YAYASAN TIFA

pr2media
PEMANTAU REGULASI & REGULATOR MEDIA
MEDIA REGULATION AND REGULATOR WATCH

Research Paper

Regulations for Moderating Illegal Content on Social Media in Indonesia

PR2Media
www.pr2media.or.id

RESEARCH PAPER
REGULATIONS FOR MODERATING ILLEGAL CONTENT ON SOCIAL MEDIA IN INDONESIA

Author(s) : Engelbertus Wendratama
Masduki
Rahayu
Putri Laksmi Nurul Suci
Reviewer(s) : Novi Kurnia
Policy and Governance Data Team, Tifa Foundation
Cover Design and Layout : gores.pena

Published by:
PR2Media
Jl. Lempongsari Raya, Gg. Masjid RT 9/RW 37 No. 88B
Jongkang Baru, Sariharjo, Sleman, DIY, 55581
Email: kontak.pr2media@gmail.com
office@pr2media.or.id

The publication of this research paper has been supported by the Tifa Foundation.

Recommended citation: Wendratama, E., Masduki., Rahayu., & Suci, P. L., (2023). Research Paper: Regulations for Moderating Illegal Content on Social Media in Indonesia. Yogyakarta: PR2Media.

CONTENTS

INTRODUCTION

This research paper consolidates civil society's aspirations to implement a fairer and more effective framework for moderating social media content in Indonesia. The formulation of these recommendations refers to three sources. First, a study by PR2Media titled "Pengaturan Konten Ilegal dan Berbahaya di Media Sosial: Riset Pengalaman Pengguna dan Rekomendasi Kebijakan" ("The Regulation of Illegal and Harmful Content on Social Media: A Study of User Experiences and Policy Recommendations", Wendratama et al., 2023). Second, a series of focus group discussions with diverse stakeholders, including legal scholars, media practitioners, and civil society organizations conducted in Jakarta between June 6 and 7, 2023. Third, a review of studies and reference materials regarding content moderation in Indonesia and elsewhere; this includes, for example, studies conducted by the Center for Digital Society, FISIPOL Universitas Gadjah Mada ("Moderating Harmful Content in Indonesia: Legal Frameworks, Trends, and Concerns", 2022) and Article 19 ("Content Moderation and Local Stakeholders in Indonesia", 2022).

This research paper advances a healthier policy framework for moderating social content in Indonesia. Its recommendations are intended to be substantive yet as easily comprehensible as possible, presented in conjunction with proposed regulatory articles that can be easily understood by policymakers.

Reflecting on Indonesia's regulatory structure at the national law (*undang-undang*) level, and considering the country's empiric political and economic conditions, the authors of this research paper have, in conjunction with civil society organizations involved in advocacy efforts, identified two desired outputs:

> First, in the short term, this research paper is intended to propose revisions to the Information and Electronic Transactions Law (UU ITE), particularly Article 15 regarding the accountability of electronic service providers, which as of August 2023 remains under discussion by the Indonesian government and Commission I of the House of Representatives. This proposal was conveyed by PR2Media during the Public Hearing held by

the Committee for the Revision of UU ITE, Commission I, House of Representatives of Indonesia, on August 23, 2023 (PR2Media, 2023).

Second, in the long term, this document is presented as part of an overall effort to reform Indonesia's internet regulation policies, seeking a total revision to UU ITE that provides a fairer and more comprehensive framework for digital platforms' moderation of illegal content.

What is the background for this proposal? Research conducted by PR2Media in 2023 found that, in Indonesia, social media companies' moderation of illegal content on social media lacks transparency and fails to take a human rights mindset, especially as compared to developed democracies. For comparison, the European Union passed the Digital Services Act (coming into effect in 2024) that sets high transparency and accountability standards for digital platforms, particularly those with large user bases, and their moderation of illegal content.

Presently, Indonesia uses several regulations to moderate internet content: Law No. 19/2016 regarding Information and Electronic Transactions (henceforth UU ITE), Government Regulation No. 71/2019 regarding the Implementation of Electronic Transaction Systems, and Regulation of the Minister of Communication and Informatics No. 5/2020 regarding the Implementation of Electronic Systems in Private Environments. However, this regulatory framework is insufficient for promoting and reinforcing the government's regulatory role, ensuring transparency and accountability in digital platforms' moderation of illegal contact, and protecting internet users' freedom of expression.

Presently, social media platforms depend heavily on artificial intelligence (AI) when ascertaining whether or not content should be deleted. Human moderators of social media content are minimal, as is transparency in the content moderation process. Campaigning for Indonesia's 2014 general election, which will be permitted beginning November 2023, has the potential to accelerate the dissemination of mis-/disinformation and hate speech, both of which are detrimental to democracy. Reflecting on the tumultuous information ecosystem that emerged during previous elections, as well as the effect of this chaos on democratic processes and social cohesion, it is increasingly urgent to prepare detailed transparency and accountability standards for the moderation of illegal content on social media.

For this study, "illegal content" refers to all forms of content that are prohibited by Indonesian law. This includes pornography (a violation of Law No. 44/2008), gambling

(*Criminal Code*, UU ITE), hate speech (*Criminal Code*, Law No. 19/2016, Law No. 40/2008, Police Circular SE/6/X/2015), and the dissemination of false information and defamation (*Criminal Code*, Law No. 19/2016). It must be recognized that illegal content remains poorly defined in Indonesian law. It is highly contextual and thus requires clear careful regulation and moderation by social media platforms.

Thanks to its solid methodology, the findings of PR2Media legitimize its endeavors to make recommendations regarding the revision of UU ITE and draft new legislation regarding the accountabilities of social media platforms in Indonesia. For instance, a survey conducted by PR2Media in early 2023 found that many of Indonesia's social media users are dissatisfied with the social media platforms' means of moderating content and addressing their concerns. This finding can also provide grounds for the government and the House of Representatives to identify better mechanisms for moderating content and addressing grievances. Further enriching this research are the perspectives of diverse stakeholders, collected through interviews, which highlight the importance of developing a better regulatory framework for the moderation of social media content.

This study has considered the current regulatory environment, as well as how existing laws—UU ITE No. 19/2016, Law No. 71/2019, and Regulation of the Minister of Communication and Informatics No. 5/2020—intersect. The proposed amendments will complement and reinforce the existing framework. Research by PR2Media has shown that only through a total revision of UU ITE will solid and comprehensive mechanisms be developed for moderating illegal content.

At the same time, the authors also emphasize that the revision of existing laws must be considered within the broader context of digital moderation. For instance, there are presently discourses regarding the revision of the Broadcast Law to encompass video-on-demand platforms (such as Netflix) and video-sharing platforms (such as YouTube). The European Union has implemented similar regulations through the Audio-Visual Media Services Act. Similar discourses have emerged regarding other elements of the current digital ecosystem, including competition between digital platforms (the Digital Markets Act in the European Union) and the use of artificial intelligence (the AI Act in the European Union).

This document elucidates three elements. First, it provides a general discussion of the moderation of illegal content on social media, including the definition and scope of social media and the definition and classification of illegal content. Second, it explores the means through which social media platforms recognize and flag illegal content, as well as the grievance and appeal mechanisms made available by these platforms. Third, it discusses the initiative for establishing a Social Media Commission,

requiring social media platforms to transparently publish annual reports regarding their moderation activities, and conducting independent audits of their compliance with the proposed rules. As part of its efforts to provide concrete policy recommendations to the government and House of Representatives, recommended revisions to Article 15 of UU ITE that emphasize the accountability of electronic service providers are appended to this study.

A. Challenges to Moderating Social Media Content

1. Distribution of illegal content on social media: the current conditions

The distribution of illegal content on social media in Indonesia has long drawn concern. According to data from the Directorate General of Informatic Applications, Ministry of Communication and Informatics, approximately 1.4 million pieces of illegal content had been reported as of March 6, 2023. Most such items are found on Twitter (now X), followed by Meta (Facebook and Instagram). Pornographic content was the most prevalent, followed by gambling and fraud (Ministry of Communication and Informatics, 2023).

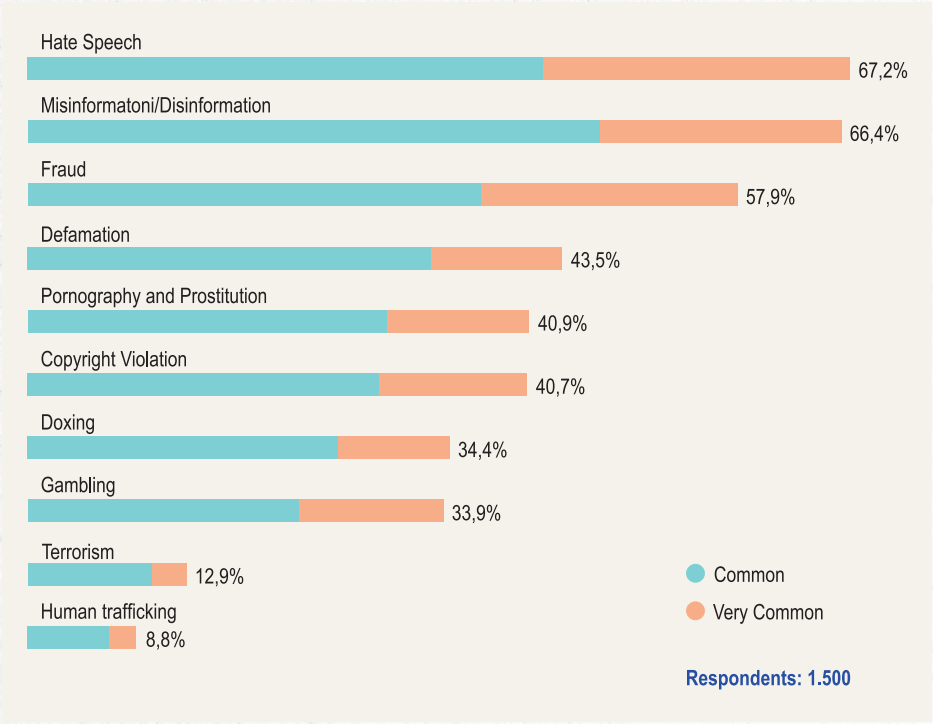
The study "Pengaturan Konten Ilegal dan Berbahaya di Media Sosial: Riset Pengalaman Pengguna dan Rekomendasi Kebijakan" ("The Moderation of Illegal and Harmful Content on Social Media: A Study of User Experiences and Policy Recommendations", Wendratama et al., 2023)—conducted in early 2023 by PR2Media—had similarly disconcerting results. All of this study's respondents (1,500 social media users in 38 Indonesian provinces) reported that they frequently saw illegal content on social media, with the most common being hate speech, mis-/disinformation, and fraud.

The most common forms of illegal content on social media in Indonesia are presented in Figure 1.

It should be noted that several of these forms of illegal content are identified as harmful (rather than illegal) by social media platforms, other nations' regulatory framework, and several United Nations instruments—for instance, misinformation, (some) hate speech, and adult (rather than child) pornography. Nonetheless, PR2Media has used the above classification to comply with the regulatory framework in Indonesia.

As the number of social media users in Indonesia has increased, and as content has become increasingly diversified, illegal content has become increasingly prevalent. However, as noted by the PR2Media study—which also involved interviews with various stakeholders—existing regulatory instruments are insufficient for overcoming these challenges. Said regulatory challenges will be explored below.

Figure 1. Most Commonly Seen Forms of Illegal Content in Indonesia



2. Regulatory challenges in Indonesia

In Indonesia, regulations regarding illegal content on social media have thus far been limited to identifying the mechanisms and authorities available to the government in preventing the use and distribution of illegal content. However, the means through which social media platforms moderate themselves remain untouched, even though said platforms are better suited to this task and have more resources available to them.

In Indonesia, social media content (and general internet usage) is regulated through Law No. 19/2016 regarding the Revision of Law No. 11/2008 regarding Information and Electronic Transactions, as implemented through Government Regulation No. 71/2019 regarding the Implementation of Electronic Transaction Systems, and Regulation of the Minister of Communication and Informatics No. 5/2020 regarding the Implementation of Electronic Systems in Private Environments.

UU ITE faces two major challenges in moderating illegal content on social media:

1. There is no detailed definition of illegal content, and thus it is common for the government, social media platforms, and internet users to differ in their understandings of prohibited content (for instance, “disturbing public order”).

2. Social media platforms' obligations in handling illegal content, as well as potential sanctions for neglect, remain unclear. All electronic system providers (search engines, digital storage services, digital marketplaces, chat applications, etc.) are treated the same under UU ITE, even though they differ significantly in their individual characteristics and user bases (social influence).

These problems are significant, given that social media platforms are the only ones capable of moderating (for instance, deleting) content on their websites. Governments and users can only report problematic content.

The regulation (moderation) of content is a process used by social media platforms to protect their users by evaluating, identifying, limiting, reducing traffic to, and/or deleting illegal and harmful content and/or the accounts that post said content. In evaluating content, social media platforms frequently employ human moderators and/or artificial intelligence (AI).

> If social media platforms refuse to respond to government grievances, there is only one extreme remedy available: completely blocking Indonesians' access to said platforms. Such an approach is authoritarian and counterproductive, as said platforms also contain legal content that benefits users.

Given this difficulty, and recognizing the importance of social media providers in content moderation and their effect on society, new regulations have been passed to regulate these providers' management of illegal content on their platforms.

For instance, the European Union passed the Digital Services Act (legislated in 2022, enacted in 2024), which outlines the obligations of internet service providers in moderating illegal content on their platforms (including search engines, marketplaces, and social media). Obligations vary, depending on platforms' user bases in Europe; the larger the platform, the greater its obligations. Very large social media platforms[1] have numerous obligations, which include reporting on their moderation mechanisms and reducing the risks derived from their design and usage (European Commission, 2023).

---

[1] Defined as platforms whose monthly active user base represents at least 10% of the population of the European Union (i.e., 45 million people). In April 2023, such platforms included YouTube, Facebook, TikTok, Instagram, Twitter, and LinkedIn (European Commission, 2023).

> Regulations in the European Union outline how platforms should moderate illegal content, ensuring that the government is not required to become directly involved. This is considered more effective, as the resources of social media providers are better suited to this purpose. The government needs only to monitor their compliance.

As with existing regulations in Germany and France, the Digital Services Act regulates how platforms should moderate illegal content, ensuring that the government is not required to become directly involved. This is considered more effective, as the resources of social media providers are better suited to this purpose. The government needs only to monitor these providers' practices—a desire also expressed by Semuel Abrijani Pangerapan, the Director General of Applications and Informatics in an interview with PR2Media researchers in March 2023 (Wendratama et al., 2023). Other nations currently debating similar laws include the United Kingdom and Singapore, both of which intend to implement online safety bills.

B. Definition and Scope of Social Media

Definition

Definitions of social media have been varied, and academics and practitioners have been unable to agree upon a singular definition. Nevertheless, drawing on research by Aichner et al. (2021) that explored the history of social media definitions between 1994 and 2019 through various academic publications, the authors define social media as "internet-based electronic systems that enable users to mutually exchange electronic information and/or electronic documents using open electronic systems that are controlled by social media providers."

Here, the emphasis is on the ability to mutually exchange content through "open systems", a definition that includes YouTube, Facebook, TikTok, Instagram, Twitter, and LinkedIn. It excludes "closed systems" that are protected by encryption, such as WhatsApp, Telegram, and Signal.

Certainly, the dissemination of illegal content in Indonesia occurs on platforms other than social media, including search engines, digital storage services, marketplaces, and internet-based chat applications. However, because the research conducted by PR2Media (2023) was limited to social media, the proposed regulations are likewise focused on open social media platforms. Ideally, any regulations that require action from electronic system providers when moderating illegal content should include all

forms of electronic systems (as seen in UU ITE). However, the obligations of service providers should reflect their own particular characteristics, as also seen in the Digital Services Act in the European Union.

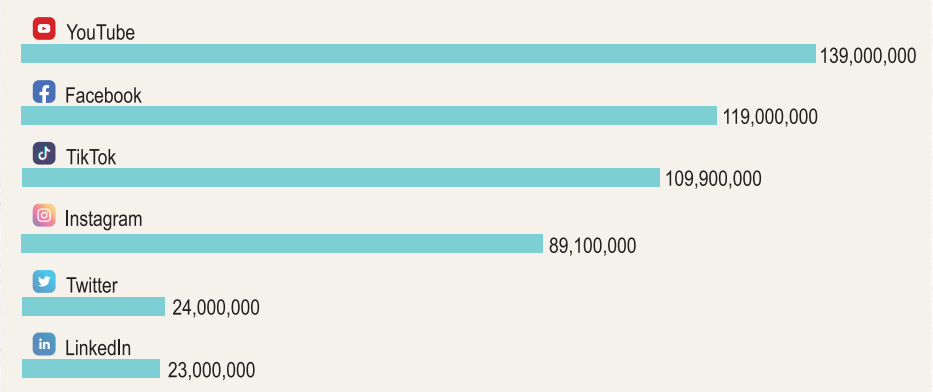
Scope

The proposed regulations here will focus on social media platforms with relatively large user bases, as these platforms are the ones that have the greatest influence on the Indonesian people. Such scope considerations are also evident in the Digital Services Act, which sets more obligations for larger platforms. As mentioned previously, the greatest obligations are borne by the largest platforms—those whose monthly active user base represents at least 10% of the population of the European Union (i.e., 45 million people). In April 2023, such platforms included YouTube, Facebook, TikTok, Instagram, Twitter, and LinkedIn (European Commission, 2023).

PR2Media argues that this threshold—10% of the population, or 27 million users (according to the 2023 Census, Indonesia had a population of 278 million)—is far too large. For example, despite its prominence, Twitter had only 24 million Indonesian users in January 2023 (Kemp, 2023).

> As such, we argue that a rational threshold is 20 million users. These proposed regulations are only intended for those social media platforms with at least twenty million monthly users. Certainly, this threshold can be adjusted, similar to how the European Commission regularly updates its list of very large online platforms (VLOP) that have the greatest obligations.

Figure 2. Number of social media users in Indonesia, early 2023



Using twenty million monthly users as its threshold, this proposal encompasses the following platforms: YouTube, Facebook, TikTok, Instagram, Twitter, and LinkedIn.

C. Definition and Classification of Illegal Content

The next challenge is the broad definition of illegal content under Indonesian law, which also includes content described as "harmful content" by social media providers. Harmful content is content that, though legal, may cause physical or psychological harm.

Examples of content that are defined as harmful, rather than illegal, by social media providers and United Nations instruments (see United Nations Human Rights Office of the High Commissioner, 2023) include mis-/disinformation, hate speech, adult pornography, and gambling. However, under Indonesian law, all such content is illegal. In other words, the publication and dissemination of harmful content is a criminal act under Indonesian law, and thus anyone distributing such content should face legal consequences. In Indonesian law, there are thus only two mechanisms for dealing with illegal and harmful content: court mechanisms (criminal charges and, in some cases, civil suits) and non-court mechanisms (conflict resolution and other administrative action) (Rahman et al., 2022).

Consequently, several sections of UU ITE have frequently been used to criminalize citizens who express their concerns and views, thereby repressing the freedom of expression that should be protected under Indonesia's 1945 Constitution.

Aside from the threat of criminal action for the distribution of harmful content on social media (an act that is quite broad), UU ITE has also been multi-interpretable and controversial due to its ambiguous and expansive terminology, including phrases such as "morality" and "public order."

Several regulations have also been created to facilitate the implementation of content moderation and "guide" law enforcement officials. These have included Joint Decision of the Coordinating Ministers of Politics, Law, and Security No. B-96/HK.00.00/07/2021 regarding the Guidelines for Implementing Certain Articles of the Information and Electronic Transactions Law (UU ITE) and Circular of the Director of the Indonesian Police No. SE/6/X/2015 regarding the handling of hate speech. Nevertheless, diverse stakeholders have indicated that the definition of illegal content remains problematic.

For example, the 2015 circular by the Director of the Indonesian Police categorizes insults, defamation, and blasphemy as hate speech, but provides no clear definition of such acts (Palatino, 2015). Meanwhile, referring to various international agreements (Gagliardone et al., 2015), hate speech is limited to "attacks against individuals or groups based on 'inherent characteristics', such as ethnicity, religion, nationality, and gender."

1. Legal considerations: hate speech, mis-/disinformation, and defamation

One of the most problematic forms of illegal content in Indonesia is hate speech. A survey of 1,500 social media users spread among 38 Indonesian provinces conducted by PR2Media (Wendratama et al., 2023) found that hate speech is the most common type of illegal and/or harmful content encountered by internet users. Such responses cannot be separated from the fact that Indonesian law defines defamation very broadly, with a corpus of case law that allows challenges against many forms of speech even as it limits Indonesians' freedom of expression.

As such, this study urges Indonesia to recognize three forms of hate speech. These definitions, which refer to the Rabat Plan of Action, adopted by the Office of the United Nations High Commissioner for Human Rights (United Nations High Commissioner for Human Rights, 2013), are as follows:

1. **Expression of views and opinions that represent criminal acts**
   Views and opinions that should be recognized as criminal acts include (1) efforts to promote discrimination, enmity, or violence against any people or religion and (2) efforts to affirm or promote any form of discrimination and hatred. These two points are ensconced in two United Nations commissions, both of which have been ratified by Indonesia: the International Covenant on Civil and Political Rights (1966) dan the International Convention on the Elimination of All Forms of Racial Discrimination (1965).
2. **Expression of views and opinions that may face administrative sanctions or civil charges**
   Hate speech in this category includes expressions and utterances that contain hated, as defined by Article 19, Paragraph 3, of the International Covenant on Civil and Political Rights (ICCPR); under this covenant, an individual's freedom of expression may be curtailed to protect the rights and reputation of others; national security; public order; public health; or moral interests.
3. **Expression of views and opinions that may not face sanctions**
   Expressions that may not face sanctions include those indicating intolerance of and displeasure with others. In such instances, education is preferred. Such programs may involve digital literacy programs by the government or by diverse stakeholders.

   As such, not all forms of hate speech are criminal (Putri, 2021). Indeed, looking at instances of hate speech in Indonesia, the majority of court cases are related to acts that fall into the latter two categories–i.e., non-criminal.

Another means of promoting shared interpretations and minimizing infringements of citizens' human rights is implementing a threshold test, one also based on the Rabat Plan of Action (United Nations High Commissioner for Human Rights, 2013). This threshold test has been designed for content containing hate speech, but can also be adopted for all forms of content, criminal or not, including mis-/disinformation and defamation. The six components of the threshold test are presented below:

SIX-POINT THRESHOLD TEST

1. **Context**
   Content must be analyzed within the socio-political context of its creation and dissemination.
2. **Speaker**
   The position/status of the speaker/actor in society must be considered, with special consideration of the position/status of the speaker/actor in the eyes of their intended audience.
3. **Intent**
   Carelessness and neglectfulness do not make for criminal intent, and neither does the act of sharing content. To show intent, there must be a desire to instigate and provoke the audience.
4. **Content and Form**
   Analysis of the content and form of the message will show its ability to instigate or provoke audiences. This includes the explicitness/implicitness of the message, as well as its means of expression (for instance, through satire).
5. **Reach**
   This refers to the size of the audience reached. The larger the audience, the greater the potential harm caused by messages.
6. **Potential risk**
   This refers to the potential extent to which a message may instigate/provoke its intent audiences, with careful consideration of directness and causality.

PR2Media holds that these six points (particularly the first five) can also be used to review other forms of illegal content. For example, when dealing with defamation and mis-/disinformation (both of which are illegal in Indonesia), the first five points may be used to reduce ambiguity and undue criminalization.

Referring to the classification and threshold models above, PR2Media recommends that illegal content deemed to be harmful content (particularly mis-/disinformation and defamation) not be always criminalized; it is necessary to consider the intensity and the scope of the content.

Social media providers' legal responsibility or liability for the content uploaded by users remains heavily debated. In the United States, for example, efforts have been made to revise Section 230 of the Communications Decency Act (1996), which waives social media platforms' liability for user content and protects them from legal action. Under this law, social media platforms cannot be treated the same as news publishers, which are liable for the content they publish. At the same time, there are some types of content for which social media providers are legally liable: copyright violations, violations of federal criminal law, human trafficking, and sex trafficking involving minors (US Congress, 2019). Responding to the proposed revisions, Mark Zuckerberg—the founder of Facebook—stated that social media providers should be liable for "some" content uploaded by users, but rejected the idea that social media providers should be responsible for all content uploaded by users. As of writing, the process is ongoing, though President Joe Biden has urged Congress to ensure that revisions are completed expeditiously (Morrison, 2023).

Generally speaking, platform liability regimes around the world follow one of four models: (1) strict liability, whereby platforms are fully liable for the content posted by third parties; (2) knowledge-based liability, whereby platforms are not liable for content unless they were aware of illegal content and failed to take action against it; (3) fault-based liability, whereby platforms may face sanctions if they are found to have failed to prevent the distribution of prohibited content; and (4) broad immunity from liability (Frosio, 2021).

This study argues that different models of liability should be implemented for illegal and harmful content. For example, the knowledge-based liability model is well-suited to dealing with illegal content in urgent situations (except for misinformation and hate speech). This is similar to the Digital Services Act. Here, the definition of knowledge is limited to reports from trusted flaggers and court decisions (to ensure that social media providers do not universally monitor or filter content, and to avoid potential abuses of reporting mechanisms).

Trusted flaggers, those organizations competent in recognizing and flagging illegal content, have an important position in the current content moderation ecosystems. Reports from trusted flaggers must be prioritized by social media providers, as such flaggers have been trained to provide comprehensive and reputable reports. Organizations must receive training from social media providers and work as their partners. For example, ECPAT Indonesia (an organization focused on eradicating sexual violence against children) has acted as a trusted flagger for Twitter.

2. Legal considerations: defamation and derogation

Cases of defamation and derogation are closely intertwined with privacy issues. As such, a different approach should be taken. This research recommends a notice-to-notice approach, whereby social media providers may provide a counter-notice after receiving a report, and then decide whether or not legal action is required.

One revision made to UU ITE in 2016 involved Article 27. To avoid ambiguity, the definition of defamation was returned to Article 310 and Article 311 of the (since replaced) *Criminal Code*, which specifies:

1. Said information must be intended as an attack on the honor or reputation of a person and be disseminated publicly (known by many people).
2. It must not be done to promote the public interest or to defend oneself (i.e., through coercion).

In other words:

1. If the accusation is true, but disseminated not to promote the public interest or to defend oneself (Article 310, Paragraph 3, *Criminal Code*), it may be deemed a criminal act under Article 310, Paragraph 1, of the *Criminal Code*.
2. If the accusation is shown to not be true, then it may be deemed a criminal act under Article 311, Paragraph 1, of the *Criminal Code*.

As such, where a statement (i.e., piece of content) is accurate or factual, the person disseminating it on electronic media may be charged under the law against defamation or derogation.

Defamation and derogation are heavily intertwined with the concept of privacy. Within the context of the dissemination of information, privacy refers to individuals' right to ensure that individuals' lives and personal information remain confidential—or at least exposed to the fewest number of people possible (Wendratama, 2021). Unfortunately, unlike in other nations, Indonesia has yet to discuss the different privacy standards expected by private citizens vis-à-vis public figures. The United States, for example, clearly distinguishes between ordinary citizens and public figures in its privacy standards. Public figures, being individuals who work in front of the public or earn a living through their interactions with the public, are expected to enjoy less privacy than the average person. Legal precedent uses four elements to identify public figures (Yanisky-Ravid & Lahav, 2006):

1. Access to and control of media (particularly relevant to politicians and others who frequently receive media coverage)

2. Involvement in public life; this includes individuals who occupy public positions and/or benefit from public financial resources in sectors such as politics, business, art, sports, etc.
3. Anyone who voluntarily involves themselves in public life, including those seeking power, influence, and popularity.
4. Anyone involved in public controversies, be it voluntarily or not; these include cases that have received media coverage.

To simplify these criteria, public figures generally include public officials, politicians, artists (painters, actors, musicians, and internet celebrities), and sportspersons. Persons in these professions have a lower expectation of privacy, as they are perceived as deliberately exposing themselves to the public and benefitting from said exposure. Ordinary citizens, such as bank tellers, have a much higher expectation of privacy and courts tend to side with them when they challenge the media/public exposure of their private information. Conversely, when politicians challenge content that they claim to be defamatory or to violate their privacy, the burden of proof is much greater; they must clearly show that the exposure of the information was deliberately intended to, and did, cause significant harm. In other words, the exposure of information regarding politicians is part and parcel of their chosen profession.

The authors agree with such a paradigm, as politicians and other public figures should have a lower expectation of privacy.

> In Indonesia, it would be better for regulations to clearly distinguish between "classes of privacy", ensuring that public figures recognize that their profession entails lower privacy expectations, and this affects their ability to challenge content as defamatory and derogatory."

D. Means used by social media to recognize and flag illegal content

Several studies (De Gregorio, 2020; Pirkova, 2022; Leong, 2022) have shown that algorithms and artificial intelligence are the backbones of social media platforms' moderation of content. This, in turn, heavily influences the means through which social media platforms recognize and evaluate harmful content.

In Indonesia, transparency has been lacking in social media providers' moderation of content for their users. For instance, there is no readily available information regarding

when content is moderated by artificial intelligence and when it is moderated by human beings. Likewise, no information has been made available as to whether Indonesian citizens are retained as content moderators or how many Indonesian moderators have been retained.

Elsewhere, Facebook has indicated that its content moderation is handled by third-party providers in India, Ireland, and the Philippines. In practice, these human moderators are only fluent in fifty languages, even though Facebook provides its services in more than one hundred languages. The number of moderators, therefore, must be increased (Barrett, 2020). In Indonesia alone, where there are hundreds of indigenous languages spread around 13,000 islands, the challenge is greater. It is therefore necessary to ensure that social media providers are transparent in their moderation, ensuring that said practices are suited to Indonesia's diversity (Article 19, 2022).

As such, PR2Media recommends regulations that require social media providers to explicitly identify the mechanisms they use to recognize and flag illegal content, be it artificial intelligence, human moderation, or a combination thereof. Section D provides a foundation for further regulation. The proposed articles can be found in Appendix 1 of this research paper, which was communicated by PR2Media during the Public Hearing held by the Committee for the Revision of UU ITE, Commission I, House of Representatives of Indonesia, on August 23, 2023 (PR2Media, 2023).

E. Means Used by Social Media to Respond to Reported Illegal Content

The moderation of illegal content by social media providers should be conducted proactively and responsively, based on reports submitted by users and government actors.

The word "proactive" means that social media providers, through their human moderators and/or automated systems, are responsible for content moderation. This should be based on clear regulations or guidelines, and include information as to the type of prohibited content, time of identification, and any actions taken against the content and/or user accounts involved. Such regulations/policies should be easily accessed by users, thereby ensuring that they understand what content is acceptable and what content is prohibited.

Meanwhile, the word "responsive" means that social media providers should take action against content and/or accounts based on user/government reports. In this, social media providers need to evaluate the flagged content. Evaluation should be

based on clear regulations/guidelines, with moderation conducted transparently, non-discriminatorily, fairly, and with respect for human rights.

Referring to several sources, including the Digital Services Act (European Commission, 2023) and Santa Clara Principles[2] (Santa Clara Principles, 2021), PR2Media proposes the following:

1. User complaint mechanisms: Social media providers must have mechanisms available for users to report potentially illegal content. Such mechanisms must be easily accessible and readily operated.
2. Notification mechanisms for users whose content has been deemed illegal/ against platforms' terms of use: Mechanisms must be designed to facilitate the communication of clear and comprehensive information to users whose content has been deemed to violate applicable laws and/or terms of use. Social media providers must ensure the communication of notices that contain the following elements:
    - An explanation of why the electronic information has been reported as illegal content, including the law(s) which it violates.
    - The accurate electronic location of the reported content, as shown (for example) through a URL, and, if necessary, additional information that makes it possible to identify illegal content in accordance with the type of content and hosting mechanisms;
    - An indication of the credibility of the individual or entity that reported the content, as well as the accuracy and integrity of the report.
3. After a report containing the electronic contact information of the individual or entity that reported the content, social media providers must, without any undue delay, communicate proof of receipt to the reporting individual or entity.
4. Social media providers must process every filed report using clear mechanisms and address every report in a timely, objective, and consistent manner.
5. Where social media providers use automated systems to process and/or assess reports, they must explicitly communicate to users whose content has been moderated that these systems are in use.
6. Systems should allow individuals to easily trace the progress of their reports.
7. Social media providers shall communicate their decisions, without any undue delay, as well as any appeal mechanisms available to users whose content has been

[2] Mechanisms proposed by human rights organizations, lawyers, and academics for digital platforms' moderation of content. Since they were first proposed in Santa Clara, California, in 2018, these principles have been supported by twelve major corporations, including Meta, Google, Apple, and Twitter.

reported. Two appeal mechanisms should be made available: (1) for flaggers, if their reports are not supported, and (2) for users whose content has been reported, if their content has been removed or their account has been suspended.

8. Social media providers must explain their decisions in detail, including:
   - An indication of whether the content will be deleted or access to said content will be limited; or
   - An indication of whether service provision will be suspended, temporarily or permanently; or
   - An indication of whether the user account will be suspended, temporarily or permanently; or
   - An indication of whether the user account's ability to monetize content will be suspended, terminated, or limited.

   With such clarification, users will have more certainty as to the status of their content and/or accounts. Such clarity will also inform their attitudes vis-à-vis the decisions made by social media providers.

Furthermore, the principles contained within the Santa Clara Principles should also be considered when drafting new regulations in Indonesia. They are:

Figure 3. Transparency and Accountability Principles for Social Media Providers

	Principle	Implementation
1. **Human Rights and Due Process**	Companies should ensure that human rights and due process considerations are integrated at all stages of the content moderation process, and should publish information outlining how this integration is made. Companies should only use automated processes to identify or remove content or suspend accounts, whether supplemented by human review or not, when there is sufficiently high confidence in the quality and accuracy of those processes. Companies should also provide users with clear and accessible methods of obtaining support in the event of content and account actions.	Users should be assured that human rights and due process considerations have been integrated at all stages of the content moderation process, including by being informed of: • How the company has considered human rights—particularly the rights to freedom of expression and non-discrimination—in the development of its rules and policies; • How the company has considered the importance of due process when enforcing its rules and policies, and in particular how the process has integrity and is administered fairly; and • The extent to which the company uses automated processes in content moderation and how the company has considered human rights in such use.

	Principle	Implementation
2. **Understandable Rules and Policies**	Companies should publish clear and precise rules and policies relating to when action will be taken with respect to users' content or accounts, in an easily accessible and central location.	Users should be able to readily understand the following: • What types of content are prohibited by the company and will be removed, with detailed guidance and examples of permissible and impermissible content; • What types of content the company will take action against other than removal, such as algorithmic downranking, with detailed guidance and examples on each type of content and action; and • The circumstances under which the company will suspend a user's account, whether permanently or temporarily.
3. **Cultural Competence**	Cultural competence 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 should publish information as to how these considerations have been integrated in relation to all operational principles. 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.	Users should have access to rules and policies and notice, appeal, and reporting mechanisms that are in the language or dialect with which they engage. Users should also have confidence that: • Moderation decisions are made by those familiar with the relevant language or dialect; • Moderation decisions are made with sufficient awareness of any relevant regional or cultural context; and • Companies will report data that demonstrates their language, regional, and cultural competence for the users they serve, such as numbers that demonstrate the language and geographical distribution of their content moderators.
4. **State Involvement in Content Moderation**	Companies should recognize the particular risks to users' rights that result from state involvement	Users should know when a state actor has requested or participated in any actioning on their content or account. Users should also know if the company believes that the

	Principle	Implementation
	in content moderation processes. This includes a state's involvement in the development and enforcement of the company's rules and policies, either to comply with local law or serve other state interests. Special concerns are raised by demands and requests from state actors (including government bodies, regulatory authorities, law enforcement agencies and courts) for the removal of content or the suspension of accounts.	actioning was required by relevant law. While some companies now report state demands for content restriction under law as part of their transparency reporting, other state involvement is not reported either publicly or to the actioned users. But companies should clearly report to users when there is any state involvement in the enforcement of the company's rules and policies. Specifically, users should be able to access: • Details of any rules or policies, whether applying globally or in certain jurisdictions, which seek to reflect requirements of local laws. • Details of any formal or informal working relationships and/or agreements the company has with state actors when it comes to flagging content or accounts or any other action taken by the company. • Details of the process by which content or accounts flagged by state actors are assessed, whether on the basis of the company's rules or policies or local laws. • Details of state requests to action posts and accounts.
5. **Integrity and Explainability**	Companies should ensure that their content moderation systems, including both automated and non-automated components, work reliably and effectively. This includes pursuing accuracy and nondiscrimination in detection methods, submitting to regular assessments, and equitably providing notice and appeal mechanisms. Companies should actively monitor the quality of their decision-making to assure high confidence levels, and are encouraged to publicly share data about the accuracy of their systems and to open their process and algorithmic systems to periodic external	Users should have confidence that decisions about their content are made with great care and with respect to human rights. Users should know when content moderation decisions have been made or assisted by automated tools, and have a high-level understanding of the decision-making logic employed in content-related automated processes. Companies should also clearly outline what controls users have access to which enable them to manage how their content is curated using algorithmic systems, and what impact these controls have over a user's online experience. Source: Santa Clara Principles (2021).

Principle	Implementation
auditing. Companies should work to ensure that actioning requests are authentic and not the result of bots or coordinated attacks. There are many specific concerns for automated systems, and companies should employ them only when they have confidence in them, and in a transparent and accountable manner.	

F. Social Media to Provide Appeals Mechanisms

The proactive or reactive moderation of content cannot always satisfy all involved parties. In certain conditions, users or government actors may be dissatisfied with providers' decisions and challenge them. In such instances, social media providers should be required to provide systems that enable users/government actors to file appeals.

However, it is difficult for social media providers to implement appeal mechanisms as they provide limited information to users. In interviews conducted by PR2Media, social media users indicated that they were able to reactivate their accounts after receiving assistance from individuals they knew, without any direct response or notification from the platform itself (Wendratama et al., 2023). Most social media users in Indonesia are not certain as to the mechanisms available for filing appeals with social media providers. Meanwhile, civil society organizations—despite being the official partners of these platforms—feel powerless in their negotiations (Article 19, 2022). This problem is thus frequently associated with the abrogation of individuals' freedom of expression as well as individual and public safety.

User appeals must receive careful consideration from platforms, as they ensure said platforms' accountability to the citizens who use their services. Such appeals are also important for preventing arbitrariness in the moderation process, thereby protecting users' freedom of expression and opinion. As recommended by researchers from CfDS UGM (Rahman et al., 2022), the appeal mechanisms used by social media platforms should be regulated by the state following the revision of the moderation mechanisms contained within UU ITE. Regulations regarding how platforms respond to appeals are therefore necessary.

The Digital Services Act requires very large online platforms, those whose monthly user bases represent at least ten percent of the population of the European Union, to provide users with at least three appeal mechanisms. These mechanisms must be made available to flaggers whose reports are not upheld as well as users whose content is moderated.

1. Internal appeal mechanisms

The availability of mechanisms for users to appeal platforms' decisions is important for all users who feel as though their freedom of expression has been abrogated by the platform or who disagree with the decision rendered.

When providing such appeal mechanisms, it is important for platforms to ensure that all mechanisms can be easily used and accessed, as well as facilitate evidence-based decisions and appeals.

Appeal mechanisms should also give users (both those who report content and those whose content is reported) a means of presenting additional information to support their appeals. This information should be considered during the review of the appeal.

One important point that should be recognized by platform providers is timeliness. Reviewing materials and rendering decisions based on the information provided should be prioritized, especially when the content itself is time-sensitive (for example, political content during elections).

The moderation of content should not be discriminatory or arbitrary. Reflecting the Santa Clara Principles, which holds that content moderation should be transparent and accountable, being objective, non-discriminatory, proportional, and just, with respect for user rights. Here, the word "proportional" means that social media providers must prioritize appeals for the most severe sanctions, such as content removal and account suspension.

Platform providers must ensure that appeals are decided by qualified staff, rather than automatic tools. All staff should have the requisite cultural competencies. As stated in the Santa Clara Principles, those who decide appeals must understand the language, culture, and socio-political context of the post being moderated.

Referring to the Santa Clara Principles, PR2Media recommends that social media providers develop appeal mechanisms that contain the following elements:

- Clear and accessible processes, including a detailed written description of timeframes, to allow users to track the progress of their reports.
- A review or assessment by an individual not involved in the original assessment, and thereby able to provide a second opinion.

- The linguistic and cultural understanding possessed by the individual involved in the appeal process.
- Opportunities available to provide additional evidence to support the appeal process.
- The results of the assessment and the considerations leading to these results, in a form that is sufficiently clear and understandable.

2. External mechanisms, outside the judicial system

In this study, PR2Media recommends the establishment of the Social Media Council, which should provide an external appeal mechanism. The form and function of the Social Media Platform shall be discussed below.

3. Judicial processes

The Digital Services Act stipulates that users can also employ judicial processes to challenge the decisions made by social media providers when moderating content. If the courts deem that a provider's decision to remove access to content has violated the terms and conditions established by said provider, social media providers shall be required to restore the affected content.

G. Social Media Council

Both the authors of this study as well as the civil society organizations involved in this research and related policy advocacy recommend that Indonesia create an independent agency for monitoring and guiding social media providers' moderation of content. Here, "independent" indicates that the agency should not be directly/indirectly related to, or involve representatives from, the government or social media platforms.

We argue that the social media councils that have been established around the world may serve as a reference model. In 2019, Stanford University's Global Digital Policy Incubator (GDPi), Article 19. and David Kaye—the United Nations Special Rapporteur on Freedom of Opinion and Expression—met to identify a solution to the challenge of illegal social media content and recommended the establishment of a multi-stakeholder agency: a social media council (SMC).

There are two major obstacles in content moderation: balancing the responsibility to uphold users' freedom of expression with the need to prevent the potentially dangerous effects of harmful content, and navigating the challenges of the privatization of digital space. An SMC is a multi-stakeholder mechanism that provides an open,

independent, transparent, and accountable forum for moderating content on social media platforms following international human rights standards. The SMC model employs a voluntary approach to monitoring content moderation: participants, who may include social media providers, government actors, scholars, and civil society organizations, may register as participants. The efficiency of the SMC depends on the voluntary compliance of social media platforms; when registering, they must commit to honoring and adhering to the decisions/recommendations of the SMC. Social media platforms would also benefit; participation would increase their credibility amongst users, as it would provide them with transparency and accountability. Government participation in the SMC would also increase platforms' legitimacy.

PR2Media recommends that, to best address the specific challenges of the Indonesian context, an SMC should have the following three roles:

- Provide general guidelines for the practice of content moderation, thereby ensuring that content moderation adheres to international standards of freedom of expression as well as values specific to Indonesia (including various local contexts, given the diversity of Indonesia).
- Become a forum for diverse stakeholders to discuss recommendations related to content moderation.
- Become an organization for administrative appeals for those who are dissatisfied with social media providers' content moderation decisions. Aside from handling appeals, the SMC should review the content moderation decisions contained within social media providers' reports.

The authors argue that the SMC should not be located solely in Jakarta. To ensure that the SMC can adequately fulfill these three roles, it should have affiliates internationally and in every Indonesian province (to better recognize the nation's diversity). In so doing, this transparent, accountable, and independent multi-stakeholder forum can ensure that freedom of expression remains protected in content moderation and dissemination. International standards can thus be integrated into the processes through which content is moderated, even as the diversity of information and ideas is accommodated.

This multi-stakeholder agency (consisting of representatives from social media platforms, government agencies, researchers/academics, civil society organizations, religious leaders, legal scholars, etc.)

will be the best option for monitoring and evaluating social media providers' compliance with regulations.

As an independent agency, the SMC shall be non-structural and accountable to the president.

As it is a multi-stakeholder platform, the SMC will use a different approach than platforms' internal monitoring mechanisms. The SMC will include representatives from marginalized groups and local/national civil society organizations across the political spectrum; this will ensure that recommendations are rooted in the lived experiences of those who are most influenced by moderation. Through routine meetings, as well as the open and participatory sharing of information, the SMC will become a powerful agency for overcoming public distrust of social media platforms.

The SMC must also be seen as an alternative solution to two problems, namely 1) platforms' self-regulation being deemed inefficient and ineffective in moderating illegal content, and 2) the ineffective, and frequently repressive, use of regulations to moderate illegal content.

The second point above is associated with the due process of law, which must be followed by the government in exercising its authority and removing access to electronic systems, as allowed by Article 40, Paragraph 2b, of UU ITE.

Article 40, Paragraph 2b, reads, "In prevention, as intended by Paragraph (2a), the Government has the authority to remove access and/or instruct Providers of Electronic Systems to remove access to Electronic Information and/or Electronic Documents that contain illegal content."

As per this paragraph, the authority to remove access lies entirely with the government; in other words, there are no mechanisms through which other entities can decide to remove access. Given this situation, the SMC may provide a solution for reducing the risk of repressive action and government overstep.

Funding Sources for the SMC

The SMC shall be established as an independent agency by law, which shall regulate the dissemination of illegal content on digital platforms. Because of its legal mandate, funding for the SMC shall also come from the State, as well as member dues (i.e., from social media providers). A similar funding model has been used by the Press Council,

which derives its funding from press companies, press organizations, government assistance, and other non-binding forms of assistance (Law No. 40/1999 regarding the Press).

The SMC should not involve itself in tax policies or matters of corporate competition, as such issues are highly business-oriented. The council should instead establish content moderation standards that are fair, reliable, transparent, and non-arbitrary. At a time when social media providers' content moderation practices are increasingly informed by their business interests, the SMC offers a relatively rapid means of coordinating urgent content moderation issues and addressing accountability issues.

Ultimately, the presence of the SMC and the proposed guidelines for its involvement in the moderation/management of social media content embraces the following two policy approaches. First, quasi-regulatory: the government promotes meetings of digital businesses whereby they establish internal regulations and compliance mechanisms but does not directly determine the characteristics of these mechanisms or their implementation. Second, co-regulation: social media providers create and implement their own regulations and standards, while the government creates regulations for enforcing these standards. As with broadcast media, the classic example of this approach, the social media industry desires self-regulation while the public seeks a collaborative approach.

H. Annual Reports from Social Media

Indonesia has three laws regarding content moderation, which also regulate social media platforms: the Law on Information and Electronic Transactions (UU ITE), Government Regulation No. 71/2019 regarding the Implementation of Electronic Transaction Systems, and Regulation of the Minister of Communication and Informatics No. 5/2020 regarding the Implementation of Electronic Systems in Private Environments.

Within the context of handling illegal and harmful content, these regulations do not specify any means of reporting the results of content moderation practices or ensuring social media platforms' transparency and accountability to their users. Information regarding social media platforms' content moderation practices tends to be difficult to obtain. As such, the processes and mechanisms used by social media platforms to moderate content are poorly understood by the average person (Rahman et al., 2022).

In interviews conducted by PR2Media with Yendra Budiana and Ajiwan Arief, respectively members of the Ahmadi community and SIGAB Indonesia (Wendratama

et al., 2023), these representatives of minority groups indicated that their efforts to flag content received no response from social media platforms, despite said the detrimental effect of this content's widespread dissemination. Damar Juniarto from SAFEnet indicated that social media platforms' handling of illegal and harmful content remains dissatisfactory, as platforms are non-responsive and non-transparent. Their reports are purely quantitative, without any clear explanation of the mechanisms used (Wendratama et al., 2023).

As such, social media platforms should provide detailed reports to ensure their transparency and accountability to their users. Referring to the Digital Services Act (DSA), the European Union has required very large online platforms such as YouTube, Facebook, Instagram, and TikTok to publish reports every six months (Nosák, 2021). Conversely, the Indonesian government has yet to require social media platforms to report their moderation practices and actions over the year. To monitor the performance of online platforms or electronic system providers within the private sphere, the Ministry of Communication and Informatics has only required providers to register themselves and comply with the guidelines set in Regulation of the Minister of Communication and Informatics No. 5/2020 regarding the Implementation of Electronic Systems in Private Environments.

PR2Media recommends that social media platforms' annual reports should contain the following information:

First, social media providers should openly communicate the mechanisms they use to identify and flag electronic information and/or electronic documents that violate applicable law, be they automatic (artificial intelligence, AI) or human. Although AI is capable of acting more rapidly than human moderators, it still has difficulty detecting context, sarcasm, and cultural meaning (Duerte et al., 2017).

In regards to this point, reports need to be communicated simply while detailing the moderation process and the policy being violated (West, 2018). Currently, only a few platforms have published information regarding the number of human moderators whom they employ: for example, Facebook employs 15,000 human moderators (Koetsier, 2020), while YouTube/Google employs 10,000 human moderators (Newton, 2019). Unfortunately, no platform has published specific information regarding the number of human moderators employed in Indonesia.

In their format, annual reports should provide quantitative and qualitative information regarding the actual results and effects of content moderation. This should encompass user reports and violations of platforms' terms of service, as well as the results of automated/human content moderation (UNESCO, 2023). Second, annual reports should communicate the number of reports received as well as the action taken

in response. As per the Santa Clara Principles (2021), which were drafted based on consultations with more than fifty organizations and individuals, social media platforms are expected to communicate the following information regarding the content and accounts affected by their sanctions:

- Total number of pieces of content actioned and accounts suspended.
- Number of appeals of decisions to action content or suspend accounts.
- Number (or percentage) of successful appeals that resulted in pieces of content or accounts being reinstated, and the number (or percentage) of unsuccessful appeals
- Number of posts or accounts reinstated by the company proactively, without any appeal, after recognizing that they had been erroneously actioned or suspended.
- Numbers related to content removals and restrictions made during crisis periods, such as during the COVID-19 pandemic and periods of violent conflict.

Meanwhile, with regards to moderating actions, Pirkova (2022) writes that the Digital Services Act (DSA) requires all platforms to publicly report their use of automated moderating tools, the accuracy levels of said tools, and the training and assistance provided by platforms to their moderators. For very large online platforms, the DSA requires reports regarding their risk evaluation activities, risk mitigation efforts, and the results of internal audits (Nosák, 2021). The public has the right to transparent information regarding reported content and the justifications for its decisions, with this information indirectly maintaining the affective relationships between users and social media platforms. Moreover, according to UNESCO (2023), significant transparency will provide stakeholders with the information necessary to make appropriate decisions. As such, it is hoped that annual reports can be effectively and clearly communicated.

To summarize, PR2Media makes the following recommendations for the content contained in annual reports.

Figure 4. Content Guidelines for Annual Reports

Aspects	Details
1. **Number and types of content reported**	• Total number of pieces of content actioned and accounts suspended. • Number of appeals of decisions to action content or suspend accounts. • Number (or percentage) of successful appeals that resulted in pieces of content or accounts being reinstated, and the number (or percentage) of unsuccessful appeals

Aspects	Details
	• Number of posts or accounts reinstated by the company proactively, without any appeal, after recognizing that they had been erroneously actioned or suspended. • Numbers related to content removals and restrictions made during crisis periods, such as during the COVID-19 pandemic and periods of violent conflict. Reference: Santa Clara Principles (2021)
2. **Reporting Bodies**	• Government • Trusted flaggers – organizations competent in recognizing and reporting illegal and harmful content (Wendratama et al., 2023) • Users
3. **Mechanisms for actioning reports**	• Automated systems (artificial intelligence/AI) - Means for using tools, including information on the tools used for certain actions - Accuracy rate of tools used, including the level of trust placed in said tools by social media platforms • Human moderators Training and assistance provided by platforms Reference: Digital Services Act (in Pirkova, 2022) and Santa Clara Principles (2021) • Risking by platforms • Risk mitigation efforts • Audit reports and implementation References: Digital Services Act, particularly those for VLOPs (in Nosák, 2021)

I. Independent Auditing of Social Media

To ascertain social media providers' compliance with applicable regulations, independent audits of their activities and management must be conducted. Such audits should not be financial but rather compliance-oriented.

> Aspects that should be audited include content curation mechanisms (for example, how content is recommended to users by automated systems), content moderation mechanisms (for example, how content that violates the terms of service is recognized and flagged/removed by automated systems, as well as the extent to which human moderators are involved), the effectiveness of appeal mechanisms, and the risk-management mechanisms implemented.

This audit must be conducted by external organizations (auditors) with technical competence and proven professional expertise in risk management, as well as a record of independent auditing. Such audits should be conducted at least once per annum, with costs borne by social media providers.

Below are the DSA's requirements for independent audits (European Commission, 2023), which may be adapted to the Indonesian context.

1. Social media providers shall cooperate with and provide assistance to the auditing institution, including by providing access to accurate data deemed necessary, including, if deemed relevant, data related to said providers' internal algorithms.
2. Audits shall be conducted in accordance with industry best practices and professional ethics, as well as a high degree of objectivity, with due consideration of the applicable code of standards.
3. Auditors shall maintain the confidentiality, security, and integrity of all information, including industrial secrets, that are learned during the course of their duties.
4. Audit reports must be evidence-based, to provide a significant explanation of the activities undertaken and the conclusions drawn. Audit reports shall provide information and, if necessary, recommend actions that may be undertaken by social media providers to ensure compliance with this Law.
5. Audit reports shall be communicated to the relevant institutions (Government and Social Media Council).
6. Social media providers shall communicate their response(s) to audit reports to the relevant institutions (Government and Social Media Council).

In the European Union, independent auditing of very large online platforms (i.e., those with at least 45 million monthly active users in the European Union) only began in late August 2023 (Deloitte Legal, 2023); as such, no independent audit has been reported. The deadline for the completion of platforms' first independent audit is August 2024, with the results published before November 2024 (Tremau, 2023). As of writing, no independent auditor has been announced as involved in any audits. Nevertheless, some auditing organizations have indicated their readiness to assess platforms' compliance with the DSA, including Deloitte (Cankett & Fackovcova, 2022) and Ernst & Young (Legat & Guzy, 2023).

How are sanctions determined?

With all of those obligations, what are the consequences for social media providers that fail to comply?

Sanctions, in our opinion, should be based on the extent of "failure" of the system in place to comply with regulations. As a result, the sanctions are determined not on a case-by-case basis, but rather by examining what procedures have been used by social media providers to detect illegal content.

In this case, regulations should establish a "threshold" or "failure" criterion as the basis for imposing sanctions. During the process, Indonesian authorities must consider the proactive measures taken by platforms to monitor and remove illegal content.

CONCLUSION

The communication of this research paper to Indonesian policymakers (the Government and House of Representatives) and its public dissemination is part of PR2Media's participation in the reform of digital media governance in Indonesia. This paper, which is intended to serve as a basis for open dialogue, will provide a historical record of public participation in the regulation of social media and the advancement of public interests. This publication advances PR2Media's efforts to ensure that digital spaces (the internet and social media) are recognized as the right of all citizens (i.e., internet constitutionalism).

We argue that social media offers a public space to openly communicate information and knowledge within the public domain, and thus requires regulation. In this context, all parties (particularly the government and social media platforms) must honor international human rights norms, including the international conventions that have been ratified by Indonesia, including the International Covenant on Civil and Political Rights (ICCPR, 1966) and the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD, 1965).

The International Covenant on Civil and Political Rights (ICCPR) is a multi-lateral agreement that requires nations to honor and uphold the civil and political rights of individuals, including their right to life, right to vote, and right to due process of law, as well as their freedom of religion, freedom of expression, and freedom of association. This covenant was ratified by Indonesia in 2005, including through Law No. 12/2005 regarding the Ratification of the International Covenant on Civil and Political Rights.

Meanwhile, the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) is a United Nations convention that requires member states to eradicate all forms of racial discrimination and to advance mutual understanding. This convention also requires member states to outlaw race-based hate speech and membership in hate-based organizations. This convention was ratified by Indonesia in 1999 (Law No. 29/1999 regarding the Ratification of the International Convention on the Elimination of All Forms of Racial Discrimination.

The framework and legislation proposed here as part of a total revision of UU ITE emphasizes the importance of periodically reviewing regulations to ensure that they remain timely, effective, and proportional. Such periodic reviews must also consider the mechanisms through which regulations are implemented, using an evidence-based approach to understand the moderation and management of illegal content on social media. Through periodic review, social media regulations can remain relevant to the dynamics of digital technology.

We wish to express our gratitude to all parties involved in the preparation of this paper. First and foremost is the preparatory team, which represents civil society organizations that advocate for human rights in the media (including SAFEnet, LBH Pers, Mafindo, ICJR, and AJI Indonesia). We would also like to thank the Tifa Foundation, which provided financial support. This research paper is open to the public, and thus we open ourselves to input and recommendations.

REFERENCES

Books, Journals, Articles, and Videos

Aichner, T., Grünfelder, M., Maurer, O., & Jegeni, D. (2021). Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, Behavior, and Social Networking*, *24*(4), 215–222. DOI: 10.1089/cyber.2020.0134

Article 19. (2022). *Content moderation and local stakeholders in Indonesia*. Article 19. https://www.article19.org/wp-content/uploads/2022/06/Indonesia-country report.pdf

Article 19. (2021, October 12). Social Media Councils: One piece in the puzzle of content moderation. *Article 19*. https://www.article19.org/resources/social-mediacouncils-moderation/

Article 20 Digital Services Act. (n.d.). Article 20, internal complaint-handling system – the Digital Services Act (DSA). *Cyber Risk GmBH*. https://www.eu-digital-services-act.com/Digital_Services_Act_Article_20.html

Barrett, P. M. (2020). *Who moderates the social media giants? A call to end outsourcing*. New York University Stern Center for Business and Human Rights. https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf

Cankett, M., & Fackovcova, L. (2022, May 18). EU Digital Services Act: Are you ready for audit?. *Deloitte*. https://www2.deloitte.com/uk/en/blog/auditandassurance/2022/eu-digital-services-act-are-you-ready-for-audit.html

De Gregorio, G. (2019). Democratising online content moderation: A constitutional framework. *Computer Law & Security Review: The International Journal of Technology Law and Practice*. https://doi.org/10.1016/j.clsr.2019. 105374

Deloitte Legal. (2023). Rules for external audit under the RU's Digital Services Act (DSA). *Lexology*. https://www.lexology.com/library/detail.aspx?g=1301fe0e-46fd-4fea-9444-a68b0bc16901

Frosio, G. (2021). Regulatory shift in state intervention: From intermediary liability to responsibility. In Edoardo, C., Amélie, H., & Clara, I. (eds.), *Constitutionalising social media*. Hart Publishing. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3850483

Kemp, S. (2023, February 9). Digital 2023: Indonesia. *Datareportal*. https://datareportal.com/reports/digital-2023-indonesia

European Commission. (2023). Digital Services Act: Commission designates first set of very large online platforms and search engines. *European Commission*. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413

European Commission. (2023). The Digital Services Act: Ensuring a safe and accountable online environment. *European Commission*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-servicesact-ensuring-safe-and-accountable-online-environment_en

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. United Nations Educational, Scientific, and Cultural Organization (UNESCO). https://unesdoc.unesco.org/ark:/48223/pf0000233231

Legat, O., & Guzy, A. (2023, January 5). Security on the internet. New European initiative – digital services act. *EY*. https://www.ey.com/en_pl/law/security-on-the-internetnew-european-initiative-digital-services-act

Leong, P. (2022, July 27). Content moderation of social media in Southeast Asia: Contestations and control. *Fulcrum*. https://fulcrum.sg/content-moderation-ofsocial-media-in-southeast-asia-contestations-and-control/

Palatino, M. (2015, November 29). Will Indonesia's police circular on hate speech suppress freedom of expression?. *Advox*. https://advox.globalvoices.org/2015/11/29/willindonesias-police-circular-on-hate-speech-suppress-freedom-of-expression/

Pirkova, E. (2022, July 6). The Digital Services Act: Your guide to the EU's new content moderation rules. *Access Now*. https://www.accessnow.org/digital-services-acteu-content-moderation-rules-guide/

PR2Media (2023). *Penyampaian usulan revisi UU ITE kepada Komisi I DPR RI*. [Video]. https://www.youtube.com/watch?v=bHp5r3wG6xE&t=6s.

Putri, D. (2021). Apakah semua ujaran kebencian perlu dipidana? Catatan untuk revisi UU ITE. *The Conversation*. https://theconversation.com/apakah-semua-ujarankebencian-perlu-dipidana-catatan-untuk-revisi-uu-ite-156132

Santa Clara Principles. (2021). *The Santa Clara principles on transparency and accountability in content moderation*. Electronic Frontier Foundation. https://santaclaraprinciples.org/open-consultation/

Tremau. (2023, May 31). The DSA & audits. *Tremau*. https://tremau.com/the-dsa-audits

United Nations Assembly Resolution 2106. (1965). *International convention on the elimination of all forms of racial discrimination* [General Assembly Resolution 2106 on 21 December 1965]. https://www.ohchr.org/sites/default/files/cerd.pdf

United Nations High Commissioner for Human Rights. (2013). *Annual report of the United*

Nations High Commissioner for Human Rights. United Nations High Commissioner for Human Rights. https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf

United Nations Human Rights Office of the High Commissioner. (2023). *Special rapporteur on freedom of opinion and expression.* United Nations Human Rights Office of the High Commissioner. https://www.ohchr.org/en/special-procedures/sr-freedom-of-opinion-and-expression

United States Congress. (2019). Liability for content hosts: An overview of the communication decency act's section 230. *Congressional Research Service*. https://sgp.fas.org/crs/misc/LSB10306.pdf

Wendratama, E., Masduki., Rahayu., Suci, P. L., Rianto, P., Aprilia, M. P., Paramastri, M. A., & Adiputra, W. M. (2023). *Pengaturan konten ilegal dan berbahaya di media sosial: Riset pengalaman pengguna dan rekomendasi kebijakan*. Pemantau Regulasi dan Regulator Media. https://pr2media.or.id/publikasi/pengaturan-konten-ilegaldan-berbahaya-di-media-sosial-riset-pengalaman-pengguna-dan-rekomendasikebijakan/

Wendratama, E. (2021). Perlindungan data pribadi: Tantangan regulasi di Indonesia. In Masduki (Ed.), *Kebijakan media dan COVID-19 di Indonesia*. Penerbit Komunikasi Universitas Islam Indonesia Program Studi Ilmu Komunikasi. https://pr2media.or.id/publikasi/kebijakan-media-dan-covid-19-di-indonesia/

Yanisky-Ravid, S., & Lahav, B. Z. (2017). Public interest vs. private lives – affording public figures privacy in the digital era: The three principles filtering model. *University of Pennsylvania Journal of Constitutional Law*, 19(5), 1–61. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2931864

Policy, Legislation, and International Instruments

Government Regulation No. 71/2019 regarding the Implementation of Electronic Transaction Systems

International Convention on the Elimination of All Forms of Racial Discrimination. (1965). https://www.ohchr.org/en/instruments-mechanisms/instruments/international convention-elimination-all-forms-racial

International Covenant on Civil and Political Rights (ICCPR). (1966). https://www.ohchr.org/en/instrumentsmechanisms/instruments/international-covenant-civil-and-political-rights

Joint Decision of the Coordinating Ministers of Politics, Law, and Security No. B-96/HK.00.00/07/2021 regarding the Guidelines for Implementing Certain Articles of the Information and Electronic Transactions Law (UU ITE)

Joint Decision of the Minister of Communication and Informatics, Attorney General, and Police Director No. 229 of 2021 and No. 154 of 2021 regarding the Guidelines for Implementing Certain Articles of Law No. 11/2008 regarding Information and Electronic Transactions, as Revised by Law No. 19/2016 regarding the Revision of Law No. 11/2008 regarding Information and Electronic Transactions.

Law No. 12/2005 regarding the Ratification of the International Covenant on Civil and Political Rights (2005). https://www.dpr.go.id/doksetjen/dokumen/-Regulasi-UU-No.-12-Tahun-2005-Tentang-Pengesahan-Kovenan-Internasional-Tentang-Hak-Hak-Sipil-dan-Politik-1552380410.pdf

Law No. 19/2016 regarding the Revision of Law No. 11/2008 regarding Information and Electronic Transactions (2016). https://web.kominfo.go.id/sites/default/files/users/4761/UU%2019%20Tahun%202016.pdf

Law No. 29/1999 regarding the Ratification of the International Convention on the Elimination of All Forms of Racial Discrimination 1965 (1999). https://www.dpr.go.id/dokjdih/document/uu/UU_1999_29.pdf

Law No. 40/1999 regarding Press (1999). https://peraturan.bpk.go.id/Home/Details/45370/uu-no-40-tahun-1999

Regulation of the Minister of Communication and Informatics No. 5/2020 regarding the Implementation of Electronic Systems in Private Environments

APPENDICES

Appendix 1

Proposed Revision to Article 15 of UU ITE to Promote the Transparency and Accountability of Social Media Providers in Moderating Illegal Content

(This proposal was presented by PR2Media during the Public Hearing held by the Committee for the Revision of UU ITE, Commission I, House of Representatives of Indonesia, on August 23, 2023.)

No.	Article	Current Wording	PR2Media Recommendations	Explanation
1.	Article 15	(1) Any Electronic System Provider must provide Electronic Systems in a reliable and secure manner and shall be responsible for the proper operation of the Electronic Systems. (2) Electronic System providers shall be responsible for their Operation of Electronic Systems. (3) The provision as intended by paragraph (2) shall not apply where it is verifiable that there occur compelling circumstances, fault, and/or negligence on the part of the Electronic System users.	Add Article (4), to read: In implementing Article (2), social media providers shall: (a) Use the precautionary principle to ensure that electronic information and/or electronic documents that contain illegal content cannot be accessed. (b) Openly and transparently present the systems through which social media providers recognize and identify electronic information and/or electronic documents as illegal electronic information and/or electronic documents, be they automated systems or involving persons employed by said social media providers, as well as the actions undertaken by social media providers to address violations and suspected violations. (c) Openly and transparently present the mechanisms used by social media providers to recommend electronic information and/or electronic documents to users, as well as the effects of said systems on the electronic information and/or electronic documents presented to users. (d) Provide all necessary information to users affected by Paragraph (a), or by the suspension of their social media accounts, including the electronic information and/or electronic documents that violate applicable law and/or platform policies, an explanation of the violation identified, the means through which social media platforms detect violations, and the appeal mechanisms available to users. (e) Respond to and follow up on public and Government grievances regarding electronic information and/or electronic documents that contain content that violates applicable law, including any evaluation and response to said content, while forefronting the principles of caution and justice.	Monitoring of the implementation of Article (4) is to be done by the Government.

	Article	Current Wording	PR2Media Recommendations	Explanation
			(f) Publish an annual report on public and government grievances regarding electronic information and/or electronic documents that contain illegal content as well as their responses to such grievances. (g) When implementing Paragraph (a), social media providers shall create appeal mechanisms for users, which shall include an evaluation by persons not involved in the initial evaluation and/or opportunities for users to provide new information. (h) Conduct audits, involving an independent auditor, at least once per annum to evaluate compliance with Paragraphs (a), (b), (c), (d), (e), (f), and (g), and openly and transparently publish the results of said audits. (i) The obligations of social media roviders, as described in Paragraphs (a), (b), (c), (d), (e), and (f) shall be proscribed through government regulation.	
2.	Addendum to Chapter 1, General Provisions, Article 1	In this Law, what is meant by: Numbers 1 through 23.	Add Number 24: Social media are internet-based electronic systems that enable users to mutually exchange electronic information and/or electronic documents using open electronic systems that are controlled by social media providers.	—
3.	Elucidation to Article 15, Paragraph (4)	—	Social media providers are entities that provide internet-based electronic systems that enable users to mutually exchange electronic information and/or electronic documents using open electronic systems, with at least twenty million active monthly users in Indonesia, and/or other social media that are proposed by society and approved by the pertinent legal body.	—
4.	Elucidation to Article 15, Paragraph (4), Point h.	—	Procedures for independent auditing shall be proscribed by Government Regulation, which shall contain: 1. Social media providers shall cooperate with and provide assistance to the auditing institution, including by providing access to accurate data deemed necessary, including, if deemed relevant, data related to said providers' internal algorithms. 2. Audits shall be conducted in accordance with industry best practices and professional ethics, as well as a high degree of objectivity, with due consideration of the applicable code of standards. 3. Auditors shall maintain the confidentiality, security, and integrity of all information, including industrial secrets, that are learned during the course of their duties.	1. This audit is not an audit of finances, but an audit of social media providers' compliance with this Law. 2. This audit shall be conducted by an independent auditor, namely an organization or enterprise with experience in independent auditing.

No.	Article	Current Wording	PR2Media Recommendations	Explanation
			4. Audit reports must be evidence-based, to provide a significant explanation of the activities undertaken and the conclusions drawn. Audit reports shall provide information and, if necessary, recommend actions that may be undertaken by social media providers to ensure compliance with this Law. 5. Audit reports shall be communicated to the relevant institution (Government). 6. Social media providers shall communicate their response(s) to audit reports to the relevant institution (Government).	

Notes:

1. Reference for the definition of social media operator: Aichner, T., Grünfelder, M., Maurer, O., & Jegeni, D. (2021). Twenty-Five Years of Social Media: A Review of Social Media Applications and Definitions from 1994 to 2019. *Cyberpsychology, Behavior and Social Networking*, 24(4), 215-222. doi: 10.1089/cyber.2020.0134
2. The use of "open electronic systems" is per Law No. 19/2016, Article 1, Paragraphs 5 and 7.
   Paragraph 5 : An Electronic System is a set of electronic devices and procedures that serves to prepare, collect, process, analyze, store, display, announce, send, and/or disseminate Electronic Information.
   Paragraph 7 : An Electronic System Network is an interlinked network of two or more Electronic Systems, which are closed or open.
3. The use of "operators of social media" is per Law No. 19/2016, Article 1, Paragraphs 6 and 6a.
   Paragraph 6 : Operation of Electronic Systems is the use of Electronic Systems by state administrators, Persons, Business Entities, and/or society.
   Paragraph 6a : Operators of Electronic Systems are all persons, state administrators, Business Entities, and/or society that provide, administer, and/or operate Electronic Systems, be it individually or collectively, to reach Electronic Systems users for their own purposes and/or the purposes of others.
4. Several details in Paragraph (4), Points (a) through (h), refer to the Santa Clara Principles (https://santaclaraprinciples.org/open-consultation/), particularly the mechanisms recommended for human rights organizations, lawyers, and academics. Since being formulated during a conference on content moderation in Santa Clara, California, in 2018, these principles have received the support of twelve large digital corporations, including Meta, Google, Apple, and Twitter.
5. Regarding the definition of "operators of social media" in the elucidation of Article 15, Paragraph 3, we cannot use the threshold "at least 10% of the population" (as in the European Union) as this would exclude platforms such as Twitter. This is problematic, and thus a threshold of "20 million" has been selected. Data on the number of social media users in Indonesia are presented below:
   a) YouTube : 139 million
   b) Facebook : 119 million
   c) TikTok : 109.9 million
   d) Instagram : 89.1 million
   e) Twitter (X) : 24 million
   f) LinkedIn : 23 million

Source: https://datareportal.com/reports/digital-2023-indonesia

Recommended Articles for Promoting Transparency and Accountability in Social Media Providers' Moderation of Illegal Content (As Part of a Total Revision of UU ITE)

A. Definition and scope of social media

1. Social media are internet-based electronic systems that enable users to mutually exchange electronic information and/or electronic documents using open electronic systems that are controlled by social media providers.
2. Social media providers are entities that provide internet-based electronic systems that enable users to mutually exchange electronic information and/or electronic documents using open electronic systems, with at least twenty million active monthly users in Indonesia, and/or other social media that are proposed by society and approved by the pertinent legal body.

B. Policies and regulations related to prohibited content on social media

1. Social media providers shall publish clear and precise rules and policies regarding when actions shall be undertaken in response to content and/or user accounts that are deemed to violate their terms of service.
2. Said regulations and policies shall make it possible for users to easily understand:
   a. What types of content are prohibited by the company and will be removed, with detailed guidance and examples of permissible and impermissible content;
   b. What types of content the company will take action against other than removal, such as algorithmic downranking, with detailed guidance and examples on each type of content and action; and
   c. The circumstances under which the company will suspend a user's account, whether permanently or temporarily.

C. Means through which Social Media Providers Recognize and Identify Illegal Content

Social media providers shall:

1. Use the precautionary principle to ensure that electronic information and/or electronic documents that contain illegal content cannot be accessed.
2. Openly and transparently present the systems through which social media providers recognize and identify electronic information and/or electronic documents as illegal electronic information and/or electronic documents, be

they automated systems or involving persons employed by said social media providers, as well as the actions undertaken by social media providers to address violations and suspected violations.

3. Openly and transparently present the mechanisms used by social media providers to recommend electronic information and/or electronic documents to users, as well as the effects of said systems on the electronic information and/or electronic documents presented to users.
4. Provide all necessary information to users affected by Paragraph (a), or by the suspension of their social media accounts, including the electronic information and/or documents that violate applicable law and/or platform policies, an explanation of the violation identified, the means through which social media platforms detect violations, and the appeal mechanisms available to users.
5. Respond to and follow up on public and Government grievances regarding electronic information and/or electronic documents that contain content that violates applicable law, including any evaluation and response to said content, while forefronting the principles of caution and justice.
6. Publish an annual report on public and government grievances regarding electronic information and/or electronic documents that contain illegal content as well as their responses to such grievances.
7. When implementing Paragraph (a), social media providers shall create appeal mechanisms for users, which shall include an evaluation by persons not involved in the initial evaluation and/or opportunities for users to provide new information.
8. Conduct audits, involving an independent auditor, at least once per annum to evaluate compliance with Paragraph (a), Paragraph (b), Paragraph (c), Paragraph (d), Paragraph (e), Paragraph (f), and Paragraph (g), and openly and transparently publish the results of said audits.
9. The obligations of social media providers, as described in Paragraphs (a), (b), (c), (d), (e), and (f) shall be proscribed through government regulation.

D. Mechanisms through which Social Media Platforms respond to Reported Content

1. Social media providers shall implement a system for handling internal reports of prohibited content, which shall be structured, accountable, easily accessed, and designed for timely moderation.
2. The aforementioned system for handling internal reports of prohibited content shall include a system for receiving reports, processing reports, communicating decisions, and appealing decisions.

3. Social media providers shall take the necessary steps to ensure that reporting mechanisms contain the following elements:
    a. An explanation of why the electronic information has been reported as illegal content, including the law(s) which it violates.
    b. The accurate electronic location of the reported content, as shown (for example) through a URL, and, if necessary, additional information that makes it possible to identify illegal content in accordance with the type of content and hosting mechanisms;
    c. The name and email address of the individual or entity that reported the content;
    d. An indication of the credibility of the individual or entity that reported the content, as well as the accuracy and integrity of the report.
4. Moderation of reported content shall be objective, non-discriminatory, proportional, and just, while respecting the rights of users.
5. Social media providers shall process every report using clear mechanisms and address every report in a timely, objective, and consistent manner.
6. After a report containing the electronic contact information of the individual or entity that reported the content, social media providers shall, without any undue delay, communicate proof of receipt to the reporting individual or entity.
7. Where social media providers use automated systems to process and/or assess reports, they shall explicitly indicate the usage of these systems in their reports.
8. Social media providers' processing and assessment of reports shall be conducted by qualified staff, and shall not rely solely on automated systems.
9. Staff responsible for handling reports and/or moderating content shall be required to understand the language, culture, and socio-political context of the postings that they moderate.
10. Any decisions regarding the moderation of illegal content shall include:
    a. An indication of whether the content will be deleted or access to said content will be limited; or
    b. An indication of whether service provision will be suspended, temporarily or permanently; or
    c. An indication of whether the user account will be suspended, temporarily or permanently; or
    d. An indication of whether the user account's ability to monetize content will be suspended, terminated, or limited.
11. Social media providers shall communicate their decisions, without any undue delay, as well as any appeal mechanisms available to users.

E. Appeal mechanisms

The internal appeal mechanisms made available by social media providers shall encompass the following:

1. Clear and accessible processes, including a detailed written description of timeframes, to allow users to track the progress of their reports.
2. A review or assessment by an individual not involved in the original assessment, and thereby able to provide a second opinion.
3. The linguistic and cultural understanding possessed by the individual involved in the appeal process.
4. Opportunities available to provide additional evidence to support the appeal process.
5. The results of the assessment and the considerations leading to these results, in a form that is sufficiently clear and understandable.

F. Social Media Council

1. To ensure that social media moderation serves the interests of general society, an independent Social Media Council shall be established.
2. The Social Media Council shall be responsible for:
   a. Providing recommendations to social media providers regarding content moderation.
   b. Providing general guidelines for content moderation that ensure that said practices serve the interests of the Indonesian people.
   c. Providing an administrative appeal body for users who are dissatisfied with the results of social media platforms' internal moderation mechanisms.
3. The Social Media Council shall be a non-structural institution accountable to the President of the Republic of Indonesia.
4. Members of the Social Media Council shall include:
   a. Representatives of the government
   b. Representatives of social media providers
   c. Representatives of civil society organizations
   d. Academics, societal leaders, content moderation experts, legal experts, and religious leaders.
5. The Chair and Deputy Chair of the Social Media Council shall be appointed by members of the Council.
6. Funding for the Social Media Council shall originate from:
   a. Social media providers
   b. State assistance and other non-binding forms of assistance

G. Annual reporting

1. Social media providers shall transparently communicate their mechanisms for recognizing and flagging electronic information and/or electronic documents that violate the law, be they conducted by automated systems (artificial intelligence) or by human moderators.
2. Social media providers shall report specific information regarding the number of human moderators employed in Indonesia.
3. Social media providers shall report the number of complaints received as well as the actions taken in response to said reports.
4. Social media providers shall report on the content and accounts sanctioned, including the following:
    a. Total number of pieces of content actioned and accounts suspended.
    b. Number of appeals of decisions to action content or suspend accounts.
    c. Number (or percentage) of successful appeals that resulted in pieces of content or accounts being reinstated, and the number (or percentage) of unsuccessful appeals
    d. Number of posts or accounts reinstated by the company proactively, without any appeal, after recognizing that they had been erroneously actioned or suspended.
    e. Numbers related to content removals and restrictions made during crisis periods, such as during the COVID-19 pandemic and periods of violent conflict.
5. Social media providers shall report their risk assessment activities to their users, as well as potential means of mitigating the risks of social media.
6. Social media providers shall communicate all audits and audit reports conducted by independent auditors.

H. Independent audits of social media

1. To evaluate social media providers' compliance with Points B through G above, social media providers shall be audited by auditing organizations and/or independent auditors.
2. Social media providers shall cooperate with and provide assistance to the auditing institution, including by providing access to accurate data deemed necessary, including, if deemed relevant, data related to said providers' internal algorithms.
3. Audits shall be conducted in accordance with industry best practices and professional ethics, as well as a high degree of objectivity, with due consideration of the applicable code of standards.

4. Auditors shall maintain the confidentiality, security, and integrity of all information, including industrial secrets, that are learned during the course of their duties.
5. Audit reports must be evidence-based, so as to provide a significant explanation of the activities undertaken and the conclusions drawn. Audit reports shall provide information and, if necessary, recommend actions that may be undertaken by social media providers to ensure compliance with this Law.
6. Audit reports shall be communicated to the relevant institution (Government).
7. Social media providers shall communicate their response(s) to audit reports to the relevant institution (Government).
8. Guidelines for independent audits shall be established through technical regulations.

PR2Media

www.pr2media.or.id