricted free speech. Section 69a: allows government to block people's access to the internet, sometimes hate speech.

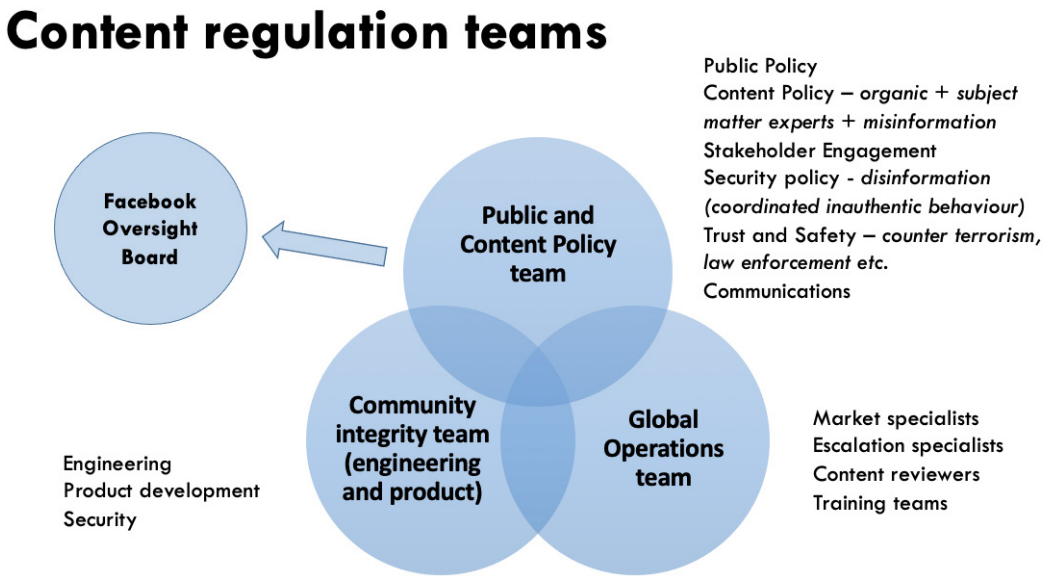
7. Internal regulatory systems analysis

Recent analysis of Facebook's content moderation policy and its development suggests that it borrows concepts and tools from the US legal system and constitutional law "to resolve tensions between regulating harmful speech and preserving free expression" (Kadri and Klonick 2019). Our study was less concerned with analysing the legal basis for its hate speech policy, than understanding the ways in which Facebook's hate speech review process was shaped by its corporate structure, roles, policies, procedures and cultural context.

Facebook's internal content review process involves three areas of the organisation: Public and Content Policy, Global Operations, and Engineering and Product (see Figure 2). These teams work cross-functionally to inform each other, and also draw on the voluntary labour of platform users who report or 'flag' what they perceive to be content violations, the outsourced services of external moderation service providers, such as Accenture and Genpact, who provide contract content review labour, and the cooperation of trusted civil society partners, who provide information and advice to policy staff about violating content trends.

While Facebook's technologies, organisational structure, strategy and policies operate largely at a global level, its procedures for identifying and moderating hate speech, alongside its educational and advocacy activities, are increasingly informed by local information and conditions via its country experts or 'market specialists', its outsourced content reviewer, and its trusted partner organisations. It has a 'glocal' corporate culture, one that operates with global aims, objectives and policies, but which is constantly being influenced and re-aligned by its investments in specific nations and regions, and its impacts on those societies. This dynamic can be seen, for example, in its responses to the 2018 independent human rights assessment it commissioned into how its platform enabled hate speech dissemination and was used to incite violence in Myanmar (Frankel 2020; Warofka 2018).

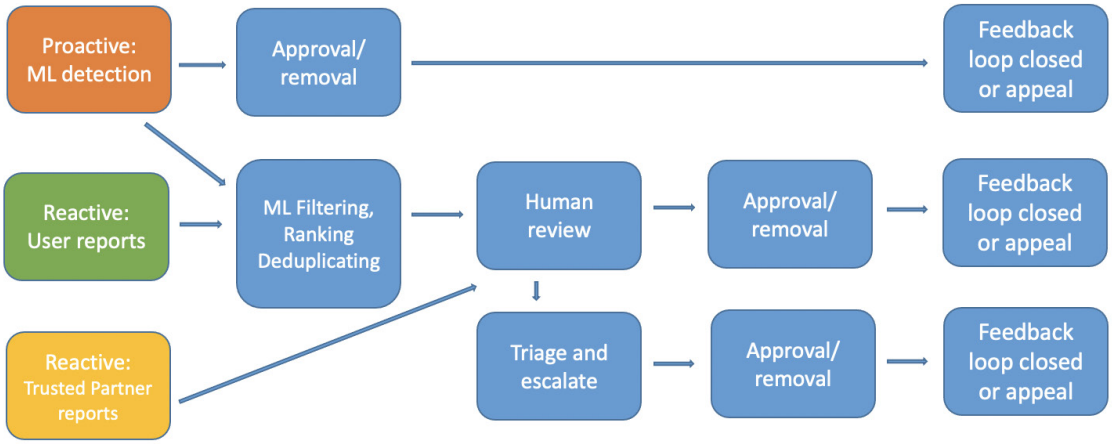
Figure 2. Facebook content regulation ecosystem



Facebook's content regulation areas were growing during the period of our study with, for example, Stakeholder Engagement expanding from four staff internationally to a dozen staff by February 2020. Similarly, new market specialists were being employed and roundtables were being held with academics and subject matter experts on how to address harmful content. This growth introduced some fluidity to the organisational structure with, for example, changes to global team titles and individual roles. While Facebook indicated that these teams work cross-functionally, we were unable to verify the scope or effect of this interaction for hate speech regulation because we did not have interview access to review experts or reviewers in Global Operations or any roles in Community Integrity.

We could not, for example, assess the efficacy of Global Operations' outsourced content moderation, which has been questioned in recent studies (Carlson and Rousselle 2020; Murphy 2020; Soundararajan et al 2019). Publicly available models of the content moderation process provided by Facebook have not included trusted partners as source of 'reactive' or post-comment reports, or indicated the possibility of appeal. Figure 3 indicates the content moderation process as indicated by our interviewees.

Figure 3. Facebook content reviewing workflow



In reviewing Facebook's moderation processes, the Yale Law School Facebook Data Transparency Advisory Group (FDTAG 2019) could not assess the accuracy of its automated detection, or the efficacy of Facebook's internal checks on the consistency and accuracy of human reviewer decision-making. Based on Facebook's description of its human error auditing process the report indicated it seemed "well designed and likely to be helpful in leading to increasing levels of accuracy over time" (15). However we agree with their suggestions that Facebook could:

1. release error rates for moderation of each type of violation,
2. release statistics on reversals of content removals after appeal,
3. make apparent the percentages of violating content detected internally via machine learning and human moderation
4. release reports on human review audits in specific languages or locations, to track improvement over time
5. systematically test and validate internal reviewer moderation judgements against those of users. (FDTAG 2019, 16-17)

As our case study analysis indicates further below, there is some user disagreement about the accuracy of internal review of hate speech. Our research indicates more work needs to be done on the consistency of hate speech policy application to moderation decision-making, and on providing more detailed feedback to those flagging violating content, and those whose content has been removed. This latter point is consistent with the European Commission's most recent Code of Conduct trial finding that "most of the IT companies must improve their feedback to users' notifications" (EC 2020, 1) in order for users to better understand the nature of Facebook's community standards, and have informed access to appeal.

For our regulatory systems analysis, our interviews were restricted to Policy team members, and data provided by market specialists. These constraints meant focusing our investigation on the problems of identifying new forms of hate speech and working with external partners to mitigate this problem.

Facebook's Asia Pacific (APAC) office is based in Singapore, and some of its key insights into the changing conditions for hate speech emergence in different national contexts are drawn from its counter terrorism team, its internal market specialists, its external trusted

partner program and its stakeholder engagement team. These roles feed into the processes of developing policy, enforcement mechanisms and products.

The counter terrorism team gathers intelligence or 'signals' about emerging threats, working with safety and security and enforcement teams to identify dangerous organisations and to map hate networks using a range of analyticals methods. Intelligence gathering is complex and politically fraught, and the company can not act prematurely to remove hate figures, networks or content or it will diminish public trust. Key problems for counter-terrorism are that bad actors:

- are adversarial and adaptive, looking to circumvent any measures that control their activities. This can affect the extent to which Facebook communicates its strategy around hate speech control.
- express 'hate' using terms which avoid policy violations and enforcement measures e.g. making indirect rather than direct threats, condemning rather than attacking. This deliberate obfuscation can make automated detection difficult.
- may post relatively benign content as a means of recruiting users to engage in more extreme offline activities.
- can mirror each other's tactics in reciprocal radicalisation; for example, terrorist actors adopting the network structures of white supremacists.

Facebook's counter-terrorism team liases with community and civil society groups, academics and law enforcement, to develop insight into bad actors, and to assemble the evidence needed to sanction these actors. It is unclear how counter terrorism identifies its external contacts, and to what extent it strategically maps the CSO groups it engages with. More public transparency about the range of bodies and groups consulted would be helpful for external stakeholders in establishing whether there are gaps in Facebook's knowledge acquisition and outreach strategy, but Facebook has indicated this is not possible in many cases due to local political sensitivities. It would, however, help potential CSO and law enforcement collaborators if Facebook made clearer the types and weight of evidence needed to take action on hate figures and groups.

Market specialists are appointed to oversee local content regulation issues in most Asian national territories, including northern and southern India. They review content which has been reported for potential abuse, investigate user account issues, gather and analyse data about emerging trends, and look for ways to improve the user experience by contributing to better policies, processes and tools. Market specialists keep an eye on events and user behaviour that could be potentially problematic, and provide early warning about the evolution of hate speech and groups promoting hate. One key issue for future appointments is to ensure candidates have sufficient insight into the scope of intersectional discrimination within their jurisdiction.

Trusted partners are politically non-aligned, not-for-profit civil society organisations, which have a trusted public profile in their countries of origin, and which do not derive any financial support from, or have any affiliation with, government. They have direct lines of communication to Facebook's market specialists and its public policy staff, which enable them to report harmful developments in social media communications. Trusted partners have become essential in identifying hate speech evolution; for example, in flagging emerging hate organisations or new trends in violating content:

> *There might be a new slur in Burmese that our systems have not picked up on. There might be a new emoji that's being used in a certain market to depict hateful ideologies or to depict a certain group...and so these trusted partners...if they report something... it will get immediate attention. (Interviewee A 2020)*

Trusted partners can also explain a rise in user reports of otherwise innocuous images or memes which have become metaphors used by hate groups or in hate speech contexts.

The Trusted Partner Channel team also handles direct email contact from organisations and institutions in urgent and serious public matters, escalating this information to the Policy team. The channel was established in 2018 in response to the Office of the High Commissioner for Human Rights' investigation into the use of Facebook in Myanmar to mobilise hate against the Rohingya Muslim minority (OHCHR 2018). It enabled the OHCHR to engage with Facebook directly and expediently about ongoing hate speech issues it was identifying on the platform.

However, there are three issues for the efficacy of the Trusted Partner scheme. The first is the difficulty of finding appropriate partners in authoritarian or highly politically polarised countries. Second, there is little public information about the program and so it is unclear to CSOs whether they meet the conditions to act as a trusted partner, even if they are willing to undertake this role and meet the qualifications. Third there is little transparency about which organisations Facebook is working with, and how everyday users or community groups can contact them in order to escalate reports of serious hate content. While some organisations publicly declare the nature of their work with Facebook, others do not wish to be identified for safety reasons in order to avoid a backlash from extremist groups, political parties or governments. This leaves a gap in the emergency response strategy for hate speech reporting in global south countries of the Asia Pacific, where there are no publicly declared trusted partners. In this respect it could be useful for Facebook to nominate international or regional trusted partners which operate at a supranational level such as Digital Rights Watch, as escalation contacts for serious public concerns.

Stakeholder Engagement is another avenue of input to formulating hate speech policy and procedures. This

team seeks out input from academics, civil society, journalists, independent research groups, and other experts on potential hate speech mitigation strategies, feeding them back into the Public Policy team, which interacts with governments, members of parliament and politicians around issues of policy and regulation. Key issues for these teams are reconciling the diverse opinions on what constitutes hate speech and how it should be moderated:

> *there's extreme freedom of expression stakeholders in this region who just think you're not there to police the internet, it is not your job to sanitise the internet, people should be able to read the most horrendous hate speech and make their own mind up about it. And then others who think we're not doing enough. (Interviewee A 2020)*

At the time of our organisational interviews in December 2019-January 2020, Facebook had not yet held an APAC roundtable on hate speech. This type of event is critical where civil society complaints and media reports suggest that particular types of hate, such as Islamophobia, are on the rise in the region (Aljazeera 2020; SBS News 2021; Shaheed 2021). More regular outreach events may help Facebook to better manage CSO concerns, and offset, for example, recent threats of legal action from Muslim groups.

Recommendations

Facebook should:

- make transparent the types and weight of evidence needed to take action on hate figures and groups, to assist law enforcement agencies and civil society groups in collating this information.
- audit the trusted partners program in the APAC region to ensure it is comprehensive and the parameters for membership are clearly stated and publicly accessible
- make public a trusted partner in each country, or nominate a supranational trusted partner for the APAC region, so that individuals and organisations have a direct hate speech reporting partner for crisis reporting issues.
- conduct an annual APAC roundtable on hate speech involving key non-governmental stakeholders from the protected groups in all countries.

8. Country case studies

In response to our request for interviews with market specialists in the case study countries Facebook provided us with an outline of LGBTQ+ initiatives in the APAC region and specific to these countries. It works with the LGBTQI+ groups to celebrate identity and engage in counterspeech efforts; to encourage the wellbeing of LGBTQI+ people; and to hear concerns and consult on platform and product policies. The company noted, for example that since January 2020 it has consulted LGBTQI+ advocacy groups throughout APAC during the development process for the following policies:

i. Conversion therapy in advertisements,
ii. Outing-risk groups,
iii. Commercial surrogacy,
iv. Slurs in a positive context,
v. Dating policies (prior to the launch of Facebook Dating), and
vi. Hate speech: concepts vs people.[12]

Facebook indicates that these consultations have led to concrete policy changes: for example, it now prohibits conversion therapy content that promotes claims to change or suppress one's sexual orientation, or to suppress one's gender identity. Other hate speech relevant activities are included in the country case studies.

Indonesia

Indonesia is home to expansive LGBTQ+ communities with a growing number of advocacy organisations and increased social and cultural engagement in recent years. Many of the organisations working on LGBTQ+ rights are mainly concerned with health issues, such as HIV/AIDS, as some of the founders of LGBTQ+ groups had personal experiences with the disease in the early 2000s and were able to receive support from international organisations such as the United Nations (UNAIDS 2006). More recently, LGBTQ+ groups have begun to focus on countering homophobia, discrimination and hate speech, and relying more on social media platforms to raise awareness of LGBTQ+ rights (Adjie 2020). In the past decade, LGBTQ+ groups have formed alliances with feminist, sexual and reproductive health and pro-democracy and human rights groups to broaden their support base. Despite the scaling-up of efforts by LGBTQ+ communities to advocate for more rights, there is increasing discrimination against them and, reportedly, significant levels of hate speech on social media (Renaldi 2021).

There has been a marked increase in hate speech against LGBTQ+ communities since 2016, as a direct result of the Indonesian Minister of Technology, Research and Higher Education, Muhammad Nasir's suggesting a ban of LGBTQ+ people on university campuses as they were seen to threaten Indonesian morals and norms (Harvey 2016). During the presidency of Joko Widodo (Jokowi), hate speech against LGBTQ+ groups has multiplied as conservative Islamic organisations have grown in power and influence. Anti-LGBTQ+ public comments by government officials, threats to LGBTQ+ Indonesians by state commissions, militant Islamists and mainstream religious organisations, have continued unabated (Human Rights Watch 2016; McDonald 2020). This shift towards a more Islamist, less tolerant Indonesia has worsened discrimination against LGBTQ+ people despite increased advocacy and activism by LGBTQ+ groups. Discrimination against the LGBTQ+ communities in Indonesia remains widespread with the Indonesian Psychiatrist Association classifying homosexuality, bisexuality and transsexualism as mental disorders that can be cured through proper treatment (Yosephine 2016).

The most relevant legislation to hate speech in Indonesia is the 2008 Electronic Information and Transaction Law, ITE (*Informasi dan Transaksi Elektronik*), which was amended in 2016. Article 28(2) prohibits the following: 'any person who deliberately and without authority disseminates information without intention for inflicting hatred or dissension on individuals and/or certain groups of community based on ethnic groups, religions, races and inter-groups" (Republic of Indonesia 2008). This law is concerned with defamation and online blasphemy but has been applied in ways that limit freedom of speech, according to some journalists and human rights advocates. Hundreds of websites have been blocked as the government deems the information 'negative content' – vaguely defined to include anything from pornography to immorality.

12 By this Facebook is referring to the way in which they remove hate speech directed at people, but generally not hate expressed towards concepts, such as a religion, corporation or the monarchy.

In 2020, Facebook noted that even though the platform has provided a safe space for Indonesia's LGBTQ+ community to communicate and organise on rights advocacy, it has also been used for "harassment, bullying, and involuntary 'outing' of LGBTQ+ users" (Facebook 2020c). In response to an independent human rights impact assessment of its Indonesian activities (Article One 2018) the company has since hired a policy lead and program manager in Indonesia, increased the number of content reviewers who speak Indonesian and Javanese, and improved its automated content detection capability in Indonesian.

The Indonesian LGBTQ+ Facebook groups we focused on in our content analysis are Perkumpulan Arus Pelangi, Suara Kita and Yayasan GAYa NUSANTARA. We manually collected all 2019 posts from their pages and analysed the content of randomly selected comments, which constituted approximately 10% of all comments. We examined 190 comments from Arus Pelangi, 40 comments from Suara Kita and 146 comments from Yayasan GAYa NUSANTARA. We interviewed three page administrators from these groups.

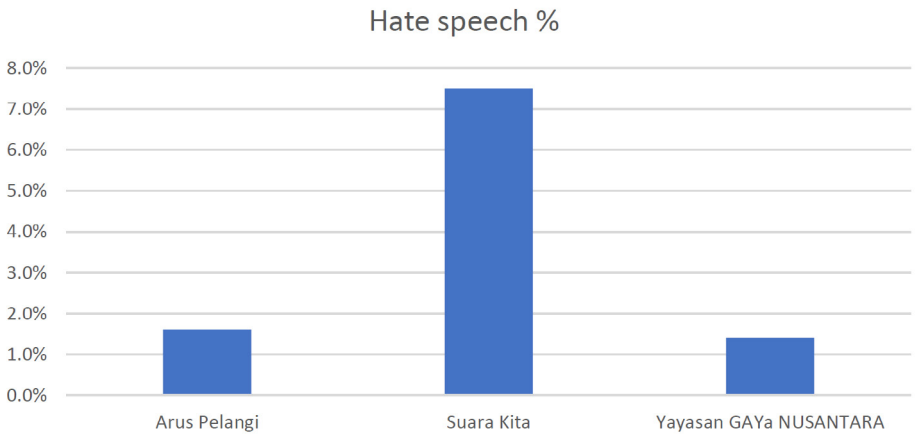
Key findings

1. The most common form of hate speech content publicly visible on these pages is 'discrimination', but the most common hate speech comments the admins receive (not publicly available data) is 'inferiorisation.'
2. Indonesian LGBTQ+ page admins are the least inclined of our case study organisations to directly engage with hate speech commenters on their pages.
3. All pages were 'attacked' by an anti-LGBTQ+ movement that provided 1-star ratings to their pages on Facebook, which the admins find demoralising
4. None of the admins of these groups receive training on how to manage hate speech either from Facebook or third-party organisations
5. When page admins contacted Facebook to report on hate speech content, they received automatic responses and no follow-up. They felt deterred from contacting Facebook further.

Results and analysis

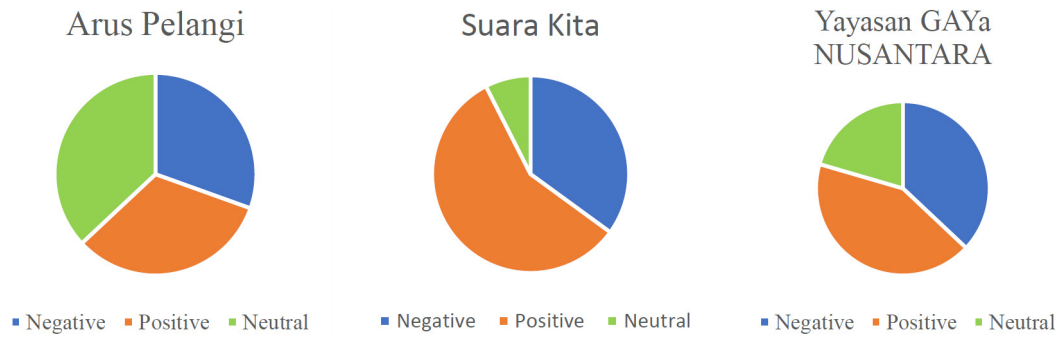
Hate speech content was publicly observable on the pages of Arus Pelangi, Suara Kita and Yayasan GAYa SUNANTARA and this constituted less than 10% of all comments (Figure 4). Suara Kita has the most hate speech comments, at 7.5% of our sample, despite being the group with the lowest relative volume of comments. Arus Pelangi and Yayasan GAYa SUNANTARA had less than 2% of comments on their pages displaying hate speech content (1.6% and 1.4% respectively). This finding is welcome considering the hostile and dangerous environment within Indonesia for LGBTQ+ groups online and offline. The admin of Arus Pelangi, one of the largest and oldest LGBTQ+ groups in Indonesia, credits the low incidence of hate speech to Facebook's word blocking and profanity filter tools which they used from the very beginning of their account. The admins of other pages also note that they are not as active on Facebook as YouTube and Twitter, where they see more hate speech content.

Figure 4. Percentage of hate speech content found in Indonesian LGBTQ+ Facebook groups' 2019 comments



Looking at the sentiment of comments, which we divide into three types – positive, negative and neutral, we do not find any consistent pattern to suggest that more positive comments on a page also results in fewer hate speech comments. Overall, negative comments represent one-third of all comments visible on the three groups' pages, but they have varying levels of hate speech content. Both Surara Kita and Yayasan GAYa SUNANTARA had a similar level of negative comments on their pages in 2019, at 35% and 37% respectively, yet the latter group has received far fewer unmoderated hate speech comments than the former. There is no correlation between the level of positivity in the comments and a lower likelihood of receiving hate speech comments.

Figure 5. Indonesia case study comment sentiment analysis, 2019 Indonesia



Examining the data by type of hate speech, we find that the most common form of hate speech was discrimination – comments that deny target victims' rights to do ordinary things and discriminate against targets based on religious beliefs (Figure 6). Examples of hate speech comments found include:

- Religious-based hate speech comments, such as, "the gays go to hell", including how they should be punished for acting against God's will.
- Some of the hate speech comments threaten direct physical violence against LGBTQ+ people such as stoning to death and beheading.
- Suggestions that being homosexual constitutes deviant behaviour, reflecting a narrative perpetuated by Indonesian political and religious elites

Interviews with the page admins of Indonesian LGBTQ+ pages revealed three important insights. First, even though they were able to articulate what hate speech is, none of them had received formal training from Facebook or other third-party organisations on how to manage hate speech. They have taken it upon themselves to interpret hate speech according to the narratives in their society, and they all expressed some difficulty in interpreting hate speech comments. They suggested it would be useful to receive training directly from Facebook so they are better equipped to deal with hate speech content.

Secondly, page admins refrained from engaging directly with hate speech commenters on their pages. They either left the comments alone or, in the case of GAYa NUSANTARA, deleted them. Admins felt that trying to engage in conversations with, or to educate, hate speakers is futile or may escalate the issue. Indeed, unlike the Yayasan GAYa NUSANTARA admin, the others were relatively non-interventionist – not deleting, blocking, or engaging with hate speech commenters.

Thirdly, page admins reported having received hate speech content that would be classified as 'inferiorisation' through other channels, such as their personal pages or in private messages. They were called, for example, pigs and dogs - abuse which in a predominantly Islamic society constitutes an act of dehumanisation. These comments conform to the pattern of religion-based hate speech seen on the groups' public pages.

Figure 6. Number of hate speech comments by type of hate speech, Indonesia

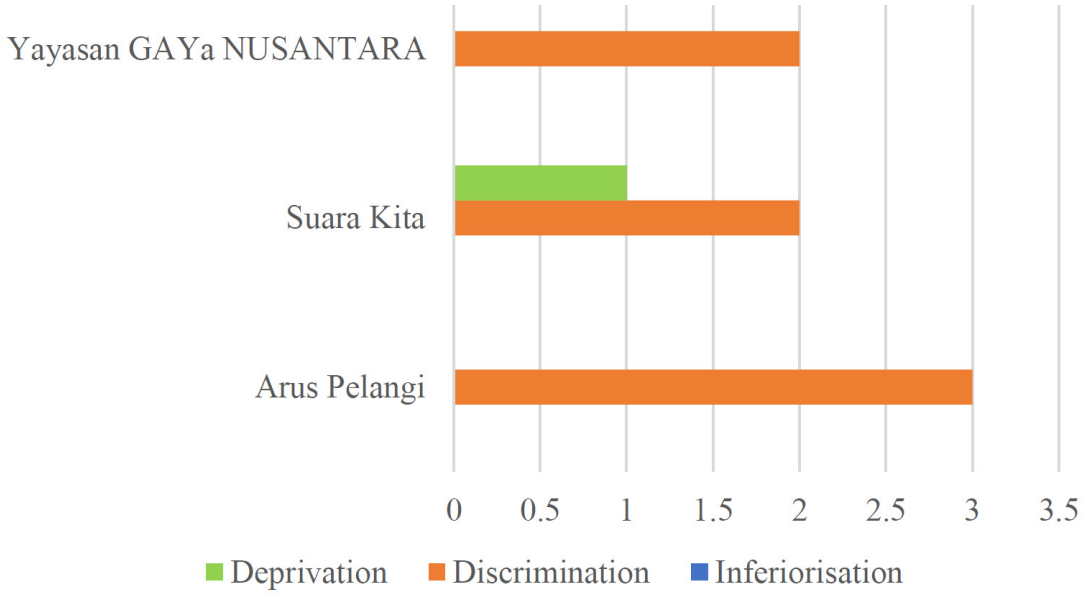


Figure 7. Hate Speech Management Actions Indonesia

Management action	Arus Pelangi	Suara Kita	Yayasan GAYa NUSANTARA
Daily post moderation	X	X	X
Familiarity with Facebook Community Standards on hate speech	X	X	X
Hiding/Deletion of hate speech posts			X
Blocking of hate speech			X
Engagement with hate speakers (replies, private messages)			
Intentionally leaving some hate speech unmoderated on the page	X	X	X
Having received direct physical threats			X
Banning of individuals due to hate speech			X
Taking screenshots of hate speech for reference	X		X
Reporting of hate speech to Facebook	X		X
Taking further action against hate speech with law enforcement or human rights bodies			

In sum, Indonesia presents a case of a society whose LGBTQ+ communities are vulnerable to hate speech on Facebook. It is a country with no legislation to safeguard the rights of LGBTQ+ people, while its political, social and religious elites publicly perpetuate anti-LGBTQ+ narratives with impunity. Our analysis of the comments on Arus Pelangi, Suara Kita and Yayasan GAYa NUSANTARA pages in 2019 reveals a generally low level of publicly visible hate speech comments, despite the admins of the pages being relatively non-interventionist in comparison to cases in other countries. This would suggest that Facebook has been partially successful in filtering out anti-LGBTQ+ words. However, the admins also asserted that they receive a much higher volume of hate speech on other social media platforms: YouTube, Twitter and Instagram. Religious-based hate speech comments are the most prevalent on the pages we examined. Two of the LGBTQ+ organisations have directly reported hate speech content to Facebook but find the process disempowering, and they are not inclined to further engage directly with Facebook when it comes to reporting.

The Philippines

The Philippines is home to vibrant and growing online LGBTQ+ communities. However, members of these groups commonly experience discrimination and bullying on social media (Human Rights Watch 2017; Judson et al. 2020).

The Philippines is in the midst of passing a highly politicised and controversial piece of legislation – the Sexual Orientation and Gender Identity (SOGIE) bill (2000). The bill aims to prevent acts of discrimination against people based on their sexual orientation, gender identity and expression. The SOGIE bill has been passed in The Philippine lower house but has languished in the Senate, and is widely believed to be supported by key members of the Liberal Party, several of whom have been subject to significant hate speech attacks on their public Facebook pages over some years (e.g. Risa Hontiveros, Bam Aquino) (Torregoza 2018). While President Rodrigo Duterte himself is not ideologically against the bill, opponents of the bill are members of both the opposition and his own administration (Rey 2019).

Currently, The Philippines does not have a specific legal framework that directly deals with hate speech. Its Constitution protects freedom of speech generally, but historically this instrument has never been applied to protect minorities based on sexual orientation. The amendment on article 3 section 4 prevents any laws being passed that could limit freedom of expression. Other relevant laws, such as the Penal Code (1930), article 4, only protect individuals' rights on the basis of race and religion, not sexual orientation.

The LGBTQ+ groups we focussed our hate speech data analysis on in The Philippines were LGBT Philippines, Bahaghari LGBT and Mindanao Pride. Together these groups have around 20,000 likes and were the three largest LGBTQ+ Facebook groups in the country at the

time of data collection. We manually collected all posts across these three groups in 2019 and analysed the content of randomly selected comments that constituted approximately 10% of all comments. There were 164 comments for LGBT Philippines, 238 comments for Bahaghari, and 155 comments for Mindanao Pride in 2019. We then interviewed page administrators from each group.

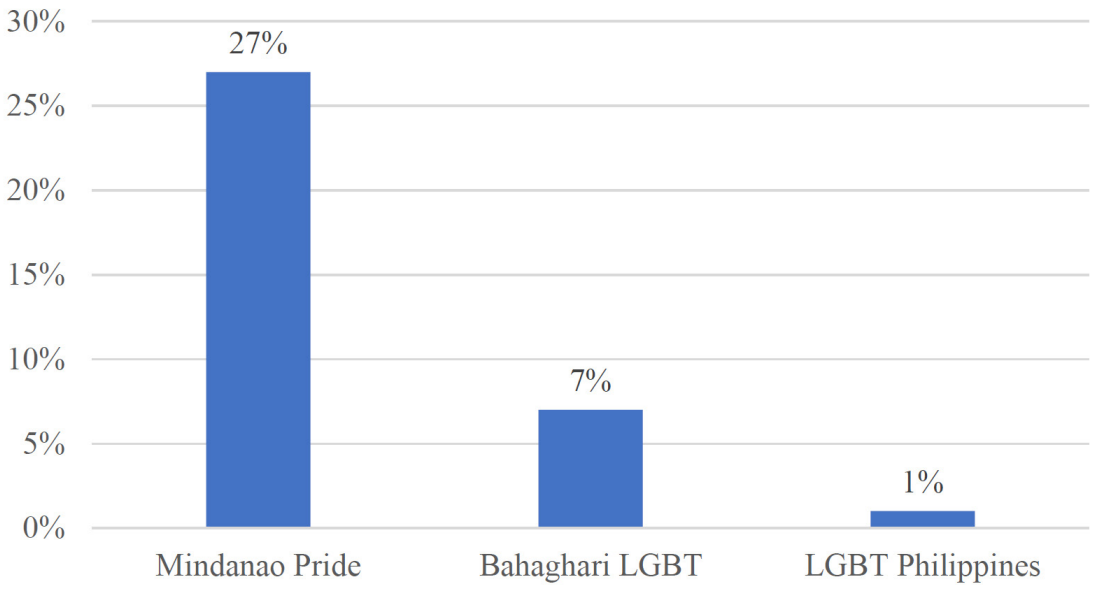
Key findings

1. The Philippine LGBTQ+ communities analysed experience the highest level of publicly observable hate speech across the 15 groups studied in the 5 case study countries.
2. The most common form of hate speech against LGBTQ+ communities is 'deprivation' – verbal actions that deny target users' rights to express their views and negate the validity of their opinions
3. The profile characteristics of the most vulnerable targets are being Muslim, male and gay
4. The politicisation of LGBTQ+ issues in The Philippines is believed to have increased hate speech incidents on Facebook, particularly from pro-Duterte supporters
5. LGBTQ+ groups find their efforts to get Facebook involved in managing hate speech on their pages futile and ineffective, and feel disempowered by the process.

Results and analysis

Analysing comments relating to all posts in 2019 across three pages, we find that Mindanao Pride has a relatively high level of hate speech content, while Bahaghari and LGBT Philippines received very few hate speech comments (Figure 8). Out of 155 posts, Mindanao Pride received 42 hate speech comments, while Bahaghari received 17 (out of 238). Indeed, Mindanao Pride has the greatest number of hate speech comments found in all of the 15 cases in 5 countries under study.

Figure 8. Percentage of hate speech content found in Filippino LGBTQ+ Facebook groups' 2019 comments



Looking at the sentiment of comments, which we divide into three types – positive, negative and neutral, we find that there seems to be a correlation between negativity and hate speech in The Philippines cases. Pages that receive a high number of negative comments also receive a higher number of hate speech comments (Figure 9). Three-quarters of comments analysed for LGBT Philippines were positive and the page received only two hate speech comments. Comments on Bahaghari's page were also overwhelmingly positive and the page received 7% of hate speech comments. Conversely, 41% of comments found on Mindanao Pride page were negative and the group received an almost equal number of hate speech comments.

Examining the data by type of hate speech, we find that the most common form of hate speech was 'deprivation' – comments aimed at existentially depriving targets of power to express their own opinion (Figure 10). One of the more frequently found examples of deprivation type of hate speech targeted at the LGBTQ+ groups in The Philippines were:

- The Philippines has too many other problems that are priorities rather than improving LGBTQ+ rights
- Other gays and lesbians are not asking for more rights, so you should not either
- You need to focus on helping the President on other more important issues rather than call for more rights for yourselves

Figure 9. Comment sentiment analysis, 2019 Philippines

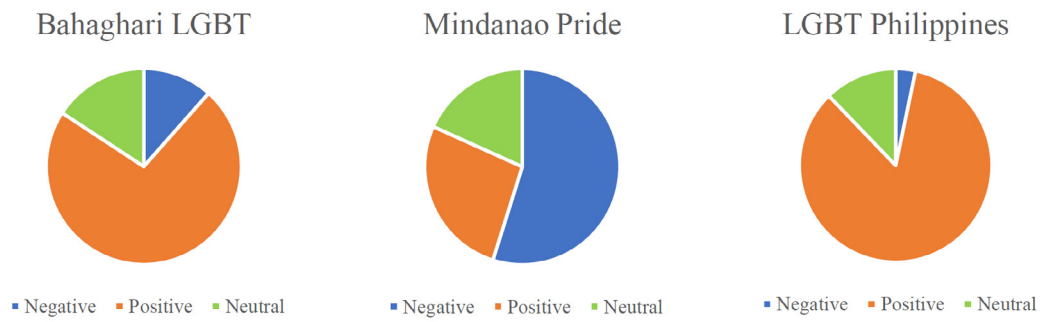
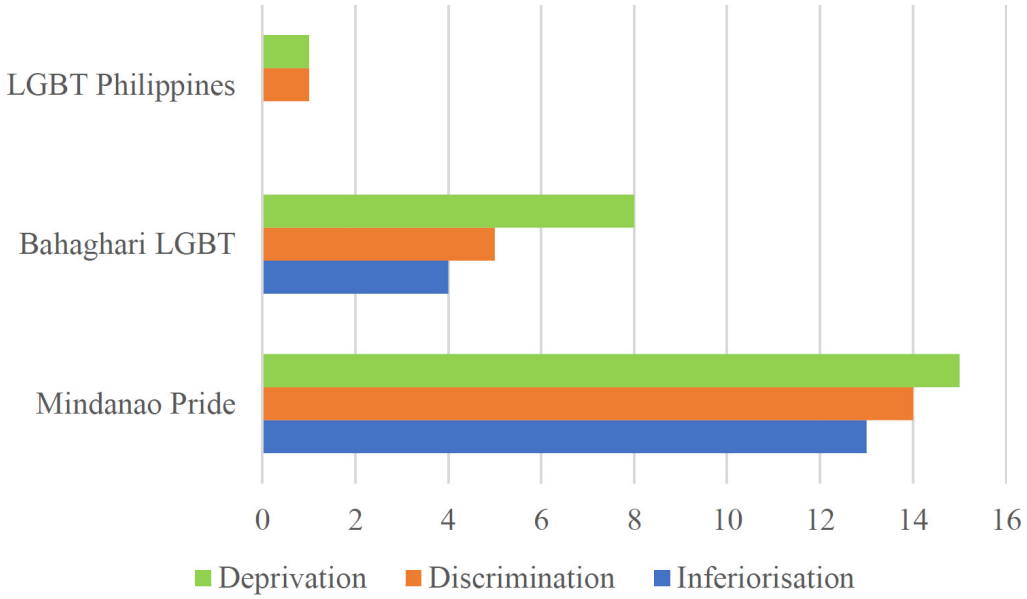


Figure 10. Number of hate speech comments by type of hate speech, Philippines



This finding is particularly important because 'deprivation' has only been recently recognised as type of hate speech under Facebook's existing Community Standards, and so may not yet be well identified by Facebook filters or human reviewers.

Note that we do find higher instances of hate speech comments in sub-national languages. In the cases of The Philippines, these include Bisaya, Maguindanao and Marano, which are languages spoken in Mindanao. As we understand that existing algorithms used on social media platforms are primarily focused on majority languages, rather than minority ones, our findings are especially pertinent to future efforts to improve text detection in code-switching or sub-national languages.

Our interviews with page administrators revealed several critical aspects of hate speech management.

Mindanao Pride regularly received hate speech comments on its page. The admin stated that initially the group sought to manage hate speech comments by engaging with the comments through replies, but they soon realised such actions did not hinder hate speech. Subsequently, the group decided to not engage with or remove the comments. However, in December 2018, Mr. Rhadem Camlian Morados, the admin of Mindanao Pride, began to receive death threats and threats of kidnapping via the group's page. He and his friends have separately reported these incidents to Facebook but received no response. As he continued to receive violent threats and intimidation, Mr Morados filed an affidavit with the National Bureau of Investigation (NBI), the Commission on Human Rights, and the Criminal Investigation and Detection Group in July 2019. The case was also filed with the United Nations in November 2019 for an Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity; Special Rapporteur on the promotion and protection of the right to freedom of peaceful assembly and association; Special Rapporteur on the situation of human rights defenders; and Special Rapporteur on freedom of religion or belief, pursuant to Human Rights Council resolutions 32/2, 34/18, 41/12, 34/5 and 40/10. To date, there has been no update on the case he filed with either The Philippines or international organisations.

Figure 11. Hate Speech Management Actions, Philippines

Management action	Mindanao Pride	LGBT PH	Bahaghari
Daily post moderation	X	X	X
Familiarity with Facebook Community Standards on hate speech	X	X	X
Hiding/Deletion of hate speech posts	X	X	
Blocking of hate speech	X	X	X
Engagement with hate speakers (replies, private messages)	X		X
Intentionally leaving some hate speech unmoderated on the page	X		X
Having received direct physical threats	X	X	
Banning of individuals due to hate speech		X	
Taking screenshots of hate speech for reference	X	X	
Reporting of hate speech to Facebook	X	X	
Taking further action against hate speech with law enforcement or human rights bodies	X		

The LGBTQ+ Philippines page admins actively remove comments and block accounts of those engaging in hate speech. They feel frustrated that when they report hate speech content to Facebook, they regularly receive what appear to be 'automatic messages' that state the content provided does not constitute hate speech according to Facebook's Community Standards. The admins have resorted to filing a complaint with the national police anti-cybercrime unit. However, there is no law in The Philippines against hate speech and therefore nothing has eventuated from the complaint. They are hopeful that the SOGIE bill will change this situation.

Bahaghari takes a different approach to managing hate speech. The page admins refrain from removing hate speech comments or blocking accounts of those engaged in hate speech. As much as possible, they leave the hate speech comments on their page and seek to fruitfully educate those who leave hate speech comments. As Bahaghari is not just an LGBTQ+ advocacy group, but a human rights organisation as well, the admins note that they receive more hate speech comments when discussing other, non-LGBTQ+ issues. On some occasions, they have been 'red-tagged' or labelled as supporters of a communist rebel group (the New People's Army), and in these instances the admins do remove comments from their page.[13]

Interviews with admins of these pages provide invaluable information about the state of hate speech experienced by these pages. Across all sites, admins confirmed they properly understand what hate speech is and are aware of the definition used in Facebook's community standards. They manage hate speech content first by trying to engage with individuals who post hate speech comments. Following this initial first step, these groups diverge in their responses, ranging from doing nothing to banning specific accounts from further engagement with the page. Two of the three pages, which receive physical threats to the organisations or individuals working for the organisations, directly communicate with Facebook to take further action. Yet they both find engaging with Facebook ineffective and disempowering. They do not feel that Facebook genuinely cares about hate speech and feel less inclined to engage with Facebook when encountering further hate speech.

Australia

Australia is the case study country with the most liberal attitudes to LGBTQ+ identifying people. Nevertheless, Australia's eSafety Commissioner (2020) estimates that around one in seven (14%) Australian adults aged 18–65 were the target of online hate speech in the year to August 2019, with LGBTQ+ and Aboriginal or Torres Strait Islander Australians experiencing this abuse at more than double the national average. Further, only

13 'Red tagging' or red-baiting are common slurs used to denounce human rights advocates, independent journalists and political activists in The Philippines (OHCHR 2020b)

6 years ago the Australian Human Rights Commission found that 91% of LGBTQ+ people in Australia knew someone who had experienced violence on the basis of their sexual orientation or gender identity (AHRC 2015).

Of all the countries in this study, Australia has the strongest anti-discrimination protections for LGBTQ+ communities. Some (but not all) of its anti-hate speech laws include incitement against sexuality/homosexuality and transgender as a basis for legal action. Additionally, other areas of law protect LGBTQ+ communities from discrimination. The Commonwealth *Sex Discrimination Act 1984* protects people from discrimination on the grounds of sexual orientation, gender identity and intersex status (ss 5A, 5B, 5C). Many states/territories also have anti-discrimination legislation that covers these grounds. In 2017 the federal government amended marriage laws to achieve marriage equality for same sex couples.

Section 18C of the 1975 federal *Racial Discrimination Act 1975* makes it unlawful for someone to commit a public act that is reasonably likely to offend, insult, humiliate or intimidate another person or group because of their race, colour or national or ethnic origin (with exemptions for art, scholarship and news reporting or genuine commentary). The Commonwealth *Criminal Code Act 1995* contains two sections that could pertain to hate speech prosecutions. Section 11.4 covers the offense of 'incitement', for example where an individual encourages people to commit a serious harm against others. Section 80.2 (1) probits urging violence against members of groups, based on the targeted person's race, religion, nationality, national or ethnic origin or political opinion. However none of these laws protect people on the basis of their sexual orientation or gender identity. However New South Wales, Australia's most populous state and home to Australia's annual Gay and Lesbian Mardi Gras, did in 2018 introduce a new section 93Z into its Crimes Act making it illegal to publicly threaten or incite violence on the grounds of race, religion, sexual orientation, gender identity, intersex or HIV/AIDS status. Several other states also have anti-discrimination or vilification laws.

As well as having a relatively robust legislative framework to prevent discrimination and hate speech, Australia also has unique social media laws, including the *Sharing of Violent Abhorrent Material Act* (2019), which is part of the federal Criminal Code. This Act requires online content, internet service and hosting providers to report to federal police, and remove violent audio-visual content accessible via their services that shows a person engaging in a terrorist act, murder, torture, rape or violent kidnap that has occurred, or is occuring, in Australia. Australia's parliament is also in the process of approving an updated version of its *Enhancing Online Safety Act 2015*. The *Online Safety Bill 2021* has several provisions relevant to managing hate speech online, in that it:

> *... specifies basic online safety expectations; establishes an online content scheme for the removal of certain material; creates a complaints-based removal notice scheme for cyber-abuse being perpetrated against an Australian adult...and establishes a power for the eSafety Commissioner to request or require internet service providers to disable access to material depicting, promoting, inciting or instructing in abhorrent violent conduct for time-limited periods in crisis situations. (Online Safety Bill 2021)*

More broadly Australia has, like other country case studies, a problem with the degree of unmoderated race hate posted on social media platforms (SBS News 2021; Ware and Seear 2020).

In the second half of 2020 Facebook established the Combatting Online Hate Advisory Group to consult on Australian hate speech trends. The group includes "multiple representatives from the LGBTQI+ community...[who] ... bring personal perspectives of gay, trans, non-binary, and Aboriginal and Torres Strait Islander LGBTQISB people to the discussion" (Facebook 2020d).

The Australian LGBTQ+ Facebook groups we focused on in our hate speech content analysis are Sydney Gay and Lesbian Mardi Gras Festival, which is liked by 410 thousand people and followed by 408,000, Australian Marriage Equality, liked by 286,000 people and followed by 282,000, and LGBT Rights Australia, liked by 316,000 people and followed by 282,000. We manually collected all 2019 posts from their pages and analysed the content of randomly selected comments, which constituted approximately 10% of all comments. We examined 9084 comments from Sydney Gay and Lesbian Mardi Gras, 2900 comments from Australia Marriage Equality and 5997 comments from LGBT Rights Australia. We spoke to one page administrator.

Key findings

1. No statistically significant capture of hate speech content against LGBTQ+ groups, despite providing the highest number of comments per page in case study groups
2. Extremely low incidence likely due to effective moderation by page admins
3. However, unmoderated racism and harmful content present which would normally have been removed by professional community managers
4. The only page admin interviewed had not had formal training in moderating hate speech
5. Australian Community Managers network suggested admins need better tools to address hate speech:
   - ability to switch off comments on risky threads (actioned by Facebook in 2021 following the ACCC's Digital Platforms Inquiry and Voller vs Nationwide case)

- page admins should also have group admin option for users to report violating posts to either admins or Facebook

Results and analysis

Despite the significantly larger number of comments examined from Australian Facebook pages than those in our other Asian region case studies, there was no statistically significant incidence of hate speech in our Australian sample. As a result we have omitted the data analysis in this case study. However, based on the results for all other case studies, and our one interview, it is likely that hate speech does escape the Facebook filters and is later removed by page administrators. Given Australia's relatively stronger speech protections for LGBTQ+ rights than other case study countries, and the sector's expertise in rights advocacy it is likely that any hate speech posted would be quickly removed by site administrators of these large public accounts. There was however, unmoderated racism and harmful content on these pages which otherwise would have been removed by professional community managers, according to advice from our Australian Community Managers network researcher.

One page admin for the Mardi Gras festival Facebook site responded to our request for an interview. They indicated that they checked the page every day for offensive speech, with alerts to their phone "all the time." They used the profanity filter on the page, hid comments "if it gets too intense" and blocked some accounts - although rarely for hate speech. They have reported posts to Facebook, although they "can't remember it being a satisfying response." They also occasionally receive requests from members of the community asking them to report posts, although they encourage the users themselves to take that action themselves.

The admin recalled that the most recent troubling response followed targeted promotion of a transgender festival event, which received a lot of negative, transphobic comments after the ad targeting did not reach the desired audience: "every time I looked at a post, someone was going 'get out of here'." This raises the issue of Facebook advertising training for marginalised groups to avoid them reaching out to ambivalent or hostile audiences.

Mardi Gras's page administrator had not had training in how to moderate a Facebook page or group, and was not familiar with Australia's hate speech laws. They had read Safe and Strong: An LGBTQ+ Guide to Facebook and Instagram, a resource produced by Facebook and Instagram in partnership with the AIDs Council of NSW and Trans Pride Australia (ACON 2020). They said they followed the community's guidelines and norms rather than trying to enforce the platform's rules.

India

LGBTQ+ rights in India have strengthened considerably in recent years, especially after a Supreme Court ruling in September 2018 that decriminalised gay sex. This judgement, which LGBTQ+ rights activists had campaigned for for decades, found that discrimination on the basis of sexual orientation is a fundamental violation of rights to privacy, liberty, equality, and human dignity (Dixit 2020). It is one of the legal advances that has laid the groundwork for better legal and political protections for LGBTQ+ communities. In 2011 transgender people were officially identified as the third sex in the national Census and in December 2019 the Indian parliament passed The Transgender Persons (Protection of Rights) Act.

However, much work on rights advocacy is still needed as education, housing, work and training opportunities are still often denied LGBTQ+ identifying people, who often face harassment, bullying and violence in these aspects of everyday life (ICJ 2019). Those LGBTQ+ people who join rights advocacy social media communities, especially those from Muslim communities, commonly experience discrimination and harassment online (Deutsche Welle 2021; Knight 2019), largely as a result of reactionary social attitudes borne of historic legal discrimination, religious and political conservatism. One social media study suggests opponents of the decriminalisation of gay sex, see LGBTQ+ people as a threat to Indian culture, family systems and marriage as an institution (Khatua et al. 2019)

Generally, the right of equality before law and equal protection under the law is guaranteed by Articles 14 and 21 of the Indian Constitution. However, the LGBTQ+ community in India has suffered legal discrimination since 1871, primarily through Section 377 of the Indian Penal Code (IPC). This colonial-era law, which was annulled in the 2018 case Navtej Singh Johar vs Union of India (Columbia Global Freedom of Expression 2018) was a ban on voluntary sexual intercourse between people of the same sex. The law referred to this as an 'unnatural offence' and those found guilty could be punished with imprisonment for life, or up to 10 years jail and a fine. University of London law lecturer Mayur Suresh argues that expressions of disgust and contempt towards LGBTQ+ people were common in 19th and 20th century legal commentary, and that the Indian government remarked in 2003 that decriminalising homosexuality would "open the floodgates of delinquent behaviour" (Suresh 2018).

Religious discrimination against LGBTQ+ people in India is a more complex phenomenon tied to political trends. Worldwide Muslim majority states and religious leaders currently regard homosexuality as unnatural and a punishable offense, as part of a fundamentalist understanding of Islam (Rehman and Polymenopoulou 2013). Since 2018, there has been an attempt by India's Hindu nationalists to revise their country's recent

aversion to homosexuality, and to promote the narrative that precolonial Hinduism was accepting of peoples of diverse genders and sexualities, so as a religion it is superior to Islam. However this position overlooks the discriminatory tendencies embedded in the Hindu caste system (Uphadhyay 2020), which underpin hate speech against Dalit and minority ethnic LGBTQ+ people (Soundararajan et al 2019).

More broadly speaking, India also has a significant political problem with anti-Muslim and misogynist hate speech (Avaaz 2019; Chaturvedi 2016; Soundararajan et al 2019). This is framed against India's introduction of its new *Citizenship Amendment Act 2019*, which allows immigrants from religious minorities in neighbouring countries eligibility for citizenship rights - unless they are Muslim. As a response to rising tides of hate speech online, Facebook has publicly promised to better limit the distribution of violent and hateful content in the lead up to 2021 state elections (Reuters 2020). While it has already banned T. Raja Singh, a member of Indian Prime Minister Narendra Modi's Hindu nationalist Bharatiya Janata Party for posting anti Muslim hate (Purnell and Horwitz 2020), his youth brigade and several fan groups remain online and he retains an Instagram account. In 2020, Ms Ankhi Das then Facebook public policy lead for South India and Central resigned following accusations that she was unwilling to support a harder moderation line against community standards violations by BJP politicians and Hindu nationalists, arguing that 'punishing violations by politicians from Mr. Modi's party would damage the company's business prospects in the country' (Purnell and Horwitz 2020).

More recently Facebook responded positively to the Indian government's introduction of a new social media law, the *IT (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021*, which imposes takedown and complaints conditions on significant social media intermediaries with more than five million users (Mendiratta 2021). This new law, part of the Information Technology Act, 2000, section 87 (2) requires Facebook and other major platform companies to take down illegal content at the request of the government or its agency within 36 hours, and non-consensual sharing of intimate images with 24 hours, to appoint Grievance Officers who are required to acknowledge complaints about content and resolve them within 15 days, and to publish a monthly compliance report on complaints actions and content takedowns. However, it is unclear how complaints can be made under the new law.

In 2019, Equality Labs' report found that Islamophobic, caste and LGBTQ+ oriented hate was the most common in their sample of hate speech posts from Facebook. This team found the majority (93%) of all hate speech posts reported to Facebook during their study remained on the platform, including Tier 1 hate speech (Soundararajan et al 2020). 11% of reports were not replied to and 43% of all initially removed posts were restored after an average period of 90 days from the date of reporting. In their advocacy they also found that "Facebook staff often lacked the necessary cultural competency and literacy in the needs of caste, religious, and gender/queer minorities" (17).

In 2019 Facebook launched a program in India called Building Social Cohesion and Inclusion online, which involved at least 10 local LGBTQI+ groups, including Tamil Nadu LGBTQ, Qknit and Point of View, in regional workshops held across 5 Indian states. The company notes that they "received training in Facebook's policies and international standards on hate speech, tools to report violating speech and leveraging campaigns tools to effectively counter negative speech." (Facebook 2020d). In 2020 it also convened an India roundtable with rights organisations and individual activists, and participated in the Queer Muslim Project's roundtable with 20 participants, discussing challenges around representation and censorship, and successful strategies for creating content and communities online.

The Indian LGBTQ+ Facebook groups we focused on in our hate speech content analysis are Queerala, a community organization for Malayali LGBTQ+ people, an ethnic group from Kerala state; the Gaysi Family began as an LGBTQ+ community group and blog over a decade ago; and Yes We Exist distributes LGBTQI+ information. Queerala is liked by 49 thousand people and followed by 51 thousand, the Gaysi Family, liked by 19.6 thousand people and followed by 20 thousand, and Yes We Exist, is liked by 39 thousand people. We manually collected all 2019 posts from their pages and analysed the content of randomly selected comments, which constituted approximately 10% of all comments. We examined comments from 404 comments from Queerala, 28 comments from Gaysi Family, and 692 comments from Yes We Exist. We interviewed page administrators from Queerala and Gaysi family.

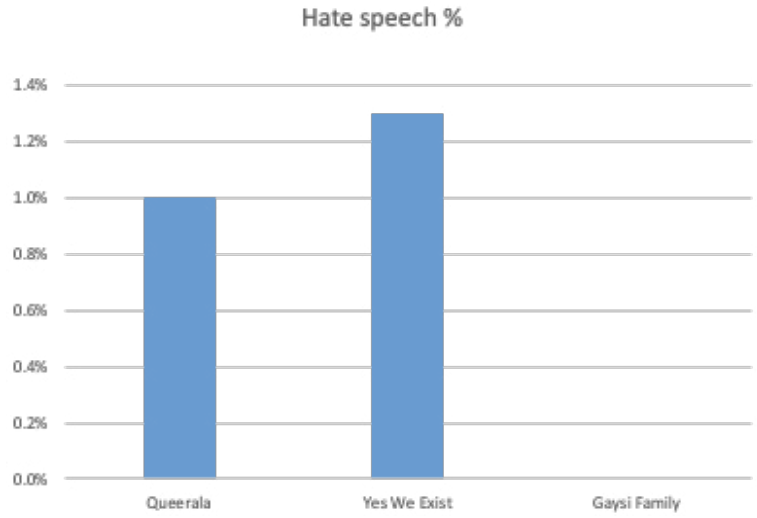
Key findings

1. Very low incidence of visible hate speech due to active page admin moderation
2. Hate speech is most often directed at lesbians and gay Muslims
3. Page admins also reported threats and hate from Hindu nationalists in reaction to posts criticising the Modi government and 'right wing' politics
4. Common use of vomiting emojis as a hate reaction e.g. to gay wedding images
5. Both groups note manipulation of individual photo posts, with hateful captions superimposed
6. Admin interviewees were disappointed that Facebook reviewers would not remove content they had flagged as discriminatory or hateful
7. Admin interviewees indicated that Facebook rarely acts on individually flagged hate speech, so they sometimes collectively report egregious comments.

Results and analysis

Analysing comments relating to all posts in 2019 across three pages, we find a very low incidence of hate speech. Of 404 Queerala comments sampled only 1% were identified as hate speech and of 692 comments on the Yes We Exist page, only 1.3% were hate speech. None of the 28 sampled comments on the Gaysi Family page were offensive. Our page administrator interviews suggest that the low incidence of visible, unmoderated hate speech is due to active management techniques. Queerala has a team of admins working on comment moderation around the clock. Gaysi Family has a team of three, one in the UK, to address comments across timezones. Their work tends to focus more on Instagram, where they have more engagement than Facebook, although their moderation practices are roughly the same.

Figure 12. Percentage of hate speech content found in Indian LGBTQ+ Facebook groups' 2019 comments



Examining the data by type of hate speech, we find no statistically significant difference in the types of hate speech although discrimination appears marginally more prevalent in our small sample (Figure 14). According to our page administrator interviews, some of the more common types of hate speech were:

- Slurs against lesbian couples and Muslim men
- Vomiting emojis in response to gay wedding images
- Accusations that homosexuality is a disease, and it can be treated - it shouldn't be celebrated

The Queerala admin noted that pictures of two women marrying receive more hate comments than pictures of two men getting married, because representations of Indian womens' sexuality are more strongly 'policed', that is surveilled and critiqued, than male sexuality. They also suggested that if they post about an individual member, that person receives more hate speech responses on their personal profile than the group page.

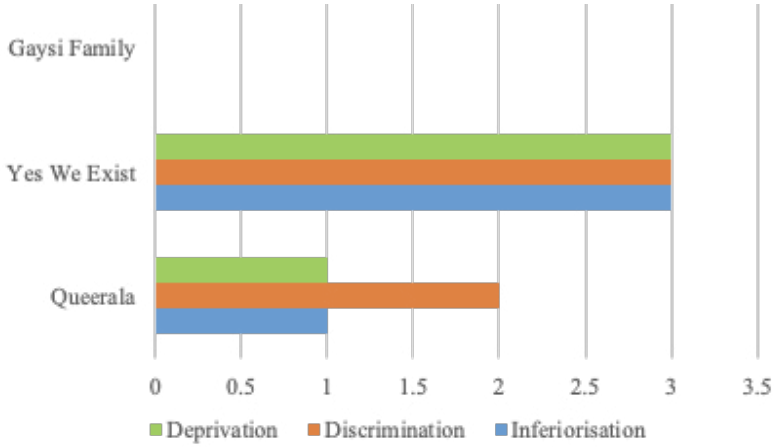
Both groups noted that images of LGBTQ+ people are sometimes copied and reposted elsewhere with hate speech superimposed, and with aspects of the image altered to avoid Facebook's automated image moderation filter. The Queerala admin noted the most difficult speech to moderate was moral blackmail or public shaming, where individuals accused LGBTQ+ people of ruining families or familial relations.

Both groups engage in counterspeech as well as hiding and sometimes removing comments. The Gaysi Family admin used to hide or delete hate comments depending on their tolerance for the content on the day in question. However, recently our interviewee noted they have realised that just deleting hate gives the perpetrators the opportunity to complain that their voice is being silenced. Instead the admins sometimes try to respond to hate speakers and to address perpetrators' bigotry directly, warning them that they will be banned from the page if they continue to post hate. However, the admin noted the emotional labour of counterspeech can be draining: "Some days I respond to them and on some days I don't have the energy to respond to someone who clearly comes from a place where they don't want to learn, they just want to tell you that your existence is wrong and you should die."

Queerala's admin has attended digital security workshops run by Point of View, a not for profit gender rights advocacy group established by filmmaker Bishakha Datta. However, they were not aware that Facebook offered information about banning and moderation of posts on its Help Centre.[14] The Gaysi Family page admin

14 www.facebook.com/help/248844142141117/banning-and-moderation/?helpref=hc_fnav

Figure 13. Number of hate speech comments by type of hate speech, India



had not had formal training in moderating hate speech and relied on their experience of marginalisation to interpret what is hateful. They did not realise that hiding hate speech did not remove it from the view of the hate speaker or their network, and when informed this was the case, called this moderation option 'useless.' They also indicated that they were 'confused' by Facebook's definition of hate speech because many times when they reported what they interpreted as hate posts, their complaints were not upheld.

The Queerala page administrator provided several examples of unmoderated hate speech, one of which had been reported to Facebook and approved by the reviewer even though it clearly violates Tier 1 hate speech rules (Figure 14).

Both page administrators complained that reporting speech to Facebook is not often effective, and the Queerala admin wondered whether Facebook takes reports seriously. The Gaysi admin wondered whether inconsistencies in Facebook review standards were due to reviewers' personal, internalised bias and whether this was an issue that training might address.

Both groups used collective strategies to organise their response to hate speech. The Gaysi community members will respond to hate comments if the page admin has not yet responded, in informal reactive moderation. Queerala uses Whatsapp groups to discuss their response to repeat offenders and will alter their community members to report the same hate comment, so that is flagged many times in Facebook's review process. The admin said that their success in getting content removed seems to depend on the scale of flagging. However sometimes posts that were removed were reposted only days later, so the group was putting a lot of energy and effort into flagging with little effect. It was unclear whether these posts had been restored after the originating poster appealed to Facebook against the removal.

Neither group had reported hate posts to the police or taken legal action to prosecute perpetrators. The Queerala moderator said that they would not report hate speech to the police, because they would not take any action unless the reporting person had influence with 'the authorities.' However, Gaysi Family has in the past,

Figure 14. Example of flagged violating content and Facebook support message

Figure 15. Hate Speech Management Actions, India

Moderation activities	Queerala	Gaysi Family
Daily post moderation	X	X
Familiarity with Facebook Community standards relating to hate speech	X	
Hiding/Deletion of hate speech posts	X	X
Blocking of hate speech	X	X
Engagement with hate speakers (replies, private messages)	X	X
Intentionally leaving some hate speech unmoderated on the page		
Having received direct physical threats	X	X
Banning of individuals due to hate speech	X	X
Taking screenshots of hate speech for reference	X	
Reporting of hate speech to Facebook	X	X
Taking further action against hate speech with law enforcement or human rights bodies		

managed to get transphobic images taken down from a violating Instagram account after they gained personal access to a member of the Instagram policy team.

In summary, these Indian LGTB+ groups on Facebook are often subject to hate speech. The low incidence of unmoderated hate speech found in this study is due to active moderation by the page administrators, including hiding, deleting and reporting of comments and hate speaker accounts to Facebook. Both groups use rotating moderation to constantly monitor comments, but Queerala also used private Whatsapp groups to collectively discuss regulatory strategy. Neither group had confidence in the efficacy of reporting hate to Facebook or trust in the accuracy and consistency of reviewer decision-making.

Myanmar

Myanmar is home to a burgeoning LGBTQ+ community, much of which exists online. A religious and conservative society, there is still a strong stigma associated with being LGBTQ+ (Colors Rainbow and Equality Myanmar 2020). In comparison to other countries in the Asia Pacific, like Thailand or The Philippines, being part of the LGBTQ+ community remains unsafe in Myanmar. But there are signs that things are changing for the better as LGBTQ+ organisations in Myanmar have become more vocal in recent years in demanding greater support and recognition, with more public activities to bring attention to LGBTQ+ rights and participation in the pro-democracy protests (Hlaing and Fishbein 2021). Although the 2008 Constitution acknowledges the right to equality before law regardless of race, sex or religion, there is no official legal status of acceptance or legal rights for LGBTQ+ individuals in Myanmar. There is no specific legislation that protects members of the LGBTQ+ community from discrimination or vilification. As such, the only legal recourse members of the LGBTQ+ community facing abuse and/or harassment is if such action constitutes defamation.

Currently, Myanmar does not have a specific legal framework that directly deals with hate speech. The most relevant legislation is Article 66(d) of the *Pyidaungsu Hluttaw Law No. 31/201 (Telecommunications Law)* (2013) and its 2017 amendment. According to Article 66(d), 'blackmailing, bullying, making wrongful restraint on, defaming, disturbing, exerting influence on or threatening a person using telecommunication network" can carry a maximum prison sentence of 2 years, a fine of one million kyats, or both. This legislation has not been used for hate speech cases in Myanmar. Human rights activists worry the law can be used to silence free speech, which has been the case in the past couple of years as activists and journalists who are critical of the government have been persecuted using this legislation (Reporters Without Borders 2017).

The use of Facebook accounts to vilify the Rohingya Muslims, and to fuel the genocide against this ethnic minority, and Facebook's inability to control this

wave of hate speech, have been widely condemned (Choudhury, 2020; Miles 2018; Venier, 2019). Part of Facebook's response to the UN's Independent Investigative Mechanism for Myanmar (IIMM) was to provide some of its data to aid the UN in its investigation and to commission its own independent human rights assessment, from Business for Social Responsibility (BSR). Among its many recommendations the BSR report (BSR 2018), argued Facebook should "Proactively draw upon local stakeholder insights to improve Community Standards enforcement" and "research the distribution characteristics of hate speech in Myanmar and act upon relevant findings" (5). It also urged Facebook to support the development of Unicode translation of Burmese, and to invest further in digital literacy and counter hate speech initiatives. In response, Facebook has expanded its expertise across engineering, product and policy to work specifically on Myanmar and to improve its proactive detection software of hate speech in Burmese and ethnic languages.

Key findings

1. The Myanmar LGBTQ+ communities experience the lowest level of publicly observable hate speech across the 15 groups under study in the case-study 5 countries
2. The most common form of hate speech against LGBTQ+ communities is 'inferiorisation' – verbal actions that dehumanise and inferiorise targets
3. There is a strong inverse correlation between positivity of content sentiment and hate speech: the higher the positive content on a group's page, the lower the observable hate speech content
4. Page admins feel that the ongoing politicisation and marginalisation against other minority groups – the Rohingyas and Muslim Myanmar people in particular – have shielded the LGBTQ+ communities from hate speech
5. None of the LGBTQ+ page admins is aware of Facebook's community standards in relation to hate speech. Their training on monitoring hate speech is delivered by third-party international organisations.

The LGBTQ+ groups we focus on in Myanmar are Colors Rainbow Yangon, And PROUD, and Diversity for Love.[15] Together these groups have around 136,000 likes and are well known community LGBTQ+ groups. We manually collected all posts across these three groups in 2019 and analysed the content of randomly selected comments that constituted approximately 10% of all comments. There were 425 comments for Colors Rainbow Yangon, 286 comments for and PROUD, and 173 comments for Diversity for Love.

Results and analysis

Analysing comments relating to all posts in 2019 across three pages, we find that there is generally a very low number of hate speech comments visible (Figure 16). Hate speech content does not constitute more than 2% of all comments analysed. Across the five countries we examine, Myanmar has the lowest incidence of hate speech content found on the pages of LGBTQ+ groups. This is welcome news as much media coverage of hate speech in Myanmar in recent years has focussed on atrocities perpetrated against the Rohingya Muslim minority. The findings also demonstrate that hate speech is not experienced equally across different types of minority groups even within the same social media ecosystem. Just because some sub-minority groups experience increased hate speech does not mean such experience is shared with other minority groups. The vast majority of comments on our case study pages were in Burmese and not in other minority languages of Myanmar.

15 And PROUD's account is no longer accessible at the time of writing.

Figure 16. Percentage of hate speech content found in 2019 comments

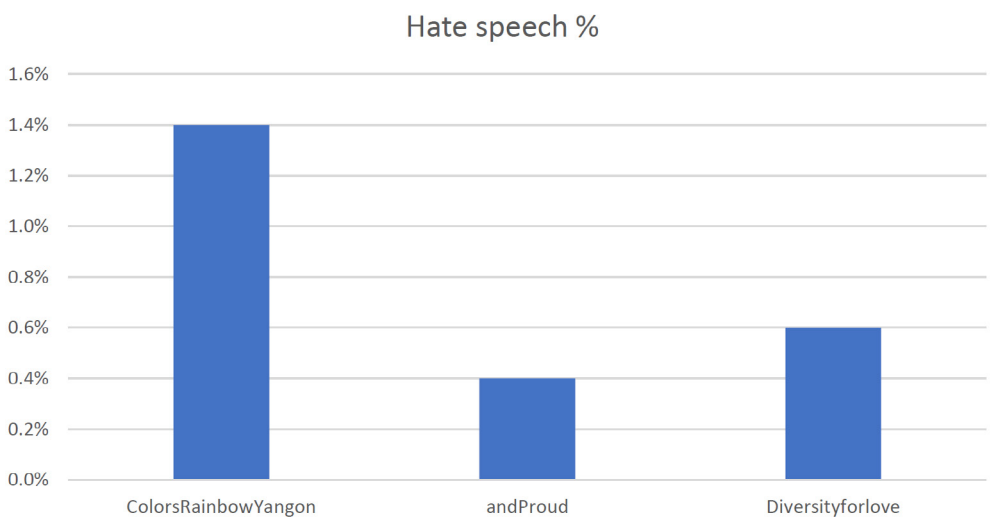
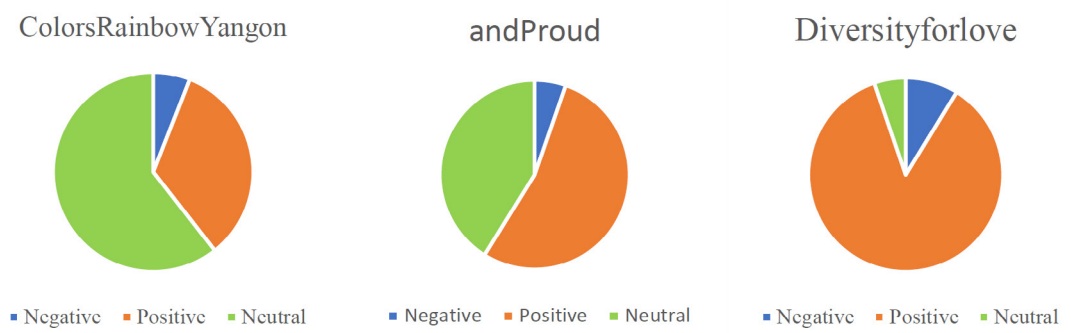


Figure 17. Comment sentiment analysis, 2019



Looking at the sentiment of comments,which we divide into three types – positive, negative and neutral – we find that there is a strong inverse correlation between positivity and hate speech in the Myanmar cases. Pages that receive a high number of positive comments also receive a low number of hate speech comments (Figure 17). Positive comments on Diversity for Love constitute 85% of all comments, while the page only registers one hate speech comment. Similarly, 52% of comments on and PROUD are positive, with another 41% being neutral, while the page shows only one hate speech comment.

Examining the data by type of hate speech, we find that the most common form of hate speech was 'inferiorisation' – comments that are intended to dehumanise and inferiorise target victims (Figure 18). Colors Rainbow Yangon had the most hate speech content: 7 comments out of the total 425 comments analysed. 71% of those hate speech comments were classified as 'inferiorisation.' Among these comments are dehumanising terms such as *kalar* and *achout* (disgusting and mentally ill). The single hate speech comments on and PROUD and Diversity of Love pages both demonstrated 'inferiorisation.'

Interviews with page admins of the three groups confirmed our analysis that there is a generally low level of hate speech content on these pages. Admins see their pages as largely a positive and supportive environment for LGBTQ+ members, noticing hate comments only on rare occasions. There have been no physical threats made that worry them to the point that they would conceive of contacting Myanmar authorities.

Across the three pages the preferred hate speech management action is to intentionally leave such content alone. Page admins note that because the volume of unfiltered hate speech is so low, whenever there is one such speech act, it is better to leave it on their pages so that other members can see. Often members themselves directly engaged with this content, seeking to redress it through explanations, which did not lead to additional hate speech comments or replies.

Overwhelmingly, the page admins of these Myanmar LGBTQ+ groups were not aware of Facebook's community standards in general, nor specifically those on hate speech. Their interpretation of hate speech is based on training they received from third-party organisations such as the International Foundation for

Figure 18. Number of hate speech comments by type of hate speech

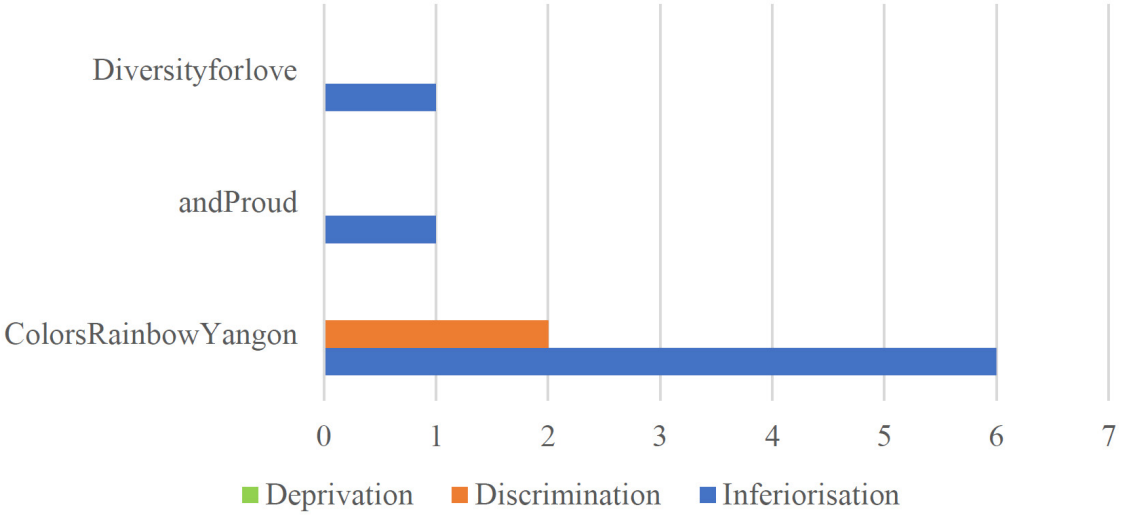


Figure 19. Hate speech management actions

Management action	Colors Rainbow Yangon	Proud	Diversity for Love
Daily post moderation	X	X	X
Familiarity with Facebook Community hate speech standards			
Hiding/Deletion of hate speech posts		X	
Blocking of hate speech			
Engagement with hate speakers (replies, private messages)	X	X	X
Intentionally leaving some hate speech unmoderated on the page	X	X	X
Having received direct physical threats			
Banning of individuals due to hate speech			
Taking screenshots of hate speech for reference			X
Reporting of hate speech to Facebook		X	X
Taking further action against hate speech with law enforcement or human rights bodies			

Electoral Systems (IFES) or the Australian Embassy. However, they are aware of the lack of legal frameworks to protect members of sexual and gender minorities and have a good understanding of Article 66(d). Their lack of awareness of Facebook's definition of hate speech did not deter them from directly reporting hate speech content to Facebook, although they said they received no response. None of the admins felt the necessity to engage further with Facebook on managing hate speech and said they would rather moderate it themselves.

One admin suggested Facebook could change the process of obtaining the blue badge verification to cover unregistered organisations like theirs. Many LGBTQ+ groups are communities on social media rather than officially registered organisations, which makes them ineligible to receive blue badge verification. The interviewee in question felt strongly that being officially verified by Facebook as authentic would help limit hate speech attacks, by signalling its platform legitimation. It would also help identify these groups as legitimate counter speech supporters, allowing them to quote attacks as part of critical counterspeech messaging without having those posts removed.

In summary, Colors Rainbow Yangon, Proud, and Diversity of Love feel that Facebook provides a critical platform for LGBTQ+ communities to flourish and prosper. Through the Facebook platform they have been able to establish online communities, provide a positive and supportive environment for members and the necessary space for advocacy of LGBTQ+ rights. Their platform presence has enabled them to grow in size and become more active online and offline in raising awareness on LGBTQ+ issues.

Case study summary

In sum, our research has shown that the LGBTQ+ group accounts we investigated have all experienced varying degrees of hate speech which was not removed by Facebook's machine learning filter, with The Philippines group experiencing the greatest number of hate comments and the Australian groups the least. Interviewees in several 'at-risk' countries also indicated they experienced allied hate speech on their personal accounts, as a result of their association with the group account.

Examining this small sample of publicly available data on Facebook pages is insightful but incomplete in providing a full picture of the actual incidence of hate speech experienced by these groups on the platform. Further as we do not have access to data on the incidence of proactively removed hate speech on these accounts, we cannot speculate on the effectiveness of Facebook's algorithmic reviewing process in the case study countries. We can however indicate that the page administrators of these groups as a whole were not satisfied by Facebook's reactive, user reporting process.

Our research has indicated that page administrators are key actors in hate speech regulation for pages and groups, but work largely as volunteers and so have no professional expertise in moderation or community management. Most were not aware of either Facebook community standards or training resources – but some had received third party training on hate speech management. No-one mentioned the Facebook Help Centre as a site for information about content moderation.

Responses to hate speech regulation showed some national commonalities. In Indonesia, Myanmar, and The Philippines page admins tended to leave hate speech alone, while Indian admins were more inclined to engage hate speakers in debate, and Indian and Australian admins actively report, hide and remove hate speech.

All page admins feel disempowered by their engagement with reporting or flagging hate speech to Facebook. In the majority of cases, when page admins have tried to report to Facebook what they believe to be hate speech comments, they have received automated messages and no follow-up. On a few occasions, when page admins have received what appear to be tailored, in person responses from Facebook, they have been told that content they identified as hate speech did not meet the criteria as set out in Community Standards. When they then asked for further clarification, they received only automated messages. These negative experiences of reporting hate speech have dissuaded them from further engagement. Where page admins felt that automated responses were becoming the only reporting feedback they received from Facebook, or where removed speech was later restored, this added to their sense of disempowerment. As a result we are signalling an issue with 'reporting fatigue', where individuals are disinclined to report as a result of perceived lack of impact on Facebook moderation practices.

As a result, we argue that Facebook needs to work with page administrators from protected category advocacy groups to improve their regulatory literacy and to invite their support in cooperative and collaborative hate speech regulation. This follows a recommendation from the 2020 Asia Pacific Regional Forum on Hate Speech, Social Media and Minorities, convened by the UN Special Rapporteur on minority issues, Dr Fernand de Varennes. That gathering suggested social media companies:

> *should engage with civil society and minorities to identify hate speech and develop lists of language that amounts to advocacy of hatred that constitutes incitement to discrimination, hostility or violence in certain contexts (APRF 2020, 6).*

To transform this proposal into practical media literacy and governance actions, we propose Facebook hold regular collaborative forums on hate speech evolution with such groups, train them in techniques of counter speech and hate speech moderation, and seek their help in identifying emerging hate speech trends.

Further in light of Facebook's historic concerns about the quality of user flagging, it is vital that the company help train users in effective moderation and reporting. We argue that making hate speech management training mandatory for all Facebook page administrators, with modules in all major languages, would further engage these key regulatory gatekeepers with the definition, forms and consequences of hate speech, as well as advice on appropriate moderation processes necessary to control this problem.

To improve the quality of hate speech reporting, users need basic instruction on content flagging before they experience hate speech in everyday or extreme contexts. If page administrators from minority groups who regularly experience hate speech, and are motivated to combat it, feel confused or disempowered by the reporting process, it suggests that other users could feel the same. We urge Facebook to make publicly transparent and portable all content regulation procedures, in an easy to follow, downloadable guide, including penalties for violations and details of the appeals process. This material should be made available in as many regional languages as possible, and be automatically recommended to all users who start a Facebook account.

Recommendations

Facebook should:

- better recognise the role of page administrators as critical gatekeepers of hate speech content, and support their improved regulatory literacy via training and education.
- improve the regulatory literacy of all page administrators by providing mandatory hate speech moderation training modules in major languages, and
- support extended training in hate speech management to APAC located page administrators working for groups in protected categories.
- make publicly transparent all content regulation procedures, in an easy to follow guide, including penalties for violations and the appeals process. The guide should be available in as many regional languages as possible and automatically recommended to all users who start a Facebook account.
- facilitate regular consultative forums with target groups and protected category page owners to discuss best practice moderation approaches.

9. Conclusion

We began this report by asking three central questions: 1) What constitutes hate speech in different Asia-Pacific jurisdictions? 2) How well are Facebook's policies and procedures positioned to identify and regulate this type of content? 3) How can we understand the spread of hate speech in this region with a view to formulating better policies to address it? We have answered these questions comprehensively, investigating hate speech regulation across our five case study countries, analysing Facebook's policy system and ways it can work with its stakeholders to better identify and regulate hate, and exploring protected group experiences of hate speech on the platform to understand how Facebook might improve its policy response to their concerns and reports. Our key recommendations are outlined at the beginning of this report.

Overall, our recommendations centre on the fact that hate speech is very context dependent, and intimate local knowledge is required to understand and address it fully. This requisite knowledge includes recognising local interpretations of hate speech, any legislation that may capture hate speech, the possibility of legislative overreach by governments that infringes on free speech, and the need for much greater collaboration and partnership with local communities to address this significant problem. In this respect we have identified the role of page administrators as critical gatekeepers of hate speech content, and support their improved regulatory literacy via training and education. We also support Facebook's continual re-evaluation of its hate speech definitions, so that it captures speech that is genuinely harmful, while not overly impacting on speech that should not be regulated, whether by government or by private entities.

In particular, it would be beneficial for the region if governments, CSOs and Facebook could collaborate on an Asia Pacific regional hate speech monitoring project like that devised by the European Commission. This would be a way of reaching agreement on a definition of hate speech and its harms, promoting better remedial responses, improving reporting, and combating the spread of hate speech in the region. Again this proposal is one supported by the 2020 Asia Pacific Regional Forum on Hate Speech, Social Media and Minorities (APRF 2020, 3). However, such a project would have to be driven by a regional political alliance such as the ASEAN Intergovernmental Commission on Human Rights (AICHR).

10. Challenges for future research

The main challenge for our project was lack of access to a significant random sample of hate speech content posted by Facebook users in the case study countries. We understand the difficulties for Facebook in providing data access to academic researchers, notably in the Social Science One project (King and Persily 2020), and the sensitivity around user privacy. However, if we had had access to examples of hate speech content that was flagged and either rejected or removed by Facebook reviewers, or which had been flagged, removed and then restored on appeal, it would have given us invaluable insight into Facebook's application of content regulation standards. One of the most common comments we received from page administrators of the LGBTQ+ groups under study is that they directly contacted Facebook about hate speech content but that "it went nowhere." Allowing researchers to gain a clearer picture of the process and efficacy of Facebook's internal hate speech review would enable them to assess its responsiveness to local context and help them convey the complexity and nuance of Facebook's decision-making to stakeholder groups. This overall would improve the trust relationship between Facebook and its consumer base, especially advocacy groups who are directly impacted by hate speech.

Australia's relatively positive LGBTQ+ experience on Facebook contrasts significantly with that of Islamic groups, who recently accused Facebook of "failing to take down hate speech against minority groups" (Truu 2021). For example, the Australian Muslim Advocacy Network (AMAN) has lodged a complaint with Australia's Human Rights Commission under section 9 and 18c of the *Racial Discrimination Act 1975*, noting a number of examples of racial vilification on Facebook in an earlier response to an Australian Human Rights Commission Issue Paper (AMAN 2019). Given the significant Muslim populations in the Asia Pacific, Facebook's response to anti-Islamic hate speech will be vital to the future of its operations in the region.

11. Reference List

Adjie, Moch. Fiqih Prawira. 2020. "Survey on Acceptance in Indonesia Gives Hopes to LGBT Community." *Jakarta Post*, June 28, 2020. https://www.thejakartapost.com/news/2020/06/28/survey-on-acceptance-in-indonesia-gives-hopes-to-lgbt-community.html

AIDs Council of NSW (ACON). 2020. *Safe and Strong: An LGBTQ+ Guide to Facebook and Instagram*. https://www.acon.org.au/wp-content/uploads/2020/02/ACON-Facebook-Instagram-LGBTQ-Guide.pdf

Al Jazeera. 2021. "Facebook Executive Who Shared Anti-Muslim Post Apologises." August 27, 2021. https://www.aljazeera.com/economy/2020/8/27/facebook-executive-who-shared-anti muslim-post-apologises-report

Alkiviadou, Natalie. 2019. "Hate Speech on Social Media Networks: Towards a Regulatory Framework?" *Information and Communications Technology Law* 28, iss. 1 (July): 19-35.

Article One. 2018. *Assessing the Human Rights Impact of Facebook's Platforms in Indonesia: Executive Summary for Facebook Inc.* https://about.fb.com/wp-content/uploads/2020/05/Indonesia-HRIA-Executive-Summary-v82.pdf

Article 19. 2017. *Myanmar: Interfaith Harmonious Coexistence Bill (3rd version)* September. https://www.article19.org/wp-content/uploads/2017/09/170907-Myanmar-Hate-Speech-Law-Analysis-August-2017.pdf

Asia Centre. 2020. *Hate Speech in Southeast Asia: New Forms, Old Rules*. https://asiacentre.org/wp-content/uploads/2020/07/Hate-Speech-in-Southeast-Asia-New-Forms-Old-Rules.pdf

Asia-Pacific Regional Forum (APRF), 2020. *Integrated Asia-Pacific Recommendations: Asia-Pacific Regional Forum on "Hate Speech, Social Media and Minorities"*, October 19-20, 2020. Minority Forum Info. https://www.minorityforum.info/page/asia-pacific-regional-forum-on-hate-speech-social-media-and-minorities

Australian Human Rights Commission (AHRC). 2015. *Resilient Individuals: Sexual Orientation, Gender Identity and Intersex Rights National Consultation Report*. Sydney: AHRC. https://humanrights.gov.au/our-work/lgbti/publications/resilient-individuals-sexual-orientation-gender-identity-intersex

Australian Muslim Advocacy Network (AMAN). 2019. *A Response to the Consultation Questions Contained in the Human Rights Commission Issues Paper, "Free and Equal: An Australian Conversation on Human Rights."* August 30, 2019. http://www.aman.net.au/wp-content/uploads/2019/10/Joint-Submission-to-the-Australian-Human-Rights-Commission-from-the-Muslim-community_reduced-size.pdf

Avaaz. 2019. *Megaphone for Hate: Disinformation and Hate Speech on Facebook During Assam's Citizenship Count*. October, 2019. https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20(1).pdf

Barker, Kim, and Olga Jurasz. 2019. *Online Misogyny as Hate Crime: A Challenge for Legal Regulation?* London: Routledge.

Barrett, Paul. 2020. *Who Moderates the Social Media Giants? A Call to End Outsourcing*. New York: NYU Stern Center for Business and Human Rights. https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version?fr=sZWZmZjI1NjI1Ng

Beltran, Michael. 2020. "Artists Rally Against Philippine Anti-terrorism Bill as Fears for Free Speech, Human Rights Increase." *South China Morning Post*, July 2, 2020. https://www.scmp.com/lifestyle/arts-culture/article/3091224/artists-rally-against-philippine-anti-terrorism-bill-fears.

Bliuc, Ana-Maria, Faulkner, Nicholas, Jakubowicz, Andrew and Craig McGarty. 2018. "Online networks of racial hate: A systematic review of 10 years of research on cyber-racism." *Computers in Human Behavior* 87 (May): 75–86.

Brison, Susan J. 1998. "Speech, Harm, and the Mind-Body Problem in First Amendment Jurisprudence." *Legal Theory* 4, iss. 1 (February): 39-61.

Brooking, Emerson T., and P. W. Singer. 2018. *Like War: The Weaponization of Social Media*. New York: Eamon Dolan.

Brown, Alex. 2015. *Hate Speech Law: A Philosophical Examination*. New York: Routledge.

—. 2017. "What is Hate Speech? Part 1: The Myth of Hate." *Law and Philosophy* 36 (February): 419-468.

Business for Social Responsibility (BSR). 2018. *Human Rights Impact Assessment: Facebook in Myanmar*. October, 2018. https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf

Carlson, Caitlin Ring, and Hayley Rousselle. 2020. "Report and Repeat: Investigating Facebook's Hate Speech Removal Process." *First Monday* 25, no. 2-3 (February). https://firstmonday.org/ojs/index.php/fm/article/download/10288/8327

Chaturvedi, Swati. 2016. *I Am a Troll: Inside the Secret World of the BJP's Digital Army*. New Delhi: Juggernaut.

Choudhury, Angshuman. 2020. "How Facebook is Complicity in Myanmar's Attacks on Minorities." *The Diplomat*. August 25, 2020. https://thediplomat.com/2020/08/how-facebook-is-complicit-in-myanmars-attacks-on-minorities/

Citizenship (Amendment) Act 2019 (Indian Cth). The Gazette of India. 71. DL—(N)04/0007/2003—19 http://egazette.nic.in/WriteReadData/2019/214646.pdf

Citron, Danielle. 2014. *Hate Crimes in Cyberspace*. Cambridge, Ma.: Harvard University Press.

Colors Rainbow and Equality Myanmar. 2020. *Human Rights Violations Against the LGBT Community Trend Analysis 2015-2018*. Myanmar: Colors Rainbow. https://www.colorsrainbow.org/wp-content/uploads/2020/01/Trend-Analysis-2015-2018-English-for-Web.pdf

Columbia Global Freedom of Expression. 2018. *Navtej Singh Johar v. Union of India*. Columbia Global Freedom of Expression (website). Columbia University. https://globalfreedomofexpression.columbia.edu/cases/navtej-singh-johar-v-union-india/

Convention on the Elimination of All Forms of Discrimination against Women, 18 December 1975, United Nations Office of the High Commission of Human Rights, 34/180. https://www.ohchr.org/documents/professionalinterest/cedaw.pdf

Convention on the Prevention and Punishment of the Crime of Genocide, 12 January 1948, United Nations General Assembly, 260 A (III). https://www.un.org/en/genocideprevention/documents/atrocity-crimes/Doc.1_Convention%20on%20the%20Prevention%20and%20Punishment%20of%20the%20Crime%20of%20Genocide.pdf

Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019 (Australian Cth). https://www.legislation.gov.au/Details/C2019A00038

Davis, Matthew C., Challenger, Rose, Jayewardene, Dharshana N. W., and Chris W. Clegg. 2014. "Advancing Socio-technical Systems Thinking: A Call for Bravery." *Applied Ergonomics: Human Factors in Technology and Society* 45, iss. 2 (July): 171-180.

Del Vigan, Fabio, Comino, Andrea, Dell'Orletta, Felice, Petrocci, Marinella, and Maurizio Tesconi. 2017. "Hate Me, Hate Me Not: Hate Speech Detection on Facebook." Paper presented at *First Italian Conference on Cybersecurity, Venice, January 17-20. 2017*.

Deutsche Welle. 2021. "South Asia's LGBT Muslims Turn to Social Media for Support." January 18, 2021. https://www.dw.com/en/south-asias-lgbt-muslims-turn-to-social-media-for-support/a-56266117

Dixit, Pratik. 2020. "Navtej Singh Johar v Union of India: Decriminalising India's Sodomy Law." *International Journal of Human Rights* 24, iss. 8 (November): 1011-1030.

Douek, Evelyn. 2020. "Australia's 'Abhorrent Violent Material' Law: Shouting 'Nerd Harder' and Drowning Out Speech." 94 *Australian Law Journal* 41 (August), https://ssrn.com/abstract=3443220

Dwoskin, Elizabeth, Whalen, Jeanne, and Regine Cabato. 2019. "Content Moderators at YouTube, Facebook and Twitter See the Worst of the Web- and Suffer Silently." *The Washington Post*, July 25, 2019. https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/

Ebner, Julia. 2020. *Going Dark: The Secret Social Lives of Extremists*. London: Bloomsbury.

Electronic Information and Transaction Law, ITE ('Informasi dan Transaksi Elektronik') (Indonesian Cth). https://jdih.kemenkeu.go.id/fulltext/2008/11tahun2008uu.htm

Enhancing Online Safety Act 2015. No 24, 2015. C2018C00356. (Australian Cth). https://www.legislation.gov.au/Details/C2018C00356

eSafety Commissioner. 2020. *Online Hate Speech: Findings from Australia, New Zealand and Europe*. January 29, 2020. https://www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf

European Commission. 2016. "European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech." May 31, 2016. http://europa.eu/rapid/press-release_IP-16-1937_en.htm.

European Commission. 2020. *5th Evaluation of the Code of Conduct*. June, 2020. https://ec.europa.eu/info/

sites/info/files/codeofconduct_2020_factsheet_12.pdf.

Facebook. 2020a. *Sharing Our Actions on Stopping Hate*. July 2, 2020. https://www.facebook.com/business/news/sharing-actions-on-stopping-hate

Facebook. 2020b. *Charting a Way Forward: Online Content Regulation*. February, 2020. https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf

Facebook. 2020c. Facebook Response: Indonesian Human Rights Impact Assessment. May 12. https://about.fb.com/wp-content/uploads/2021/03/FB-Response-Indonesia-HRIA.pdf

Facebook 2020d. "Facebook's Approach to Combatting Online Hate". Letter to Fiona Martin. December 1, 2020.

Facebook. 2021. *Facebook Community Standards: Recent Updates*. 2021. https://www.facebook.com/communitystandards/recentupdates/

Facebook Data Transparency Advisory Group (FDTAG). 2019. *Report of The Facebook Data Transparency Advisory Group*, Justice Collaboratory, Yale Law School, April, 2019, https://law.yale.edu/system/files/area/center/justice /document/dtag_report_5.22.2019.pdf, 39

Fick, Maggie, and Paresh Dave. 2019. "Facebook's Flood of Languages Leave It Struggling to Monitor Content." *Reuters*, April 23, 2021. https://www.reuters.com/article/us-facebook-languages-insight-idUSKCN1RZ0DW

Fisher Max. 2018. "Inside Facebook's Secret Rulebook for Global Political Speech." *New York Times*, December 28, 2018. https://www.nytimes.com/2018/12/27/world/facebook-moderators.html

Foxman, Abraham, and Christopher Wolf. 2013. *Viral Hate: Containing Its Spread on the Internet*. New York: Palgrave Macmillan.

Frankel, Rafael. 2020. "How Facebook Is Preparing for Myanmar's 2020 Election". *Facebook Newsroom*. August 21, 2020. https://about.fb.com/news/2020/08/preparing-for-myanmars-2020-election/

Freedom House. 2019. *Freedom on the Net 2019: The Crisis of Social Media*. London: Freedom House. https://freedomhouse.org/sites/default/files/2019-11/11042019_Report_FH_FOTN_2019_final_Public_Download.pdf

Gelber, Katharine. 2019. "Differentiating Hate Speech; A Systemic Discrimination Approach." *Critical Review of International Social and Political Philosophy* (February). https://doi.org/10.1080/13698230.2019.1576006.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. New Haven: Yale University Press.

Goggin, Gerard, Vromen, Ariadne, Weatherall, Kimberlee, Martin, Fiona, Adele, Webb, Sunman, Lucy and Francesco Bailo. 2017. "Digital Rights In Australia". Sydney Law School Research Paper No. 18/23. November 27. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3090774

Hao, Karen. 2020. "Facebook's New Polyglot AI Can Translate Between 100 Languages." *MIT Technology Review*, October 19, 2020. https://www.technologyreview.com/2020/10/19/1010678/facebook-ai-translates-between-100-languages/

Harvey, Adam. 2016. "Affectionate Gay Students Should be Banned from University Campuses, Indonesian Minister Says." *ABC News*, January 27, 2016. https://www.abc.net.au/news/2016-01-27/indonesia-lgbtiq-support-group-under-attack/7117446

Heldt, Amelie. 2019. "Let's Meet Halfway: Sharing New Responsibilities in a Digital Age." *Journal of Information Policy* 9: 336-369. https://doi.org/10.5325/jinfopoli.9.2019.0336

Hlaing, Kyaw Hsan, and Emily Fishbein. 2021. "'We Are Like One Group.' How Myanmar's Pro-Democracy Protests Are Giving a Voice to LGBTQ+ People." *Time*, March 5, 2021. https://time.com/5944407/myanmar-democracy-protests-lgbtq/

Home Affairs Committee (UK). 2017. *Hate Crime: Abuse, Hate and Extremism Online*. United Kingdom: United Kingdom Parliament. https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/60902.htm.

Human Rights Watch. 2016. *These Political Games Ruin Our Lives' Indonesia's LGBT Community Under Threat*. August 10, 2016. https://www.hrw.org/report/2016/08/10/these-political-games-ruin-our-lives/indonesias-lgbt-community-under-threat

Human Rights Watch. 2017. *'Just Let Us Be': Discrimination Against LGBT Students in The Philippines*. June 21, 2017. https://www.hrw.org/report/2017/06/21/just-let-us-be/discrimination-against-lgbt-students-philippines

International Commission of Jurists (ICJ). 2019. *Living with Dignity: Sexual Orientation and Gender Identity-Based Human Rights Violations in Housing, Work, and Public Spaces in India*. June, 2019. https://www.icj.org/wp-content/uploads/2019/06/India-Living-with-dignity-Publications-Reports-thematic-report-2019-ENG.pdf

International Convention on the Elimination of All Forms of Racial Discrimination (CERD), 20 November 1965, United Nations Office of the High Commission of

Human Rights, 2106 (XX). https://www.ohchr.org/EN/ProfessionalInterest/Pages/CERD.aspx

International Telecommunications Union (ITU). 2020. *The World Telecommunication/ICT Indicators Database, 23rd edition*. https://www.itu.int/pub/D-IND

Jane, Emma. 2018. "Gendered Cyberhate: A New Digital Divide?" In Theorizing Digitial Divides [italics], edited by Massimo Ragnedda and Glenn W. Muschert, 186-198. Oxon: Routledge.

Judson, Ellen, Atay, Asli, Krasodomski-Jones, Alex, Lasko-Skinner, Rose, and Josh Smith. 2020. *Engendering Hate: The Contours Of State-Aligned Gendered Disinformation Online*. London: Analysis and Policy Observatory.

Kadri, Thomas E., and Kate Klonick. 2019. "Facebook v Sullivan: Public Figures and Newsworthiness in Online Speech." *Southern California Law Review* 93, iss. 1 (November) 37-99

Khatua, Aparup, Cambria, Erik, Ghosh, Kuntal, Chaki, Nabendu, and Apalak Khatua. 2019. "Tweeting in Support of LGBT?: A Deep Learning Approach." In *CoDS-COMAD '19: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, India, January 2019*, 342-345. https://doi.org/10.1145/3297001.3297057

King, Gary, and Nathaniel Persily 2020. "Unprecedented Facebook URLs Dataset now Available for Academic Research through Social Science One." Social Science One (blog), February 13. https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one

Klein, Adam. 2017. *Fanaticism, Racism and Rage Online: Corrupting the Digital Sphere*. Cham, Switzerland: Palgrave Macmillan.

Knight, Kyle. 2019. "Section 377 is History but Young LGBT Indians Need Concrete Policies to Protect them from Bullying." *Human Rights Watch*, June 24, 2019. https://www.hrw.org/news/2019/06/24/section-377-history-young-lgbt-indians-need-concrete-policies-protect-them-bullying

Langton, Rae. 1993. "Speech Acts and Unspeakable Acts." *Philosophy and Public Affairs 22*, no. 4 (Autumn): 293-330.

—. 2012. "Beyond Belief: Pragmatics in Hate Speech and Pornography." In *Speech and Harm: Controversies Over Free Speech*, edited by Ishani Maitra and Mary Kate McGowan, 72–93. Oxford: Oxford University Press.

Lawrence, Charles. 1993. "If He Hollers Let Him Go: Regulating Racist Speech on Campus." In *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, edited by Mary J. Matsuda, Charles R. Lawrence III, Richard Delgado and Kimberlé Williams Crenshaw, 53-88. Boulder: Westview Press.

Lee, Yanghee. 2019. *Situation of Human Rights in Myanmar*. Report of the Special Rapporteur of the Human Rights Council to the United Nations General Assembly, Seventy-fourth Session. August 30, 2019. A/74/342

Leets, Laura. 2001. "Responses to Internet Hate Sites: Is Speech Too Free in Cyberspace?" *Communication Law and Policy* 6, iss. 2 (June): 287-317.

Levmore, Saul, and Martha Nussabaum, eds. 2011. *The Offensive Internet: Speech, Privacy and Reputation*. Cambridge, Ma.: Harvard University Press.

Maitra, Ishani, and Mary Kate McGowan. 2007. "The Limits of Free Speech: Pornography and the Question of Coverage." *Legal Theory* 13, iss. 1 (May): 41–68.

—. 2012. "Introduction and Overview." In *Speech and Harm: Controversies Over Free Speech*, edited by Ishani Maitra and Mary Kate McGowan, 1–23. Oxford: Oxford University Press.

Matamoros-Fernández, Ariadna, and Johan Farkas. 2021. "Racism, Hate Speech, and Social Media: A Systematic Review and Critique." *Television and New Media* 22, iss. 2 (January): 205-224.

McDonald, Joshua. 2020. "LGBT Community Targeted by Police in Indonesia." *The Diplomat*, September 18, 2020. https://thediplomat.com/2020/09/lgbt-community-targeted-by-police-in-indonesia/

McEvoy, Thomas Richard, and Stewart James Kowalski. 2019. "Deriving Cyber Security Risks from Human and Organizational Factors – A Socio-technical Approach." *Complex Systems Informatics and Modeling Quarterly.* Article 105, iss. 18 (March/April): 47–64 https://doi.org/10.7250/csimq.2019-18.03

McGoogan, Cara. 2017. "Germany to Fine Facebook and YouTube €50 if They Fail to Delete Hate Speech." *Telegraph*, June 30, 2017. http://www.telegraph.co.uk/technology/2017/06/30/germany-fine-facebook-youtube-50m-fail-delete-hate-speech/.

McGowan, Mary Kate. 2009. "Oppressive Speech." *Australasian Journal of Philosophy* 87, iss. 3 (July): 389-407.

Mendiratta, Raghav. 2021. *Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021*. WILMap, Stanford Centre for Internet and Society, Stanford University. https://wilmap.law.stanford.edu/entries/information-technology-intermediary-guidelines-and-digital-media-ethics-code-rules-2021

Miles, Tom. 2018. "U.N. investigators cite Facebook role in Myanmar crisis." Reuters, March 13, 2018. https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN

Ministry of Foreign Affairs and Trade. 2019. *Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online*. May 15, 2019. https://www.christchurchcall.com/call.html

Molina, Kristo. 2016. "Indonesian Electronic Information and Transactions Law Amended." White and Case (Client Alert), December 15, 2016. https://www.whitecase.com/sites/whitecase/files/files/download/publications/indonesian-electronic-information-and-transactions-law-amended.pdf

Munn, Luke 2020. "Angry by design: toxic communication and technical architectures." *Humanities and Social Sciences Communications* 7, iss.1 (July): 1–11. https://doi.org/10.1057/s41599-020-00550-7.

Murphy, Laura.W. 2020. *Facebook's Civil Rights Audit–Final Report*. July 8, 2020. https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Office of the High Commissioner for Human Rights (OHCHR). 2018. *Terms of Reference (TOR). Human Rights Research Consultant*. South-East Asia Regional Office.

Office of the High Commissioner for Human Rights (OHCHR). 2020a. *Report on Conversion Therapy*. Human Rights Council, Forty-fourth Session. May 2020. A/HRC/44/53. https://www.ohchr.org/EN/Issues/SexualOrientationGender/Pages/ReportOnConversiontherapy.aspx

Office of the High Commissioner for Human Rights (OHCHR). 2020b. *Situation of Human Rights in The Philippines*. Report of the United Nations High Commissioner for Human Rights. June 29. A/HRC/44/22. https://www.ohchr.org/Documents/Countries/PH/Philippines-HRC44-AEV.pdf

Online Safety Bill 2021 (Parliament of Australia). https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=r6680

Parekh, Bhikhu. 2012. "Is There a Case for Banning Hate Speech?" In *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, edited by Michael Herz and Peter Molnar, 37-56. Cambridge: Cambridge University Press.

Perrigo, Billy. 2019. "Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch." *Time*, November 27, 2019. https://time.com/5739688/facebook-hate-speech-languages/

Purnell, Newley, and Jeff Horowitz. 2020. "Facebook's Hate-Speech Rules Collide With Indian Politics." *Wall Street Journal*, August 14, 2020. https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi

Pyidaungsu Hluttaw Law No. 31/201 (Telecommunication Law) 2013 (Burmese Cth). https://www.myanmar-law-library.org/IMG/pdf/2013-10-08-communication_law_66-bu.pdf

Radics, George Baylon. 2019. "Human Rights in Asia: The History and Current State of LGBTQ Rights in South Asia, East Asia, and Southeast Asia." In *Global Encyclopedia of Lesbian, Gay, Bisexual, Transgender, and Queer (LGBTQ) History*, edited by Howard Chiang, 784-789. Farmington Hills, Michigan: Charles Scribner's Sons.

Rehman, Javaid, and Eleni Polymenopoulou. 2013. "Is Green a Part of the Rainbow? Sharia, Homosexuality and LGBT Rights in the Muslim World." *Fordham International Law Journal* 37, no. 1 (November): 1-52.

Renaldi, Adi. 2021. "Indonesia's LGBTQ Community Angry at Rise of Conversion Therapies." *Nikkei Asia*, March 14, 2021. https://asia.nikkei.com/Spotlight/Society/Indonesia-s-LGBTQ-community-angry-at-rise-of-conversion-therapies

Reporters Without Borders. 2017. "Burma Urged to Free Journalists, Amend Telecommunications Law." Updated August 23, 2019. https://rsf.org/en/news/burma-urged-free-journalists-amend-telecommunications-law

Reporters Without Borders. 2018. *Online Harassment of Journalists: Attack of the Trolls*. https://rsf.org/sites/default/files/rsf_report_on_online_harassment.pdf

Reuters. 2019. "Facebook Takes Down Hundreds of Indonesian Accounts Linked to Fake News Syndicate." February 1, 2019. https://www.reuters.com/article/us-facebook-indonesia-idUSKCN1PQ3JS

Reuters. 2020. "Facebook to Curb Hate Speech as Indian States go to Polls." March 31, 2020. https://www.reuters.com/article/india-election-facebook-idUSKBN2BN0OT

Rey, Aika. 2019. "Sotto: At Least 15 Senators Oppose SOGIE Equality Bill." *Rappler*, September 30, 2019. https://www.rappler.com/nation/sotto-says-senators-oppose-sogie-equality-bill

Rieder, Bernhard, Matamoros-Fernández, Ariadna ,and Oscar Coromina. 2018. "From Ranking Algorithms to 'Ranking Cultures': Investigating the Modulation of Visibility in YouTube Search Results." *Convergence: The International Journal of Research into New Media Technologies* 24, iss. 1 (January): 50-68.

Robertson, Geoffrey. 2012. *Crimes Against Humanity: The Struggle for Global Justice*. London: Penguin.

Rochefort, Alex. 2020, "Regulating Social Media Platforms: A Comparative Policy Analysis."

Communication Law and Policy 25, iss. 2 (April): 225-260.

Rodriguez, Axel, Argueta, Carlos, and Yi-Ling Chen. 2019. "Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis." Paper presented at *2019 International Conference on Artificial Intelligence in Information and Communication, Japan, February 11-13*. doi: 10.1109/ICAIIC.2019.8669073.

Rolph, David. 2010. "Publication, Innocent Dissemination and the Internet After '*Dow Jones and Co. v Gutnick*.'" *University of New South Wales Law Journal* 16, no.1: 84-94.

SBS News. 2021. "Australian Group Takes Fight to Facebook, Saying Platform is 'Awash' With Hateful Islamophobia." March 21, 2021. https://www.sbs.com.au/news/australian-group-takes-fight-to-facebook-saying-platform-is-awash-with-hateful-islamophobia

Sexual Orientation and Gender Identity and Expression (SOGIE) Bill 2000 (The Philippines Cth). https://www.congress.gov.ph/legisdocs/first_17/CR00101.pdf

Shaheed, Ahmad. 2021. *Countering Islamophobia / Anti-Muslim Hatred to Eliminate Discrimination and Intolerance Based on Religion or Belief*. Report of the Special Rapporteur on Freedom of Religion or Belief. Human Rights Council, Forty-sixth Session. February 25. A/HRC/46/30

Sloan, Luke, and Anabel Quan-Haase. 2017. *The SAGE Handbook of Social Media Research Methods*. Thousand Oaks: Sage Publications.

Soundararajan, T., Kumar, A., Nair, P., and Greely, J. 2019. *Facebook India Towards The Tipping Point Of Violence: Caste And Religious Hate Speech*. Equality Labs. https://www.equalitylabs.org/facebookindiareport

Statista. 2020. "Facebook Quarterly Revenue APAC 2010-2019 by Segment." June 11, 2020. https://www.statista.com/statistics/223282/facebooks-revenue-in-asia-since-1st-quarter-2010-by-segment/

Stecklow, Steve. 2018. "Hatebook: Why Facebook is Losing the War on Hate Speech in Myanmar." *Reuters*, August 15, 2018. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

Suresh, Mayur. 2018. "This is the Start of a New Era for India's LGBT Communities." *The Guardian*, September 7, 2018. https://www.theguardian.com/commentisfree/2018/sep/06/india-lgbt-homophobia-section-377

Suzor, Nicholas. 2019. *Lawless: The Secret Rules that Govern Our Digital Lives*. Cambridge: Cambridge University Press.

Teeni-Harari, Tali and Sharon Yadin. 2019. "Regulatory Literacy: Rethinking Television Rating in the New Media Age." University of Missouri-Kansas City Law Review, Vol. 88, iss 1: 101-114. https://ssrn.com/abstract=3457766

Tirrell, Lynne. 2017. "Toxic Speech: Toward an Epidemiology of Discursive Harm." *Philosophical Topics* 45, no. 2 (Fall): 139-161.

Torregoza, Hannah. 2018. "Sen. Bam Aquino to Seek Reelection, Says Unfazed by Social Media Attacks." *Manila Bulletin*, October 12, 2018. https://mb.com.ph/2018/10/12/sen-bam-aquino-to-seek-reelection-says-unfazed-by-social-media-attacks/

Truu, Maani. 2021. "Australian-Muslim Rights Group Lodges Hate Speech Complaint Against Facebook." *SBS News*, April 22, 2021. https://www.sbs.com.au/news/australian-muslim-rights-group-lodges-hate-speech-complaint-against-facebook

UNAIDS. 2006. *HIV and Men Who have Sex with Men in Asia and Pacific*. https://data.unaids.org/publications/irc-pub07/jc901-msm-asiapacific_en.pdf

United Nations Education, Scientific and Cultural Organization (UNESCO). 2016. *Countering Online Hate Speech: Unesco Series On Internet Freedom*. https://unesdoc.unesco.org/ark:/48223/pf0000233231_eng

Upadhyay, Nishant. 2020. "Hindu Nation and its Queers: Caste, Islamophobia, and De/coloniality in India." *Interventions: International Journal of Postcolonial Studies* 22, iss. 4 (April): 464-480.

Vaidyanathan, Siva. 2018. *Anti-Social Media: How Facebook Disconnects Us and Undermines Democracy*. New York: Oxford University Press.

Venier, Silvia. 2019. "The Role of Facebook in the Persecution of the Rohingya Minority in Myanmar: Issues of Accountability Under International Law." *The Italian Yearbook of International Law Online 28*, iss.1 (October): 231-248 https://doi.org/10.1163/22116133_02801014

Vilk, Viktorya, Vialle, Elodie, and Matt Bailey. 2021. *No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users*. PEN America. https://pen.org/report/no-excuse-for-abuse/

Ware, Shelley, and Seear, Kate. 2021. "Open Letter to the Commonwealth Attorney General." Outer Sanctum Podcast (blog), September 1, 2020. https://www.outersanctum.com.au/updates/2020/9/1/open-letter-to-the-commonwealth-attorney-general

Warofka, Alex 2018. "An Independent Assessment of the Human Rights Impact of Facebook in Myanmar". Facebook Newsroom. November 5,

2018, Updated August 6, 2020.https://about.fb.com/news/2018/11/myanmar-hria/

We Are Social. 2021. *Digital 2021 Global Overview Report*. We Are Social and Hootsuite, March 31, 2021. https://wearesocial.com/digital-2021

Wijeratne, Yudhanjaya. 2020. "Facebook, Language and the Difficulty of Moderating Hate Speech." *Media@LSE* (blog), July 23, 2020. https://blogs.lse.ac.uk/medialse/2020/07/23/facebook-language-and-the-difficulty-of-moderating-hate-speech/

Williams, Matthew L., Burnap, Pete, Javed, Amir, Liu, Han, and Sefa Ozalp. 2020. "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime." *The British Journal of Criminology* 60, iss. 1 (January): 93-117.

Xu, Teng, Goossen, Gerard, Cevahir, Huseyin Kerem, Khodeir, Sara, Jin, Yingyezhe, Li, Frank, Shan, Shawn, Patel, Sagar, Freeman, David, and Paul Pearce. 2021. "Deep Entity Classification: Abusive Account Detection for Online Social Networks." Paper presented at *The 30th Security Symposium (USENIX Security 21), USENIX Association Conference, Vancouver, B.C., August 11-13, 2021*. https://www.usenix.org/conference/usenixsecurity21/presentation/xu

Yosephine, Liza. 2016. "Indonesian Psychiatrists Label LGBT as Mental Disorder." *Jakarta Post*, February 24, 2021. https://www.abc.net.au/news/2016-01-27/indonesia-lgbtiq-support-group-under-attack/7117446