ARTICLE 19

In partnership with UNESCO

Funded by the European Union

# Content Moderation and Local Stakeholders in Indonesia

June 2022

SOCIAL MEDIA 4PEACE

**ARTICLE 19**
**T:** +44 20 7324 2500
**F:** +44 20 7490 0566
**E:** info@article19.org
**W:** www.article19.org
**Tw:** @article19org
**Fb:** facebook.com/article19org

This publication was produced with the financial support of the **European Union** and **UNESCO**.

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO or the European Union concerning the legal status of any country, territory, city or area or its authorities, or concerning the delimitation of its frontiers or boundaries.

The authors are responsible for the choice and the presentation of the facts contained in this publication and for the opinions expressed therein, which are not necessarily those of UNESCO or the European Union and do not commit the Organizations.

## Acknowledgements

Many thanks to Sherly Haristya, who conducted this research and authored this report in close consultation with ARTICLE 19.

A Data Protection Officer certified by the Association of Data Privacy Professionals Indonesia, Sherly earned her PhD in Communication Studies from the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. Her research has been bridging the global and national (Indonesia) debates on Internet governance, content moderation and personal data protection issues. During her PhD, she also had the opportunity to serve on the national multi-stakeholder organising committee of the UN Internet Governance Forum in Bali, Indonesia, in 2013 and on the Secretariat of Internet Assigned Numbers Authority Stewardship Transition Coordination Group.

Many thanks to all the persons who have shared their expertise and views and contributed to the research process.

Thanks also to our partners at UNESCO who have provided input and feedback at various stages of the development of this report.

# Contents

# Executive summary

- The population of Indonesia is characterised by social, cultural, and economic diversity. Against a background of deep historical roots of segregation, there are efforts to spread false information, hate speech, and violent extremist materials on social media that have been damaging to individuals and civil society.

- The problem is particularly acute in relation to 'grey-area' speech, which is speech that falls under categories whose prohibition is not mandated by, or not compatible with, international standards on freedom of expression and/or community standards, but whose amplification can result in real-world violence. In this regard, content moderation conducted by social media platforms can play a critical role in preventing the spread of such problematic content from turning into real-world dangers.

- While there are efforts made by platforms to conduct content moderation in line with the local context in Indonesia, the research found a disconnect between the global community rules of platforms and their enforcement at the local level, especially in tackling the 'grey-area' speech.

- Effective content moderation in a big and diverse country such as Indonesia requires a transparent and sustainable dialogue between platforms and local civil society groups. The leading civil society groups in Indonesia, particularly those who are the trusted partners of platforms, have a special line of communication to address urgent problematic content and explain the local context to platforms. However, the partnerships should be improved, and local partners should be empowered to create a more inclusive, credible, and meaningful dialogue between Indonesian civil society groups, individual users, and platforms.

- The report tested the idea that a local Coalition on Freedom of Expression and Content Moderation could play a role to fill in the gap in current content moderation practices. Most interviewees responded positively to this suggestion and the report concludes with recommendations on how to facilitate the creation of such a civil society Coalition on Freedom of Expression and Content Moderation in Indonesia. By reinforcing the capacity of local actors, the coalition would act as a bridge to develop and nurture relationships and effective dialogue with social media companies on content moderation issues. The coalition would contribute to the development of content moderation practices that uphold international standards on freedom of expression while duly taking into consideration the local context.

# Introduction

This publication has been produced as part of the United Nations Educational, Scientific and Cultural Organization's (UNESCO) project **Social Media 4 Peace** funded by the European Union (EU).

## About the project

This report is part of the **Social Media 4 Peace** project that UNESCO is implementing in Bosnia and Herzegovina (BiH), Kenya, and Indonesia, with support of the EU. The overall objective of the project is to strengthen the resilience of civil society to potentially harmful content spread online, in particular hate speech and disinformation, while protecting freedom of expression and contributing to the promotion of peace through digital technologies, notably social media. ARTICLE 19's contribution to the project focuses on concerns raised by the current practices of content moderation on dominant social media platforms in the three target countries.

ARTICLE 19 considers that social media companies are, in principle, free to restrict content on the basis of freedom of contract, but that they should nonetheless respect human rights, including the rights to freedom of expression, privacy, and due process. While social media platforms have provided opportunities for expression, a number of serious concerns have come to light. The application of community standards has led to the silencing of minority voices. The efforts of tech companies to deal with problematic content are far from being evenly distributed: for instance, it has been shown that '87% of Facebook's spending on misinformation goes to English-language content, despite the fact that only 9% of its users are English speaking.' It has also been revealed that most resources and means in terms of content moderation are being allocated to a limited number of countries. Generally speaking, the transparency and dispute resolutions over content removals have so far been inadequate to enable sufficient scrutiny of social media platforms' actions and provide meaningful redress for their users. Finally, it is doubtful that a small number of dominant platforms should be allowed to hold so much power over what people are allowed to see without more direct public accountability.

This report specifically looks at the situation of local actors who, while they are impacted by the circulation of harmful content on social media or the moderation thereof, often find themselves unable to take effective action to improve their situation in that respect. They may feel frustrated by the inconsistencies of platforms' application of their own content rules; they may feel that global companies ignore their requests or misunderstand the current circumstances of the country or region. Some may lack understanding of content rules or of content moderation, but that is not the case of all local stakeholders.

The research then seeks to test, through the views of local stakeholders, the assumption that a local Coalition on Freedom of Expression and Content Moderation could play a role to fill the gap between the realities of local actors and companies that operate on a global scale. The idea for such a coalition is based on ARTICLE 19's work on the development of Social Media Councils, a multi-stakeholder mechanism for the oversight of content moderation on social media platforms. ARTICLE 19 suggested that Social Media Councils should be created at a national level (unless there was a risk that it would be easily captured by the government or other powerful interests) because this would ensure the involvement of local decision-makers who are well-informed of the local context and understand its cultural, linguistic, historical, political, and social nuances. While the development of a self-regulatory, multi-stakeholder body such as a Social Media Council is a long-term and complex endeavour, a local Coalition on Freedom of Expression and Content Moderation would be a lighter approach that could be supported within a shorter timeframe. Basing its work on international standards on freedom of expression and other fundamental rights, such a coalition could provide valuable input to inform content moderation practices, notably through its knowledge and understanding of the local languages and circumstances. As a critical mass of local stakeholders, it could engage into a sustainable dialogue with social media platforms and contribute to addressing flaws in content moderation and improving the protection of fundamental rights online. The coalition could provide training and support on freedom of expression and content moderation to local civil society actors that are impacted by content moderation. Finally, it could possibly pave the way to the creation of a Social Media Council in the country at a later stage. Through this research, at the initial stage of the **Social Media 4 Peace** project, the idea of a local Coalition on Freedom of Expression and Content Moderation was submitted to local stakeholders, whose views have enabled the formulation of recommendations on how to approach the facilitation of a pilot coalition in the specific context of Indonesia. In order to guarantee the effective ownership of the coalition by its members, the process facilitating its creation will necessarily include a validation exercise that ensures that potential members have the opportunity to discuss the findings of the research.

For the purposes of this report, we rely on the following definitions:

- **Content moderation** includes the different sets of measures and tools that social media platforms use to deal with illegal content and enforce their community standards against user-generated content on their service. This generally involves flagging by users, Trusted Flaggers or ‘filters’, removal, labelling, down-ranking or demonetisation of content, or disabling certain features.
- **Content curation** is how social media platforms use automated systems to rank, promote, or demote content in newsfeeds, usually based on their users’ profiles. Content can also be promoted on platforms in exchange for payment. Platforms can also curate content by using interstitials to warn users against sensitive content or applying certain labels to highlight, for instance, whether the content comes from a trusted source.

## Methodology and structure of the report

This research is based on a combination of desk research and qualitative interviews with 26 key informants (representatives from various local stakeholders). The researcher also had the opportunity to present preliminary findings and gather further feedback and contributions from stakeholders during two focus group discussions organised by Perludem and one public event organised by SAFEnet.

The desk research allowed for the identification of issues linked to the circulation of problematic content on social media in Indonesia. The identified content moderation issues were then discussed with multiple stakeholders during interviews. The interviews aimed to understand the experiences and challenges of Indonesian groups in dealing with platforms on content moderation issues.[1] The interviewees also conveyed their reflections on the idea of a local coalition on content moderation and freedom of expression. The potential structures, members, roles, and dynamics of the coalition were also discussed.

The researcher presented the findings of this research in two focus group discussions conducted by the election watchdog Perludem (the Association for Elections and Democracy). Perludem held a series of discussions to prepare a roadmap for secure elections in 2024 in Indonesia. The first focus group discussion was on 29 December 2021 and involved participants from multiple stakeholders (including civil society groups and electoral bodies; social media representatives were unable to attend the meeting). The second focus group discussion took place on 18 February 2022, and only civil society actors were invited. Overall, Perludem and the participants welcomed the idea of forming a local civil society coalition on content moderation and freedom of expression.

The researcher was also invited to discuss the findings during the public launch of SAFEnet's 2021 Digital Rights in Indonesia Situation Report on 2 March 2022. The four presenters from civil society groups and a representative of Meta commented on the current state of digital rights in Indonesia and discussed the potential way forward and the role that civil society could take in responding to the situation. Around 65 participants attended the live webinar. The researcher explained the findings of this ongoing research and discussed with the other presenters and meeting participants the need to form a local civil society coalition on content moderation and freedom of expression. The panel concluded with a shared view on the need to unite and empower Indonesian civil society groups to ensure their participation in the development and enforcement of Internet-related policies in Indonesia that balance freedom of expression with the safety of individuals and the public.

At the time of finalising this report, the dialogue between the researcher, public authorities, social media companies, and civil society groups continues, particularly on the preparation of peaceful elections in 2024 and how a future coalition on content moderation and freedom of expression could play a role in securing peaceful elections.

The structure of the report is as follows:

The Introduction highlights the diversity and complexity of Indonesia society. It further presents how the deep conflicts that exist within society have at times been exploited for political and economic profit.

The first chapter, The state of content moderation in Indonesia, describes the landscape of social media platforms and explores the dynamics and issues related to the use of social media and the practices of content moderation in the country. One striking issue is the proliferation of 'grey-area' content (content that contains what can be described as 'the seeds of hate' but does not necessarily amount to explicit incitement; the key concern with such content is that its massive amplification could lead to real-world violence). This chapter looks at the need for social media to understand and appreciate the local context when applying the global community standards, and concludes that while there are a number of initiatives and discussions involving social media with civil society and state actors, there is room for more meaningful dialogue within the development and enforcement processes of platforms' community standards.

The second chapter, Analysis of stakeholders, provides an analysis of the relevant stakeholder groups that deal with or are impacted by the content moderation practices.

Based on this analysis, the third chapter includes recommendations on how to facilitate the formation of a civil society coalition on content moderation and freedom of expression in Indonesia to bridge the dialogue between social media and local civil society.

## Indonesia at a glance

Many of the 'largest and the most' designations are bynames for Indonesia. It is the world's largest island country located in the Southeast Asian region. Spanning across 1,905 square kilometres, the country is divided into Western, Central, and Eastern parts of Indonesia. It is comprised of five main islands (Sumatera; Kalimantan; Java; Nusa Tenggara and Bali; and Sulawesi, Maluku and Papua) that are further divided into 34 provinces and over 17,500 islands. With a population of over 277 million people, it is the fourth most populous country and the third largest democracy in the world. In terms of economy, Indonesia is the largest economy in Southeast Asia and is predicted to be the fifth largest economy in the world by 2024. While its economy has been impacted by the pandemic, Indonesia is in the process of economic recovery to a pre-pandemic level, with a gross domestic product growth rate expected of about 4% this year.

The population of Indonesia is characterised by social, cultural, and economic diversity. While recognised as a country with the largest number of Muslims (231 million people in 2021), Indonesia also acknowledges five other official religions, namely Protestantism, Catholicism, Hinduism, Buddhism, and Confucianism. There are more than 300 ethnic groups living together in the country. Besides some of the largest ethnic groups, Javanese (40% of the total population), Sundanese (15.5%), Malay (3.7%), Bataknese (3.6%), Madurese (3%), and Betawi (2.9%), there are also other smaller ethnic groups, such as the Buginese, Balinese, Acehnese, Papuan and Chinese Indonesians. The national language is Indonesian. However, only 7% of the total population speak Indonesian as their mother tongue: the rest use it as their second language. There are more than 300 different languages and local dialects spoken in Indonesia.

The pandemic has widened the economic and educational disparity in the country. There are increasing numbers of people living in poverty today. The Central Statistics Agency (BPS) released the level of inequality in Indonesia's population expenditure as measured by the Gini Ratio, which was 0.384 in March 2021. The country has not been able to return to September 2019's Gini Ratio of 0.380. The government has been struggling to close the disparity in education services and standards, and the gap has even widened during the pandemic. Only those with sufficient facilities, access to electricity, and the Internet could adapt. However, Internet connectivity is highly concentrated in the Western part of Indonesia, particularly on the more urbanised island of Java. Meanwhile, the eastern provinces, such as Maluku, West Sulawesi, North Maluku, East Nusa Tenggara, and Papua, are struggling with Internet access.

Prior to its declaration as an independent country, there was already social, political, and cultural segregation during colonial times in the Dutch East Indies (as Indonesia was known at the time). The segregation between Muslims and Christians was reinforced by the discriminatory policies of the colonial regime against certain groups for economic and political gains. A cooperative yet competitive relationship between the ethnic Chinese and the Dutch colonial government was also formed during that period. The Dutch needed to cooperate with the Chinese to keep the Dutch East Indies economy alive. However, the Chinese were gradually considered a threat as the income of the colony was declining at that time. This resulted in the killing of Chinese people in 1740 (known as 'Geger Pacinan' or 'Tragedi Angke').

The founding father of Indonesia, Sukarno, realised the opportunities and challenges that cultural diversity represented for the country. Sukarno formulated the official state philosophy, known as Pancasila (Five Principles), to serve as the foundation of the country:

1. Belief in the one and only God
2. Just and civilised humanity
3. The unity of Indonesia
4. Democracy guided by the inner wisdom in the unanimity arising out of deliberations amongst representatives
5. Social justice for all of the peoples of Indonesia

These principles are often generalised in terms of religious devotion, humanitarianism, nationalism, consultative democracy, and social justice.

As Indonesia was established as a sovereign nation in 17 August 1945, Pancasila became the foundational element of the 1945 Constitution of the Republic of Indonesia. Section XA of the Constitution identifies 'human rights' as recognised and protected by the government. The need to unify such a diverse country also led to the formulation of the official national motto, Bhinneka Tunggal Ika ('Unity in Diversity'), which is inscribed on the national emblem of Indonesia, the Garuda (Eagle) Pancasila. It declares the essential unity of its inhabitants regardless of ethnic, regional, social, or religious differences.

Following the 1965 incident,[2] Suharto took the leadership of the country. There is a long list of human rights violations in his era: during Suharto's 32 years of military dictatorship, there was a cultural and political repression in the country. The government controlled the press and circulation of information in the country and restricted freedom of expression. Accordingly, people had to rely on rumours and non-official sources in order to get information about, and understand, the current events of the day. The government also introduced the official use of the discriminatory labels of pribumi (indigenous) and non-pribumi, which was lifted after the post-Suharto era. The ethnic Chinese were not recognised as ethnic groups and all Chinese-related culture was banned (the celebration of Chinese New Year, the use of Chinese names and the Mandarin language). Fuelled by the government and military officials, there was strong anti-Chinese sentiment across the country, expressed in the May 1998 riots that led to the resignation of President Suharto.

The fall of Suharto in 1998 brought reforms and democratic gains to Indonesia. The reformation in 1998 was a milestone for the acknowledgement of human rights principles as the country adopted a comprehensive human rights law (Law No. 39/1999 on Human Rights). Pluralism in politics and the media also took hold in the country.[3] Acknowledging the impressive democratic gains since 1998, Freedom House describes Indonesia as partly free in their 2021 report, noting that the 'the country continues to struggle with challenges including systemic corruption, discrimination and violence against minority groups, conflict in the Papua region, and the politicised use of defamation and blasphemy laws'.

# The state of content moderation in Indonesia

## Social media landscape in Indonesia

Internet penetration is steadily increasing in Indonesia. In 2021, data from the Indonesian Internet Service Provider Association (APJII) and HootSuite placed Indonesia's Internet penetration rate at 73.7% (202.6 million people). The penetration is due to the rapid growth in the number of mobile Internet subscriptions. There were over 345.3 million subscriptions in 2021, an increase of 1.2% since 2020. Accordingly, 96.4% of total Internet users (195.3 million people) use mobile Internet in the country.

The first reason why Indonesians use the Internet is to allow them to use social media. There are at least 16 social media platforms and chat apps being used in the country. They are YouTube (93.8% of usership), Instagram (86.6%), Facebook (85.5%), Twitter (63.6%), Line (44.3%), TikTok (38.7%), LinkedIn (39.4%), Pinterest (35.6%), WeChat (26.2%), Snapchat (25.4%), Tumblr (18.4%), and Reddit (17.1%). Meanwhile, chat apps used in the country are WhatsApp (87.7%), Facebook Messenger (52.4%), Telegram (28.5%), and Skype (24.3%).

The research from APJII confirmed that there are five main social media platforms used in Indonesia: YouTube, Facebook, Instagram, Twitter, and LinkedIn. Meanwhile, data from HootSuite also shows an increasing number of TikTok users. With regard to chat apps, APJII found that WhatsApp is predominantly used in the country (93.7%).

YouTube has to a large extent replaced the television industry in the country. There are around 190 million Indonesian Internet users ranging from 15–64 years old that use YouTube. This is illustrated by the fact that the popular YouTube channel of Indonesian artist Deddy Corbuzier named #CLOSEDTHEDOOR[4] put up billboards in main streets of some big cities in Indonesia to promote the channel and equate it to a television show. The billboard challenged the public with the question: 'STILL WATCHING TV?'.

With more than 173 million users in Indonesia, Facebook is dominated by the younger segments of the adult population. There are 33.6% of users from the millennial generation (25–34 years old) and 30.2% from generation Z (18–24 years old), 14.3% are in the age range of 35–44 years old. Instagram is more popular among young adults and teenagers aged 18–24 years old (36.4%), 25–34 years old (31.6%), and 13–17 years old (12.9%).

Indonesian users have been especially dependent on big social media platforms during the pandemic. They go to YouTube and Facebook for entertainment as well as educational content, news, and political and social content. Indonesian users access Twitter if they want to look for information from experts in various fields, including information related to Covid-19. Indonesian users use TikTok and Instagram mostly for entertainment, life style, and e-commerce.

While they use social media for various purposes, several studies indicate that there is a key similarity in the use of social media and chat apps among Indonesian users – receiving and sharing information with friends and relatives.[5]

## Overview: Impact of content moderation on peace and stability

This chapter discusses categories of problematic content (disinformation, hate speech, online harassment, terrorism and radicalism, and online gender-based violence), how such content is amplified by malicious actors, the relationships between stakeholders and platforms, the availability of content rules in local languages, the effectiveness of remedies provided by platforms, the problems related to polarisation/conflicts in society, the marginalisation of certain groups, and the impact on freedom of expression and media freedom.

There are efforts in Indonesia to exploit conflicts that exist in society for political and economic gain. The spread of false information and hate speech on social media platforms is particularly intensified during election periods. Political candidates hire social media campaign strategists to manage their online election campaigns: this strategy often involves hiring ‘buzzers’ and the use of chatbots to spread disinformation and steer public debates.

The use of social media in Indonesian politics started during the 2012 Jakarta gubernatorial election. In the final round of votes, two candidates remained for the positions of governor and vice governor each: the incumbent Fauzi Bowo and Nahrowi Rahmi and the challenging pair Joko Widodo (Jokowi) and Basuki Tjahaya Purnama (commonly called Ahok). The language of the Bowo–Rahmi campaign heavily emphasised the individual traits of Jokowi and Ahok. They called for people to choose a candidate of the same faith and deliberately insinuated that Jokowi's mother was Catholic. Governor Jokowi and Vice Governor Ahok ended up winning the election.

The centrality of social media in election campaigns was even more apparent in the 2014 Presidential election and the 2017 Jakarta gubernatorial election. In 2014, the clash between supporters of two candidates, Jokowi and Prabowo, was visible on social media platforms. As Jokowi was elected to be the President of Indonesia, Ahok filled the position of Jakarta Governor. In 2017, Ahok ran for re-election in the Jakarta gubernatorial election. Following the second round of voting, there was increasing polarisation between the supporters of the two remaining candidates, Ahok and Anis. The supporters of one candidate labelled the supporters from the other camp as the enemy.

An outspoken figure with a double minority background (Chinese Indonesian and Christian), Ahok has been subject to racist comments. In a campaign speech, Ahok criticised his political opponents for using Islam as a campaign tool. A man named Buni Yani edited the speech and added a caption that made Ahok seem like he was insulting the Quran. The edited video was uploaded to Yani's social media accounts and went viral, which led to a flare up. Massive protests from Indonesian Muslim groups took place on 4 November 2016 (Action 411) and 2 December 2016 (Action 212). These huge pressures forced Jokowi to decide that Ahok must be charged and prosecuted. Ahok did not win the election, and eventually he was jailed for blasphemy.

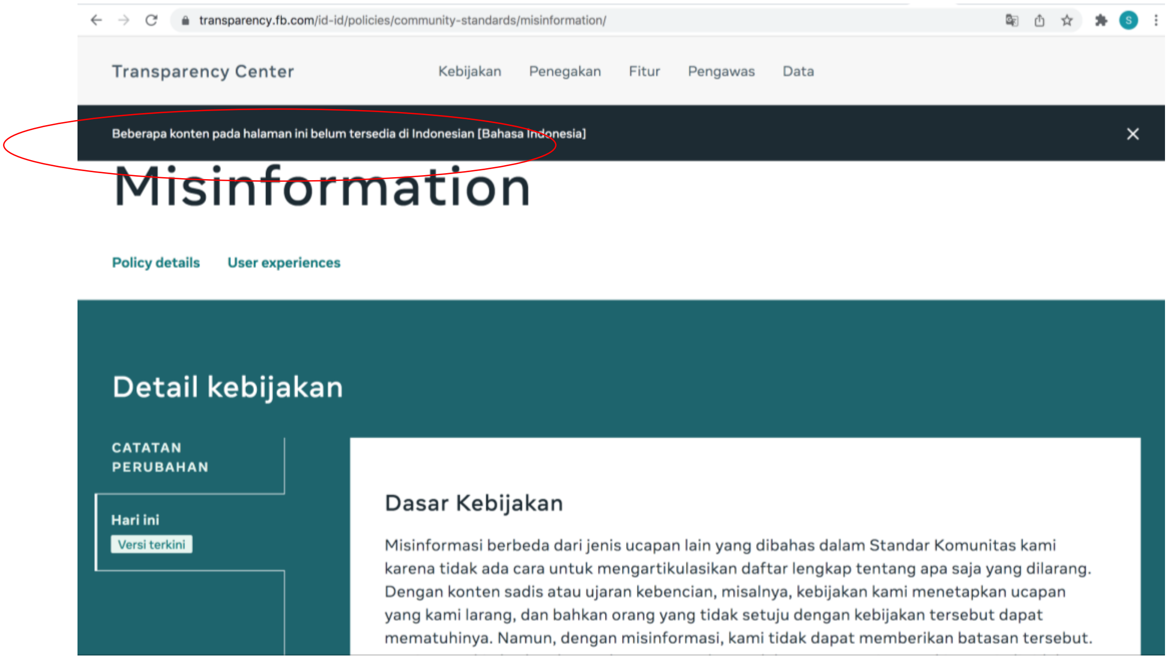
The circulation of disinformation that sought to delegitimise the election process and results continued into the Presidential election in 2019. Building on the sentiment against minority groups developed in the previous elections, there was a circulation of hoaxes on social media that claimed some protestors were shot by Chinese police during the post-election demonstrations in May 2019. While this message contained no explicit statement that promoted violence, anti-Chinese sentiment triggered chaos (this is a clear example of the risks linked to 'grey-area' content, as shall be further explored in this report). The incident prompted the government to block access to social media platforms and chat apps to curb the spread of such hateful disinformation. However, the spread of anti-Chinese disinformation found its way to the encrypted platform Telegram.

In August 2019, the government ordered further Internet restrictions during violent protests in Papua. Despite the decision of the Jakarta State Administrative Court in 2020 that there is no legal basis in the Electronic Information and Transactions (ITE) Law that regulates Internet shutdowns, Internet users in Papua experienced issues in accessing the Internet in April 2021.[6]

## Social media platforms operate globally

This section examines the disconnect between the global community rules of platforms and their enforcement, particularly in tackling 'grey-area' speech in the specific context of Indonesia. To do so, this section focused on examining the hate speech policies of platforms as they are particularly relevant to understanding the complexity of 'grey-area' content.

For several reasons, the global content rules appear to be problematic.[7] A first serious flaw in the system is that the level of commitment to provide the full details of community standards in the Indonesian language varies from one platform to another. As the community standards are living documents that are regularly updated, the availability of the Indonesian translations appears not to be updated. For example, as shown in the Figure 1 captured on 4 March 2022, the Indonesian version of Facebook's misinformation guidelines announces that 'Some of the content on this page is not yet available in Indonesian language.' In the meantime, in its English Community Standards, Facebook states 'Please note that the US English version of the Community Standards reflects the most up-to-date set of the policies and should be used as the master document.'



Beberapa konten pada halaman ini belum tersedia di Indonesian [Bahasa Indonesia]

Misinformation

Detail kebijakan

Dasar Kebijakan

Misinformasi berbeda dari jenis ucapan lain yang dibahas dalam Standar Komunitas kami karena tidak ada cara untuk mengartikulasikan daftar lengkap tentang apa saja yang dilarang. Dengan konten sadis atau ujaran kebencian, misalnya, kebijakan kami menetapkan ucapan yang kami larang, dan bahkan orang yang tidak setuju dengan kebijakan tersebut dapat mematuhinya. Namun, dengan misinformasi, kami tidak dapat memberikan batasan tersebut.

*Figure 1: Screenshot of Facebook's misinformation guidelines in Indonesia language.*[8]

Moreover, while global content rules elaborated by social media companies typically explain the types of content that are prohibited on their platforms along with some specific examples to help their users understand the scope of the restrictions, they do not seem to take into consideration 'grey-area' content and how, in the context of Indonesia, such content can lead to real-world harm. Here are some examples:

- **Twitter**

  'We prohibit content that makes violent threats against an identifiable target. Violent threats are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm, where an individual could die or be significantly injured, e.g., 'I will kill you.''[9]

- **TikTok**

  'Do not post, upload, stream, or share:

  - Hateful content related to an individual or group, including:

    – calling for or justifying violence against them.'[10]

- **YouTube**

  'Don't post content on YouTube if the purpose of that content is to do one or more of the following.

  - Encourage violence against individuals or groups based on any of the attributes noted above. We don't allow threats on YouTube, and we treat implied calls for violence as real threats. You can learn more about our policies on threats and harassment.

  - Incite hatred against individuals or groups based on any of the attributes noted above.'[11]

- **Instagram**

  'We remove content that contains credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages.'[12]

- **Facebook**

  'We aim to prevent potential offline harm that may be related to content on Facebook. While we understand that people commonly express disdain or disagreement by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence.'[13]

In addition to difficulties related to the definition of categories of prohibited content,[14] it would be helpful to know how platforms assess 'grey-area' speech in relation to the potential harms that are likely to occur in light of the sensitivity of certain issues in the local context. This may include, for example, the motives of the poster, the frequency of past violations made by the poster, the profile of the victim, the power relationship between the poster and the victim, and the harm likely to result in relation to the sensitivity of the issue from the local perspective. Furthermore, clear indications are necessary about the appeals mechanisms, appropriate content moderation measures that companies may take (such as the types of sanctions), and the timeframe within which companies intend to manage the content in question. In that respect, for example, Facebook briefly explains its considerations for deciding if a threat is credible and needs to be removed:

> *We remove content, disable accounts and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety. We also try to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety. In determining whether a threat is credible, we may also consider additional information like a person's public visibility and the risks to their physical safety.*[15]

In response to a question on whether it is possible for community guidelines to specify how to handle 'grey-area' speech, a platform representative in Indonesia underlined how it is not always straightforward to translate the policy texts into action particularly when an abundance of the content requires a comprehensive understanding of many contexts. 'Grey-area' content is, consequently, difficult to define and therefore content moderation in this area could be complemented by a dialogue with locally-relevant multi-stakeholder expert groups to guide the processes.

This finding highlights the gap in the current state of content moderation that is primarily conducted by platforms. There is a need for the community standards of platforms to provide a reliable and consistent baseline to guide content moderation decision-making processes. However, given the prevalence and complexity of 'grey-area' content in Indonesia (see below), community standards may not be able to capture all the considerations and enable effective responses. This means that platforms should work hand-in-hand with the relevant local stakeholder groups to enhance content moderation mechanisms. Academic research funded by Facebook came to a similar conclusion of a need to complement community standards with guidance from local groups in order to fully grasp the complexity of the local context and the degree of harm experienced by the related groups.

## Problematic content in Indonesia

This section addresses the complexity of several categories of content that are distributed massively in Indonesia. These include disinformation, hate speech, online harassment, terrorism, radicalism, and online gender-based violence. While these are different categories of problematic content, they all present the same difficulty related to what this study describes as ‘grey-area’ content – ‘seeds of hate’ directed to certain individuals or groups, which, although it does not amount to an explicit incitement to discrimination, hostility, or violence, may be amplified through social media and lead to serious violence.

The case studies below showcase the gap between the global community rules of platforms and their enforcement at the local level.

### Case study 1

This example looks at the absence of shared understanding between a platform and an election authority regarding content moderation measures.

As explained in the ‘Introduction to the context of the country,’ there are deep historical social, political, and cultural divides in Indonesia, particularly between Muslims and Christians. Since 2012, societal tensions, particularly through social media, have intensified. In 2014, the clash between supporters of two Presidential candidates, Jokowi and Prabowo, gained visibility on social media platforms. In 2017, there was increasing polarisation between the supporters of the two final candidates, Ahok and Anis Baswedan. Ahok was jailed for blasphemy after a manipulated video gave the impression that he was insulting the Quran. The circulation of disinformation using anti-Chinese and anti-communist sentiments continued into the Presidential election in 2019.

Observing such a critical situation, a representative of an election authority argued with the local representatives of a platform that the proliferation of anti-Chinese and anti-Communist content during the campaign period for the 2019 Presidential election was dangerous. They argued that such content might lead to violence and therefore requested the platform to moderate that type of content. However, the platform was consistently resistant to taking any actions because the content was perceived as not containing incitement to violence. The consistent refusal frustrated the election authority and eventually stopped them from flagging that 'grey-area' content to the platform.

### Case study 2

This case study shows how a platform refused requests to take down online gender-based violence with no explanation for the factors that determine how the platform assess the severity of a 'grey-area' speech, the appropriate content moderation measures, and the approximate timeframe to manage the content in question.

Two Muslim women reached out to Andreas Harsono, the Indonesian researcher of Human Rights Watch, asking for help. They had been subjected to online bullying, intimidation, and death threats on their social media accounts after one of them had talked in a webinar, arguing that Islam allows Muslim women decide how they want to dress, including whether they want to wear a hijab (headscarf and long-sleeve shirt) or not. She had talked about how women should be able to control their own bodies. Her friend and another male colleague in Cairo, Egypt, had defended the same position during the webinar. Due to the online bullying and later the threats of being hacked and poisoned to death, the women decided to move to Jakarta, and they reported the threats to the National Police.

Harsono read more than 60 pages of material, checked the online material and corresponding hyperlinks, and developed a legal analysis of the case, which he emailed to the headquarters of the social media platform. He provided an analysis of the overall context, such as the context of the webinar (which had been discussing a new government regulation to allow state-school female students and teachers to choose to wear the hijab or not). He also described the multiple posts against the three victims.

Harsono had just written a report for Human Rights Watch on the bullying, intimidation, discrimination, and sometimes violence to force Muslim girls and women to wear the hijab in Indonesia, a predominantly Muslim country: *'I Wanted to Run Away – Abusive Dress Codes for Women and Girls in Indonesia'* was published in March 2021. The email concluded with a request to the platform to take down the death threats messages.

Harsono sent the email on 27 April 2021 and a follow-up on 15 June, copying several other people at Human Rights Watch. A representative of the platform eventually replied on 4 August. They wrote, 'There is real nuance to our hate speech and bullying policies, however, and a lot of what folks experience as threatening or unpleasant doesn't meet the (very classic human rights oriented) standard.' The decision came as a huge disappointment for the three victims.

### Disinformation

There are concerted efforts in Indonesia to produce and spread disinformation online through paid commenters and bots. The spread of disinformation is used to deepen the existing social, racial, and religious divisions in the country, and such efforts are more aggressive during election periods.

This growing problematic trend is reflected in the latest reports from the Oxford Internet Institute released in 2020 and 2019. Back in 2017 and 2018, specific groups designed to destabilise the country with disinformation and hate speech had been identified, such as the Saracen and Muslim Cyber Army movements. Today, an increasing number of paid commenters, called 'buzzers', as well as automated accounts, seek to manipulate the online information landscape on behalf of political parties and private contractors. The Oxford Internet Institute found that such teams work to support certain narratives and attack their opposition by exploiting the prevailing divisions and delegitimising the electoral process.

Election watchdog Perludem further emphasised that the absence of regulation on paid political advertising and the lack of advertising transparency on social media platforms have enabled the distribution of disinformation as paid political advertisement targeted at specific users. Scholarly research has showed that this practice is dangerous for democracy. Another study conducted by Perludem in collaboration with Facebook more specifically reflects that the spread of disinformation, which aims to blur the public information on technical election procedures and to delegitimise the election processes, have the potential to eliminate a person's right to vote.

### Hate speech

Not all hate speech that circulates in Indonesia can be clearly identified.[16] The less severe forms of hate speech (which this study defines as the 'grey-area') are difficult to identify. However, past incidents in the country show that such 'grey-area' content contributed to deepening the polarisation in society and in some cases even resulted in riots.

The National Hate Speech Dashboard by think tank Centre for Strategic and International Studies (CSIS Indonesia), an initiative to monitor and visualise hate speech online until the 2024's Presidential Election and beyond, is monitoring the general trends of hate speech on Twitter in Indonesia. Currently, the initiative is focused on examining tweets that target vulnerable communities in Indonesia, namely Ahmadiyyah, Shi'a, and Chinese Indonesians. The researchers noted that there is also significant hate speech in Indonesia's online sphere directed towards other religious and racial minority groups, as well as LGBTIQ+ communities (lesbian, gay, bisexual, transgender, intersex, and questioning).

The 'grey-area' type of hate speech does not necessarily include explicit incitement to hatred or discrimination, but its amplification promotes division in society. In the long run, it could turn into problematic speech and could potentially lead to violence.[17] This is an area in which 'buzzers' have become expert: they carefully craft content that claims to promote democracy, freedom of expression, and the Unity in Diversity motto of Indonesia, but in effect belittles opposition groups and nurtures polarisation in society.

Mixed with disinformation, polarisation is then leveraged and used to trigger animosity. An exemplary case is the circulation of hoaxes on social media that claimed some protestors were shot by Chinese police during the post-Presidential election demonstrations in 2019.

While this message contains no explicit statement that promoted violent actions, a sensitive anti-Chinese issue was able to trigger racial hatred and resulted in riots.

The issue of hate speech in the country is also characterised by the overly broad definition stipulated in the 2008 ITE Law. An overbroad definition of hate speech is often misused to restrict free speech and criminalise those who are critical of power-holders. Indeed, civil society organisations have been criticising the law.

### Online harassment

There are increasing efforts in Indonesia to threaten journalists and activists through online harassment. The Alliance of Independent Journalist (AJI) noted that worrying doxing incidents directed against journalists started to happen back in 2018. There were at least three online persecutions at that time. In 2020, Southeast Asia Freedom of Expression Network (SAFEnet) reported that there were 13 cases of doxing towards journalists, human rights activists, and citizens. This is double the number of cases occurring in 2019.

These attacks are typically conducted after the involved actors have voiced critical comments towards authorities through their social media accounts or media outlets. The attacks have also been directed against activists and citizens who have joined anti-government demonstrations. Death threats and harassment were directed at them and their family members. Their social media accounts and chat apps were also hacked.

While death threats and harassment might not always result in real-world violence, the act of doxing that publicly reveals someone’s personal data on social media and chat apps (such as a home address, family photos, telephone number, and even more concerningly their location) may lead to escalation.

For example, Freedom House highlighted that student and activist protestors who attended a demonstration against the Omnibus Law in Yogyakarta in October 2020 received online threats. Student organisers and participants of online discussions with topics that were critical of the authorities were also subjected to online harassment. Their personal data, including their location, was doxed. While such cases may not always cause violence, they sometimes do. For instance, Victor Mambor (a well-known West Papua journalist, the founder of independent online media Jujur Bicara Papua (Jubi)\Honestly Speaking about Papua), faced a series of online threats, digital attacks, and doxing, and subsequently his car was vandalised by unknown people. The consequences of online threats against less well-known journalists and activists in the depths of Indonesia could be more severe.

### Terrorism and radicalism

Social media and private chat groups are intensely used in Indonesia to disseminate radical ideology, to recruit members and terrorist fighters, and to promote violent extremism.

Research from the Institute for Policy Analysis and Conflict (IPAC) showed that one of the most serious challenges to combat violent extremism content online is the absence of shared understanding on how to differentiate between intolerant but legitimate political opinion and ethno-religious hate speech that contains potentially terrorist content and violent extremism.[18] There is no shared understanding among civil society groups, the government, and platforms. According to the study, while there could be agreement around what constitutes violent extremist content, it remains much more challenging to draw a clear line between ethno-religious hate speech and violent extremism.

For example, research from Bhinneka Kultura Nusantara, a research group focused on the diversity of society in Indonesia, identified 37 social media accounts that are mostly affiliated with religious groups that produce and promote narratives of exclusionary and discriminatory parenting. The two main narratives promoted by those religious groups are ‘masuk surga sekeluarga’ (reach the heaven with all family members) and ‘bangun peradaban Islam’ (build Moslem civilisation). While the first narrative insists on traditional roles within the family, the second leaves no room for tolerance towards other diverse religious groups. While focusing on the family unit, these two narratives contradict the principles and values of democracy. They might barricade rooms for dialogue between diverse groups and shut down the voices of minorities. While such narratives do not explicitly encourage violent extremism, the researchers argued that such content nurtures violence at the family level in Indonesia. In addition, family and religion-based violence are potentially interconnected. This can be seen in the cases of family suicide bombings that have happened several times in Indonesia. The latest was completed by a couple of spouses at Katedral Makassar Church, South Sulawesi, in March 2021.

### Online gender-based violence

The pandemic has forced millions of people to go online. However, a spike in online violence towards individuals based on their gender identity or sexual orientation corresponds to the growing number of Internet users in Indonesia. The National Commission on Violence Against Women (Komnas Perempuan) recorded 940 reported online gender-based violence cases in 2020, an increase of 241 cases from 2019. The Legal Aid Foundation of the Indonesian Women's Association for Justice (LBH APIK) dealt with 307 cases in 2020, while before the pandemic, it handled 17 cases in 2019 alone. Moreover, during 2019, the Digital At-Risks (DARK) Subdivision of SAFEnet assisted 45 victims of online gender-based violence, whereas it received 169 filed cases from March–June 2020. Meanwhile, many other online gender-based violence victims are reluctant to report their cases – some might not know the procedures, while others feel uncomfortable or traumatised, and can even distrust law enforcement agencies if they need to report their cases.

A report from SAFEnet specifies the various categories and activities that can be included as online gender-based violence, namely privacy infringement, surveillance and stalking, reputation damage, online and offline harassment, online and offline threats, and cyber attacks. These activities are typically directed at someone who is involved in an intimate relationship, public profiles (such as activists, journalists, researchers, artists), and also survivors of physical attacks.[19]

While online gender-based violence content may not always cause casualties, the victims may experience immaterial losses, such as the loss of privacy or self-confidence. The victims also have to bear the stigma that can cause them to lose their jobs, relationships, and future lives. ARTICLE 19 has put forward recommendations on how social media platforms have a role to play in addressing gender-based harassment and violence against women on their platforms.

## Content moderation dynamics in Indonesia

### Content moderation decisions in Indonesia

The previous section showed that there are serious content moderation concerns in Indonesia. Building upon this finding, this section highlights that the government and civil society groups have been battling to push platforms to take local context into consideration in their content moderation decisions. The government acted through dialogue with social media companies, also sometimes blocking platforms, and generally through content-related regulations.[20] Leading civil society organisations have facilitated a dialogue between platforms and the wider society.

To explain the above point, it is helpful to explore the dynamics of content moderation practices in Indonesia.

After the 2017 Jakarta gubernatorial election and in preparation for the 2019 Presidential election, the circulation of hoaxes, hate speech, and extremist content had become a matter of serious concern. Previously, the government had relied on public authorities and the public to report problematic content, while also blocking websites that were deemed to be harmful to the public. As the role of social media became more central in Indonesia, the government started to reach out to platforms to ask for their help in tackling problematic content.

It was difficult for the Ministry of Communication and Information Technology (MCIT) of the Republic of Indonesia to reach out to US-based social media offices and to establish a dialogue and a speedy review process. The Communication Minister Rudiantara publicly complained that Facebook was slow in responding, as it needed to discuss requests coming from the government with its legal team at headquarters. Additionally, he doubted that the US-based policies of the platform could understand the national context. In an interview with the media, Rudiantara stated, ‘I respect their policies, but this is happening in Indonesia, so you should follow the rules here.’

However, in its requests to platforms, the government made no difference between legitimate (although offensive) political opinion and criminal incitement and violent extremism in the context of Indonesia – and platforms were only willing to comply with requests that were aligned with their global community standards.

MCIT then leveraged its efforts to tackle the proliferation of problematic content through blocking. Rudiantara said, 'Actually, Indonesia doesn't intend to block, but if social media are outrageous, then we can close.' This step was finally taken by the government in July 2017. It blocked the encrypted chat app Telegram after the owners did not respond to MCIT's request for removal. Many private extremist groups make use of Telegram, especially during the 8–9 May 2018 prison riots during which terrorist suspects detained at the Brimob headquarters went on a rampage, killing at least five police investigators and one prisoner. The blocking acted as a wake-up call for other tech giants to respond to the government's content moderation requests.

In order to facilitate coordination between the government and the social media headquarters, MCIT also pushed big platforms to establish local representation in the country. Additionally, platforms and MCIT launched the national trust-related programmes that involved civil society groups: YouTube's Trusted Flagger, the trusted partners initiative of Facebook, and Tik Tok's Child Safety Partner (see further explanation on the involved civil society groups in the Analysis of stakeholders). Twitter's Global Trust and Safety Council invited two Indonesian civil society organisations (ICT Watch and the Wahid Institute) to join the Council in 2016 (SAFEnet and ECPAT[21] Indonesia joined the Council later). This initiative ensured a special line of communication for civil society groups to flag any problematic content to the social media representatives in the country.

Following the 2018 Cambridge Analytica scandal that reportedly impacted Indonesia, there has been increasing pressure from the authorities to hold Facebook accountable. In 2018, the Parliament and the government called Facebook executives to testify in a public hearing. Since then, the government has become more aggressive in regulating social media content. While the spirit is to protect society from the proliferation of problematic content, the content-related regulatory efforts in Indonesia have not been able to strike the balance between the protection of freedom of expression and the safety of individuals and the public. As a result, online content and Internet-related regulations in Indonesia have failed to uphold international standards on freedom of expression.

Generally, international standards stipulate that any restriction must be:

- **Prescribed by law**: any restriction must be formulated with sufficient precision. Overbroad restrictions are not allowed.
- **In pursuit of a legitimate aim**: restrictions shall only be permitted for (a) the respect of the rights or reputation of others and (b) the protection of national security or of public order, of or public health or morals.
- **Necessary and proportionate**: restrictions must have a direct and immediate connection between the expression and the protected interest. Moreover, proportionality means that the restrictions must be specific, tailored, and the least intrusive means to achieve the same limited result.

A study by the Institute for Policy Research and Advocacy (ELSAM) noted that the regulations to limit online content in Indonesia have not been aligned with the international standards on limitations on freedom of expression.[22] Indonesia's regulatory efforts on the limitations on the right to freedom of expression typically contain an overly broad definition of negative content and the absence of detailed procedures to implement the restrictions in Indonesia.

Articles 27–29 of the ITE Law defined in broad terms the categories content that are prohibited online.[23] The Regulation of the Government of the Republic of Indonesia Number 71 of 2019 on Electronic Systems and Transactions (PP 71\2019) and the Communications and Information Ministerial Regulation No. 5/2020 on Private Electronic System Operators (MR 5\2020) are further attempts to regulate online content. However, those regulations are characterised by the overbroad definition of prohibited online content and the absence of any detailed procedures.

The overly broad definition of prohibited online content in Indonesia's Internet-related regulations, along with the compliance of platforms with the government's requests for securing their presence and expansion in the country, may further undermine the protection of freedom of expression in the country. For example, in the first year of TikTok's presence in Indonesia, MCIT blocked eight domain name system (DNS) servers of TikTok on 3 July 2018, on the basis that the service hosted pornography, immorality, religious harassment, and other negative content. TikTok immediately met with MCIT and Indonesian child protection officials on 4 July and pledged to collaborate with the Indonesian Government on removing 'negative content'. After the block was lifted on 10 July, the next day TikTok stated its commitment to provide a secure, healthy, and good quality platform for Indonesia, a very important market for TikTok. Meanwhile, Indonesian users and society remain uninformed about the definition and scope of 'negative content' that is prohibited by the government and social media.

MR 5\2020 in particular requires digital intermediaries, such as social media platforms, to ensure that their 'electronic systems' do not contain prohibited electronic information or documents or facilitate the spreading of such prohibited content (Article 9(3)). In the meantime, Indonesian users are uninformed about how and to what extent platforms rely on international standards on freedom of expression and on a satisfying assessment of the local context in their content moderation decisions.[24]

The lack of transparency and accountability of social media platforms are apparent in the interactions between civil society groups and platforms. During the interview process, civil society groups who are the trusted partners of platforms, criticised the platforms' responses to their requests. They viewed these platforms as being more responsive to disinformation and hate speech content, but not to online gender-based violence. The first two categories of problematic content are more pervasive, and platforms have received a lot of pressure from governments and the world to take responsibility for tackling them. Meanwhile, platforms have not put similar attention and effort into the emerging issue of online gender-based violence. So, the weight of one problematic piece of content against another is not the same in the eyes of platforms.

Ellen Kusuma, the Head of Gender-Based Violence Online Subdivision of SAFEnet, argued that platforms did not have consistent mechanisms in addressing the concerns raised by her team. While on one occasion, platforms might attend to their requests nimbly, on another it could take more than a week for platforms to respond to their requests. Furthermore, there are no clear rules on the types of problematic content and appropriate content moderation measures to be taken by platforms on the issue. Meanwhile, she emphasised that the victims of online gender-based violence need a quick and appropriate response from platforms before the problematic content that endangers them escalates.

Furthermore, as previously presented in the case studies, some interviewees conveyed that platforms sometimes showed some resistance to their requests, particularly on 'grey-area' content. Platforms often argue that there is no evidence that the relevant problematic content is escalating. However, in consideration of the quick amplification from 'grey-area' content to real-world violence, as evidenced in case study 1 (the post-Presidential election riots in 2019), the question arises of whether platforms are able to listen to signals and recommendations from local actors to assess the local context.

Another case happened to the fact-checkers of the Indonesian fact-checking organisation Mafindo (the Indonesian Anti-Slander Society) who criticised the decisions of a platform not to close social media accounts that spread false information on a continuous basis. The founder of Mafindo, Harry Sufehmi, conveyed that as an organisation that supports freedom of expression and after carefully paying attention to the repeated violations made by the same accounts, Mafindo saw the potential damage of the hoaxes produced by those accounts towards the polarisation in society which in the end may cause casualties. They then contacted the related platform several times, only to receive a diplomatic reply that restated the platform's decision. Moreover, while Indonesian fact-checkers have urged the platform to display clarified facts side by side with hoaxes, it preferred to reduce the visibility of those identified hoaxes. Fact-checkers wanted to empower the public to differentiate reliable from false information, but they presumed that the platform was reluctant because such actions might hurt the platform's public image and reputation.

A recurring pattern identified in all of these above views and experiences of civil society actors dealing with platforms is that platforms largely hold the decision-making power in the negotiation process. The following examples show that a closed coordination and dialogue between platforms and Indonesian stakeholder groups is needed in order to make content moderation decisions that appreciate the local context in Indonesia.

A lack of understanding about Indonesia is apparent in the content moderation decisions of platforms that use mass flags from individual users. Whenever there are mass requests from users directed towards the same account, platforms tend to move quickly, responding to and even agreeing to these requests. For example, in June 2021, civil society groups noted that Instagram took down content from at least two accounts of activists that advocated for corruption eradication efforts in Indonesia based on users' reports. Those activists received a notification that their posts contained incitement to violence and thus infringed on community guidelines. Civil society groups highlighted that these mass user reports and hacking attempts against activists and journalists were part of the counterattack by corruptors to prevent the strengthening of the Corruption Eradication Commission (KPK).

On another occasion, platforms also responded to mass flagging by closing social media accounts of paid pro-government influencers. In August 2021, the Twitter account of Ade Armando was suspended twice without a clear explanation. While there is no evidence, this communication lecturer is perceived by society as one of the key figures of the pro-government influencers as he often posts content that could steer public debates.[25] Interviewed by the press, Armando stated that he did not know which of his Twitter posts were reported, but he suspected that there were cyber teams that reported his account to Twitter.

Siti Cotijah from the Society Participation Division of the National Commission on Violence Against Women (Komnas Perempuan) complained that the anti-violence education material that the organisation livestreamed on YouTube was taken down by the platform. She explained to the researcher:

> *"The live streaming was 2 hours long from 10:00 to 12:00. However, in the first hour, YouTube accidentally cut off the live streaming and deleted the content of the first hour. Then we continued our live streaming again by changing the title of the streaming event using alphabets mixed with numbers (note: from violence into v10l3nc3 or k3k3r454n in the Indonesian language). After the live event was done, we renamed the title using the correct spelling."*

Cotijah said that they tried to contact YouTube, but even an established public authority like themselves found the reporting and appeal mechanisms challenging and one-sided. They did not know the reason behind the take-down decision, but they presumed it was because their video used the word 'violence' in Indonesian language ('kekerasan'). Consequently, the automated content moderation system 'thought' that the video promoted violence.

On the other hand, the Head of External Communication of Arus Pelangi, a civil society organisation that promotes the protection of LGBTIQ+ rights, mentioned during the interview their observations about content posted on Instagram by an organisation that advocates against the protection of these minority groups: those posts stayed on the platform, presumably because the uploader typed 'violence' as 'v10l3nc3' ('k3k3r454n' in the Indonesian language) in the content.

More importantly, all of these examples point out that platforms should be in close and meaningful coordination and dialogue with Indonesian stakeholder groups in order to make content moderation decisions that take Indonesia's language and context into consideration in their content moderation decisions. This need is apparent and important in the protection of credible online social movements and actors for strengthening public participation and democracy in Indonesia.

This can be seen in the case of attacks towards online social movements in Wadas Village in the context of a conflict between the police and residents who reject the Bener Dam construction and mining plans in Wadas Village in the Bener District of the Purworejo Regency, Central Java. The General Secretary of SAFEnet Anton Muhajir reported that SAFEnet received reports from Wadas residents and youth activists who sided with the residents that their Twitter accounts were suspended due to mass flags. They appealed to Twitter for the decisions. Meanwhile, SAFEnet supported them by corresponding with Twitter to explain the credibility of their profiles and digital activism activities. Within a few days, Twitter reactivated their accounts and even verified the account of Wadas Melawan\Wadas Fights Back with a blue sign.

**Actors who initiate requests for content moderation**

The practice of filing content moderation requests has been coming from all stakeholder groups in Indonesia. The Indonesian Government is active in sending content moderation requests to social media platforms. Some leading civil society organisations are trusted partners of several social media platforms, allowing them a direct communication channel to raise urgent issues to platforms.

MCIT, particularly through its Directorate of Informatics Application Control under the Directorate General of Informatics Application (Aptika), is the lead ministry in dealing with platforms with regard to content moderation. The Directorate of Informatics Application Control works on the basis of reports on problematic content produced by its Information Crawler Machine and team, as well as reports filed by the public and governmental institutions.

MCIT proactively detects content violations using the Information Crawler Machine 'Cyber Drone 9'. It is a crawler system driven by artificial intelligence (AI) to detect content violations. A team of around 100 people monitors the system and reviews the material it flags for blocking. The blocking is then done by electronic service providers, including social media platforms. Civil society groups have criticised the legitimacy and accuracy of the Cyber Drone 9 because it may result in the blocking of legitimate content and over-blocking.

The team also works based on reports filed by the public on a site called aduankonten.id. The public could report alleged problematic content they encounter on sites, social media, mobile apps, and software. Furthermore, in September 2021, MCIT released a site, instansi.aduankonten.id, to enable governmental agencies to be more effective and coordinated in reporting problematic content.[26] Upon receiving requests and recommendations from the public and state agencies, MCIT together with other governmental bodies will verify the reports. MCIT will file content moderation requests to social media platforms or block the sites that are verified to infringe laws and regulations in Indonesia.

During election periods, MCIT works particularly closely with the Elections Supervisory Agency (Bawaslu) and the General Elections Commission (KPU) to govern content infringements related to elections. They signed the Memorandum of Action (MoA) in 2018, 2019, and 2020 to combat disinformation and hate speech. The agreements address the responsibilities and roles of each institution and also the coordination and exchange of data and information on online content among the three entities to tackle problematic content related to elections. The agreements also stated the need to increase the capacity of each institution to monitor online content and to empower the public to use the Internet wisely. Following the signing of the MoA, each year, social media providers in the country make joint declarations of support for the government's efforts to eradicate hoaxes and hate speech.

Bawaslu became involved in social media monitoring after the signing in 2018. Bawaslu operates in collaboration with MCIT and KPU to proactively monitor online content related to elections. Bawaslu also opened a helpline for the public to raise their concerns via email and WhatsApp over alleged problematic online content related to elections. Additionally, Bawaslu can directly request platforms to take down problematic content or accounts and also ask the police to enforce any online content with election offences.

In particular, Bawaslu holds bilateral partnerships with Facebook and Google to counter hoaxes. With Facebook, Bawaslu has developed a special line of communication to flag infringements to the platform. Facebook and Bawaslu also conducted a series of digital literacy training activities for election supervisory agencies at the provincial and regional levels. Furthermore, Bawaslu and Facebook also organised a roundtable discussion with KPU and MCIT to agree on shared perceptions in handling 'negative content' on social media in the 2019's election.

Google also trained Bawaslu's officials to do reporting on problematic content on YouTube. Bawaslu's YouTube account is synchronised for reporting directly to Google. Bawaslu and Google set up public service announcements to campaign against disinformation and hate speech in the YouTube ad. In addition, Bawaslu and KPU together with Google and some civil society organisations developed a programme called 'Pintar Memilih' (clever to choose). They developed a website (pintarmemilih.id) that contains information related to elections and delivered digital literacy campaigns to eight campuses in Indonesia.

Non-governmental organisations and actors also play an active role as content flaggers dealing directly with platforms. As there are trust-related programmes and partnerships between social media platforms and Indonesian civil society groups (the safety trusted partner initiative of Facebook for online safety issues, YouTube's Trusted Flagger, Twitter's Trust and Safety Council, and Tik Tok's Child Safety Partner), the latter have special lines of communication to flag any problematic and urgent content to the social media companies.

The involved civil society organisations are mostly those who work on digital rights in the country, namely ICT Watch, SAFEnet, and Mafindo. There are also civil society organisations who work with groups impacted by content moderation issues that were invited to join social media trust-related initiatives, such as LBH APIK, the Wahid Institute, Yayasan Cinta Anak Bangsa (YCAB) Foundation, and ECPAT Indonesia (see the Analysis of stakeholders for further explanation on the social media affiliations of these trusted partners).

Although a Meta representative claimed in the public discussion conducted by SAFEnet that they have 12 trusted partners in Indonesia, the researcher was only able to identify and reach out to some, since the list of trusted partners is not publicly available. This confirms the findings of academic research with regard to the transparency and inclusiveness issues of trust-related partners of social media.

Aside from the content flagging requests made by each stakeholder group, Indonesia also has a multi-stakeholder WhatsApp group to manage Covid-19 related disinformation. This group consists of governmental actors, representatives from social media platforms, civil society groups, and health communities. Whenever there is any substantial disinformation regarding Covid-19 in Indonesia, this group will discuss and propose any necessary actions to manage that content. One of the group's members conveyed that the group was able to manage individual cases of Covid-19 hoaxes. This is a loose initiative and more should be done, for example, to have clear rules and governance structures, to enable the group to tackle hoaxes in Indonesia more massively while also ensuring compliance with international standards on freedom of expression.

Individual users in Indonesia also actively file reports whenever they find problematic content to platforms. However, some incidents show that the reporting mechanism can be misused by some parties, both pro- or anti-government volunteers or paid influencers or entities, to silence others who hold different views.

### Access for individual users to an internal complaint mechanism

Most of the interviewees agreed on the need for platforms to provide simple and interactive complaint mechanisms to protect the freedom of expression and safety of individual users at large and, more importantly, of the vulnerable groups.

As previously explained, even an established public body such as the National Commission on Violence Against Women (Komnas Perempuan) found that it took a lot of effort and time to understand the process and be able to file a complaint.

Ellen Kusuma of SAFEnet mentioned in the interview that users who experienced online gender-based violence struggled to navigate the internal complaint mechanisms of platforms. Leading civil society organisations working on digital rights play a key role in empowering and directing victims to report their cases to platforms.

Platforms should do more to provide accessible complaint mechanisms and they should provide more educational materials for users to understand how to use these complaint mechanisms.

### Facebook’s Oversight Board

Some leading civil society actors during the interviews stated that they are aware of the existence of Facebook’s Oversight Board. They are hopeful because the Board has the mandate to review Facebook’s decisions. However, they asserted that they have not filed any cases with the Board.

Meanwhile, the first transparency report from the Board shows that less than 8% of submissions to the Oversight Board came from the Asia-Pacific region. This small number may reflect the level of awareness of users in the region and, particularly in Indonesia, on their rights to freedom of expression and online safety in relation to the responsibilities of platforms to ensure those rights through appropriate content moderation mechanisms.

Furthermore, there is a widespread perception in the country that platforms have the legitimacy, authority, and right to govern their spaces with their own rules. A study, for example, captures how the public sector in Indonesia relies heavily on social media to spread information to the public, and therefore they must follow the rules of social media. Meanwhile, users, especially those with a relatively large number of followers, show an attitude of compliance with the rules of platforms because they do not want their accounts to be suspended. So, whenever there are users impacted by the content moderation decisions of platforms, there are only a few that know they can appeal to those platforms, not to mention appeal to a more high-level entity like the Facebook Oversight Board.

### Statistics and data on content moderation

There are several studies that highlight concerns and challenges with regard to content moderation practices in Indonesia. For example, the study from the think tank Center for Indonesian Policy Studies (CIPS) examines the impact of Indonesia’s content moderation regulations on freedom of expression. Moreover, the study by IPAC noted the challenges in governing and managing online terrorism and radicalism content. Civil society organisations, such as ELSAM and SAFEnet, have issued joint statements criticising the mass flaggings and content moderation decisions of platforms that stifle legitimate content advocating for corruption eradication efforts in Indonesia.

However, there is limited to no data or research by Indonesian or international actors that monitor online content moderation practices in Indonesia, except for the transparency reports released by social media platforms (see the Recommendations: this report proposes a research capacity-building programme for potential members of the coalition to be able to monitor content on social media).

### Level of information in the transparency reports

Contrary to the recommendations from the international civil society,[27] transparency reports from social media platforms have so far not been able to reassure users that platforms include human rights and local context considerations in their content moderation decisions and practices.

Transparency reports typically provide large amounts of information on community standards enforcement, government requests for content restrictions, intellectual property removal requests, and government requests for user data. However, the policy brief produced by UNESCO to promote transparency and accountability in the digital age shows that transparency reports produced by Internet companies contain significant gaps and cover different issues in different ways.

This report found at least two gaps. First, transparency reports only portray the end results of content moderation decisions but do not show how platforms conduct content moderation processes, especially regarding the relationship between 'grey-area' content and the local context. That way, users cannot be assured that platforms have carefully undertaken an assessment of human rights implications in the local context. How do platforms incorporate an understanding of a local context in their AI-based content moderation systems, the intervention of human content moderators, and their decision-making processes? With whom do platforms consult in order to understand the local context before making content moderation decisions?

Second, some civil society actors interviewed for this project also urged platforms to provide transparency on campaign advertising, in terms of disclosing information about what kind of advertisements appear and are distributed to which users, who pays for them (advertiser's name, phone number, email, website, address), and also the commitment from platforms not to display paid campaign advertisements that do not contain disclaimer information.

In responding to the various expectations and needs of wider users that have not been reflected in the current state of transparency reports, a representative of digital platforms in Indonesia highlighted that transparency reports are useful for certain ends and that other tools, particularly sustained dialogue with local trusted partners and multi-stakeholders, can be better suited to achieve other objectives. In their view, due to their quantitative nature, transparency reports are best to gauge the volume and characteristics of the content removal requests and, in their current shape, not the broader contextual challenges surrounding the policy enforcement. They considered that having a sustained dialogue with a locally-relevant multi-stakeholder group could complement the efforts in providing transparency to the public.

### Humans or machines: Who is in charge of content moderation?

All of the civil society actors interviewed in this research stated that they have only been in contact with representatives of social media companies in the country and or in the Asia-Pacific region. As described in previous sections, they have been corresponding with representatives whenever there are urgent content moderation requests that need immediate support and responses from platforms.

The Executive Director of SAFEnet, Damar Juniarto, noted that there are three companies outsourced to undertake the content moderation review processes globally. He conveyed that there should be more engagement between Indonesian civil society groups and the human content moderators of the country so that they could inform and empower the moderators with information on the local context behind any emerging problematic content.

A similar conclusion is in relation to automated content moderation. As research shows, the opacity of an AI-based content moderation system creates **Error! Hyperlink reference not valid.**. This report finds there is room for more participation of Indonesian civil society groups in the development of automated processes, notably in relation to the assessment of the impact of automated mechanisms of content moderation.

All of the interviewees stated that they have never been informed or involved in the development process of an algorithmic-based content moderation system of platforms. The impact of such systems upon Indonesia is generally unknown.

This point was apparent in an interview with the representative of a local trusted partner of a social media platform. They stated that there was a period of time when their team was burdened with the need to deal repeatedly with the same problematic content being circulated through different channels. The related platform then accommodated their complaints through developments of its automated content moderation mechanisms, but the platform did not consult civil society and never revealed the content being moderated by such mechanism. The Indonesian civil society groups were thus not able to assess the impact of such mechanisms on online content. They could only observe that content from mainstream media outlets was prioritised in the following days after the changes in the automated system. Meanwhile, content from less credible channels appeared to be less visible.

## Interim conclusion

This research has shown that in the midst of the diverse cultural context of Indonesia, growing misuse of social media, and the complexity of ‘grey-area’ content in the country, there are several main flaws in the current practices of content moderation.

To address these issues, ARTICLE 19 and other civil society organisations have developed recommendations based on international standards on human rights.[28] Of particular importance for this study is the principle of Culture Competence, set forth in the Santa Clara Principles, which:

> *"…requires, among other things, that those making moderation and appeal decisions understand the language, culture, and political and social context of the posts they are moderating. Companies should ensure that their rules and policies, and their enforcement, take into consideration the diversity of cultures and contexts in which their platforms and services are available and used (…), and companies should ensure that reports, notices, and appeals processes are available in the language in which the user interacts with the service, and that users are not disadvantaged during content moderation processes on the basis of language, country, or region."*

In addition, flaws in content moderation identified in this study should be addressed through the following recommendations:

- Companies should ensure that their content rules are sufficiently clear, accessible, and in line with international standards on freedom of expression and privacy. It is of key importance that social media companies' content rules be made accessible and available in local languages.

- Companies should also provide more detailed examples or case studies of the way in which their community standards are applied in practice and conduct reviews of their standards to ensure human rights compliance.

- Companies should be more transparent about their decision-making processes, including the tools they use to moderate content, such as algorithms and Trusted Flagger schemes.

- Companies should ensure that sanctions for non-compliance with their Terms of Service are proportionate.

- Companies should put in place internal complaint mechanisms, including for the wrongful removal of content or other restrictions on their users' freedom of expression. In particular, individuals should be given detailed notice of a complaint and the opportunity to respond prior to content removal. Internal appeal mechanisms should be clear and easy to find on company websites.

- Companies should publish comprehensive transparency reports, including detailed information about content removal requests received and actioned on the basis of their Terms of Service. Additional information should also be provided in relation to appeals processes, including the number of appeals received and their outcome.

- Companies should collaborate with other stakeholders to develop new independent self-regulatory mechanisms, such as a Social Media Council, modelled on effective self-regulation archetypes in the journalism field.

# Analysis of stakeholders

This chapter describes the relevant stakeholder groups in Indonesia that work at the intersection of online freedom of expression and content moderation. They include civil society groups, media industries, journalists, content creators, social media companies, public authorities, academics, and think tanks. While it would require further research to comprehensively map the full list of stakeholders, this section aims at providing a general overview of the various sizes, capacities, and needs of the initial potential coalition members, as necessary to support the report's recommendations.

Overall, one of the key findings from this research is that all of the identified stakeholder groups and actors would benefit from support in terms of training and information sharing and also access to international networks. Such support would contribute to empowering these stakeholders, notably through an update of their expertise on freedom of expression and content moderation.

## Civil society organisations

As an initial remark, it is important to keep in mind that Indonesian stakeholders have generally been closely engaged with social media platforms. Social media platforms funded a national digital literacy initiative called Siberkreasi, research activities, fact-checking initiatives, and educational activities of both civil society organisations and the government. However, Indonesian civil society actors explained in the interviews that funding from social media platforms is not their only source of income. Most of the interviewees also claimed that although they are funded by platforms, they are still able to maintain a critical distance from the platforms and express their criticism. The critical views delivered by the interviewees in this report, particularly those who are the trusted partners of platforms, support this position.

Indonesia has a large number of civil society actors and coalitions spanning from those who work at the national level to groups operating at the level of villages. Research from 2011 by Nugroho has shown that there are more than 250 civil society organisations and groups actively using the Internet and social media. These numbers would most likely have increased in 2022.

There are civil society organisations focused on digital and human rights issues and those who work with groups that are impacted by content moderation issues. The first group is skilled in digital rights and issues, and they also have a lot of experience engaging with state actors and social media representatives (in this study, we call this group the leading civil society organisation actors). On the other hand, the actors in the second group are best positioned to provide expertise on the local contexts and understanding the impact of content moderation on population (for the purposes of this study, we call these organisations peripheral).

In the first category, there are leading civil society organisations that work on digital rights, who are the official trusted partners of social media platforms, such as ICT Watch, SAFEnet, and Mafindo. ICT Watch is a Trusted Flagger for YouTube[29] and part of Twitter's Trust and Safety Council. SAFEnet is a trusted partner of Facebook and also joined the Twitter's Trust and Safety Council. Mafindo is a Trusted Flagger for YouTube and third-party fact-checker for Facebook.

Based in Jakarta and Bali, **ICT Watch** is a civil society organisation that focuses on the issues of digital literacy skills, online expression, and cyber governance, and works on those issues in collaboration with other stakeholders. Since 2002, ICT Watch has been active in various initiatives to promote these issues in the country. Together with other national stakeholders, it initiated the Indonesia Internet Governance Forum (ID-IGF) in 2012 and co-hosted the UN Internet Governance Forum in 2013 in Bali. It has participated in the initiation of the Digital Literacy National Movement (Siberkreasi). ICT Watch has received various awards, for example from the UN World Summit on the Information Society in 2016 and 2017 for its Internet Sehat programme. It also took part in the formation of SAFEnet.

Established in June 2013, **SAFEnet** is a network of digital rights defenders in Southeast Asia. It focuses on three digital rights as its working areas, namely the right to access, the right to expression, and the right to safety. Its members are spread in 19 cities in

Indonesia. It monitors cases related to digital rights violations in their respective area, conducts digital rights campaigns, and delivers digital rights capacity-building for wider civil society actors in Indonesia. It partners with regional and international civil society organisations and also Internet-related companies and associations.

**Mafindo** is an independent fact-checking organisation. It started as a grassroots community in 2013 and since then has been growing until being formalised into a registered civil society organisation in 2016. Besides having full-time staff, Mafindo also has more than 95,000 members and over 1,000 individual fact-checking volunteers from across all regions in Indonesia.

There are also leading civil society groups who focus on digital issues but are not trusted partners of any social media platforms. They are ELSAM, Tifa Foundation, Human Rights Watch–Indonesia, and Engage Media–Indonesia.

**ELSAM** is a human rights organisation based in Jakarta. It focuses on the topics of human rights and technology, business and human rights, fundamental freedoms, transitional justice, eco-social justice, and human rights education. It conducts research, advocacy, and training to mainstream the protection of human rights in Indonesia.

Established in 2000, **Tifa Foundation** is a civil society organisation that promotes the embodiment of open society in the areas of natural resource governance, human rights, democracy and social movement, and transparency and accountability in digital data ecosystem. It has been serving as a grant-making organisation and aims to be a civil society hub in Indonesia.

**Human Rights Watch** is an international civil society organisation that investigates and reports on human rights abuses in various parts of the world. Its local researcher, Andreas Harsono, has covered Indonesia for Human Rights Watch since 2008.

**Engage Media** is an Asia-Pacific non-profit that promotes digital rights, open and secure technology, and social issue documentaries. It has offices in Australia and Indonesia and several staff in Bangkok, Kuala Lumpur, and Manila.

In the second category (civil society groups who work with groups that are impacted by content moderation issues), this research identified that only some of the groups are aware of the issues of content moderation. They may be invited to join social media trust-related initiatives, which means they are prioritised in reporting problematic videos to the related platforms. For example, LBH Apik is the trusted partner of Facebook. Moreover, the Wahid Institute and YCAB Foundation are also in the Twitter's Trust and Safety Council. ECPAT Indonesia is the partner of several social media safety initiatives: Twitter Trust and Safety Council, Facebook Safety Trusted Partner, TikTok Child Safety Partner, YouTube Trusted Flagger, and Internet Watch Foundation.

Since 1995, **LBH Apik** has worked to promote the realisation of a gender-just legal system and strengthen the women's movement in empowering gender-just laws. It has been providing free legal aid to women and children and assisting cyber-crime victims. It has 16 offices spread in various cities in Indonesia.

Founded in 2004 by the former President of Indonesia Abdurrahman Wahid, the **Wahid Institute** is a Jakarta-based civil society organisation that advances the development of progressive Muslims to promote the creation of democracy, multiculturalism, and tolerance throughout Indonesia and the world. The organisation is currently led by his daughter, Yenny Wahid.

The **YCAB Foundation** is a non-profit social foundation founded in 1999 in Jakarta. Its mission is to improve welfare through education and inclusive financing by providing education and economic empowerment for underprivileged youth and mothers.

**ECPAT Indonesia** is a national network of 22 organisations and two individual members from 11 provinces in Indonesia. Since 2000, it has been tackling child prostitution, porn, and trafficking for sexual purpose in Indonesia. In 2005, it collaborated with ECPAT International, an international civil society organisation operating in 98 countries, on the issue of child sexual exploitation. In 2012, ECPAT Indonesia became an official member of ECPAT International.

There are also other civil society groups whose focus intersects with online problematic content and who showed an interest on the topic and being involved in the coalition. This includes general human rights organisations or organisations focused on LGBTIQ+ rights, terrorism and radicalism, religious issues, and elections. They are Arus Pelangi; Amnesty International–Indonesia; Warga Muda\Young Citizens; Paparisa Ambon Bergerak; the Institute for International Peace Building\Yayasan Prasasti Perdamaian; Democracy and Electoral Empowerment Partnership (DEEP) Indonesia; Komite Independen Sadar Pemilu (KISP)\Independent Committee Aware of Elections; and Perludem.

**Arus Pelangi** is an organisation that encourages the realisation of a society that respects the rights of LGBTIQ+ people as human rights.

**Amnesty International** is an international civil society movement with more than 10 million people in various countries campaigning to end human rights abuses. It has a national section in Indonesia with around 60 staff working on various human right issues, including freedom of expression online.

**Warga Muda** is a youth collaborative network across various ethnic groups, religions, races, classes, and professions spread across 34 provinces in Indonesia. It aims to create a structurally and culturally friendly ecosystem for the participation and representation of young people in the public and private sectors at both the regional and national level.

Started in 2010, **Paparisa Ambon Bergerak** is an initiative in the form of a basecamp for various creative communities across racial and religious backgrounds in Ambon, Maluku, in the eastern part of Indonesia, to gather and share ideas that promote the development of Ambon. It has received widespread praise for its contributions to the peacebuilding movement in Ambon through the use of digital and social media to counter false information.

The **Institute for International Peace Building** is a non-governmental organisation established in January 2008 that works to develop an integrated policy and national strategy to reduce the level of threats from violent groups through dialogue. It focuses on developing and deepening understanding of peace and conflict, political violence, terrorism and other transnational crimes.

There are at least three election watchdogs in different parts of Indonesia who showed interest in the development of a local content moderation coalition. Together with eight civil society organisations they developed the Coalition for Social Media Ethics in Indonesia, an initiative for producing ethical recommendations to guide the roles of election candidates\parties\campaign teams, social media platforms, civil society organisations, and media to ensure responsible election campaigns on social media.

The first election watchdog, **DEEP Indonesia**, is based in Depok, West Java. It works to strengthen the principal values of democracy for the realisation of good quality elections. It has 20 representatives across various districts in West Java. Located in Yogyakarta, Central Java, **KISP** is a youth association that promotes democratic values to the public and monitors election-related issues. It has been operating since 2018. Lastly, **Perludem** is an election watchdog based in Jakarta and founded in 2005. It carries out research, advocacy, monitoring, education, and training in the field of elections and democracy for policymakers, organisers, participants, and voters. Perludem is currently preparing a roadmap to secure a peaceful election in the country in 2024.

Most of the local civil society organisations who were interviewed were aware of the impact of content moderation issues at the local level, but they were not aware of any means to engage with platforms and seek solutions to the problematic content they are faced with. The interviews with Moch Edward Trias of Yogyakarta-based election watchdog KISP, and with Pierre Ajawaila of Paparisa Ambon Bergerak, provided the opportunity to discuss the situation at the local level. Both of them shared a similar perception that civil society actors have no bargaining power and that, as a consequence, platforms will, for instance, take a long time to respond to their urgent requests. They also conveyed that civil society organisations and actors that are not directly dealing with platforms on content moderation issues may have less awareness and understanding of the issues, as well as less resources to work on content moderation issues. Those organisations and actors might not have enough resources in terms of staff and capacity to constantly monitor online content. They therefore often seek help from the key civil society actors active on content moderation issues whenever they encounter content that requires moderation by platforms.

## Media industries and journalists

There are at least six associations in the media industry and five journalists' associations[30] in Indonesia. The Alliance of Independent Journalists (AJI) and the Indonesian Cyber Media Association (AMSI) are the two leading media industry and journalists' associations in the field of fact-checking. **AMSI** was initiated by 26 Indonesian online media leaders in April 2017 with a determination to strengthen the professionalism, trustworthiness, and independency of online media in Indonesia. **AJI** is a journalists' association that promotes press freedom in Indonesia. It was founded in 1994 as a resistance from the Indonesian press community against the arbitrariness of the regime. Together with Mafindo and Google, AJI and AMSI supported the formation of a collaborative fact-checking project among several online media companies, named **cekfakta**. The proposed coalition could engage with AMSI and AJI in its initial formation as they have experience dealing with false information, and later seek the interests of other media and journalists' associations to join the coalition.

## Academia and think tanks

There are some think tanks working on content moderation, hate speech, and violent extremism content, such as **CIPS**, **CSIS**, **IPAC**, and the **Indonesian Institute**.

**CIPS** is an independent, non-profit, and non-partisan think tank that promotes social and economic reforms through research and policy engagement on the topics of food security and agriculture, community livelihoods, and education. It has published research on digital-related issues, such as the impact of content moderation regulation on freedom of expression in Indonesia, the responsibility of digital platforms in Indonesia over online content, the rights of digital consumers, and personal data protection.

Established in 1971, **CSIS** is an independent and non-profit organisation that focuses on policy-oriented studies on domestic and international issues. It delivers research, dialogue, and public debate in the areas of disaster management, economics, international relations, and also political and social change. It has produced research reports and policy briefs on digital issues, including online hate speech, digital literacy, data governance, and the digital economy, to name just a few.

Founded in 2013, **IPAC** studies the dynamics and prevention of six types of conflict in Indonesia: communal, land and resource, electoral, vigilante, extremist and insurgent, and various forms of dispute.

The **Indonesian Institute** is a centre for public policy research on the topics of economic, social, politics, and law. Since 2004, it has been operating for the realisation of public policies that uphold human rights, law enforcement, the participation of various stakeholders, and the application of democratic governance principles.

These think tanks could assist the proposed coalition with research and monitoring on content moderation and hate speech trends online. They could also empower the research capacity of the coalition members to identify any patterns of online problematic content at the local level. The coalition could also involve Drone Emprit, a big data technology developed by Ismail Fahmi to monitor and analyse social media traffic.

Furthermore, there are also at least three leading universities that focus on the issues of human rights and digital technologies. They are the Center for Human Rights Law Studies of Airlangga University, Surabaya, East Java; the Center for Cyber Law and Digital Transformation of Padjajaran University, Bandung, West Java; and the Center for Digital Society of Gadjah Mada University, Yogyakarta, East Java.

Formed in 2009, the **Center for Human Rights Law Studies** is a unit under the Faculty of Law, Airlangga University, which conducts research and assessment activities in the fields of business and human rights, natural resources rights, minority rights and religious freedom, freedom of expression, academic freedom, and local government and human rights.

The **Center for Cyber Law and Digital Transformation** is a research centre under the Faculty of Law, Padjajaran University. It conducts research, education, advocacy, and socialisation relating to cyber law.

The **Center for Digital Society** is a research centre under the Faculty of Social and Political Sciences, Gadjah Mada University. It conducts research, publication, education, and policy advocacy related to the issues of digital society in Indonesia.

## Content creators and users

As illustrated, for instance, by a YouTube video discussing the blocking of TikTok accounts, it is fairly evident that users generally do not know how to appeal content moderation decisions or do not have the power to negotiate with platforms: a lot of the comments under the video aired confusion and frustration as users felt that their TikTok accounts were blocked without clear explanations.[31]

This study identified that there are at least two emerging content creators' associations in the country that could be a potential member of the proposed coalition. **Indonesian YouTuber Association** (AYI) is an organisation for all content creators\YouTubers in Indonesia. As of February 2022, it had 759 members. **Indonesian Santri YouTuber Association** (AYSI) is a creative santri community initiated in June 2021 to develop digital da'wah (preaching of Islam) through YouTube channels and other social media.[32] Their focus so far is on how to increase their capacity as content creators. However, this research suggests that there can be an opportunity to involve those associations to introduce them to content moderation issues and to empower them to join the debates.[33]

## Social media platforms

While this study is focused on content moderation of the five largest social media platforms in the country, the proposed coalition could also consider engaging with other social media platforms and chat apps available in the country (see The state of content moderation in Indonesia). The participation of these players may provide insights on the various challenges, capacities, and resources they face in conducting content moderation in Indonesia. An engagement with chat app companies would also shed some light on the intricacies of conducting content moderation on personal chat groups. The companies could also learn from the needs and challenges of civil society in conducting content moderation at the local level.

## Public authorities

There are several key state organisations that deal with content moderation in Indonesia, namely MCIT, National Cyber and Crypto Agency (BSSN), cyber police together with the ministerial and state organisations that intersect with the content being moderated (for example, health, medicine, finance, etc).

Furthermore, there are also independent public authorities that are involved in content moderation issues, such as the National Commission on Human Rights (Komnas HAM), National Commission on Violence Against Women (Komnas Perempuan), and election providers (Bawaslu and KPU).

BSSN is a government agency that handles information and cyber security issues, namely the security of Internet protocol-based telecommunication networks and infrastructures. Komnas HAM is an independent national human rights institution of Indonesia carrying out studies, research, counselling, monitoring, and mediation related to human rights issues in Indonesia. Komnas Perempuan is an independent state institution for the enforcement of human rights of Indonesian women. Bawaslu is the election supervisory board in Indonesia. KPU is the body that organises elections in Indonesia.

# Conclusion

This research has shown that social media companies have not conducted their content moderation practices on the basis of an informed understanding of the diversity of Indonesian society and the rich history and context of the country. Indonesian users and civil society groups have not been given enough information, space, or power to contribute to the decision-making processes of content moderation. This is especially important in relation to the ‘grey-area’ content, that is, messages that might not, per se, fit within the categories of prohibited content under the global community standards of platforms, but nonetheless their massive amplification could lead to real-world violence. Indeed, the recent history of Indonesia suggests that such ‘grey-area’ content has contributed to further polarisation and ultimately to violence in the country.

Because of the lack of a transparent and sustainable dialogue between civil society groups and social media companies, there is no shared understanding of the consequences that certain content can have in the context of Indonesia and of what the appropriate content moderation measures should be.

The dynamics of content moderation practices in Indonesia show that the enforcement processes of community standards should be informed by a solid understanding of both international standards of freedom of expression and the local context.

Most local civil society organisations interviewed for this research see favourably the idea of forming a civil society Coalition on Freedom of Expression and Content Moderation in Indonesia. To guarantee the effective ownership of the coalition by its members, the process facilitating its creation would have to start with a validation exercise that ensures potential participants have the opportunity to discuss the findings of the research. The coalition would act as a bridge to develop and nurture relations and dialogue with social media companies, certain state actors, and even international organisations on content moderation and freedom of expression issues. This coalition should consist of various civil society groups that can provide a balanced judgement about protecting freedom of expression and the safety of individuals and public in the specific context of Indonesia.

# Recommendations

This research has shown that in the midst of the complex and diverse cultural context of Indonesia, growing use and misuse of social media in the country, and the complexity of 'grey-area' problematic content in the country, there has been a lack of meaningful and continuous dialogue between platforms and leading and peripheral civil society groups.

Civil society groups and lay users have been battling individually, instead of coordinating, against the content moderation decisions of platforms. Most of them do not know how to appeal against the platform's decisions. Meanwhile, the leading civil society groups in their capacity as the official partners of platforms have often felt powerless in the negotiation process with platforms. Platforms usually hold the final decision-making power, while not displaying sufficient understanding of the complexity of the local context. Accordingly, there have been cases of over and under content moderation in the country, that either hurt freedom of expression or the safety of individuals and public.

When we submitted the idea of a local Coalition on Freedom of Expression and Content Moderation to the interviewees, most of them responded positively.

To be clear, there is already a number of multi-stakeholder groups and civil society alliances working on issues of Internet governance, freedom of expression, and social media ethics in the country, but only few have shown interest, resources, and commitment to develop work on the issue of the contribution of local actors to content moderation on social media.

The multi-stakeholder national committee on the ID-IGF is a yearly Internet-related dialogue forum in Indonesia inspired by the UN IGF. ID-IGF has a multi-stakeholder committee who is responsible for preparing the conduct of the forum every year. While the members of ID-IGF work on various digital-related issues ranging from infrastructure, law, and economics, to social issues related to the Internet, very few of them pay specific attention to the issue of content moderation.

Moreover, civil society coalitions on freedom of expression in the country typically focus on advocating for government action to protect freedom of expression. The closest one to the topic of this research is the Coalition for Social Media Ethics in Indonesia (see the Analysis of stakeholders, particularly civil society groups), but the interview process has shown that not all organisations in this group are interested in the topic of content moderation: some members of this coalition preferred to address disinformation on social media through digital literacy and the creation of counter narratives.

This situation suggests the creation of a new coalition, but one that includes civil society groups, academia, think tanks, and associations of media industries, journalists, and users from the existing coalitions that show interest and commitment to focus on the topic of content moderation. This **Coalition on Freedom of Expression and Content Moderation** could start by working with some key actors and organisations from the existing coalitions while leaving the door open for the other members of those coalitions to join the new coalition.

The coalition should be designed to **unite and strengthen the various civil society organisations and actors** who have been working on, or are concerned with, content moderation issues. In addition, the coalition should aim at **engaging in a continuous and meaningful dialogue with the more powerful stakeholder groups, state actors, and social media companies**.

In relation to civil society actors who have worked closely with social media platforms, there is a clear power imbalance between them and the tech companies. Furthermore, most of the civil society actors interviewed in this research also conveyed their concerns in relation to the participation of state actors and the potential abuse of their authority to pursue their political interests in the coalition. Some expressed a clear preference for the coalition not to include state actors, to ensure that civil society groups can more freely express their views in the coalition. They also observed that the bureaucratic processes of governmental institutions would risk slowing down the working tempo of the coalition.

While there is a power imbalance between civil society groups and other more powerful actors, there is also a culture of dialogue among state and non-state stakeholder groups in the country. This can be seen in the existence of multi-stakeholder groups such as ID-IGF and Siberkreasi (the national multi-stakeholder digital literacy movement in Indonesia), in which state and non-state actors work together. Additionally, some civil society interviewees also mentioned the need to involve governmental actors to increase and nurture their relationships with the government, as well as to increase the legitimacy of the coalition and its capacity to be heard by platforms. Some of the informants perceived that it is beneficial to keep governmental actors informed on the projects of the coalition and on platform governance debates at large. This would contribute to encouraging state actors to gain a better understanding of the need to formulate content moderation and digital platform related regulations that uphold international human rights standards.

The above findings and reasons have led this research to consider the importance for civil society groups to build a critical but cooperative relationship with the more powerful stakeholder groups in order to nurture a culture of dialogue, but without being co-opted or captured by those stakeholder groups. A coalition that consists only of civil society groups but engaged in close coordination with state actors and social media companies appears to be the most appropriate approach.

As explained above, the coalition would seek to act as a bridge to develop sustainable relations between various civil society groups with social media platforms and possibly governmental actors. This would enable the coalition to play a role in conducting dialogues to ensure that content moderation decisions will comply with international standards on human rights while taking into consideration the multi-dimensional complexity of the local context. The coalition could also seek to help users impacted by the content moderation decisions of platforms.

In its approach to public actors, the coalition could first explore the development of relationships with independent public authorities, such as Bawaslu, KPU, National Commission on Human Rights (Komnas HAM), National Commission on Violence Against Women (Komnas Perempuan), or from the ministries or state organisations that intersect with the issue of content moderation (such as MCIT and BSSN). The coalition should seek to stabilise its relationship with those public authorities to ensure the continuity of institutional cooperation even when progressive individuals leave their posts.

The interviewees also mentioned that the role of the coalition could be expanded to:

- Cooperating with existing coalitions and networks to **create a more holistic approach and ecosystem of a healthy digital sphere in Indonesia**. For example, the proposed coalition could work together with digital literacy groups and initiatives in the country;
- Contributing to **more responsible social media campaigns for the upcoming General Election in 2024**;
- Encouraging the **adoption of laws and regulations related to content moderation and social media** in Indonesia that comply with international standards on freedom of expression and other fundamental rights;
- **Developing relations between Indonesian users and the Facebook Oversight Board** and its members to bring cases related to Indonesia to the Board;
- **Strengthening exchanges and cooperation** between Indonesian stakeholder groups and influential international organisations and key actors working on platform governance and content moderation related issues;
- In the longer term, the development of the coalition could lead to the **creation of a Social Media Council** that would promote the development and enforcement of community standards of social media platforms and uphold international standards on freedom of expression while duly taking into consideration the voices of local actors and the local context in Indonesia.

### Sustainability of a civil society coalition

To enable the coalition to fulfil all or some of the above-mentioned potential roles, the prerequisite is to develop and strengthen the internal structure, knowledge, and coordination capacity of civil society groups to increase their deliberation and bargaining capacity so they can coordinate their efforts and gain effective bargaining power to deal with platforms and state actors.

The credibility of the civil society coalition among all stakeholder groups is key for its success. As a kickstart, this research suggests bringing together all of the civil society groups and actors identified in this research. ARTICLE 19 together with UNESCO and the researcher will submit the final results of the study to a validation exercise by the stakeholders who, by their participation and contribution, will reconfirm their interests and commitment to join the coalition.

The members of the future coalition could then appoint a leader. Based on the analysis of stakeholders, SAFEnet, Tifa Foundation, and Perludem appear as having the potential to lead the coalition. They are all well-regarded local civil society organisations that have a good track record of engaging with platforms and state actors. SAFEnet has been acting as a civil society hub and engaging with issues of online freedom of expression with various lay users, national and international partners. Tifa is progressing with its work on intermediary liability and data governance. Of particular note, the Executive Director of Tifa, Shita Laksmi, was an expert staff at MCIT. Perludem has been working on the issues of election-related disinformation together with election providers and social media platforms, and also initiated the Coalition for Social Media Ethics in Indonesia.

The coalition would also need constitutive documents, such as a Memorandum of Understanding (MoU), charter, governance structure, and workplan, to be drafted and adopted in a participative and transparent manner.

The coalition will need to have the capacity for internal coordination to manage the diversity of actors, goals, and strategies. Along the way, the coalition should increase its inclusiveness by involving more civil society actors in order to be perceived as legitimate and effective. In other words, the internal legitimacy (inclusiveness and representativeness) of the coalition will result in the perceived external legitimacy and efficacy of this coalition to the intended audience. In order to manage the potential tension between inclusiveness and efficacy of the coalition, the application of a new civil society organisation should be supported by two endorsements from existing coalition members.

This is a proposed selection mechanism to ensure that future members have critical and cooperative engagement with current members. Some interviewees mentioned that they may decide to join the coalition after seeing the composition and reputation of the coalition members. Some stated they would be reluctant to join and associate themselves with the coalition if there are any civil society groups that prefer a more confrontative strategy.

Subsequently, the capacity of the members of the coalition needs to be reinforced in terms of knowledge on content moderation systems, governance of platforms, international standards on freedom of expression, as well as the capacity to research the needs and challenges of conducting content moderation in the various regions in Indonesia. As this research has found, there is still limited awareness of content moderation issues in Indonesia. While the leading civil society groups have knowledge on digital rights and capacity to research the trend of problematic online content at the national level, they may not have the expertise on the local context of content moderation issues. Meanwhile, civil society organisations at the provincial level are familiar with the local patterns of amplification of problematic speeches, but they may not have the capacity to turn their tacit experiences into explicit knowledge that could reinforce their position in engagement with platforms on content moderation.

The coalition and its constitutive process should look at developing the capacity of civil society organisations that do not work directly on digital rights to monitor the amplification trends of problematic content ('grey-area' content) in their local area, and to engage with platforms in order to contribute to the adoption of content moderation practices that respect freedom of expression. The coalition should also strive to link leading and peripheral civil society groups and support them to develop research on the progression of 'grey-area' content from legitimate (i.e. protected by freedom of expression) to illegitimate online content, as well as the design of content moderation practices that are compatible with international standards on freedom of expression. Coalition members would also benefit from training on strategic communication, engagement, and negotiation skills in presenting their research and joint positions, so that they can have a stronger position and a more meaningful dialogue with platforms and state actors.

## Creating a coalition

The following sequence of steps can be used to create a **Coalition on Freedom of Expression and Content Moderation**:

1. **Hold an inception meeting to conduct outreach and to build trust** among the potential members of the coalition:

    a. ARTICLE 19, UNESCO, and the research consultant to submit the findings of the research for validation by potential members of the coalition. By potential members of the coalition, we refer to all of the identified civil society groups and actors in this research, especially those who showed concerns on the topic of content moderation during the interviews.

    b. The potential members to introduce and share their concerns and works with regard to content moderation issues in Indonesia and to confirm their interest and commitment to join the coalition.

    c. The potential members to discuss the vision, mission, and goals (short, medium, and long-term), and the benefits they might gain from the coalition.

    d. The potential members to appoint a specific individual and organisation to be a focal point (a leader and secretariat) in order to get the coalition up and running. The appointed leader and secretariat will be in charge of the coordination, engagement, and the related administrative duties of the coalition (manage documentation of meetings, manage members database, establish a mailing list or other communication channels, prepare follow-up meetings, finance, etc).

    e. The inception meeting is to conclude the agenda and action plans for the next follow-up meeting. The leader and secretariat are responsible for the preparation of the next meeting, such as preparing drafts of an MoU, governance structure, coalition charter, and working plan of the coalition based on the discussion at this meeting and then circulate the drafts to the mailing list to seek input, which will be further discussed at the subsequent meeting.

f. The potential members may provide recommendations for other potential key members that have not been identified in the research. The secretariat shall reach out to those recommended potential members for the next meetings.

2. **Convene follow-up meetings to confirm commitments** from potential members and further discuss and conclude an MoU, charter, governance structure, and workplan of the coalition.

   a. The potential members to develop and discuss the MoU for civil society organisations or individuals joining the coalition. The MoU will provide the terms of reference of the coalition and may include the expected objectives and deliverables, the roles and responsibilities of members, the resource and financial plans (or how contributions to the coalition can be made by members), work breakdown structure and schedule, the codes of conduct, independence and conflicts of interests, procedures for joining and departing, and dispute resolution.

   b. The coalition charter may cover the names and branding of the coalition, purposes and key principles, and vision, mission, and goals of the coalition.

   c. The governance structure may include the structure of the coalition (the leader, steering committee, secretariat, working groups); decision-making processes and procedures; an dcriteria for leadership, membership, and appropriate rules.

   d. The workplan: the participants may discuss a workplan and a roadmap for the coalition, including a few key issues and objectives to be addressed by the coalition, and a financial plan. Additionally, they may consider discussing working processes, including communications and documentation, frequency, and location of meetings for their workplan.

   e. Those organisations or individuals who express an interest to join the coalition, could then agree on and sign the MoU and other constitutive documents.

3. **Announce the coalition**: invite platforms and (progressive) state actors in the formal launch of the coalition. Once a critical number of members have joined, the formation of the coalition can be made public through press releases (including on its own website

and social media handles) and goals communicated widely. This could generate more public interest and desire for more organisations to join the coalition.

4. After the formation process, the coalition is then ready to **undertake capacity-building and run their agreed working programmes** (such as developing joint research and engagement strategies), and then to monitor and evaluate the progress and impact of their works.

# Annex A: Risk analysis

The **Coalition on Freedom of Expression Online and Content Moderation** emerges as a unique opportunity for participation and contribution by all the actors and as a mechanism for meaningful change. The coalition offers a path to consensus on key content moderation issues – and opportunities to address them. The following table provides an overview of the potential risks related to the formation and functionality of the coalition, identified by the respondents, including potential ways to overcome and mitigate them.

| Risk type* | Description of risk | Likelihood** | Impact*** | Monitoring and mitigation |
|---|---|---|---|---|
| Finance | The funding and sustainability of the coalition in the future | Possible | **Major** | • The coalition members to agree on the roadmap of the coalition, including plans on funding sources |
| Political | Government adopts new restrictive laws and regulations | Possible | **Major** | • Joint funding applications and participation of coalition members in donor meetings and agenda-setting process<br>• The coalition could play a role in advocating against such bill |
| Reputation | The failure of the coalition would reflect badly on the reputation of ARTICLE 19 in the country | Unlikely | **Minor** | • The coalition to have appropriate membership selection mechanisms to ensure inclusiveness and effectiveness of the coalition (critical yet cooperative behaviour of members) |
| Safeguarding | Participants in the coalition could be harassed by social groups | Possible | **Moderate** | • Liaise with community leaders to prevent harassment<br>• The coalition to propose and advocate for legal protection for human rights defenders in Indonesia |
| Stakeholder | Some civil society organisations may lack knowledge of freedom of expression standards | Likely | **Moderate** | • Provide training |

|  |  |  |  |  |
|---|---|---|---|---|
| Stakeholder | Civil society actors are already very busy and spread thin over multiple projects | Likely | **Severe** | • Ensure that the coalition finds a balance between being a light-weight approach and effectiveness |
| Stakeholder | Civil society organisations may become tired in the journey and lack commitment and interest to join the coalition | Almost certain | **Major** | • Discuss the shared vision and mission, goals, working agenda and benefits for the involved participants<br>• Keep communication flow alive in the coalition |
| Stakeholder | Distrust among civil society organisations may prevent an effective development of the coalition | Almost certain | **Major** | • Design the appropriate governance structure that balances representation of all groups |

*Notes:*

* The risk type is pre-classified in the following categories: Political, Safeguarding, Stakeholder, Finance, Compliance, Reputation, Other, Covid-19.

** The risk likelihood is presented on the scale: Unlikely, Possible, Likely, and Almost certain.

*** The risk impact is presented on the scale: Minor, Moderate, Major, and Severe.

# Annex B: Potential members of the coalition

In spite of the researcher’s best efforts, interviews could not be arranged within the timeframe of this research.

| Organisation | Category |
|---|---|
| Center for Cyber Law and Digital Transformation of Padjajaran University in Bandung, West Java | Academia and think tanks |
| Center for Digital Society of Gadjah Mada University | Academia and think tanks |
| Center for Human Rights Law Studies of Airlangga University in Surabaya, East Java | Academia and think tanks |
| House of Pancasila Nationality\Rumah Kebangsaan Pancasila | Civil society organisation |
| ICT Volunteers (RTIK) | Civil society organisation |
| Indonesian Cyber Media Association (AMSI) | Media organisation |
| Indonesian Institute | Academia and think tanks |
| Indonesian Medical Association (IDI) | Civil society organisation |
| Indonesian Santri YouTuber Association (AYSI) | Content creators |
| Indonesian Women Coalition\Koalisi Perempuan Indonesia | Civil society organisation |
| Indonesian YouTuber Association (AYI) | Content creators |
| Institute for Policy Analysis of Conflict (IPAC) | Academia and think tanks |
| Institute for Policy Research and Advocacy (ELSAM) | Civil society organisation |
| Legal Aid Foundation of the Indonesian Women’s Association for Justice (LBH APIK) | Civil society organisation |
| Papua-Fransiscan Justice, Peace, and Integrity of Creation (JPIC) | Civil society organisation |
| Protection Desk Indonesia\Yayasan Perlindungan Insani Indonesia (YPII) | Civil society organisation |
| Remotivi | Civil society organisation |
| Wahid Institute | Civil society organisation |
| Women’s Crisis Center in Jombang | Civil society organisation |
| Women’s solidarity for humanity and human rights (Spekham) | Civil society organisation |
| Yayasan Cinta Anak Bangsa (YCAB) Foundation | Civil society organisation |

# Annex C: Interview sheet

The researcher held interviews with representatives from the following organisations:

| Organisation | Category |
|---|---|
| Amnesty International – Indonesia | Civil society organisation |
| Arus Pelangi | Civil society organisation |
| Association for Elections and Democracy (Perludem) | Civil society organisation |
| Democracy and Electoral and Empowerment Partnership (DEEP) | Civil society organisation |
| ECPAT Indonesia | Civil society organisation |
| Engage Media – Indonesia | Civil society organisation |
| Gaya Nusantara | Civil society organisation |
| Human Rights Watch – Indonesia | Civil society organisation |
| ICT Watch | Civil society organisation |
| Independent Committee Aware of Elections (KISP) | Civil society organisation |
| Indonesian Anti-Slander Society (Mafindo) | Civil society organisation |
| Institute for International Peace Building\Yayasan Prasasti Perdamaian | Civil society organisation |
| Maarif Institute for Culture and Humanity | Civil society organisation |
| Paparisa Ambon Bergerak | Civil society organisation |
| Partisipasi Muda\Generasi Melek Politik | Civil society organisation |
| Southeast Asia Freedom of Expression Network (SAFEnet) | Civil society organisation |
| Tifa Foundation | Civil society organisation |
| Warga Muda | Civil society organisation |
| Alliance of Independent Journalists (AJI) | Media organisation |
| Center for Indonesian Policy Studies (CIPS) in Jakarta | Academia and think tanks |
| Centre for Strategic and International Studies (CSIS) in Jakarta | Academia and think tanks |

In addition, the researcher had the opportunity to present and discuss preliminary findings with stakeholders:

- At two focus group discussions organised by Perludem on 29 December 2021 and 18 February 2022 as part of a series of discussions to prepare a roadmap for secure elections in 2024.
- At the public launch of SAFEnet's 2021 Digital Rights in Indonesia Situation Report on 2 March 2022.

# Bibliography

AJI, The Alliance of Independent Journalists 2018 Year-End Note: Persecution and Violence Threaten Journalists, 2018.

Alexandra, L., Satria, A., Suryahudaya, E. G., and Krisetya, B., CSIS National Hate Speech Dashboard, 2021.

Amalia, M., Esti, K., and Camil, M. R., The industry of political buzzing in Indonesia and its impact on social media governance: Examining viral tweets, 21 June 2020.

APJII, APJII's Internet Survey Report 2019 – 2020 (Q2) [Indonesian], 2020.

ARTICLE 19, *Hate Speech Explained: A Toolkit*, 2019.

ARTICLE 19, *Indonesia: Ministerial Regulation 5 will exacerbate freedom of expression restrictions*, 29 September 2021.

ARTICLE 19, *Online Harassment And Abuse Against Women Journalists And Major Social Media Platforms*, 2020.

ARTICLE 19, *Side-stepping Rights: Regulating Speech by Contract, Policy Brief*, 2018.

ARTICLE 19, *Social Media Councils: One Piece in the Puzzle of Content Moderation*, 2021.

ARTICLE 19, *Watching the Watchmen: Content Moderation, Governance, and Freedom of Expression, Policy Brief*, 2021.

Bailey, H. and Howard, P. N., Country Case Studies Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation, 2020.

Bawaslu, Bawaslu, KPU, and MCIT sign MoA to tackle hoaxes and negative content at regional election, [Indonesian], 31 January 2018.

Bawaslu, The 2020 online election campaign will be supervised more strictly, [Indonesian], 2020.

Bawaslu, Together with Facebook and Google, Bawaslu undertook these efforts to tackle hoaxes, [Indonesian], 29 March 2019.

Beatty, A., Berkhout, E., Bima, L. Pradhan, M. and Suryadama, D., 'Schooling progress, learning reversal: Indonesia's learning profiles between 2000 and 2014', *International Journal of Educational Development*, 85, 2021.

Bhinneka Kultura, Parenting narratives to promote tolerance in families in Indonesia [Indonesian], 4 December 2021.

Bradshaw, S. and Howard, P. N., *The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation*, Oxford Internet Institute/University of Oxford, Computational Propaganda Research Project, 26 September 2019.

CSIS Indonesia, *Fire in the Husk: The Phenomenon of Hate Speech in Indonesia*, 18 August 2021.

ELSAM, *Internet Content Governance in Indonesia: Policies, Practices and Problems*, 2017.

Fealy, G., 'Bigger than Ahok: Explaining the 2 December mass rally', Indonesia at Melbourne, 7 December 2016.

Fitriani, A., Satria, P. P., Nirmalasari, and Adriana, R. *The Current State of Terrorism in Indonesia: Vulnerable Groups, Networks, and Responses*, Centre for Strategic and International Studies, Jakarta, 2018.

Freedom on the Net 2021. Indonesia, 2021.

Gorwa, R., Binns, R., and Katzenbach, C., 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', *Big Data & Society*, 7(1), 2020.

Haristya, S. 'The efficacy of civil society in global Internet governance', *Internet Histories,* 4(3), 2020.

Harsono, A. and McMinn, T. *I Wanted to Run Away: Abusive Dress Codes for Women and Girls in Indonesia*, Human Rights Watch, 2021.

IPAC, Indonesia and the tech giants vs. ISIS supporters: Combating violent extremism online, [YouTube], [Indonesian], July 2018.

Johansson, A. C. *Social Media and Politics in Indonesia*, Stockholm School of Economics Asia Working Paper Series 2016-42, Stockholm School of Economics, Stockholm China Economic Research Institute.

Kartikawangi, D., 'Focus group-based evaluation of social media usage in Indonesia's digital government', *Asian Journal for Public Opinion Research,* 8(1), 2020.

Khoirina, M. M. and Sisprasodjo, N. R., 'Media social – Instagram usage and performance benefit (Case study on housewives online seller in Indonesia)', *International Journal of Entrepreneurship and Business Development,* 2(1), 2018.

Komnas Perempuan, Churches against sexual violence, [YouTube], [Indonesian], 3 July 2020.

Komnas Perempuan, *Records of Violence against Women in 2020*, 5 March 2021.

Lim, M., 'Freedom to hate: social media, algorithmic enclaves, and the rise of tribal nationalism in Indonesia', *Critical Asian Studies*, 49(3), 2017.

Maharddhika and Salabi, N. A. Interference to vote rights: Phenomenon and responsibility [Indonesian], 21 September 2021.

Nugroho, Y, Siregar, M. F., and Laksmi, S., *Mapping Media Policy In Indonesia*, CIPG, HIVOS, University of Manchester Business School, Ford Foundation, 2012.

Nugroho, Y., *Citizens in @action: Collaboration, participatory democracy and freedom of information – Mapping contemporary civic activism and the use of new social media in Indonesia*, University of Manchester and HIVOS Regional Office Southeast Asia, 2011.

Organisation for Economic Co-operation and Development, Education at a glance, 2019.

Perludem, Political party finance reform in Southeast-Asia, 12 September 2021.

Roeslie, C. K., Data stolen and online gender based violence: The voices of victims that were underestimated by the police, 5 November 2021.

SAFEnet, *2021 Digital Rights in Indonesia Situation Report*, February 2022.

SAFEnet, Public discussion and launch of Digital Rights in Indonesia Situation Report, [YouTube], [Indonesian], 2 March 2022.

Sastramidjaja, Y., Berenschot, W., Wijayanto, and Fahmi, I., The threat of cyber troops, Inside Indonesia, 2021.

Sholeh, B., 'The dynamics of Muslim and Christian relations in Ambon, Eastern Indonesia', *International Journal of Business and Social Science*, 4(3), 2013.

Sinpeng, A., Martin, F., Gelber, K. and Shields, K., *Facebook: Regulating Hate Speech in the Asia Pacific*, Department of Media and Communications, University of Sydney, 2021.

Siregar, F. E., 'The role of the elections supervisory agency to contend hoax and hate speech in the course of 2019 Indonesian general election', *Padjajaran Journal of Law,* 7(2), 2020.

Slama, M., 'Practising Islam through social media in Indonesia', *Indonesia and the Malay World*, 46, 2018.

Suwana, F., Pramiyanti, A., Mayangsari, I. D., Nuraeni, R., and Firdaus, Y., 'Digital media use of Generation Z during Covid-19 pandemic', *Jurnal Sosioteknologi* 19(3), 2020.

United Nations Educational, Scientific and Cultural Organization, UNESCO initiates global dialogue to enhance the transparency of Internet companies, with release of illustrative high-level principles, 3 May 2020.

van Bruinessen, M., Post-Suharto Muslim engagements with civil society and democratization, Yogyakarta, Pustaka Pelajar, 2004.

Widyaningsih, R. and Kuntarto, K., 'Family suicide bombing: A psychological analysis of contemporary terrorism', *Journal of Social Religious Research,* 26(2), 2018.

Zuiderveen Borgesius, F. J. et al., Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review,* 14(1), 2018.

# Endnotes

[1] An initial list was elaborated through desk research and in consultation with UNESCO with the aim of speaking with a broad and representative range of relevant stakeholders, including representatives from civil society, the private sector, public actors, and social media companies. However, in spite of the researcher's best efforts, it was not possible within the timeframe of this research to arrange interviews with all the organisations identified originally.

[2] On 30 September 1965, there was a coup attempt; six senior generals and a lieutenant were kidnapped and murdered. The coup was countered under the direction of General Suharto. There was then widespread antipathy towards the rise of communism, especially toward the Indonesian Communist Party (PKI). As a result of this sentiment, then-President Sukarno was marginalised and Suharto came to power. After these violent events, there was a perception of mass conversions to Christianity and thus fears of Western efforts to roll back communism and also to combat the political strength of Islam in Indonesia. The events also deepened the divide between Islam and Communism in Indonesia. The government promoted an official version of the 1965 events, that PKI was the main actor and a traitor to the country that needed to be eliminated.

[3] However, research shows that Indonesia has failed to regulate the media, in particular in relation to mitigating the profit-driven logic of the industry. See Y. Nugroho, M. F. Siregar, and S. Laksmi, *Mapping media policy in Indonesia*, 2012.

[4] It is a YouTube channel that regularly conducts a talk show on various popular topics by inviting guests ranging from lay people, artists, experts in various fields to high profile guests like Ministers Sri Mulyani, Budi Sadikin, Luhut Binsar (the last is known to be selective in accepting interview requests from the press).

[5] Chat and messenging apps are not covered in this research.

[6] M. K. Alfarizi, Internet in Papua is off, causing a digital immigrants phenomenon, who are they?, 8 June 2021.

[7] The Santa Clara Principles on Transparency and Accountability in Content Moderation, an initiative by global civil society organisations, provide a helpful reference to assess the content rules and content moderation practices of social media companies in light of international standards.

[8] Meta, Misinformation. Accessed 4 March 2022.

[9] Twitter, Hateful conduct policy. Accessed 23 December 2021.

[10] TikTok. Community guidelines. Accessed 23 December 2021. *Note:* there are four other prohibited contents, but this study quotes only this type of prohibited content to compare and contrast the rules related to violence threats in different platforms.

[11] YouTube. Hate speech policy. Accessed 23 December 2021.

[12] Instagram, Community guidelines. Accessed 23 December 2021.

[13] Meta, Violence and criminal behaviour. Accessed 23 December 2021.

[14] For an elaboration, see the 2018 analysis by ARTICLE 19 of the terms of service and community guidelines of Facebook, Twitter, and YouTube in light of international standards on freedom of expression and other fundamental rights.

[15] Meta, Violence and criminal behaviour. Accessed 23 December 2021.

[16] For a definition of hate speech under international standards on freedom of expression, see ARTICLE 19, Hate speech explained: A toolkit, 2019.

[17] In the toolkit on hate speech, ARTICLE 19 provides guidance on what policy measures state and non-state actors can undertake to create an enabling environment for freedom of expression and equality that addresses the underlying causes of 'hate speech' while maximising opportunities to counter it.

[18] For an analysis of international standards on freedom of expression in relation to national security, see the Johannesburg Principles on National Security, Freedom of Expression and Access to Information, adopted in 1995, and the Tschwane Principles on National Security and the Right to Information, adopted in 2013. In particular, these Principles state that restrictions to freedom of expression on the ground of national security can only be justified if their genuine purpose and demonstrable effect is to protect the country's existence or its territorial integrity against the use or threat of force, or its capacity to respond to the use or threat of force. The restrictions should never serve as a pretext for protecting the government from embarrassment or exposure of wrongdoing, to conceal information about the functioning of its public institutions, or to entrench a particular ideology.

[19] See also ARTICLE 19's briefing papers and recommendations on tackling online abuse and harassment against women journalists.

[20] See ARTICLE 19's analysis of such regulations at Blog: Indonesia's intermediary regulation imperils Internet freedom; and Indonesia: Ministerial Regulation 5 will exacerbate freedom of expression restrictions

[21] ECPAT (End Child Prostitution in Asian Tourism) is a global network and campaign against the sexual exploitation of children.

[22] The right to freedom of expression is guaranteed in Article 19 of the International Covenant on Civil and Political Rights.

[23] Article 27 prohibits users to distribute and/or transmit electronic information with contents of pornography, gambling, defamation, and extortion or threats. Article 28 forbids people to disseminate false information that disadvantage consumers in electronic transaction and also to spread hate speech. Article 29 disallows people to distribute violence threats or scares aimed personally.

[24] International standards, the Manila Principles on Intermediary Liability and the Updated Santa Clara Principles, states that intermediaries, including social media platforms, should not be held liable over the content produced by their uses. However, they are responsible for undertaking content moderation practices that uphold international standards on freedom of expression with sufficient understanding of the related local contexts and culture.

[25] I. Arsyam, Who is Ade Armando? US graduated University of Indonesia's Lecturer but often accused of being more like a buzzer, 30 June 2021; M. Amalia, K. Esti and M.R. Camil, The Industry of Political Buzzing in Indonesia and Its Impact on Social Media Governance: Examining Viral Tweets, 21 June 2020.

[26] T. Gobel, MCIT's cyberdrone now detects illegal selling in e-commerce, [Indonesian], 24 February 2020; A. Yuliani, Get to know cyber drone 9, Indonesia's Internet police, [Indonesian], 5 January 2018.

[27] See the Santa Clara Principles (version 2.0).

[28] ARTICLE 19's views on the regulation of platforms are presented in: *Side-stepping Rights: Regulating Speech by Contract*, 2018; *Watching the Watchmen: Content Moderation, Governance and Freedom of Expression*, 2021; and *Taming Big Tech*, 2021. See also the Santa Clara Principles (version 2.0).

[29] Internetsehat, The role of YouTube Trusted Flagger in cleaning YouTube (part 1 of 2), [Indonesian], 5 November 2018; The role of YouTube Trusted Flagger in cleaning YouTube (part 2 of 2), [Indonesian], 5 November 2018;

[30] AWPI, Profile, [Indonesian]; Press Council, Journalists organisations, [Indonesia].

[31] See the public comments (in Indonesian language) in this video – S. Alrenzha, 3 reasons and threads of blocked TikTok accounts ? Tik Tok Block, [YouTube], [Indonesian], accessed 20 January 2022.

[32] Santri is a term for someone who attends Islamic education in Islamic boarding schools.

[33] See the public comments (in Indonesian language) in this video – S. Alrenzha, 3 reasons and threads of blocked TikTok accounts ? Tik Tok Block, [YouTube], [Indonesian], accessed 20 January 2022.